AI PIONEER

EXCELLENT AI APPLICATION CASES FOR SMES IN SHANGHAI

Guided by

Shanghai Municipal Commission of Economy and Informatization

Jointly Compiled by

Shanghai Coordination Office for Promoting the Development of Small and Medium-sized Enterprises

Shanghai Small and Medium-sized Enterprise Development Service Center

Shanghai Artificial Intelligence Industry Association

CGAIG
CENTER FOR GLOBAL AI INNOVATIVE GOVERNANCE

Editor's Note

The world today is undergoing a profound transformation driven by artificial intelligence (AI). With continuous breakthroughs in algorithms, computing power and data - the three fundamental elements of AI - AI has moved from theoretical exploration in laboratories into a stage of large-scale and in-depth industrial application. It has become a core driving force for the global technological revolution and industrial transformation.

China attaches great importance to the development of AI and has clearly proposed adopting multiple measures to support all types of enterprises, including private enterprises, in deeply participating in the "AI Plus" initiative. By giving full play to market forces, China aims to promote the deep integration of AI with the real economy and foster new drivers of development. This has pointed the way and provided strong policy momentum for the application of AI technologies across industries, especially among the vast number of dynamic small and medium-sized enterprises (SMEs).

As a frontier of China's reform and opening-up and a highland of technological innovation, Shanghai has consistently regarded AI as a leading industry for priority development. The city has actively implemented national strategic plans and introduced a series of implementation measures related to AI. By closely aligning with the requirements for empowering SME development, Shanghai is accelerating AI-enabled new quality productive forces and building a world-class AI industrial cluster.

To this end, the Shanghai Municipal Commission of Economy and Informatization has made systematic arrangements and taken action in areas such as optimizing the policy environment, opening application scenarios, strengthening the supply of public services, and promoting integrated development among large enterprises and SMEs. These efforts fully support the intelligent transformation of SMEs, lower the threshold for AI application, and unleash innovation potential. To systematically sort, summarize and present Shanghai SMEs' active explorations and fruitful achievements in AI application, AI Pioneer - Outstanding Artificial Intelligence Application Cases of Shanghai Small and Medium-sized Enterprises has been compiled under the guidance of the Shanghai Municipal Commission of Economy and Informatization. Through a group of representative and demonstrative cases, this volume aims to provide more SMEs standing at the crossroads of intelligent transformation with referenceable paths, usable models and replicable experience.

Looking ahead, the development and application of AI technologies will continue to advance rapidly. Under the combined influence of policy guidance, technological drive and market vitality, Shanghai's SMEs will be better positioned to seize historical opportunities, stand at the forefront and shine brightly in the broad arena of "AI Plus". They will contribute an indispensable "pioneer" force to Shanghai's development into an international digital city and AI highland with global influence, and jointly write a splendid new chapter in the era of the intelligent economy.

Content

Part I Introduction

I. China's AI Industrial Policy Layout and Shanghai's Mission

(I) China's AI Industrial Policy Layout

China is now facing a historic opportunity in which AI empowers thousands of industries. To advance the "AI Plus" initiative in an all-round way, accelerate the deep integration of AI with economic and social sectors, reshape modes of production and daily life, stimulate a leap in productivity and a transformation of production relations, and actively build a new form of intelligent economy and intelligent society featuring human-machine collaboration, cross-sector integration, co-creation and sharing, "AI Plus" is being promoted as a new engine for high-quality development.

In this process, a number of new types of infrastructure, technological systems, industrial ecosystems and employment positions have emerged. These developments are accelerating the cultivation of new quality productive forces, enabling all people to share the development dividends of the intelligent era, and injecting strong intelligent momentum into Chinese modernization.

By 2027, AI will have achieved deep integration in six key areas. The penetration rate of new-generation intelligent terminals and agents will exceed 70 percent, the core industries of the intelligent economy will see significant growth, AI will play a more prominent role in public governance, and an open cooperation system will continue to improve.

By 2030, AI will comprehensively empower high-quality development. The penetration rate of intelligent terminals and agents will exceed 90 percent, the intelligent economy will become an important pillar of economic growth, and technological inclusiveness and the sharing of achievements will be further advanced.

By 2035, China will have fully entered an intelligent economy and an intelligent society, providing strong support for the basic realization of socialist modernization.

China's AI industrial layout should focus on six major directions: scientific research innovation, industrial upgrading, consumer innovation, public services, governance coordination and global connectivity. It should build a full-chain development pattern from technological breakthroughs to ecosystem co-construction, thereby empowering overall economic and social development.

In scientific research innovation, a new research paradigm combining data-driven methods with intelligent simulation should be developed to accelerate breakthroughs in frontier technologies, promote interdisciplinary integration, and improve the efficiency of basic research and achievement transformation. In industrial upgrading, intelligent-native business forms should be deployed, intelligent transformation should be promoted across the entire industrial process, a digital production system for agriculture should be developed, and a new ecosystem of intelligent services should be created for the service sector, forming a full-chain AI empowerment pattern. In consumer innovation, intelligent service scenarios should be expanded, consumer content should become more personalized and experiences more immersive, and a new integrated consumer format of "product + service + data" should be developed. In public services, workflow automation, intelligent learning resources and precision life services should be promoted to build an AI-enabled public service system covering the entire life cycle. In governance coordination, a smart governance system led by government, supported by enterprises and participated in by the public should be established, so as to promote human-machine collaborative decision-making in urban management, ecological protection, public safety and other fields.

(II) Shanghai's Mission in AI Policy

To implement the national strategic plan for the development of next-generation AI, accelerate the "AI Plus" initiative, advance Shanghai's "Modeling Shanghai" project, further expand AI applications, speed up the development of Shanghai as an "AI highland", and build a world-class AI industrial cluster, Shanghai has formulated the Several Measures of Shanghai Municipality for Further Expanding AI Applications, which reflects its mission in four aspects.

From the perspective of the innovation environment, Shanghai will build a first-class AI innovation environment, establish industrial innovation service platforms, and support leading talents in establishing new research and development (R&D) institutions. By building high-density and low-cost industrial clusters, Shanghai will promote the agglomeration of elements and coordinated development. It will improve an investment and financing system guided by government funds and joined by social capital, and strengthen long-term support for early-stage projects. At the same time, relying on open platforms such as the World Artificial Intelligence Conference, Shanghai will deepen international cooperation, integrate into the global open-source ecosystem, and build an open and shared innovation pattern.

In terms of innovative applications, Shanghai will use scenario-driven development and product promotion as key drivers, support the accelerated R&D and large-scale application of AI software and hardware products, and reward innovative products such as embodied intelligent robots that demonstrate strong market performance. It will actively build benchmark demonstration scenarios, work with industry authorities to collect application demands in key fields, and accelerate the formation of typical solutions that can be replicated and promoted, using scenario demonstrations to drive technological iteration and industrial clustering.

In terms of innovation factors, Shanghai will systematically reduce the cost of enterprise R&D and application of large models by issuing “computing power vouchers”, “model vouchers” and “corpus vouchers”, and by promoting the construction of computing power scheduling and corpus service platforms. High-level AI talents and teams will receive long-term and stable support based on the principle of “tailored support, one policy for one person”, thereby consolidating the foundation of innovation factors.

In terms of original innovation, Shanghai will focus on frontier areas such as basic AI theory, artificial general intelligence, intelligent chips and embodied intelligence, and systematically deploy key core technology research. It will promote a transformation of scientific research paradigms, implement the Scientific Intelligence “Hundred Teams and Hundred Projects” initiative, facilitate cross-organizational collaboration among AI talents, scientists and engineering teams, produce a number of original achievements, and build a solid technological foundation for high-quality development.

II. Four Major Trends in AI Technological Innovation

AI is reshaping the global economy and industrial landscape at an unprecedented speed. Technology is no longer merely a laboratory breakthrough; it has become a core force that penetrates thousands of industries and empowers the real economy. Taking the collection of outstanding AI application cases in Shanghai as an opportunity, we can clearly see that the future development of AI is showing four distinctive and closely connected trends. Together, they will define the next intelligent era.

(I) Multimodal Large Models: From “Understanding Text” to “Perceiving the Whole Domain”

Multimodal large models represent the next technological peak after language models. They are no longer limited to processing text, but integrate text, images, audio, video and other forms of information to achieve true cross-modal understanding and generation.

AI is moving from “reading thousands of books” to “observing all things, listening to all voices and analyzing all phenomena”, breaking down the perceptual barriers between the digital and physical worlds. In practical applications, AI will be able to handle more complex and realistic scenarios. For example, in “visual quality inspection”, future systems will not only identify defects, but also combine audio from equipment operation, maintenance log texts and real-time video streams to make comprehensive judgments. In “market insight”, AI will be able to analyze social media images, video content and user comments at the same time, accurately capture consumer emotions and preferences, and generate strategic suggestions with depth. This is not only a technological upgrade, but also a cognitive revolution - AI is understanding the world in a way closer to human cognition and becoming the “super brain” beside enterprises.

(II) Super Agents: From “Executing Commands” to “Taking Initiative”

Large-model-based agent technology is driving AI to evolve from a passive tool into a goal-driven “digital colleague”. It can understand complex instructions, plan tasks independently, call tools and execute actions, truly realizing “AI agency”.

The mode of human-machine collaboration is being reshaped. In SMEs, super agents will become ubiquitous “digital employees”. In “process automation” scenarios, they can not only identify invoices, but also proactively trace missing information, check financial rules, and push approval processes without human intervention. In “supply chain planning", they can monitor external data in real time, assess supplier risks, independently generate and execute alternative plans, and even negotiate with supplier AI systems.

Super agents will integrate dispersed AI capabilities into complete business flows, greatly improve organizational efficiency, and enable SMEs to manage complex tasks with ease.

(III) Embodied Intelligence: From the "Digital World" to the "Physical World"

Embodied intelligence gives physical entities such as robots and intelligent vehicles the ability to perceive, learn and act. It is a key bridge for AI to move from the virtual world to reality, and it is also the most direct expression of "AI + manufacturing".

The core of this trend is the deep integration of AI with the real economy. For example, in "intelligent warehousing", robots can not only execute commands, but also understand vague tasks such as "find the best-selling product", navigate autonomously in complex environments, grasp objects flexibly and collaborate with humans. In "production and manufacturing", industrial robots can complete precision assembly through visual and tactile perception, and even learn from demonstrations to adapt quickly to changes in production lines.

Embodied intelligence combines the decision-making power of AI with the execution capability of the physical world, truly promoting the intelligent upgrading of real-economy industries such as manufacturing and logistics.

(IV) AI Infrastructure: From "Technical Barriers" to "Inclusive Services"

The popularization of any technology depends on infrastructure support. AI infrastructure is developing rapidly toward intensification, inclusiveness and integration, including high-performance computing power, easy-to-use development platforms and compliant data markets.

Its core value lies in substantially lowering the threshold for AI application, turning the technology from a "strategic national instrument" into a "tool available to SMEs". For example, a "cloud-based collaborative simulation platform" enables SMEs to use powerful computing and simulation capabilities on demand without heavy investment. More "small, fast, lightweight and precise" AI applications will become as readily available as water, electricity and gas. AI infrastructure is the soil in which the first three trends take root, and it is the fundamental guarantee for AI to move from "showrooms" to "commodity housing".

The four trends do not exist in isolation, but form an organic whole that promotes and evolves in coordination. Multimodal large models are the "brain" of perception and cognition, super agents are the "hands and feet" of planning and action, embodied intelligence is the "extension" of the brain, hands and feet in the physical world, and AI infrastructure is the "body and bloodline" that supports all efficient operation. Together, they outline the complete picture of the future intelligent economy. Built on strong and inclusive AI infrastructure, multimodal models deeply understand the world, super agents autonomously manage processes, and embodied intelligence ultimately creates real value in the physical world.

To seize these four trends is to seize the future of AI. Shanghai, together with its vast number of SMEs, is standing at the forefront of this transformation.

III. Reflections and Prospects on AI Empowering SME Development

To comprehensively present the achievements of Shanghai in using AI to empower SMEs, the Shanghai Municipal Commission of Economy and Informatization has guided and launched the compilation of AI Pioneer - Outstanding AI Application Cases of Shanghai SMEs. This case collection brings together excellent practices from many SMEs and enabling institutions. It is not only a summary of phased achievements, but also provides a replicable "Shanghai model" for the deep integration of AI technology with the real economy. Looking ahead, AI empowerment of SMEs will develop as follows.

(I) Implementing Intelligent Applications to Solve Enterprise Development Bottlenecks

The core value of the case collection lies in its solid practical foundation. All selected cases are not conceptual ideas on paper, but real projects that have been implemented and have generated quantifiable benefits. They accurately target the practical pain points and bottlenecks faced by SMEs in daily operations, and use deep AI applications to reduce costs, improve efficiency and enhance core competitiveness.

The case collection focuses on showing how SMEs, as “users” of AI technology, transform frontier technologies into powerful tools for solving their own development challenges. These applications do not pursue large and comprehensive systematic transformation; instead, they make precise breakthroughs in key links in a “small, fast, lightweight and precise” way. This reflects a high degree of practicality and operability, and provides highly valuable implementation paths for similar enterprises.

In production and manufacturing, the application of “visual quality inspection tools” effectively addresses problems in traditional quality inspection, such as low efficiency, fatigue and inconsistent standards. Enterprises can train defect recognition models independently by uploading product images, enabling fast, accurate and automated inspection of product appearance defects. This not only reduces the missed inspection rate to an extremely low level and greatly improves consistency in product quality, but also frees human workers from repetitive labor and allows them to engage in work of higher value, directly responding to the urgent needs of manufacturing SMEs for higher production efficiency and better quality control.

In supply chain management, the “supplier dynamic rating system” directly addresses risk management and optimization in procurement. Through AI algorithms, the system comprehensively considers multiple dimensions such as supplier delivery punctuality, quality qualification rate and response speed, and establishes a dynamic and quantitative evaluation system. It can automatically identify and eliminate low-performing suppliers at the bottom of the ranking, help enterprises optimize supplier structures, enhance supply chain resilience and reliability, and effectively resolve procurement risks caused by information asymmetry and rough management in SMEs.

(II) Empowering Industries with Digital Intelligence to Stimulate SME Innovation

The case collection builds a diversified and collaborative empowerment ecosystem. It shows how technology enterprises, research institutions and other enabling parties provide advanced and applicable AI products and services to SMEs. These enablers are not only technology suppliers, but also “partners” in the digital transformation of SMEs. By providing lightweight solutions, platform-based technical tools and professional consulting services, they greatly reduce the technical threshold and financial cost for SMEs to apply AI, thereby stimulating the internal innovation momentum of a vast number of SMEs. This “teaching people how to fish” model promotes the intelligent upgrading of the entire industrial ecosystem and forms a sound pattern of integrated innovation and common development among large, medium-sized and small enterprises.

In product R&D and design, the “cloud-based collaborative simulation verification” platform is a typical enabling tool. For SMEs with limited resources, purchasing expensive simulation software and computing resources is a major expense. Based on cloud platforms, this platform provides AI tools such as finite element analysis, allowing SMEs to conduct in-depth simulation verification of design schemes at lower cost, automatically identify stress concentration points and areas of material waste, and offer optimization suggestions. This is equivalent to equipping each SME manufacturer with a top-level expert team, greatly enhancing R&D innovation capability and product reliability.

Part II
AI Empowerment

The dozens of selected cases in this part jointly depict a magnificent picture of deep AI empowerment. They vividly demonstrate the value leap from data insight to business decision-making and from technical tools to core competitiveness. Here, AI is no longer an external add-on tool, but an "intelligent hub" embedded within enterprises, driving systematic reconstruction of product R&D, production and manufacturing, supply chain management, and even business models. We can see that, in high-end manufacturing, the integration of AI with industrial knowledge has given rise to a new model of predictive maintenance services.

- Market Insight and Demand Analysis (Cases 1-11)
- Product Design and R&D (Cases 12-17)
- Supply Chain Planning and Procurement (Cases 18-19)
- Production, Manufacturing and Quality Inspection (Cases 20-26)
- Market Sales and Channel Management (Cases 27-28)
- After-Sales Service and Customer Operations (Cases 29-31)
- Human Resource Management and Process Optimization (Case 32)
- Others (Cases 33-41)

Part II AI Empowerment

The dozens of selected cases in this part jointly depict a magnificent picture of deep AI empowerment. They vividly demonstrate the value leap from data insight to business decision-making, and from technical tools to core competitiveness. Here, AI is no longer an external tool, but is internalized as the “intelligent hub” of enterprises, driving systematic reconstruction in product R&D, production and manufacturing, supply chain management and even business models. We can see that in high-end manufacturing, the integration of AI with industrial knowledge has given rise to a new model of predictive maintenance services.

- Market insight and demand analysis (Cases 1-11)
- Product design and R&D (Cases 12-17)
- Supply chain planning and procurement (Cases 18-19)
- Production, manufacturing and quality inspection (Cases 20-26)
- Marketing, sales and channel management (Cases 27-28)
- After-sales service and customer operations (Cases 29-31)
- Human resource management and process optimization (Case 32)
- Others (Cases 33-41)

Case 1

A High-End Commercial Film and Video Production System Based on Intelligent Large-Model Scheduling and Workflow Optimization

Shanghai Micrease Intelligent Technology Co., Ltd.

I. Basic Enterprise Information

Headquartered in Shanghai, Shanghai Micrease Intelligent Technology Co., Ltd. is an AI technology enterprise established in March 2025. Since its establishment, the company has focused on the R&D of its AI platform “Raise AI”. Its team includes R&D professionals who graduated from top universities at home and abroad, senior technical experts in the AI industry, and experienced producers and directors in the film and video industry.

Through its original combined service model of “AI market insight + AI film and video production platform + industry experts”, Micrease not only ensures technological frontier value, but also guarantees that AI-generated content can serve enterprise business objectives with precision, forming an AI application closed loop from “market insight” to “content production”.

Led by technological innovation, Micrease is effectively solving core pain points faced by SMEs in digital transformation, including high marketing costs, difficulties in creativity and slow growth. In the future, Micrease will build an industry-university-research integrated platform ecosystem around “Raise AI”, support the deep integration of Shanghai’s digital economy with the real economy, and contribute to the vigorous development of Shanghai’s AI industry.

II. Case Introduction

(I) Background and Needs

Since its emergence, high-end commercial film and video production has been an important part of brand storytelling. Its traditional model is a heavy industry based on physical execution and intensive resource input. As market competition enters an agile era, this traditional paradigm is facing inherent structural contradictions, namely the “impossible triangle” formed by the conflict among time, cost and creativity.

1. Time constraints: Traditional production is linear and process-based. From location scouting, scheduling and shooting to post-production, the production cycle is often measured in months. Meanwhile, an extremely long production feedback chain makes traditional production processes unable to meet the rapid iteration and agile marketing needs of the modern market.

2. Cost barriers: Cinematic visual quality also means a high resource threshold. As a result, high-end commercial film and video production has long been dominated by large enterprises. At the same time, the heavy-resource input model makes high-quality content difficult to scale.

3. Physical bottlenecks in production: Traditional production can only film “what can be filmed”, and cannot transform abstract concepts into concrete visual language. For example, in traditional live-action shooting, it is difficult to express molecular rupture scenes in the microscopic world, which prevents creators from fully exercising their imagination.

Therefore, in the AI era, the market needs a new content production paradigm to break the “impossible triangle” and achieve three major goals in high-end commercial film and video production:

1. Efficiency revolution: Compress the production cycle from quarters to weeks, and synchronize creativity with production.

2. Cost revolution: Compress production costs exponentially. By using artificial intelligence-generated content (AIGC) to replace costly live-action shooting, part of computer-generated imagery (CGI) production and long post-production processes, exponential cost reduction and efficiency improvement can be achieved.

3. Creativity revolution: Use AI to visualize creativity. With the imagination and visual generation capabilities of AI, macro or micro imagined worlds can be presented as visual spectacles.

(II) Implementation Plan

The project is built on the self-developed "Raise AI" platform. It adopts a technical architecture of "large-model aggregation scheduling + human-machine collaborative feedback", and solves the pain points of non-standardization and difficulty in scaling in traditional film and video production through application programming interface (API) integration and private algorithm optimization. The core logic of the system includes the following three layers:

1. Underlying system architecture: heterogeneous large-model aggregation and intelligent routing. To address the problem that a single model cannot cover different visual styles such as realism, art and anime, the platform builds a heterogeneous model aggregation layer at the bottom.

Model integration: Through API interfaces, the platform integrates global mainstream generative models including Volcengine, Kling AI, Midjourney, Stable Diffusion and Runway.

Intelligent routing mechanism: The system has a built-in natural language processing (NLP) semantic analysis module that structurally parses input requirements, including scene, lighting, motion and physical laws. Based on the parsing results, the scheduling algorithm automatically matches the most suitable model combination. For example, the system automatically distributes "abstract art style" tasks to models specializing in artistic generation, and "physically realistic" tasks to video generation models, thereby realizing task decomposition and parallel processing.

2. Middle layer: retrieval-augmented generation (RAG)-based prompt engineering and knowledge base enhancement. To strengthen the understanding of vague instructions by general large models, the system builds a middleware layer that serves as a "translator" between human natural language and machine visual language.

Structured prompt conversion: Photography, film lighting and composition rules such as the golden ratio and Rembrandt lighting are encapsulated as parameters to ensure the professionalism of generated content.

Industry knowledge base mounting: RAG technology is introduced to supplement vague input instructions with precise descriptions, ensuring the quality of output content.

3. Business logic layer: a reinforcement learning from human feedback (RLHF) closed loop. To address the "blind box" uncertainty of AI-generated content, the project establishes a data feedback loop based on expert aesthetic judgment.

Process data collection: The system fully records behavioral data from professional directors when they select, discard or retain generated materials. Aesthetic model fine-tuning: Expert decisions are transformed into preference datasets for continuous supervised fine-tuning of the model. By learning how professionals judge lighting and composition, the system builds a dynamic "aesthetic model". As the number of delivered projects increases, the usability and accuracy of generated materials improve nonlinearly.

4. Production workflow reconstruction (AI standard operating procedure, SOP): Based on the above technical architecture, the project reconstructs the SOP from strategy to delivery.

Strategy and script stage: AI assists in generating storyboards, directly transforming text scripts into visualized finished-film views and replacing traditional hand-drawn storyboards.

Asset generation stage: The intelligent scheduling engine generates high-fidelity images and video clips in batches, enabling cross-modal generation from static visuals to dynamic videos.

Refinement and delivery stage: Human experts retain decision-making authority over final editing, color grading and emotional expression, forming an efficient collaborative model in which "AI diverges and mass-produces, while humans converge and decide".

(III) A New AI-Based SOP for Film and Video Production

The project breaks the traditional linear serial process of "planning - location scouting - shooting - post-production" in film and video production, and builds a parallel and iterative new production process based on the Raise AI platform.

Figure 1 Comparison between the Raise AI production process and the traditional process

1. Creative visualization stage (from "tacit understanding" to "what you see is what you get"): In the traditional model, scripts and storyboards often remain in text or simple hand-drawn form, creating a huge communication gap. In this project, the team uses the platform to transform abstract creative scripts into high-fidelity "finished-film-level" storyboard views in real time. This change shortens the creative confirmation cycle from several weeks to several hours, realizing "communication means confirmation, and confirmation means production".

分镜描述	时长s	画面文案	分镜画面
黑色开场画面逐渐过渡到科技蓝色光线划破黑暗，一条未来感十足的冷色高科技光流隧道逐步亮起。隧道壁面由无数个蓝色微光粒子组成。隧道顶部形成涟漪状的发光波纹结构，象征着科技的脉动与智能的脉络。	3	——	
道路中心铺设着清晰的发光标线，车道边缘的一道高亮蓝色光轨持续流动，给人一种速度与未来交织的视觉冲击。隧道漩涡尽头幻化出东风双飞燕logo	2	2014-2024 十年造梦 乘风向上	
logo双飞燕元素化成两只燕子(AI生成)，飞入清华标志建筑(AI风格转化)	3	2014创绿·润未来 梦想之车由此启程，探寻汽车绿色未来	
未来感极简高端的黑色通道，首届冠军作品车以第一视角飞速驶过，黑色立柱飞掠而过，LED灯光如脉冲流动，最终停驻舞台中央展示。	5	——	
画面转移来到城市，高架桥上车流如织，呈现出车水马龙、活力蓬勃的城市交通景象。	5	2016车社会·润未来 人车文明共生，探访和谐汽车社会	

Figure 2 Film and video script co-written by Raise AI and industry experts

2. Asset generation stage (from "physical acquisition" to "digital synthesis"): For the most expensive links in traditional shooting, namely live-action location scouting and CGI production, the project uses a large-model intelligent scheduling engine as a substitute. Through structured prompt engineering, hundreds of visual asset clips that conform to physical lighting rules are generated in batches. The system supports concurrent processing and realizes a production capacity leap from "single-point shooting" to "matrix generation".

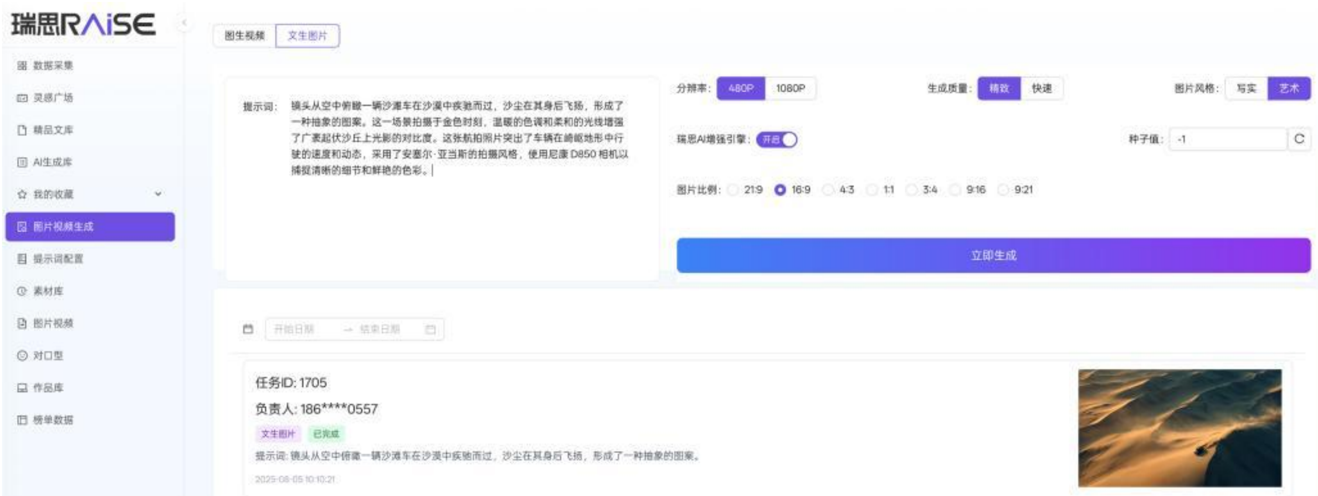


Figure 3　Raise AI creation page

Raise AI can match the most suitable model according to user needs and its self-developed intelligent scheduling technology. For example, Volcengine is better at generating artistic-style images, while Kling AI is better at generating realistic-style images. After suitable parameters are selected on the text-to-image page of Raise AI, the platform intelligently analyzes and optimizes the prompt, and intelligently selects suitable large models to generate images.

Figure 4　Main visual generated by Raise AI

3. Human-machine collaborative refinement stage (from “fully manual” to “AI generation + expert decision-making”): In post-production, a collaborative closed loop of “AI-generated material library - human expert aesthetic screening - professional software refinement” is established. AI undertakes repetitive and time-consuming material production, while human directors return to the more central work of artistic decision-making and emotional editing, ensuring that the final delivered film combines the efficiency of AI with the artistic appeal of human creators. The case can be viewed by scanning the QR code in the original document or searching for the “Micrease AI” WeChat video account.

Figure 5 Raise AI works (scan to view the complete work)

III. Project Highlights

(I) Innovation in Production Paradigm: Pioneering a Parallel Industrial Pipeline Based on "Human-Machine Coupling"

Facing the shortcomings of traditional film and video production, which relies heavily on linear processes and physical shooting, the project builds an end-to-end industrial pipeline different from the single AIGC auxiliary tools available in the market. By processing "creativity - storyboard - asset generation" in parallel, it realizes a paradigm shift in film and video production from a "handicraft workshop" model to an "intelligent factory" model. It enables the implementation of a full-process SOP for AIGC technology in vertical film and video scenarios.

(II) Innovation in Technical Route: A "Middle-Platform" Architecture that Breaks Through the Bottleneck of a Single Model

To address the problems of limited style and poor controllability in the commercial delivery of AIGC content produced by a single general large model, such as relying only on Kling AI or Jimeng, the project proposes the innovative technical route of a "heterogeneous large-model scheduling middle platform". Instead of depending on the capability ceiling of a single model, it uses an intelligent scheduling algorithm to aggregate the advantageous capabilities of top global models. Through an architecture that draws on the strengths of many models, the system ensures optimal delivery capability for complex commercial needs while maintaining technical compatibility and scalability.

(III) Innovation in Implementation Capability: An Expert Feedback Loop that Solves the "Uncontrollability" Problem

AIGC-generated content is highly random, often fails to meet commercial standards and is difficult to use for accumulating industry experience. In response, the project is among the first to introduce a human-machine collaborative mechanism into commercial content production. By using expert experience to correct the immediate output of AI, and by transforming implicit "director aesthetics" into explicit "data assets", it enables continuous model evolution and private aesthetic alignment, providing a feasible technical path for AIGC to enter high-end commercial application fields.

IV. Implementation Results

Since its formal commercialization in April 2025, the Raise AI platform has focused on content digital transformation in high-end manufacturing and the fashion industry, and has gradually expanded to e-commerce and other fields. This section uses Dongfeng Motor Corporation's "Dream Car Night" project as a typical sample. Based on actual execution data, it explains the practical delivery capability of the system in industrial-grade application scenarios from three dimensions: economic benefits, production efficiency and model transformation.

(I) Measured Benefits of a Typical Case: Brand Film Production for Dongfeng Motor's "Dream Car Night"

In October 2025, during the production of the brand film for the seventh "Dream Car Night" of Dongfeng Motor Corporation, the client required both "international-level audiovisual standards" and "agile delivery". The project replaced the traditional live-action shooting and CGI production model with the Raise AI film and video production process. By comparing actual execution data with traditional production quotas, the project achieved significant quantitative breakthroughs in cost control and delivery cycle.

1. Financial cost dimension: marketing budget savings exceeded 80 percent.

Under the traditional model, producing a high-end promotional film of the same level for an automobile enterprise requires high costs for venue rental, fleet scheduling, physical set construction and CGI special effects production. In this project, hundreds of high-fidelity virtual visual assets generated by AI completely replaced physical shooting and part of CGI modeling, directly cutting more than one million yuan in rigid production expenditure and reducing direct costs for a single project by more than 80 percent.

2. Time cost dimension: the delivery cycle was shortened by 80 percent.

According to industry standards, producing a commercial-grade film of the same scale usually takes 10 to 12 weeks, or about 70 to 84 days. Thanks to the concurrent generation capability of the heterogeneous large-model platform and the parallel workflow, this project took only 18 days, or 2.5 weeks, from creative script confirmation to final film delivery. Compared with the traditional model, delivery efficiency increased fivefold, effectively supporting the client's agile marketing needs in intense market competition.

(II) Industry Demonstration Value: Establishing a Standardized Paradigm for Commercial Delivery in the AI Era

The project not only completed the delivery of a single commercial project, but also explored and verified a replicable digital production standard for the film and video production industry in the AI era, with significant industry demonstration value.

1. Defining a new "human-machine collaboration" production relationship: The project verifies the feasibility of transformation from a "labor-intensive" model to a "computing-power-intensive" model. In the new production relationship, the functions of traditional film and video roles such as cinematographers, lighting engineers and modelers are partially replaced by algorithmic models, while the role of the creative director is transformed toward that of an "AI prompt engineer" and "aesthetic decision-maker". This model shows the industry how AI toolchains can free humans from complicated repetitive labor and enable them to focus on core creativity and emotional expression.

2. Establishing an SOP for AIGC commercial delivery: To address the common "blind box randomness" and "non-standard product" problems of AIGC content in commercial applications, the project has developed a complete SOP covering "prompt structuring standards", "asset generation quality specifications" and "human-machine interaction collaboration flows". This provides a practical reference for transforming AIGC technology from an auxiliary "toy" into an industrial-grade "productivity tool", and promotes the standardization and scaled production of non-standard creative content.

(III) Social Benefits and Industrial Strategic Significance

1. Practicing "new quality productive forces" and promoting the integration of the digital and real economies.

This project is a typical practice of using AI technology to empower the real economy, namely the automobile manufacturing industry. By applying frontier generative AI technology to the brand marketing link of traditional high-end manufacturing, it effectively reduces the marketing costs and digital transformation threshold of real-economy enterprises. It not only responds to the national call for the "AI Plus" initiative, but also vividly demonstrates how new quality productive forces enhance high-value-added links in the industrial chain through technological innovation.

2. Making technology inclusive and stimulating the innovation vitality of SMEs.

Although this project is applied to a large enterprise, the low-cost production model verified by Raise AI has more far-reaching significance for SMEs. It greatly lowers the threshold for producing high-quality video content in terms of both capital and technology, enabling SMEs and start-up brands to obtain international-standard brand content with very low budgets. This inclusive application of technology helps optimize the business environment, stimulate the innovation vitality of market entities, and promote balanced development of Shanghai's digital economy and creative industries.

3. Promoting the digital upgrading of modern services.

Film and video production, as a typical knowledge-intensive and creative service industry, has long faced problems such as low digitalization and excessive dependence on manual experience. The successful implementation of this project provides a reference path for the digital transformation of modern services such as cultural creativity, advertising and media, and helps improve the production efficiency and international competitiveness of China's cultural service trade.

(IV) Conclusion

The commercialization of the Raise AI platform is not only a technical reconstruction of the traditional film and video production process, but also powerful evidence that generative AI is moving from "experimental exploration" to "industrial-grade delivery".

Through engineering practice, this case proves that the systematic architecture of "heterogeneous large-model scheduling" and an "expert feedback loop" can fully solve the "non-standard" and "uncontrollable" problems faced by AIGC in commercial applications, and can transform the previously irreconcilable "impossible triangle" of efficiency, cost and quality in traditional production into quantifiable production dividends.

Looking ahead, Micrease will continue to consolidate this digital content infrastructure and promote the evolution of AI technology from a simple "efficiency tool" to a deep "industry-native capability". While continuously lowering the threshold for high-quality content production, it will provide deterministic and inclusive technical support for the deep integration of the digital economy and real industries in Shanghai and across China.

Case 2

From Passive Defense to Active Security: Practical Application of Combat-Oriented Intelligent Security Checkup Capabilities

Shanghai ASants Technology Co., Ltd.

I. Basic Enterprise Information

Shanghai ASants Technology Co., Ltd., referred to as ASants, was founded in April 2021 and is an innovative intelligent security enterprise. The company provides innovative security products, services and solutions based on practical attack-defense scenarios. ASants advocates AI-driven active security, helping enterprises gain comprehensive insights into their exposure in the digital era, establish a security validation framework from the attacker's perspective, infer the path of least resistance, build a "point-line-surface" attack graph covering digital assets, risks and threats, proactively foresee risks, and drive effective security risk governance.

ASants mainly serves industries such as finance, telecommunications operators, energy, government, science, education, culture, health and enterprise sectors. It has established support and cooperation with institutions including the Cyberspace Administration of the CPC Shanghai Municipal Committee, the Shanghai Municipal Public Security Bureau, the Shanghai Communications Administration, Shanghai Computer Network Emergency Response Technical Team/Coordination Center (SHCERT), Jiangxi Computer Network Emergency Response Technical Team/Coordination Center (JXCERT) and the China National Vulnerability Database of Information Security (CNNVD).

ASants is a national high-tech enterprise and a Shanghai "specialized, refined, distinctive and innovative" SME. It has been included in the 2025 Shanghai Cybersecurity Industry Innovation Breakthrough Achievement Catalogue of the Shanghai Municipal Commission of Economy and Informatization. It was named a representative vendor for adversarial exposure validation (AEV) in the Gartner 2025 Hype Cycle for Security Technologies in China, and a representative vendor in the field of breach and attack simulation (BAS) in the Gartner 2024 and 2023 Hype Cycle for Security Technologies in China. It was also recommended as a BAS vendor in IDC TechScape: 2024 China Cybersecurity Software Technology Roadmap, and received five-star ratings in BAS and external attack surface management (EASM) in the 2024 China Attack Surface Management Vendor Assessment, among other authoritative domestic and international honors in cybersecurity. The company has completed two rounds of financing, namely angel round and Pre-A round, totaling tens of millions of yuan. Its investors include Technology Innovation Center for Startups, TusStar, Weizhi Digital Technology and Xiaomiao Fund.

II. Case Introduction

(I) Background and Needs

In the wave of digital transformation, cybersecurity threats are changing significantly. Traditional passive defense methods can no longer cope with increasingly complex cyberattacks. Cybersecurity is now facing new challenges as AI technology is being abused by attackers. A Gartner survey of 286 senior enterprise risk managers revealed the five major emerging risks in the third quarter of 2024, and 80 percent of respondents expressed concern about cyberattacks that AI may bring. The ransomware group FunkSec emerged at the end of 2024 and disclosed more than 85 victims in December 2024, surpassing many traditional extortion organizations.

AI has already become one of the most powerful weapons for attackers, and the external threats faced by enterprises are constantly evolving. At the same time, enterprises are continuously advancing digital construction, and new exposures keep appearing, creating a dual impact on enterprise security development. How to use AI to improve enterprise security capabilities, build a new generation of active security system featuring dynamic perception and intelligent confrontation, and effectively respond to current and future cybersecurity challenges has become an urgent issue.

Against this background, a large financial technology enterprise worked with ASants to build an AI security checkup agent and, together with its existing active security management platform, establish a regular and efficient security operation system.

By building an active defense system with the AI security checkup agent, the project adopts exposure analysis and harmless attack simulation under an EASM + BAS integrated architecture. It shifts from “passive security” to “active security”, combines the reasoning analysis and content generation capabilities of AI large models, empowers multiple links of active security, and moves the security defense line to a point before attacks occur.

(II) Implementation Plan

The ASants AI security checkup agent mainly focuses on critical infrastructure customers in industries including finance, telecommunications operators, energy, government, science, education, culture, health and enterprises. Based on AI, it provides risk perception capabilities ahead of threats and verification capabilities for security protection effectiveness. It ensures that data security is verifiable and quantifiable, and that the efficiency of security operations is effectively improved. Through continuous attack surface discovery and vulnerability assessment, and by simulating and identifying the path of least resistance from the attacker’s perspective, the system uses authoritative evaluation standards as benchmarks, covers the complete attack chain, and combines highly customizable attack templates to assess target network data assets.

Its core service concept is “AI active security: an AI security agent ahead of threats”. It moves from Software as a Service to Result as a Service, delivering results rather than tools.

Figure 1 AI security agent ahead of threats

III. Project Highlights

(I) Key Technological Breakthroughs

1. Attack-defense knowledge base

Unstructured attack-defense knowledge, such as disposal reports and analysis reports, is stored in vector databases, knowledge graphs and other systems. Retrieval-augmented generation (RAG) technology is used so that large models can understand this knowledge and recall relevant information according to prompts to enrich the context of the chain of thought, thereby improving the accuracy and relevance of answers.

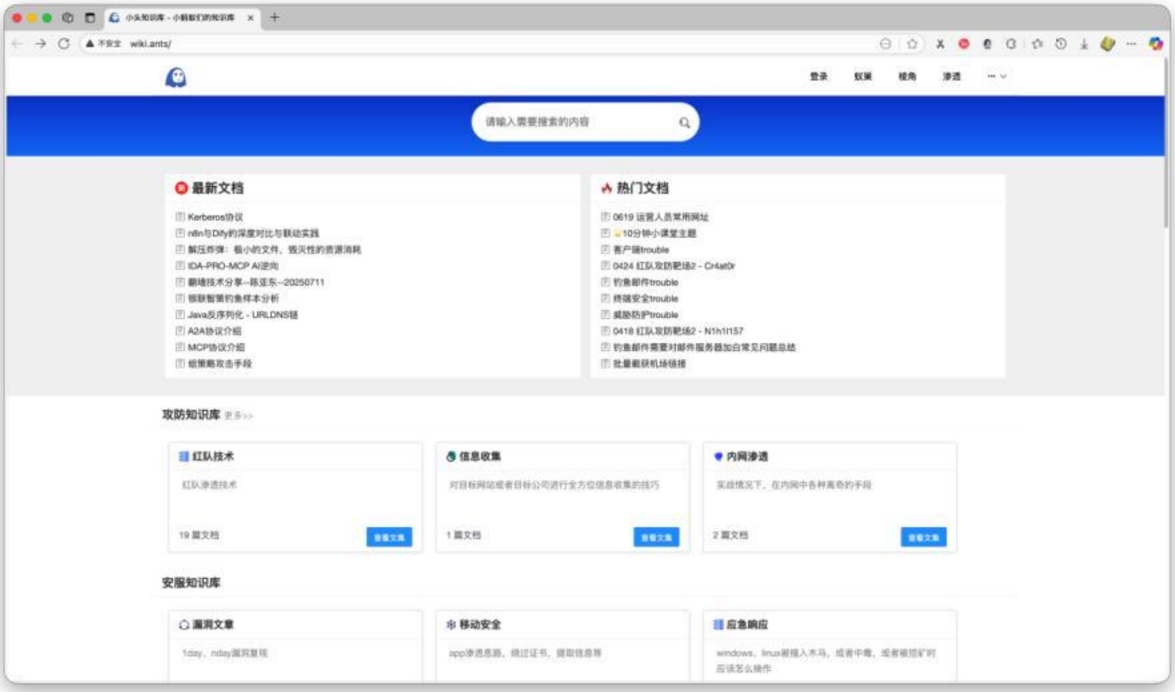


Figure 2 Attack-defense knowledge base and AI Q&A assistant

Traditional security defense has long faced problems of fragmented knowledge and isolated experience. Large volumes of unstructured data, such as security incident disposal reports, vulnerability analysis documents and attack traffic logs, are scattered across different systems and depend on manual experience for retrieval and connection. By building a multimodal attack-defense knowledge base and combining vector storage with knowledge graph technology, the AI system can break through the cognitive limitations of human experts. In the vector database, text data such as security reports, vulnerability characteristics and attack chain descriptions are transformed into high-dimensional vectors through semantic embedding models, forming a quantifiable and retrievable "memory network". Knowledge graphs extract entity relationships from massive data, such as attacker organizations, vulnerability exploitation methods and defense strategies, and construct a logical association network of attack-defense tactics.

2. Security checkup agent assistant

In complex attack-defense scenarios, the core value of the agent assistant lies in transforming fragmented clues into structured decision-making paths. Through conversational interaction, the system can dynamically decompose the problem space, guide defenders along the optimal path for analysis, and explain judgment logic in real time, forming a traceable reasoning chain. In multimodal intent understanding, the conversational agent assistant engine converts natural language queries into structured tasks. For example, when an enterprise proposes to "assess ransomware risk", the system activates multiple dimensions such as vulnerability exposure detection and security protection effectiveness, and can also enable ransomware drill plans to verify and improve employees and security teams. The system then dynamically adjusts the weight of each dimension according to the dialogue content.

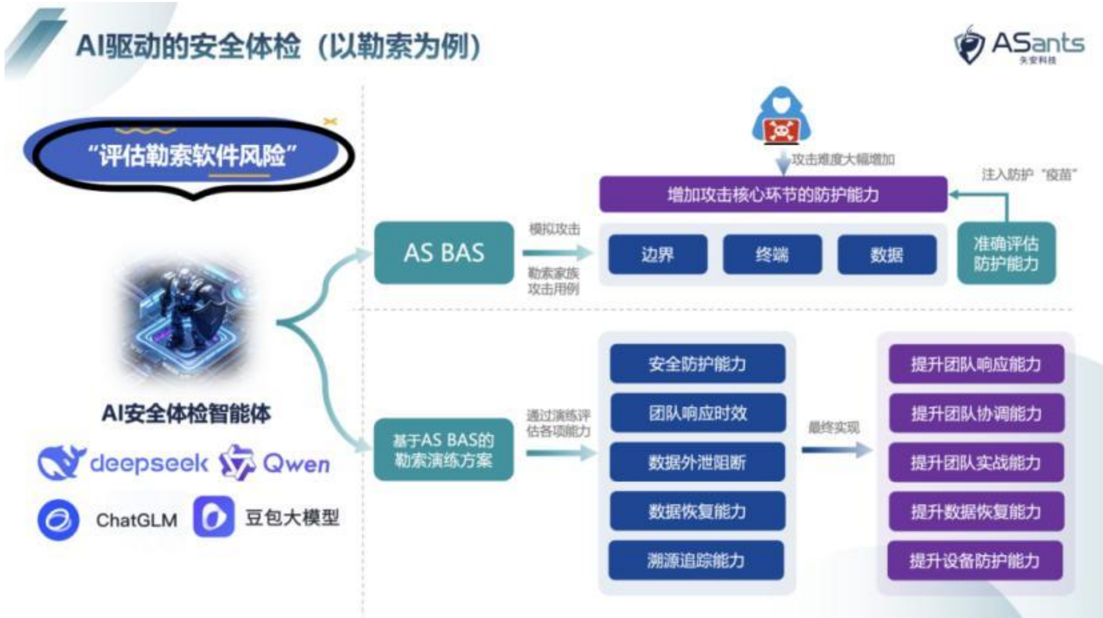


Figure 3 AI security assistant calling BAS capabilities for ransomware assessment

The essence of this dynamic guidance mechanism is to transform the implicit experience of human experts into machine-executable decision flows. Through the progressive advancement of conversational interaction, the agent assistant not only outputs conclusions, but also helps operators iteratively improve their defense capabilities and become “stronger through every battle” when confronting attacks.

3. Attack orchestration AI agent

External attack methods are rapidly emerging under the empowerment of AI and are constantly challenging enterprise security systems. Attack orchestration uses the same approach. Based on harmless atomic attack use cases that have already been processed on the ASants BAS platform, and combined with the user network environment, it quickly generates effective attack kill chains to verify enterprise security capabilities, using an AI shield to counter an AI spear.

The attack orchestration AI agent needs network environment perception, strategy generation and dynamic evolution capabilities. Its breakthrough lies in upgrading the template-based configuration, delivery and execution of traditional security products into real-time game-based decision-making capabilities according to changes in the user network environment. It decomposes complex attack-defense scenarios and attack chains into atomic action sequences and intelligently orchestrates them.

Figure 4 Schematic diagram of the attack orchestration AI agent

Through the capabilities of EASM and BAS products, combined with historical data, the system automatically detects the user network architecture. Attack path deduction is carried out based on the user network architecture, mainly through the coordination of asset mapping, knowledge-base-enhanced learning and environmental perception. The internet-exposed surface information discovered by EASM and the attack use case knowledge base of BAS together constitute the basic input layer for attack path deduction. The agent models asset nodes, such as servers, APIs and databases, and risks as a dynamic attack graph, and uses reinforcement learning to simulate multi-objective decision-making by attackers, such as the shortest attack path.

For example, in a cloud scenario, when EASM discovers an unpatched unauthorized access vulnerability in a Kubernetes cluster, the agent can call related attack patterns from the BAS knowledge base, generate potential attack paths based on network topology, calculate the priority of each path, and verify them in sequence according to priority.

4. Other AI-empowered technical links

In AI + BAS products, the capabilities of AI large models play a role in many technical links. In particular, the reasoning and content generation capabilities of AI large models can greatly improve BAS product capabilities and the operational efficiency of attack use cases. For example, in data denoising for asset discovery, a large number of false positives may appear during external asset scanning and may bury truly sensitive documents. AI large models can analyze the sensitivity of document content and prioritize sensitive documents. For vulnerability prioritization, self-developed AI models can dynamically evaluate the real risk of each vulnerability by combining multidimensional information, including Common Vulnerability Scoring System (CVSS) scores, reachable exploit paths and the business importance of assets. For AIGC phishing and voice scenarios, multimodal AI large models can automatically generate phishing copy and voice content; AI coding capabilities can be invoked to quickly generate imitation phishing websites; and automated phone calls or text messages can be sent to targets for validation.

(II) Technological Innovation

1. Attack surface management technology

Using attacker reconnaissance technology and AI monitoring capabilities, the project builds a continuously monitored external attack surface system based on an AI risk data lake. By entering the corporate entity name, the system can complete reconnaissance of externally exposed digital assets, such as domains, subdomains, IP addresses, leaked data, email accounts, social media, associated threat intelligence, supply chains and third-party services, in a short time, and update the results daily. At the same time, the system builds powerful and comprehensive risk scanning technology, which can identify risks related to vulnerabilities, weak passwords, configurations, social engineering and sensitive information, thereby achieving comprehensive management.

Innovation: Using attacker reconnaissance technology and AI monitoring capabilities, the project builds a continuously monitored external attack surface system based on an AI risk data lake.

By using large models for deep semantic analysis of massive unstructured data, such as network traffic, SSL certificates, DNS records and code repository logs, the system automatically identifies unregistered assets, such as shadow APIs and abandoned subdomains, and intelligently distinguishes asset categories, such as IoT devices, cloud services and container instances. This realizes panoramic asset visualization and continuous updating. By leveraging the advantages of large models in data analysis and understanding, the system discovers unknown assets from unstructured data, improving the discovery rate by more than 10 percent; automated classification can reduce the manual cost of asset operations by 80 percent.

By combining CVSS scores of vulnerabilities, asset exposure data and real-time threat intelligence, the system uses reinforcement learning algorithms to dynamically calculate the probability of vulnerability exploitation and automatically generate risk heat maps and security rating scores. This helps security teams quickly focus on high-risk assets and vulnerabilities, and achieve risk prioritization. Through reinforcement learning and real-time integration of intelligence data, the system dynamically evaluates the actual exploitation probability of vulnerabilities, substantially improving the timeliness and accuracy of risk detection.

With the help of a large-model knowledge base and real-time online updating capability, detected vulnerabilities are automatically aggregated and deeply analyzed. The system extracts actionable remediation strategies and optimization suggestions, and supports one-click generation of remediation plans, assisting enterprises in rapid response and shortening remediation cycles.

The system uses large models to rapidly analyze collected vulnerability traffic and code document content, filter false positives caused by limitations of plug-in rules, and perform graded data denoising to highlight key security information, further improving detection accuracy and operational efficiency.

2. Automated security validation technology

Automated security validation is a technology that automatically simulates known real-world hacker attack methods, namely tactics, techniques and procedures (TTPs), analyzes defensive responses, and identifies which defenses are effective or ineffective, thereby helping organizations improve their overall defense capabilities.

ASants pioneered harmless attack simulation technology. Through a method of simulated attack training combining human expertise and AI, it preserves real threat characteristics while ensuring that the attack simulation process is harmless and non-destructive to the business environment. This technological breakthrough gives automated security validation the necessary conditions for commercial deployment.

Innovation: AI-based simulated attack training.

Large language models (LLMs) are used to automatically generate phishing email content and phishing pages according to user scenarios. Compared with traditional page cloning and fraudulent content generation, efficiency is improved by at least 80 percent.

LLMs are used to automatically interpret key defects and problems in task data.

Based on key information such as samples and proof of concept (PoC), the system can automatically generate corresponding attack use cases according to the attack use case format. In operations and attack use case generation, AI analyzes the attack methods of advanced persistent threat (APT) organizations, extracts and summarizes key attack techniques, and automatically generates attack use cases. Compared with traditional manual scripting, efficiency is improved by at least 80 percent.

According to the log formats of target security devices, the system automatically matches corresponding fields for log matching. AI can automatically identify the correspondence between security device log fields and BAS log fields, improving efficiency by more than 80 percent.

3. Attack path assessment technology

Attack path assessment technology uses AI recognition capabilities to build an asset graph. Based on data such as asset risk vulnerabilities and security defense defects, it establishes attack paths from the initial intrusion point to the final attack target, evaluates the exploitability, success probability and potential damage of each attack path, and calculates the difficulty and cost of carrying out attacks. At the same time, it continuously monitors network changes and updates attack path models in real time to respond to newly emerging security vulnerabilities and network configuration changes, ensuring that path management always reflects the current security posture.

Innovation: With generative AI at the core, the technology evaluates the effectiveness of responses to real-time threats and enables an in-depth understanding of enterprise cybersecurity risks and necessary rectification measures. By combining the difficulty and benefits of implementation from the "attacker" perspective, it derives the path of least attack resistance and realizes risk-driven vulnerability management.

IV. Implementation Results

(I) Economic Benefits

In the financial cloud platform of a state-owned financial group, ASants used the method of "attack surface management + active security validation + attack path assessment + continuous rectification closed loop" to help build a unified security evaluation system and achieved significant results.

In the financial cloud platform, the effective protection rate of security protection products increased from less than 30 percent to 80 percent. Three truly exploitable attack paths were identified and remediated in full-link attack simulation. MITRE ATT&CK detection coverage increased by about 80 percent, forming linked detection. Multiple rounds of link reproduction improved overall defense-in-depth capability by 80 percent. Relying on a unified quantitative system, more than 70 percent of subsidiaries shifted from passive security to active security, forming a group-level governance closed loop. Through efficiency improvement and reduction of redundant investment, security operations and service costs were reduced by about 40 percent, saving about 8 million yuan. With improved protection rates reducing potential losses from security incidents, the overall economic value reached the level of tens of millions of yuan.

For enterprise customers with limited security budgets, the implementation of the AI security checkup agent will greatly reduce the need for professional personnel, making security checkup services more inclusive. Cybersecurity checkup centers jointly operated with supporting units of regulatory authorities have already launched pilots in three cities, directly delivering checkup results and improvement recommendations to enterprises. After promotion, each major city is expected to serve 500 to 1,000 enterprises annually, corresponding to operating revenue of 10 million to 20 million yuan. At the same time, the coverage of cybersecurity checkup centers will improve the overall security level of the industry and reduce economic losses caused by cybersecurity incidents.

(II) Social Value

1. Improving the overall security level of society and strengthening national security and social stability.

The AI security checkup agent can actively and in real time identify and respond to potential threats. Especially in key infrastructure fields such as finance, electricity and transportation, compared with traditional manual methods that are limited in capability and time, AI technology can more accurately discover vulnerabilities through data analysis and self-learning, reduce the disruption of cyberattacks, prevent data leakage and system paralysis, lower enterprise security losses, and reduce threats posed by security incidents to national critical infrastructure. Compared with traditional defense, the AI security checkup agent has intelligent identification and dynamic adjustment capabilities, effectively supports the national cybersecurity protection system, reduces potential losses, and significantly improves the overall cybersecurity level of society.

2. Improving social emergency response capability.

The AI security checkup agent can detect and rapidly respond to cyber threats in real time. Unlike traditional passive defense, it helps enterprises and governments respond quickly by proactively identifying risks, preventing incidents from expanding and reducing losses. Especially in high-risk industries such as finance and healthcare, it improves the overall emergency response capability of society and reduces the impact of cybersecurity incidents on social order.

Case 3

A Banking Intelligent Operations Agent Based on Hierarchical Causal Inference and Decision AI

Shanghai Datatist Information Technology Co., Ltd.

I. Basic Enterprise Information

Shanghai Datatist Information Technology Co., Ltd. (hereinafter referred to as "Datatist") was founded in 2016 with a registered capital of RMB 45 million. It is a private, unlisted enterprise. After its Series B financing, its current valuation is RMB 640 million. Its controlling shareholder is Dr. Song Bilian, and it currently has approximately 142 employees. The company's core founding members have top-tier international doctoral research backgrounds, nearly 30 years of hands-on experience in overseas frontline Internet and financial industries, and practical experience in AI projects at leading international companies such as eBay and LinkedIn. Its main business is to provide enterprises in the government, financial, retail and corporate sectors with intelligent operations services covering the full customer life cycle, including AI customer acquisition, AI replication, AI promotion, AI benefit optimization and AI cost optimization, so as to reduce costs, increase efficiency, improve operational return on investment (ROI), and respond to national calls for digital transformation. To empower industries, promote the development of new quality productive forces and support relevant industry practices, the company has implemented a number of model cases, including the intelligent operations project for a major state-owned joint-stock commercial bank from early 2024 to the end of 2025, the intelligent operations project for a major state-owned commercial bank in 2025, the digital business platform project for a securities firm by the end of 2025, the digital operations project for a Shanghai branch of a major state-owned bank by the end of 2024, and the intelligent operations platform project for a joint-stock commercial bank by the end of 2025. It has successfully empowered more than 600 leading enterprises across industries in China and abroad.

At present, Datatist owns more than 60 invention patents and software copyright certificates. It has been recognized as a national-level specialized, refined, distinctive and innovative "Little Giant" enterprise, and has obtained multiple authoritative certifications in China and abroad, including National High-Tech Enterprise certification, Capability Maturity Model Integration (CMMI) Level 3 certification, Shanghai specialized, refined, distinctive and innovative SME certification, Shanghai Science and Technology Little Giant cultivation enterprise status, participation in Institute of Electrical and Electronics Engineers (IEEE) AI standard formulation, dual software certification, information-technology application innovation equipment adaptation certification, and AAA enterprise credit certification. It has also been listed by People's Daily as one of the top three smart-city AI technology service providers, and has been included in China's Top 50 big-data government procurement white paper and the Top 100 Chinese government big-data development report.

II. Case Introduction

(I) Background and Needs

At present, large state-owned financial institutions, including banks, insurance companies and securities firms, have already built a certain foundation in model capabilities. However, they still face problems such as insufficient systematization, limited intelligence and difficulties in large-scale application when it comes to enterprise-level intelligent decision support for business scenarios. The absence of a unified strategy center and an automated marketing decision-making system has led to low decision-making efficiency and difficulty in continuously improving business outcomes. To address increasingly complex market competition and customer needs, financial institutions urgently need to build a new-generation intelligent operations system that integrates AI prediction, strategy optimization and automated execution.

In key projects such as improving financial asset customer groups, four major challenges are commonly seen: insufficient AI prediction capabilities, inadequate operational strategy granularity, a lack of scientific support for decision-making processes, and the absence of an effective closed loop in existing systems. For this reason, Datatist cooperated with large financial institutions to promote the development of an intelligent operations system around scenarios such as customer asset growth. Through in-depth business insight, data modeling analysis and intelligent strategy design, the project built an enterprise-level AI decision brain and an intelligent marketing strategy middle platform, supporting banks in achieving full-process digital upgrading from customer identification and strategy generation to execution feedback.

During model-system construction, several core difficulties are common: the existing model categories are incomplete, some business scenarios lack suitable expert models, feature identification and model development still rely heavily on manual work, efficiency is low and the development cycle is long, and an automated and platform-based support capability has not yet been formed. Meanwhile, engineering and productization capabilities are insufficient. Many models remain at the experimental stage and cannot be deployed at scale or embedded into business systems. The lack of visual application platforms also raises the threshold for business users and limits the implementation of intelligent operations capabilities in frontline scenarios.

(II) Implementation Plan

Overall architecture design

The project adopts a layered architecture design with “data layer - algorithm layer - application layer” as its core framework. It supports modular expansion and multi-tenant isolation, ensuring stable operation and efficient processing in high-concurrency and big-data environments. The data layer builds an offline data warehouse based on Hadoop + Spark and introduces Flink for real-time data stream processing, enabling the handling of billions of events per day and providing high-quality, multi-source heterogeneous data support for upper-layer applications.

In terms of system scalability and compatibility, the project supports distributed deployment and component-based design, with loose coupling among service modules to facilitate secondary development and functional expansion. The system is compatible with domestic software and hardware environments, including Kylin OS and UnionTech UOS operating systems, EverDB and GaussDB databases, as well as middleware and protocol standards that meet the information-technology application innovation requirements of the financial industry. It also supports IPv4/IPv6 dual-stack network protocols and meets the relevant compliance requirements of the People’s Bank of China.

To ensure high availability and disaster recovery, the project builds cross-region active-active clusters based on Kubernetes, enabling second-level failover and automatic recovery. System availability reaches over 99.99 percent. Through hybrid-cloud deployment and elastic scaling, the system can dynamically adjust resources according to business loads and support real-time responses for hundreds of millions of users and terabyte-level data processing needs.

Key technology implementation

With “third-generation artificial intelligence technology” as its core, the project breaks through multiple key technical bottlenecks and realizes standardized, productized and large-scale intelligent operations in financial scenarios. At the algorithm layer, it integrates advanced methods such as causal inference, deep reinforcement learning and transfer learning, and combines industry knowledge to build data- and knowledge-driven decision models. This significantly improves model interpretability and robustness. By building a financial feature library covering over 100,000 features and a pre-trained model library covering more than 100 scenarios, the system can quickly match and generate suitable models for different financial scenarios.

In model training and deployment, the project adopts hybrid parallel training technology, namely data parallelism plus model parallelism, to support the efficient training of hundred-billion-parameter models on domestic computing clusters. Through model distillation and quantization, AI models are compressed to one-tenth of their original size, meeting the low-power deployment requirements of edge terminals such as customer-manager tablets. A complete model life-cycle management mechanism is also established, including version control, effect monitoring, ethics review and risk warning, to ensure that models remain stable, compliant and controllable during large-scale use.

In terms of data security and compliance, the project uses national cryptographic algorithms for data transmission encryption. Through data desensitization, hierarchical authorization and privacy computing technologies such as homomorphic encryption and differential privacy, it provides full-life-cycle data protection. The system has passed ISO 27001 information security management system certification, supports secure cross-department data integration and joint modeling, and has increased data utilization from 15 percent to 60 percent in projects for multiple financial institutions, with model area under the curve (AUC) improved by 15 to 30 percent.

Figure 1 Datatist Automated Algorithm 4.0

Core module development

The project builds three core modules - an enterprise-level AI decision brain, a department-level “small brain” platform and an intelligent marketing strategy middle platform - forming an intelligent operations system covering the full user life cycle. The enterprise-level AI decision brain includes a financial decision model factory, a feature library management platform and life-cycle analysis management products. It supports model development and strategy generation for financial scenarios such as loans, credit cards and wealth management. This module is equipped with a ready-to-use pre-trained model library and a real-time feature center, enabling full-process automation from data input, feature extraction and model training to strategy output.

The department-level “small brain” platform serves two types of terminal users: mobile banking users and customer managers. It provides intelligent traffic distribution and personalized recommendation capabilities. The platform integrates real-time data collection, a real-time feature center, real-time model training and traffic management functions. It supports A/B testing, negative feedback strategies, cold-start strategies and other intervention mechanisms, achieving real-time response and precise reach for hundreds of millions of users. Through customer-manager portraits, product portraits and an empowerment model center, the system can provide customer managers with 360-degree views and intelligent lead recommendations, thereby improving business development efficiency and customer satisfaction.

The intelligent marketing strategy middle platform integrates a data center, an analysis center and automated marketing products, realizing full-process automation from data access and strategy design to execution feedback. The data center supports unified management and standardized processing of various types of business data, including loans, transactions, benefits and campaigns. The analysis center provides visualization tools such as data screens and dashboards, supporting business personnel in conducting multidimensional analysis and effect tracking independently. Automated marketing products automate marketing processes and strategies through the marketing strategy canvas, realizing a closed-loop operations mechanism of “user insight - intelligent decision-making - one-click action - continuous optimization”, and significantly improving operational efficiency and ROI.

Figure 2 Comparison of Value Creation Before and After Implementation

III. Project Highlights

(I) Business Level

The project focuses on the pain points of intelligent operations in the financial industry. With “third-generation artificial intelligence technology” as its core, it creates an intelligent operations solution for the financial industry driven by the integration of commercial AI decision large models.

By building an intelligent decision-making system covering the full user life cycle and integrating three core modules - the enterprise-level AI decision brain, the intelligent marketing strategy middle platform and the department-level “small brain” platform - the project realizes comprehensive digital reconstruction of financial business processes.

(II) Technical Level

First, it builds a domestically developed, independently controllable commercial AI decision large model for the financial industry, as well as an intelligent operations solution for the financial industry driven by the integration of this model. The system supports real-time responses for hundreds of millions of users and achieves availability of 99.99 percent or above.

Second, it establishes an intelligent decision brain covering a feature library of over 100,000 features and a pre-trained model library of more than 100 models, supporting ready-to-use application in all scenarios.

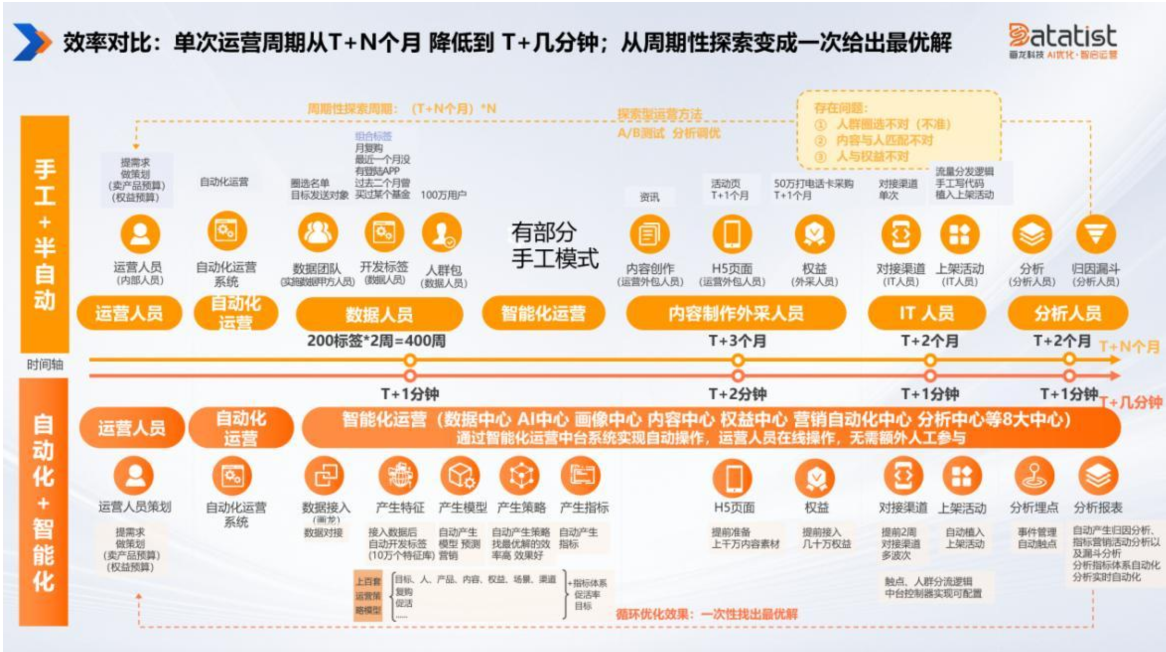


Figure 3 Comparison of Model Development Efficiency

(III) Business Model Level

The business model is innovated at two levels: intelligence, and automation plus integration.

1. Intelligence: The project establishes multidimensional machine learning and embeds hundreds of independently developed AI machine-learning algorithms, including prediction, recommendation and clustering. Relying on the standardized research of commercial AI algorithms, the intelligent decision brain automatically applies commercial AI algorithms in different operational scenarios and configures optimal operational decisions, including audience groups, content, benefits and channels.

2. Automation and integration: From data cloud platform architecture, data analysis, user portraits, opportunity prediction, marketing campaign push and A/B testing to marketing effect tracking and effect analysis, the entire process is automated. This breaks down silos between data and operations, realizes one-stop closed-loop global operations optimization, moves beyond partial optimization, integrates internal and external data and connects internal and external marketing. It truly achieves closed-loop one-stop operations optimization, improves overall operating ROI, and solves the existing problems of poor strategy timeliness and delayed execution in enterprise operations.

IV. Implementation Results

(I) The project began construction in September 2024 and has already achieved initial results after pilot implementation.

In terms of the effect of Datatist's commercial decision AI large model in empowering enterprises' digital-intelligent transformation, one branch of a state-owned bank achieved a monthly net increase of several hundred billion yuan in assets and an annual cumulative increase of several trillion yuan. It helped a major state-owned bank generate hundreds of billions of yuan in wealth-management transaction volume in a single campaign, with a model conversion rate of 7.87 percent. In an asset enhancement project for a provincial branch of a state-owned bank, after AI operations were adopted, the bank's national ranking rose to second place in the entire system, and the project won a national-level award for excellent financial technology cases.

In terms of the efficiency of digital-intelligent transformation, overseas enterprises usually take 10 to 20 years to complete digital transformation. By contrast, a municipal rural commercial bank, relying on Datatist's intelligent middle platform, completed a three-year transformation target within six months. Its operating cycle was compressed from months to minutes, increasing speed by more than five times and reducing operating costs by 97 percent. In empowering a provincial branch of a state-owned bank in personalized customer operations, the system changed the traditional state of “customers waiting for services” and “services unable to find customers”. It automatically matched customer groups, content and channels, significantly improving campaign response rates and marketing conversion rates.

Figure 4 Empowering More Than 600 Leading Enterprises

(II) Benefits Generated After Other Project Implementations

1. Cost reduction and efficiency improvement: Through automated and intelligent operations processes, labor and operating costs are significantly reduced and work efficiency is improved. For example, in a bank’s digital transformation project, operating costs were reduced by 30 percent. For a retail giant, the project built a full-domain intelligent engine that leveraged RMB 300,000 in SMS costs to generate RMB 300 million in new revenue, achieving an ROI of 800. A single livestream generated sales of RMB 60 million within six hours. For a beauty industry client, customer acquisition costs were reduced by more than 40 percent through precise customer grouping and channel matching.

2. Revenue growth: Through precision marketing and customer insight, the project improved product sales conversion rates and customer retention rates, directly driving revenue growth. In multiple projects, revenue growth rates reached 20 to 30 percent. With the support of this system, the annual number of operational campaigns at a large securities company expanded from several hundred to nearly 100,000, equivalent to adding a robot team of several hundred people. For a provincial branch of a state-owned bank, the project supported the completion of an asset enhancement target of over RMB 100 billion.

3. Improved customer satisfaction: Through personalized services and product recommendations, the project improves customer satisfaction and loyalty and strengthens the bank’s brand image.

Case 4

Innovative Practice of a Full-Stack AI Integrated Service Platform for the Semiconductor Industry

JW Insights (Shanghai) Technology Co., Ltd.

I. Basic Enterprise Information

Founded in 2008, JW Insights is a leading domestic technology service platform for the semiconductor industry. Its core product platforms include the JW Insights App and the JW Data Middle Platform. It provides five core services: consulting, brand marketing, information services, intellectual property services and industrial-park empowerment. Deeply rooted in China's three major semiconductor industry clusters - the Beijing-Tianjin-Bohai Rim, the Yangtze River Delta and the Pearl River Delta - JW Insights serves integrated circuit enterprises, investment institutions, government parks and upstream and downstream partners across the industrial chain.

With natural language processing, knowledge graphs and machine learning as its core technologies, JW Insights independently developed China's first vertical large model for the integrated circuit industry - JiweiGPT, an AI digital consulting expert. JiweiGPT enables intelligent analysis, interactive text generation and decision assistance, helping enterprises optimize operational processes and strategic planning. In addition, the company has built the JW Data Middle Platform, which aggregates massive industrial data, realizes full-chain industry data collection, cleaning, integration and value mining, and provides enterprises and governments with efficient data intelligence services.

The company is a National High-Tech Enterprise and a Shanghai specialized, refined, distinctive and innovative enterprise. It has passed ISO 9001 quality management system certification. Its self-developed "interactive text content generation algorithm" was among the first to complete the deep synthesis service algorithm filing with the Cyberspace Administration of China. It has also been selected for projects such as Shanghai's large-model empowerment of productive Internet service platforms and typical application scenarios for vertical large models in Pudong.

II. Case Introduction

(I) Background and Needs

Against the backdrop of rapid global semiconductor industry development, enterprises face multiple challenges. Traditional information retrieval models have obvious defects, massive fragmented data makes it inefficient to capture key information, and general-purpose large models lack specialized knowledge bases in the semiconductor field, resulting in insufficient credibility of output content. Relying on hundreds of millions of industrial data chains accumulated through years of deep cultivation in the semiconductor sector, JW Insights built an AI integrated service platform for this field. The platform aims to solve the industry's pain points of high information barriers, difficult decision-making and insufficient service efficiency, while providing users with full-stack intelligent services from data insight to decision support.

(II) Implementation Plan

1. System architecture design

With data-driven development and AI capabilities as its core, the project builds a complete closed loop from underlying data governance to upper-layer business implementation. It reflects a three-in-one design logic of "data - service - application". The architecture is shown below:

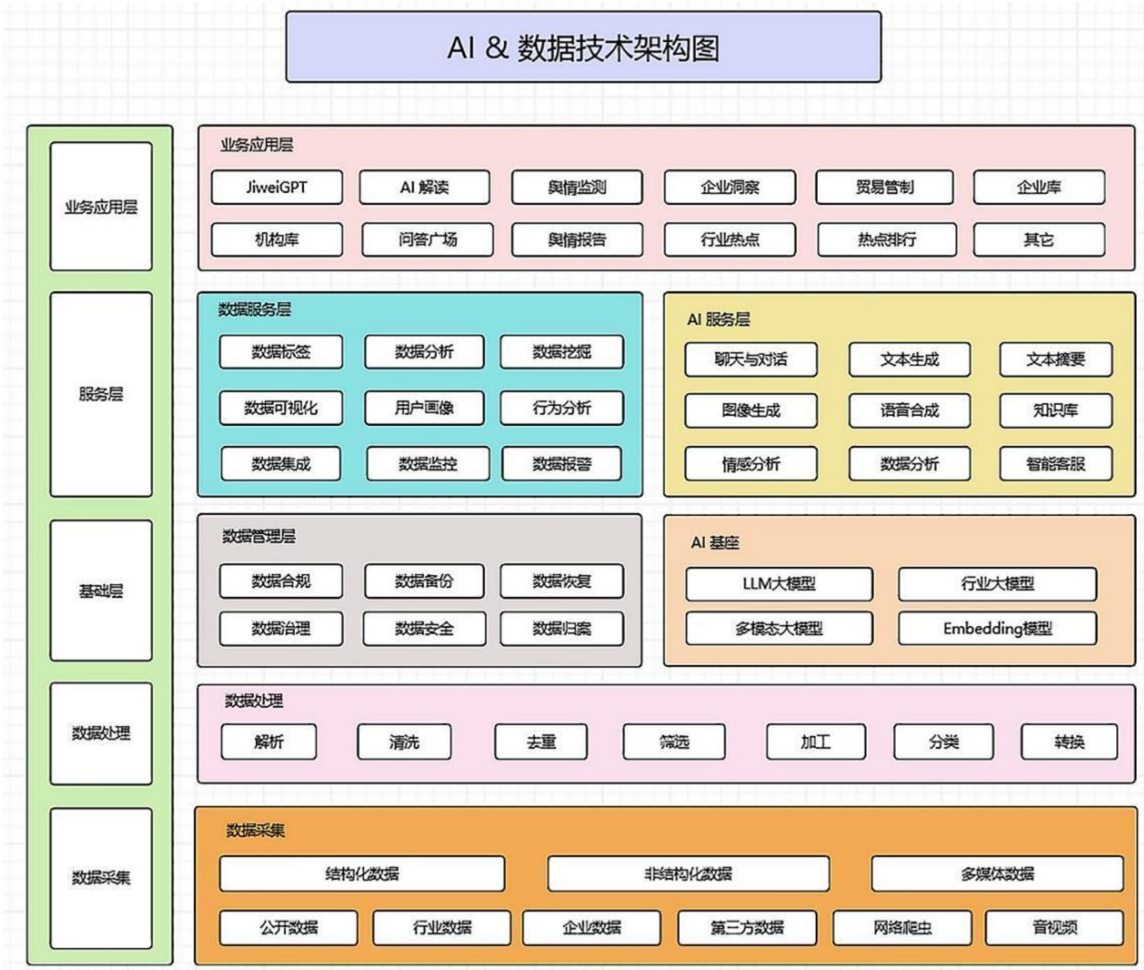


Figure 1 AI and Data Technology Architecture

The business application layer is oriented toward user value and focuses on vertical scenarios such as public-opinion monitoring, enterprise insight and trade control. It directly transforms core AI capabilities into practical business tools such as JiweiGPT conversational interaction and automated report generation, meeting enterprises' key needs in decision support and risk warning. At the same time, through modular design, the enterprise database, question-and-answer system, consulting system and report-generation system can be flexibly combined according to different scenarios, improving the efficiency of business implementation.

The service layer abstracts general technical capabilities such as text generation and user profiling into independent services, avoiding repetitive development. For example, core services such as sentiment analysis can support both public-opinion monitoring and intelligent customer service. At the same time, through data integration, monitoring and warning mechanisms, this layer realizes closed-loop management from data input to feedback application, continuously driving dynamic knowledge-base updating and model optimization.

As the "data source" of the entire technical architecture, the data collection layer is responsible for acquiring raw data from diversified channels. The data it collects is highly diverse, covering structured data, unstructured data, multimedia data, public data, industry data, enterprise data and third-party data. To acquire these data efficiently, the layer adopts highly adaptive collection methods and provides a stable data foundation for upper-layer intelligence.

The foundation layer is committed to infrastructuralizing data and models. From collection of structured and unstructured data sources, to processing such as cleaning and classification, and then to governance such as compliance and security, it forms a standardized data pipeline that ensures data usability and credibility. In terms of model capabilities, it adopts a hierarchical construction strategy: foundation models provide general intelligence, while industry large models are deeply optimized for vertical fields, ensuring both generality and professionalism.

As the core hub for refining data value, the data processing layer undertakes the key task of "refining and processing" raw data and transforming it into high-quality, usable data assets. Together with the data collection layer, it forms the "foundation" of the technical architecture. Its processing capability and quality directly determine the upper limit of upper-layer AI application performance.

The project architecture embodies a spiraling logic of "data - intelligence - business": the foundation layer ensures the continuous supply of data fuel, the service layer refines the intelligent engine, and the application layer converts it into real productivity. It realizes a seamless connection from "raw data input" to "high-value knowledge output". With breakthroughs in large models and multimodal technologies, the platform is expected to evolve from "tool empowerment" toward "ecosystem-level intelligence".

2. Main construction content

Integrated data middle platform construction: The project builds a "nine-dimensional" core database covering the entire industrial chain, including enterprise information, projects and scientific and technological achievements, investment and financing institutions, policies and regulations, market information, technologies and products, industry reports and news, expert databases and image databases. This forms a solid data foundation. The data middle platform not only provides precise data support for industrial development, but also lays the groundwork for intelligent applications.

Agent cluster and application development: By building platform pages and information platforms, the project integrates the independently developed JiweiGPT industry large-model agent and develops AI-driven advanced applications such as intelligent question answering, trend analysis and report generation. It realizes an intelligent upgrade from "people looking for information" to "information finding people". It understands and organizes users' vague consulting needs, guides users through multi-turn interaction to clarify their objectives, and automatically matches them with corresponding professional agents.

Personalized customized service design: The platform automatically matches enterprise consulting needs and calls corresponding professional agents. For scenarios such as production and operations, investment and financing, it generates multiple sets of optimization plans and calculates expected benefits. It supports real-time human-machine interaction, iteratively optimizes decision plans, and supports enterprise development. For policy formulation, park planning and industrial investment attraction, it provides customized functions such as data visualization and industrial risk warning.

Comprehensive commercialization promotion: The project formulates and implements a multi-level market strategy, promoting the platform to industrial enterprises, government departments at all levels, industrial investment institutions, industry research institutions, industrial parks and high-tech enterprises. By demonstrating data cases and service effects, it establishes market leadership, achieves user growth, and maximizes both commercial and social value.

III. Project Highlights

(I) Key Technological Breakthroughs

1. Natural-language interactive AI consulting technology: Based on a large language model (LLM) deeply optimized for professional consulting dialogues, the system achieves an accuracy rate of 90 percent or above in understanding industry consulting questions. It can accurately identify eight major categories of vague needs, such as "how to improve production capacity" and "cost optimization plans". Through a multi-turn dialogue questioning mechanism, the agent can actively clarify user intent. For example, after a user proposes "reducing losses", the agent can automatically ask whether the focus is on raw material loss, process yield or logistics loss.

2. Multi-source data AI collaborative fusion technology: By adopting an intelligent data adaptation model based on the Transformer architecture, the platform automatically identifies more than 20 heterogeneous data formats and completes data parsing without manual configuration, improving adaptation efficiency by 80 percent. Meanwhile, federated learning is introduced to enable cross-enterprise data collaboration while ensuring data privacy, supporting the simultaneous access of more than 10 partner organizations.

3. Scenario-based AI agent decision-making technology: With five major categories of built-in professional agents, including product consulting and public-opinion monitoring, the platform can automatically call corresponding agents according to user needs and respond within seconds. It adopts reinforcement learning algorithms and continuously

learns from more than 3,000 historical enterprise consulting cases, enabling the agent's decision logic to be constantly optimized. The recommendation adoption rate has increased from an initial 75 percent to over 92 percent. This technology realizes the transition from passive consulting to active decision support.

(II) Core Advantages of the Project

Compared with general-purpose large models, the core advantage of this project lies in its deep understanding of the vertical field. JiweiGPT is trained on an exclusive semiconductor industry database. When analyzing professional issues such as process technologies, supply-chain relationships and policy impacts, its accuracy and contextual understanding far exceed those of general-purpose models. For example, when addressing a complex question such as "the impact of 28 nm capacity expansion on domestic design companies", JiweiGPT can combine regional policies, enterprise capacity layouts, technology roadmaps and market demand to provide an analytical result that is both professional and practical.

At present, the platform has taken a leading position in the segmented market of intelligent information services for the semiconductor vertical field. By deeply integrating high-quality industry knowledge with large-model technology, it breaks through the limitation of general-purpose models that are "broad but not deep", and realizes full-chain innovation from data fusion and intelligent decision-making to natural interaction. It not only significantly improves the efficiency of professional consulting and lowers the threshold for use, but also continuously enhances the level of industrial intelligence through data accumulation and model iteration.

IV. Implementation Results

The implementation of this project will create China's first vertical AI integrated service platform for the semiconductor industry. Based on a three-in-one model of "data + AI + scenarios", the platform is committed to breaking down barriers to professional information. It empowers the entire industrial chain through intelligent technologies and becomes a digital infrastructure that drives industry innovation and efficient decision-making.

(I) Construction Results

Based on fifteen years of deep industry cultivation, the company has established the most comprehensive professional database for the semiconductor industry, forming a unique data asset that supports AI applications. The database fully integrates enterprise core data, covering more than 35,000 domestic and more than 5,000 overseas integrated circuit enterprises; industrial dynamic data, including 1.2 million pieces of industry news, 100,000 conferences and more than 10,000 policies; and specialized data such as investment and financing, intellectual property and trade control. It provides a solid data foundation for AI applications.

On this basis, JW Insights deeply integrates "JW big data" with AI technology to build the semiconductor industry data platform "JW Information Platform", which has full-chain AI capabilities from data parsing, cleaning and analysis to application. Its built-in public-opinion module is equipped with a 7 × 24-hour full-time monitoring network, continuously tracking more than 10,000 industry public accounts and more than 1,000 professional websites. It can generate automated analysis reports and early-warning information, providing enterprises with real-time and reliable decision support.

(II) Wide Range of Application Scenarios

At present, the project platform has entered a phase of accelerated commercialization. It has reached forward-looking cooperation with some customers and has initially formed demonstration cases covering different parties in the industrial ecosystem.

Empowering leading enterprises (demonstration cases: a leading domestic wafer foundry and China's leading image-sensor company): The platform is used for global trade compliance warning. The results are significant: export-control research and matching work that previously required legal and R&D teams several days to complete has been shortened to minutes, realizing an efficiency leap from "days" to "minutes".

Supporting chain-owner enterprises and terminal giants (demonstration case: a well-known domestic mobile phone manufacturer): Through the platform's supply-chain resilience analysis function, users can quickly simulate supply shortage risks for specific chips and screen investment targets. In practice, after preliminary plans are generated, they can be adjusted through real-time interaction with enterprise engineers, improving plan iteration efficiency by 60 percent.

Coordinating local-government industrial decision-making: For local-government needs such as policy formulation, industrial-park planning and industrial investment attraction, the platform provides customized functions such as data visualization and industrial risk warning. At present, the "Digital Foundation Project for Semiconductor Industry Development" developed in cooperation with local industrial parks covers thousands of enterprises and updates industrial dynamics weekly.

(III) Social Benefits

As a digital service product for the semiconductor vertical field, the project provides enterprises with professional, accurate and convenient consulting services. By deeply integrating high-quality industry knowledge in the vertical field with AI large-model technology, the platform helps enterprises capture the most valuable information in the sea of information and obtain professional consulting services from JiweiGPT anytime and anywhere, thereby improving the efficiency and quality of decision-making.

The company will also leverage its local advantages to lower the threshold for SMEs to use AI technologies. It will provide lightweight AI tools to help SMEs improve operational efficiency and innovation capabilities, and promote inclusive application of technology. It will establish close cooperative relationships with upstream and downstream enterprises and research institutions to build a data-driven industrial innovation network. Through data sharing and joint R&D, the project will promote the coordinated development of the industrial chain and support China's semiconductor industry in moving toward higher-quality development.

Case 5

A Group-Level Social Media Listening System Driven by Both Large and Small Models for One-Click Tracking of Social Media Trends

Target Social Technology (Shanghai) Co., Ltd.

I. Basic Enterprise Information

Target Social Technology (Shanghai) Co., Ltd. is a specialized, refined, distinctive and innovative SME and a high-tech enterprise, certified in 2023. Founded on November 11, 2011, the company is headquartered in Huaihai International Plaza in Xuhui District, Shanghai, and currently has branches or offices in Beijing, Guangzhou and Changsha. Target Social has a professional management and business team composed of senior professionals from information technology (IT), customer relationship management (CRM) and digital fields, as well as experts from public relations (PR), media and social sectors. Its business currently covers online public opinion monitoring and multimodal open-source internet data insights, enabling it to provide diverse clients with services including online public opinion monitoring, big-data governance and brand social media intelligence.

II. Case Introduction

(I) Background and Needs

As digital transformation accelerates, social media platforms have become core data carriers that aggregate user needs, market feedback and consumer trends. The massive volume of unstructured data contained in social media has become a key asset for enterprises seeking to improve the scientific basis of decision-making. Large groups usually cover multiple business segments, including branding, marketing, IT, R&D, and sales. In core tasks such as new product development, marketing promotion and competitor analysis, each business unit needs real-time, comprehensive and accurate social media data support to upgrade the entire business chain digitally.

Traditional social media data processing models generally face three major pain points. First, the scope of data collection is limited, and it is difficult to connect data interfaces across multiple platforms to form a full-domain data view. Second, the efficiency of unstructured data parsing is low, and manual processing cannot meet real-time analysis needs at the level of tens of millions of data items. Third, data is disconnected from business scenarios, and the intelligent conversion path from data to strategy is insufficient, making it difficult to accumulate data value.

To solve these problems and build a standardized data asset system, this project focuses on the core technical difficulties of intelligent social media data processing and develops a self-developed social media data intelligence system fully based on AI technology. The system's core capability modules are all supported by AI. The intelligent insight module uses a self-developed and optimized NLP model to achieve accurate text parsing, while the strategic analysis module is developed based on a customized AI large model to adapt to multiple business scenarios. Ultimately, the system realizes the intelligent upgrading of the full chain from social media data collection to application.

(II) Implementation Plan

1. System architecture

The system adopts a layered architecture with "data-driven operation and AI empowerment" as its core concept. It builds a three-tier technical architecture consisting of the data layer, the AI capability layer and the application layer. Each layer can be iterated independently while working in coordination with the others, ensuring system scalability and stability.

Data layer: full-domain data collection and asset-based foundation. As the data foundation of the system, the data layer undertakes the collection, cleaning, storage and management of full-domain social media data. Through standardized interfaces, it adapts to mainstream social media platforms such as WeChat, Weibo, Douyin and Xiaohongshu. It uses distributed crawler technology and real-time message queues (MQ) to support parallel collection of more than ten million data items per day, ensuring data timeliness. In the data processing stage, extract, transform and load (ETL) tools and customized cleaning algorithms are introduced to filter invalid data through deduplication, desensitization and format standardization. A hybrid storage architecture based on Hadoop and MySQL is built: unstructured text data is stored in a distributed file system, while structured tag data is stored in relational databases, forming a standardized social media data asset base.

AI capability layer: the core intelligent engine. As the core intelligent carrier of the system, the AI capability layer includes three modules: intelligent insight, strategy analysis and dynamic optimization. Each module is encapsulated as a model service so that its capabilities can be reused. The intelligent insight module integrates optimized NLP models and trend mining models, and is responsible for text parsing and trend prediction. The strategy analysis module is built on a customized AI large model to generate scenario-based strategies. The dynamic optimization module builds closed-loop iteration capabilities through real-time feedback algorithms. The AI capabilities produced by these modules are made available to the application layer through API calls and can flexibly adapt to multiple business scenarios.

Application layer: the outlet for business implementation. The application layer provides modular and configurable intelligent services for the needs of different business lines within the group. Through a visual workstation, it offers dedicated functional entrances for R&D, marketing, branding and other departments, such as a “demand mining module” for the R&D business unit and a “marketing optimization module” for the marketing business unit. It also supports customized report generation and data visualization, enabling precise alignment between AI capabilities and business needs and ensuring that technological value directly reaches business scenarios.

2. Business logic

The system builds an end-to-end closed-loop mechanism based on the business logic of “data circulation - intelligent analysis - business empowerment - data accumulation”. The specific process is as follows:

Data access: The system collects full-domain social media data in real time through data-layer interfaces. After cleaning and desensitization, the data is transformed into standardized data resources and synchronized to the data asset database for tag-based storage.

Intelligent processing: The AI capability layer calls NLP models to parse sentiment tendencies and pain points in text data, identifies market hot spots through trend mining models, and generates scenario-based plans through strategy analysis models in combination with historical business data.

Business application: The application layer pushes AI-generated insights and strategic plans to the corresponding business departments, supporting specific work such as R&D direction decisions and optimization of marketing activities.

Closed-loop iteration: The dynamic optimization module collects business execution performance data in real time, such as content interaction rates and R&D hit rates, and feeds it back into AI model iterative training and data asset database updates, realizing continuous optimization of “data - intelligence - business”.

III. Project Highlights

(I) Key Technological Breakthroughs and Scenario Application Value

Efficient processing of massive unstructured data: With the help of AI technologies such as NLP, the project builds a full-domain social media data parsing engine that accurately processes unstructured data, including user posts, images, videos and comments across platforms such as Douyin, Xiaohongshu and Weibo. It identifies user sentiment tendencies and demand pain points, solves the problem that manual methods cannot efficiently process massive text data, improves the accuracy and efficiency of data parsing, and simultaneously conducts multidimensional judgments on content sensitivity risks, public opinion tendencies, industry attributes and regional characteristics.

The uniqueness of this scenario lies in connecting full-domain social media data interfaces and realizing real-time collection and synchronized parsing of data across multiple platforms. This breaks the limitations of traditional tools that rely on “single-platform collection + static analysis” and adapts to enterprises’ immediate data needs in an omnichannel social media layout.

Mining the deeper value of data: Machine learning algorithms are used to analyze massive social media data. By mining hidden information such as user behavioral preferences, content interaction features and changes in keyword popularity, the system accurately identifies potential consumer demand and predicts category consumption hot spots in advance, providing data support for new product R&D and marketing direction selection. This solves the problem that traditional methods cannot deeply mine the value of data.

This technological breakthrough overcomes the limitations of traditional data statistics that emphasize results while neglecting correlations. It upgrades data value from “describing the past” to “predicting the future”, and is highly suitable for business scenarios in fast-moving consumer goods, beauty and other fashion categories that require rapid iteration and precise layout.

Automated and precise strategy generation: Based on AI large models and combined with historical data and real-time insights, the system generates multiple executable strategic plans tailored to different business scenarios. It can match the optimal communication channel combination based on target user profiles, such as “Xiaohongshu influencer seeding + Douyin live-stream conversion”, and quickly output differentiated platform operation and marketing strategies to meet business units’ needs for efficient and precise strategy output.

The technical advantage lies in “one-to-one” strategy customization. It solves the inefficiency of using one single plan throughout traditional marketing and adapts to enterprises’ multi-category and multichannel business expansion scenarios.

Building a dynamic optimization mechanism: AI technology is used to make dynamic adjustments to ongoing marketing activities and product strategies based on real-time data feedback, such as content click-through rate, conversion rate and user feedback. This realizes a closed-loop iteration of “data - strategy - results” and solves the problem that traditional strategies are difficult to adjust according to real-time market changes.

This technology breaks the separation between “formulation - execution - review” in traditional strategies. It is especially suitable for the scenario characteristics of social media platforms, where hot topics change quickly and user demand is highly immediate, making enterprise marketing decisions more flexible.

Creating an intelligent competitor monitoring system: AI is used to capture and analyze competitors’ social media communication data in real time, build multidimensional monitoring models, identify competitors’ new moves promptly, and generate strategy recommendations, helping the group take the initiative in market competition.

The scenario value lies in helping enterprises grasp market competition trends in real time, solve the problems of delayed information and single-dimensional perspectives in traditional competitor analysis, and meet enterprises’ need for rapid response and proactive layout in intense market competition.

(II) Summary of Scenario Uniqueness

The project’s core scenario advantage lies in “full-domain social media coverage + full-business-chain adaptation”. On the one hand, it connects data from mainstream social media platforms such as Douyin, Xiaohongshu and Weibo and processes the data in real time, solving the data-island problem in multi-platform operations. On the other hand, its technical capabilities run through the entire process of new product R&D, marketing promotion, competitor analysis and performance optimization. It adapts to social media operation needs in multiple industries, including fast-moving consumer goods, beauty and 3C, and forms a complete closed-loop service capability of “data collection - insight analysis - strategy implementation - dynamic optimization”, distinguishing itself from traditional tools that support only a single technical link.

IV. Implementation Results

(I) Economic Benefits

1. Improved new product R&D efficiency and reduced costs: Based on AI clustering analysis of tens of millions of user demand data points, the system accurately identifies core functional demands and experience pain points. As a result, the hit rate of new product R&D directions increased by 30 percent, effectively reducing resource waste caused by deviations in R&D direction and lowering new product development costs. At the same time, early market acceptance after product launch improved significantly, and first-month sales of new products increased by 17 percent compared with previous levels, bringing direct economic benefits to the group.

2. Optimized marketing promotion results and improved ROI: During the new product launch stage, the AI model verifies the conversion rate of selling points from multiple dimensions and dynamically adjusts communication weights,

increasing marketing budget efficiency by 8 percent. At the same time, it automatically generates standardized seeding content that fits the tone of each platform through a content management system (CMS), improving content production efficiency by 80 percent, saving labor costs in content creation, and increasing interactive conversion results by 30 percent compared with manual creation. This drove product sales growth and increased marketing ROI by 30 percent.

3. Strengthened market competitiveness and expanded market share: By using AI to monitor competitor dynamics in real time and generate strategy recommendations, the project helped the group adjust business strategies in a timely manner and continue taking the initiative in market competition. Market share increased by 3 percent compared with before project implementation, further expanding the group's market influence and laying a solid foundation for subsequent business growth.

(II) Social Value

1. Improving response efficiency to user needs and optimizing user experience: Through the social media listening system, the group can rapidly capture user pain points and feedback and adjust product and service strategies in a timely manner. This better satisfies user needs, improves user satisfaction and loyalty, and builds a positive corporate brand image.

2. Promoting industry digital transformation and setting a benchmark: The successful implementation of the group social media listening system provides enterprises in the same industry with a practical example of "AI + social media data" driving business decisions. It demonstrates the application value of digital technologies in enterprise operations and helps promote digital transformation across the industry, raising the overall level of industry development.

3. Promoting the rational use of data resources and releasing data value: The project builds a group-exclusive social media data asset database, enabling effective accumulation and reuse of data resources. It not only supports the group's own business decisions, but can also, under compliance conditions, share part of its non-sensitive data with industry partners in the future, promoting the circulation and rational use of data resources and releasing the economic and social value of data elements.

(III) Classic Case Sharing

A leading Fortune Global 500 enterprise in the daily chemical industry, as a full-category household consumer service provider, has established a stable long-term partnership with our company and is a core benchmark user of our social media data intelligence system. The group's business covers multiple core categories, including beauty and skincare, home cleaning, and maternal and infant care. Through deep customization, our system perfectly adapts to the business needs of its multi-category collaborative development.

In response to the differentiated needs of the group's three core departments - R&D, marketing and branding - the system builds a full-chain intelligent service system. On the R&D side, it accurately mines user demand and analyzes user psychology based on mainstream social media platforms, providing a basis for new product formula design. On the marketing side, it links e-commerce traffic pools and advertising traffic pools in real time and dynamically optimizes omnichannel marketing strategies. On the branding side, it monitors major competitors' social media dynamics 24 hours a day, 7 days a week, and synchronously outputs strategy analysis reports. The cooperation has formed a closed-loop system of "customized upgrading + data support + operation and maintenance assurance". With "responding to needs + solving problems + enhancing value" as its core, the technical team has completed high-frequency functional iterations and scenario optimization, continuously supporting business departments in improving decision-making efficiency and operational results.

Case 6

AI-Empowered ShengliangTong for Insights into Online New Media Reputation

Shanghai Newrank Info & Tech Co., Ltd.

I. Basic Enterprise Information

Shanghai Newrank Info & Tech Co., Ltd. is a leading domestic internet content technology company headquartered in Shanghai, with offices in Shanghai, Beijing, Chengdu, Guangzhou and Changsha. The company has more than 600 full-time employees. Its core technical team accounts for 38 percent of the workforce, employees with bachelor's degrees or above account for 88 percent, and annual R&D investment exceeds 30 percent. In the first three quarters of 2025, total revenue reached nearly RMB 1.36 billion, representing year-on-year growth of nearly 19 percent, and the company ranked among the top three in the industry by scale. Newrank serves the content industry and empowers industries with content. It provides enterprises with solutions for traffic acquisition and content asset management. Its main business includes intelligent content marketing services, such as social media influencer marketing, content seeding, brand account operations and live-streaming operations, as well as enterprise-level content asset management products and services.

Newrank is committed to enhancing the level of technological innovation and exploring the integration of AI with the content industry. It has promoted the intelligent upgrading of enterprise content marketing and provided full-chain integrated intelligent marketing solutions for tens of thousands of brands, including L'Oréal, Procter & Gamble, Ping An, Alibaba and Huawei. It actively helps enterprises establish digital content asset management standards, providing intelligent content products for nearly 400 group enterprises, including SAIC Motor, Beike and Trip.com, to rapidly build content asset management middle platforms and improve content management efficiency and operational capabilities. With "content + technology" as its core, Newrank focuses on digital marketing, builds content assets, creates content value for all industries, and contributes to the intelligent transformation of Chinese enterprises' digital content services.

II. Case Introduction

(I) Background and Needs

Against the background of full-domain penetration of social media and the popularization of generative AI technology, enterprise brand management is undergoing structural change. Mainstream social platforms have formed ultra-large-scale user ecosystems. More than 200 million works and one billion comment interactions are generated every day, and the massive interaction of user voices is reshaping the underlying logic of enterprise decision-making.

Traditional manual monitoring models have essential limitations. Cross-platform data is scattered, updates are frequent and real-time, and unstructured information is complex. As a result, enterprises cannot cover the full-domain voice, accurately identify weak points and strengths in brand communication, or monitor competitors' marketing moves and viral content strategies without blind spots. The risk of delayed market decisions increases accordingly. When negative comments reach the million-level in a single day, traditional methods cannot quickly locate core risk points or establish an effective crisis response mechanism.

Against this background, Newrank launched the full-platform new media voice AI monitoring system "Newrank ShengliangTong", based on its mature social media data monitoring technology and AI analysis capabilities. By building a three-dimensional system of "measuring voice - listening to voices - managing reputation", it systematically solves core pain points in enterprise social media insight and helps form a data-driven brand decision-making hub.

(II) Implementation Plan

"Newrank ShengliangTong" builds a full-domain social media voice monitoring system through AI technology. It addresses core issues such as multi-platform data integration, semantic parsing, demand insights, competitor tracking and public opinion early warning. It deepens the capability of user demand insight, improves the risk early-warning response mechanism, and comprehensively empowers digital and intelligent brand management.

1. System architecture:

The Newrank ShengliangTong application consists of three major modules. The “rapid trend inquiry” module combines basic lightweight applications such as social media quick search, popularity comparison, share of voice (SOV) query, search index query and real-time hot topic viewing with an AI assistant. It calls system-wide dashboard data for AI analysis and judgment. The “subscription tracking” module provides customized tracking, analysis and judgment after full-platform data collection for specified keywords, supporting more complex and refined voice trend analysis. The “reputation management” module helps users accumulate and manage analyzed content assets and media resources, making it easier to further optimize reputation and public opinion voice management.

2. Business logic:

Newrank ShengliangTong addresses the practical need to process ultra-large-scale content data across multiple platforms. First, it integrates massive cross-platform data: user feedback and brand voice are scattered across large volumes of content and comments on multiple platforms, making manual integration difficult. AI is needed to capture and aggregate cross-platform data in real time, breaking data silos. Second, it conducts multimodal and complex semantic parsing and sentiment tendency judgment. AI is required to overcome limitations in text, image and video semantic analysis, penetrate complex expressions such as internet slang, dialects and abbreviations, filter irrelevant information, accurately parse users’ real intentions, and achieve positive and negative sentiment recognition accuracy of more than 95 percent. Third, it identifies user needs and intention signals, generating demand analysis Sankey diagrams by extracting keywords and sentiment tendencies. AI large models are used to support multimodal sentiment analysis across text, images and videos, penetrate internet buzzwords and sarcastic semantics, and build a closed loop of “data collection - analysis - strategy”. Through voice penetration rate and user tag mining, the system identifies emerging track opportunities and improves the timeliness of product iteration decisions.

3. Key function descriptions:

Newrank ShengliangTong demonstrates Newrank’s strong data integration and analysis technology and its industry-leading AI-powered social media insight capability. It helps brands solve pain points and difficulties in social media marketing, including the difficulty of quantifying brand performance, capturing user feedback and following up on public opinion in a timely manner. It comprehensively realizes brand voice monitoring, public opinion risk management, competitor analysis, user demand insight and market trend research, and actively promotes the full AI-based intelligent upgrading of brand social media marketing.

(1) Cross-platform brand voice analysis: AI helps comprehensively assess brand influence

The system monitors brand voice trends across seven major platforms, including Xiaohongshu, Douyin, WeChat and Channels, in real time. It supports one-click query and analysis of brand SOV and search indexes, quantifies indicators such as brand reputation, brand communication effectiveness and brand voice scale from multiple dimensions, and uses AI to generate multidimensional analysis reports. This helps enterprises comprehensively assess brand social media influence and industry position, deeply analyze reasons for changes in brand communication trends, and provide more intuitive and convenient decision-making references for optimizing brand strategies.

(2) Real-time competitor monitoring and one-click AI analysis to quickly identify competitive advantages

The system monitors competitors’ social media movements in real time, compares an enterprise’s own and competitors’ social media data from multiple dimensions such as content, channels and placements, and supports one-click AI analysis report generation. This helps enterprises quickly identify competitive advantages, optimize strategies in a targeted manner, achieve competitive breakthroughs and capture more market share.

(3) AI accurately identifies user emotions and uncovers real user needs

By aggregating and analyzing multi-platform content and comments, the system accurately extracts semantic keywords and clustered themes from massive data. AI automatically extracts key information and precisely identifies positive and negative user emotional feedback, enabling enterprises to intuitively understand users’ evaluations, core concerns and potential needs regarding brands, campaigns and products, and to precisely adjust marketing directions, optimize product iterations and reach potential users.

(4) Rapidly discovering negative social media public opinion and notifying responsible personnel for timely handling

The system dynamically monitors negative social media public opinion in real time, promptly identifies key points of public opinion fluctuation, and conducts rapid source tracing. Unlike traditional public opinion monitoring tools that can generally locate negative public opinion only based on keywords or simple word frequency and part-of-speech features, ShengliangTong uses AI sentiment recognition technology to help brands more accurately identify social media public opinion and notify corresponding responsible personnel at the earliest possible moment, minimizing negative impact.

(5) AI “Social Media Quick Search”: cross-platform voice results available with one search

AI intelligently identifies user question intent and automatically matches event voice data sources across seven major platforms. It supports linked analysis from 30-day macro trends to microscopic attribution of viral posts, and provides five preset AI analysis scenarios: content analysis, speaking platform analysis, account analysis, hot topic trend tracking and user comment analysis. Users can obtain professional conclusions and data chart reports directly with one click, eliminating manual analysis and typesetting and greatly improving work efficiency.

III. Project Highlights

Based on multi-platform data integration, with AI intelligent analysis as the core and visual interaction as the window, Newrank ShengliangTong establishes a full-platform new media voice AI monitoring system that integrates “data monitoring - intelligent decision-making - risk prevention and control”. It enables quantifiable tracking of brand marketing, analyzable conversion of user feedback, forward-looking prediction of market trends and closed-loop management of public opinion risks, comprehensively improving market competitiveness and brand resilience.

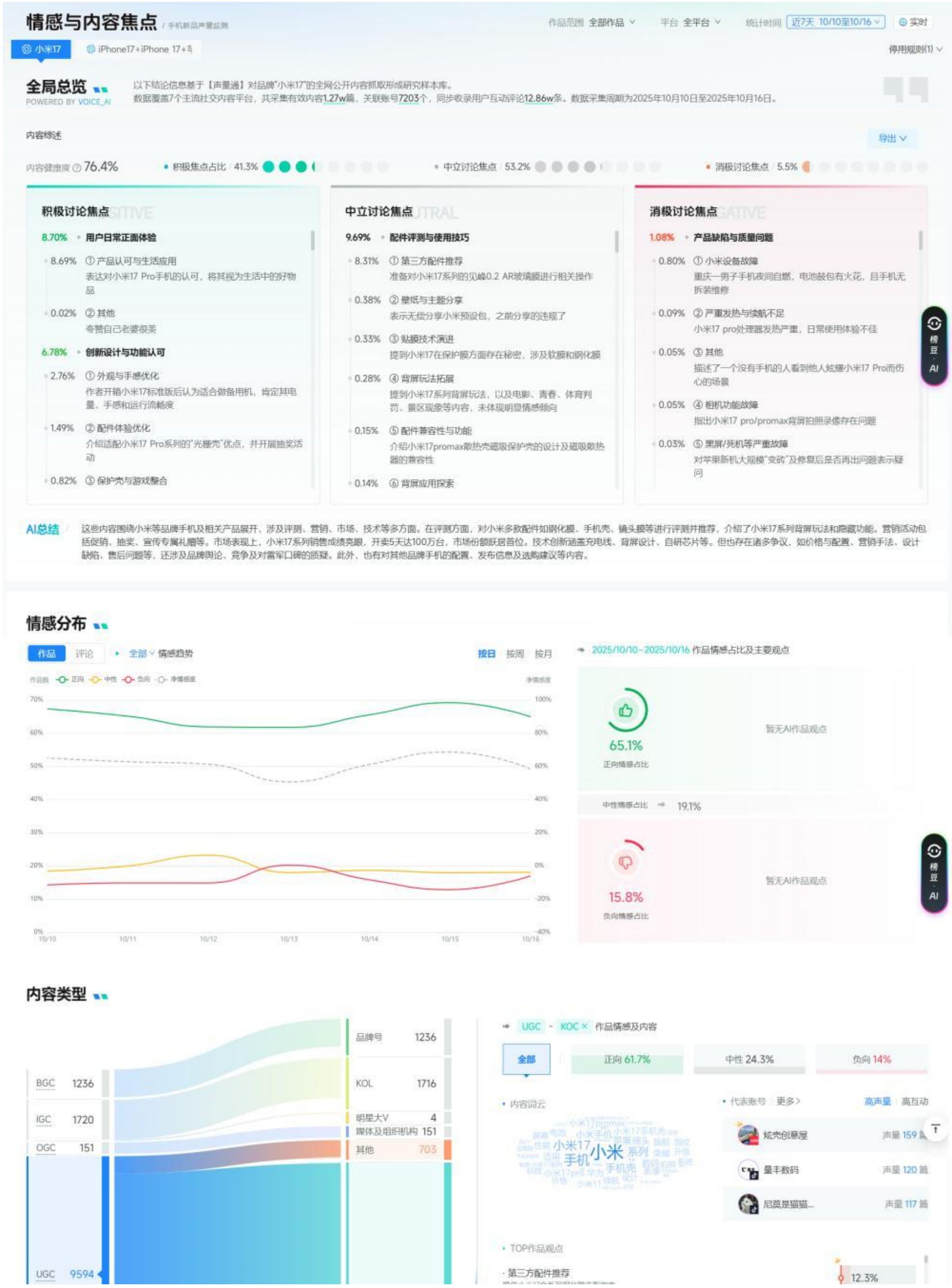
(I) Full-Domain Voice Monitoring and Evaluation System

The system supports real-time monitoring of voice data from seven platforms: Douyin, Kuaishou, WeChat Official Accounts, Channels, Weibo, Xiaohongshu and Bilibili. It monitors tens of millions of new works and hundreds of millions of comments every day, visually presents brand voice communication trends and channel distribution, and uses cross-platform aggregated data to identify content hot spots and turning points in consumer trends.

Technical advantage: The system covers data interfaces of seven major platforms, systematically collecting brand voice, competitor dynamics, user needs and industry hot spots. It supports multidimensional voice queries for brands, categories and campaigns, providing precise support for full-chain decision-making.

(II) User Demand Insight System

By using AI to parse massive user content and comments, the system extracts real consumer demands, identifies high-frequency keywords and sentiment tendencies, and generates demand analysis Sankey diagrams. AI large models support multimodal sentiment analysis across text, images and videos, penetrate internet buzzwords and sarcastic semantics, and build a “data collection - analysis - strategy” closed loop. The system accurately extracts user interests and pain points, identifies emerging track opportunities through voice penetration rates and user tags, and improves the timeliness of product iteration decisions.



ShengliangTong 1: Newrank ShengliangTong AI analysis module for sentiment and content focus

Technical advantage: The built-in AI denoising algorithm filters irrelevant content and ensures sentiment accuracy of more than 95 percent. Natural language processing technology is used to realize semantic clustering of content and support customized tag systems. A “public opinion brief” is generated automatically every day.

(III) Reputation Risk Management System

The system improves the identification and response to negative information in massive user-generated content (UGC), proactively implements crisis prevention and control, and solves the inefficiency of hot topic tracking caused by single-platform information silos. It establishes a full-process control mechanism of “early warning - handling - review”, monitors sensitive content at high frequency, supports customized negative public opinion grading responses and retention of handling records, and improves the reputation protection network.

Technical advantage: Multiple categories of early-warning rules are set up, and DingTalk or email alerts can be triggered on an hourly basis. A built-in “public opinion handling” dashboard automatically records the handling and circulation process.

(IV) Competitor and Industry Evaluation and Analysis System

The system develops an SOV calculation tool and voice comparison function, supporting cross-platform SOV and voice comparison between brands and competitors. It monitors brand and competitor voice trends, marketing moves and viral content tracking in real time, generates communication effect evaluation reports, and predicts market concentration.

Technical advantage: The system integrates core indicators such as breadth of content dissemination, depth of interaction and level of commercialization, and supports one-click export of data in PDF or Excel formats.

(V) Automated Report System

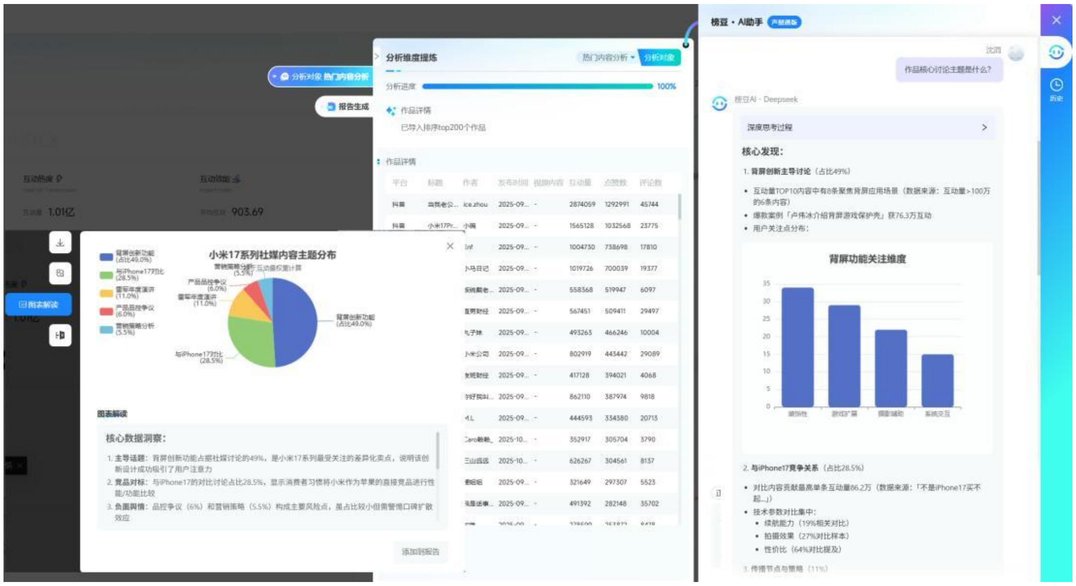
The system extracts and aggregates a series of tags from captured data, including entity relationships, sentiment analysis, thematic viewpoints and communication characteristics. It generates brand analysis reports on a daily, weekly or monthly basis, covering modules such as voice trends, competitor dynamics and demand insights.

ShengliangTong 2: Newrank ShengliangTong brand influence AI analysis report module

Technical advantage: Data is automatically updated without manual intervention. Report content can be pushed to designated email addresses or office automation (OA) groups.

Bangdou AI - Brand AI Analysis Assistant

The assistant supports customized questions, understands data retrieval intent and generates corresponding visualized analytical conclusions. It helps brands capture industry trends and quickly obtain optimization suggestions or conclusion support. Through customized questions and intent-based data retrieval, it generates corresponding visualized analytical conclusions, helping brands capture industry trends and rapidly obtain optimization suggestions or conclusion support.



ShengliangTong 3: Intelligent Q&A report generated by the Newrank ShengliangTong Bangdou AI assistant

Technical advantage: The system integrates Newrank's ten years of data monitoring capabilities with DeepSeek's deep analysis capabilities. It intelligently identifies the corresponding content data behind user questions, batches specified content data into an AI large model for analysis, breaks the traditional text-only mode, obtains illustrated analytical reports, and realizes efficient integration between content data and real-time AI Q&A.

IV. Implementation Results

Overall, the full-platform new media voice AI monitoring system "ShengliangTong" delivers multiple social and economic benefits, including promoting digital transformation, driving economic development, increasing tax contributions, improving innovation capabilities and enhancing regional services. It plays an important role in promoting the development of local industries and improving enterprise competitiveness. At the same time, ShengliangTong has also provided Newrank itself with a new path for enterprise AI integration. Newrank ShengliangTong was launched in July 2024 and achieved RMB 5 million in single-product revenue in the same year, with a profit margin exceeding 20 percent. It became profitable in its first year and will become an important growth pole for Newrank's AI applications in the future.

(I) Improving Industrial Efficiency and Regional Competitiveness, and Supporting Enterprise Transformation and Upgrading

Newrank ShengliangTong can help enterprises make new media digital assets widely open and shared across the entire company, and enable these assets to be used as valuable organizational assets in business, management and strategic decision-making, thereby exerting the commercial value of new media digital assets. It optimizes business processes and resource allocation, connects all links of data management, realizes a complete data governance process, improves business management capabilities, and thus optimizes industrial structure. It promotes industries toward greater intelligence and digitalization, further improving enterprise market competitiveness. With its advantages in big data and AI technologies, Newrank ShengliangTong can provide enterprises with more precise data support and help them discover market opportunities, avoid public opinion risks and improve operational efficiency.

(II) Improving Efficiency in the Content Industry and Supporting the Healthy Development of Industry Resources

Newrank leverages its communication advantages, resource advantages and data capabilities in the new media industry to play an active role in regulating the development of the new media industry, exploring new trends for new media practitioners and serving regional social and economic development. As a leading cultural and creative enterprise in Jinshan District, Newrank hopes to empower both content production and content consumption markets through two forms of enablement, invigorate cultural creativity and related industries, and further drive the healthy development of the content industry in Jinshan District and even across Shanghai. Newrank ShengliangTong has already provided services to multiple media institutions, including China Media Group News Center, Shanghai Media Group (SMG), Global Times Online, Jiemian News and Caixin, as well as universities such as Shanghai University of Finance and Economics, Renmin University of China and Communication University of China.

(III) High Replicability and Scalability, with Strong Fit for Platform Technology Applications

The capability of Newrank ShengliangTong to capture public social media information and conduct AI-based deep semantic analysis, including sentiment and content focus, is one of the most effective ways to obtain real, spontaneous and large-scale user feedback. This capability applies to any industry that needs to drive products, marketing or services through user voices, especially fields with rapidly changing demand and active user expression. The methodology of identifying high-frequency discussion topics, sentiment tendencies, pain points and emerging demand points from massive UGC is universal. Different industries only need to adjust monitored keywords and focus dimensions to quickly replicate the demand insight process.

The automated data collection, AI semantic analysis engine and visual report modules provided by Newrank ShengliangTong constitute an out-of-the-box market insight platform that requires no complex self-built technology stack. Enterprises only need to configure target brands, competitors and keywords to get started quickly. This aligns with the national policy orientation of “unlocking the potential of data elements”, transforms public user discussions into decision-making assets, significantly reduces operating costs and risk losses, and helps enterprises improve quality and efficiency. In the future, Newrank ShengliangTong can also be expanded into a “regional industrial monitoring platform” to help local governments grasp user feedback, innovation needs and competitive dynamics in advantageous industries such as new energy vehicles in real time, providing data support for industrial policy formulation.

Case 7

Intelliotspotgent Media H Discovery and Communication Analysis Platform

Based on a Full-Stack Self-Developed Algorithm System

Midu Technology Co., Ltd.

I. Basic Enterprise Information

Founded in 2009, Midu is an AI enterprise focused on vertical applications of large models. Deeply engaged in the AI + office sector, the company uses data and algorithms as dual engines to empower digital and intelligent transformation across industries. Based on its self-developed vertical large models, Michao and Wenxiu, Midu applies three core capabilities - intelligent retrieval, intelligent proofreading and intelligent generation - to provide secure, efficient and intelligent products, services, cloud capabilities and solutions for core office scenarios such as official document writing, content proofreading and public opinion analysis. The company has been dedicated to AI-enabled digital transformation for 15 years. With strong technological capability and innovation capacity, it has repeatedly received honors such as “Information Technology Application Innovation Unicorn” and “AI Leading Enterprise”, and has also been recognized as a Shanghai municipal enterprise technology center and a technology innovation enterprise.

II. Case Introduction

(I) Background and Needs

Against the background of media convergence and digital transformation, news production and communication methods are undergoing profound changes. Faced with information overload, complex public opinion and dispersed user attention, media organizations urgently need to improve their ability to capture news leads, understand public sentiment and predict communication trends, while continuously producing high-quality content. This poses comprehensive challenges to information processing, response speed and content innovation.

The traditional model that relies on reporter networks and professional experience can no longer cope with the accelerated and diversified information environment. Real-time, whole-network and all-media data now emerges at scale. Old models cannot fully capture or deeply analyze information, nor can they grasp public sentiment and communication trends in a timely manner. Therefore, the industry urgently needs to use big data and AI technologies to intelligently empower the full news workflow of planning, gathering, editing, reviewing and publishing.

For this reason, using AI to intelligently reconstruct business processes has become critical to strengthening core competitiveness. Natural language processing (NLP) can support news generation and summarization and improve efficiency and quality. Machine learning can analyze user behavior and support personalized recommendation. Image and video recognition technologies can support multimedia content processing. Workflow automation and collaboration can reduce manual intervention and improve work efficiency and accuracy. Only with AI empowerment can media organizations stand out in fierce competition and achieve sustainable development.

(II) Implementation Plan

Sina Public Opinion Monitor is a data intelligence software product that integrates internet information aggregation, text mining, intelligent retrieval and large-model applications. It is used to discover public opinion information on the internet and to conduct automatic classification, intelligent filtering, automatic clustering, topic detection and statistical analysis. It enables intelligent identification and targeted tracking of hot topics and emergencies, helping governments, media organizations, enterprises and public institutions keep abreast of online communication trends and providing decision-making support for effective response to online public opinion events. By deeply integrating Midu’s self-developed “Michao” large model and the mature functional system of Public Opinion Monitor, the project builds an intelligent media platform for hotspot discovery and communication analysis that combines hotspot discovery, public opinion analysis, communication tracking and audience insight, comprehensively improving enterprise efficiency in content production and communication decision-making.

The system architecture of Sina Public Opinion Monitor mainly consists of the data collection layer, data processing layer, data analysis layer and application layer.

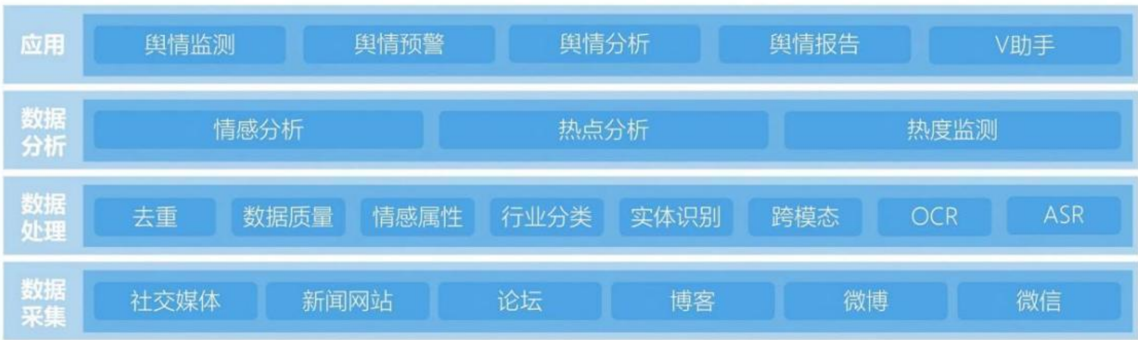


Figure 1 System Architecture of Public Opinion Monitor

The data collection layer retrieves public opinion data in real time from multiple information sources, including social media, news websites, forums, blogs, Weibo and WeChat. The data processing layer performs preprocessing operations such as cleaning, deduplication and classification on the collected data. The data analysis layer applies NLP and machine learning technologies to conduct in-depth data analysis, including sentiment analysis, hotspot analysis and popularity monitoring. The application layer provides functions such as public opinion monitoring, early warning, analysis and reporting, and supports users in operating the system and viewing results through a visual interface.

This project deeply integrates the core public opinion capabilities of Sina Public Opinion Monitor to build hotspot discovery and communication analysis functions.

1. Real-time public opinion monitoring

Public Opinion Monitor can retrieve public opinion data in real time from multiple sources, including social media, news websites, forums, blogs, Weibo and WeChat, ensuring that users can understand public opinion dynamics at the earliest possible time. The system automatically captures information from major platforms and updates it in real time, preventing information lag and helping users grasp the pulse of public opinion.

2. Public opinion analysis and mining

Through NLP technology, Public Opinion Monitor can automatically identify the sentiment orientation in content, including positive, negative and neutral sentiment, helping users judge the emotional tone of public opinion. Based on big data analytics algorithms, the system can identify current hot topics, keywords and trends, helping users understand changes in public opinion in real time. The platform tracks changes in topic popularity in real time, detects trends of public opinion diffusion in a timely manner and supports advance response preparation. For financial-industry scenarios, the platform strengthens automatic retrieval and correlation analysis of multidimensional information, including macroeconomic data, industry policies, financial reports and capital market fluctuations. It can rapidly identify potential hotspots and provide editorial teams with in-depth topic directions and data support.

3. Public opinion event tracking and early warning

Public Opinion Monitor can conduct full-process tracking of brand volume, campaign releases, product updates and negative public opinion around a specific event or competitor. It analyzes public opinion fluctuations across different periods and platforms, helping enterprises adjust their content and market strategies in a timely manner. Based on the analysis results, the system can issue early warnings about potential public opinion risks and help decision-makers respond promptly.

4. Public opinion report generation

Public Opinion Monitor automatically generates detailed public opinion analysis reports according to user needs, covering hot topics, sentiment distribution, public opinion trends, risk analysis and other content. It presents public opinion data through charts and data visualization, making analytical results clear at a glance and convenient for decision-making and reporting.

5. Public opinion response and management

Based on the nature and development trend of public opinion, the system automatically generates response strategies and suggestions, helping users take prompt measures and guide public opinion appropriately. It also provides integrated public opinion management tools that support users in issuing statements, responses and clarifications, as well as conducting public opinion guidance and interaction.

6. Historical public opinion data analysis

Public Opinion Monitor provides a historical public opinion data query function, enabling users to compare public opinion changes across different periods and events, identify trends and improve response capabilities. The system includes a rich case library of public opinion handling. Users can refer to past successful experiences and lessons to avoid repeated mistakes.

III. Project Highlights

As an enterprise centered on AI technology, Midu has more than 70 patents across multiple technological fields, including NLP, big data processing, intelligent text detection, cross-modal information retrieval and large models. Public Opinion Monitor deeply integrates core technologies such as NLP, big data analytics and machine learning, realizing intelligent empowerment across the full news workflow of planning, gathering, editing, reviewing and publishing. With a full-stack self-developed algorithm system and vertical large models, it integrates small models with large models as well as general-purpose large models with vertical large models, enabling comprehensive intelligentization of public opinion analysis.

Figure 2 Full-Stack Self-Developed Algorithm System and Vertical Large Models

The system uses NLP technology to conduct sentiment analysis, topic extraction and keyword recognition on textual information, significantly improving the accuracy and processing efficiency of public opinion monitoring. With big data analytics capabilities, it efficiently stores, processes in real time and analyzes massive information from multiple dimensions, providing users with timely and comprehensive public opinion dynamics. Through machine learning algorithms, it also supports intelligent prediction of hot topics and in-depth analysis of public opinion trends, continuously enhancing system automation and intelligence.

(I) Modular Architecture Design

Public Opinion Monitor adopts a flexible modular architecture, allowing core functions such as hotspot discovery, public opinion monitoring, competitor analysis and communication tracking to be deployed independently or used in combination.

First, according to the business scale, budget range and specific needs of different media organizations, such as vertical fields including sports and entertainment, the platform can be quickly customized and adapted. Second, the modular design reduces coupling among functional modules, facilitating subsequent secondary development and upgrades.

(II) Accumulated Vertical Industry Solutions

During project implementation, Public Opinion Monitor has accumulated standardized solutions for common pain points in the media industry. These solutions cover the entire chain, including data access, model configuration, workflow orchestration and report generation. They feature full-process solution coverage as well as standardization and reusability. From data access to report generation, the platform provides a complete solution. Media organizations can use these solutions directly without building systems from scratch, and the standardized solutions enable rapid reuse by different organizations, shortening implementation cycles.

(III) Continuous Iteration and Service Support

Relying on the large-model technology platform of Midu, Public Opinion Monitor has established a continuous data iteration mechanism to ensure that models and strategies are continuously optimized as time passes and industries evolve. By continuously collecting and analyzing new public opinion data, the platform can optimize models and strategies and improve the accuracy of public opinion monitoring and the reliability of early warnings.

The platform also provides long-term and reliable technical support and service assurance. It offers 24/7 services to ensure that user issues can be resolved promptly during use. At the same time, the platform regularly provides users with training and guidance to help them use the system more effectively.

(IV) Application Scenarios

1. Public opinion monitoring and demand forecasting analysis

Through NLP technology, Public Opinion Monitor can capture e-commerce platform reviews, hot words on industry forums and social media public opinion, and automatically generate product pain point analysis reports, providing enterprises with accurate market demand forecasts.

The system can retrieve product reviews on e-commerce platforms in real time, apply NLP technology for sentiment analysis and keyword extraction, and identify consumer satisfaction, dissatisfaction and improvement suggestions for products.

The system can also monitor hot words and topic trends in industry forums, helping enterprises understand industry dynamics and changes in market demand. For example, before an industry summit, the system detected a continuous rise in discussion about the application of AI in the media field. The enterprise accordingly adjusted its reporting strategy, focused on related content and attracted significant attention. At the same time, the system can monitor public opinion dynamics on social media in real time, and identify public sentiment orientation and points of concern regarding an event or product through sentiment analysis and keyword extraction.

The system can also automatically generate product pain point analysis reports based on captured data, providing valuable market feedback for enterprises and supporting product iteration and stronger market competitiveness.

2. Competitor dynamics tracking

Public Opinion Monitor can automatically parse new product information, patent layouts and promotional strategies from competitor official websites, generate visual competitive landscape maps and provide enterprises with comprehensive competitor analysis.

The system can retrieve new product information from competitor official websites in real time, including product features, functional descriptions and release dates, and conduct in-depth analysis through NLP and machine learning technologies to identify innovation points and market positioning of competitors.

The system can also monitor competitor patent layouts, including the number of patent applications, patent types and patent technology fields, and use big data analytics to identify competitors' technological innovation directions and R&D priorities.

Based on parsed competitor information, the system can generate visual competitive landscape maps that intuitively display key information such as market performance, technological innovation directions and promotional strategies of competitors, providing comprehensive support for targeted market competition strategies.

IV. Implementation Results

(I) Performance Increment Analysis

Since the implementation of Public Opinion Monitor, enterprises have achieved significant growth in multiple key performance indicators. Through intelligent tools, content output has increased and the efficiency of capturing news leads has improved substantially, with average daily content output increasing by 25 percent.

The system enables minute-level discovery of hotspots, allowing enterprises to respond quickly to emergencies. According to statistics, the average response time to hot events has been reduced from two hours before implementation to within 15 minutes, significantly improving market competitiveness.

Through personalized content recommendation and precise communication, user interaction rates, including likes, comments and shares, have increased by approximately 18 percent. For example, after the system was introduced to a social media account, the average interaction volume of a single piece of content increased from 500 to 600.

With improvements in both content quality and communication effects, enterprises have gained significantly more advertising cooperation opportunities. According to statistics, advertising revenue increased by 15 percent year on year, and advertising fees for some high-quality content cooperation projects doubled.

(II) Effects of Tax and Fee Reduction

The application of Public Opinion Monitor not only improves enterprise performance, but also plays a positive role in reducing taxes and fees.

The application of automation tools reduces the workload of manual monitoring and analysis, lowering labor costs in public opinion monitoring and content editing by approximately 12 percent. For example, after the system was introduced, one department reduced demand for two public opinion analyst positions.

Through intelligent analysis, the system reduces the production of ineffective content and avoids resource waste. At the same time, precise communication strategies reduce marketing costs, bringing overall operating costs down by approximately 8 percent year on year.

As enterprises advance digital transformation, some regions provide tax incentives to companies undergoing such transformation. By introducing Sina Public Opinion Monitor, enterprises successfully applied for and enjoyed relevant tax incentives, further reducing their tax burden.

(III) Efficiency Improvement Data

In terms of efficiency improvement, Sina Public Opinion Monitor has also achieved remarkable results.

Improved information processing efficiency: The application of intelligent algorithms has greatly increased information processing speed. According to statistics, the processing time for a single piece of information has been reduced from an average of five minutes to within one minute, improving overall information processing efficiency by approximately 80 percent.

Improved public opinion early warning accuracy: Through the real-time public opinion monitoring and sentiment analysis system, the accuracy of public opinion early warnings has exceeded 90 percent. For example, during a public opinion crisis, the system issued an early warning and accurately judged the nature and development trend of public opinion, providing strong support for timely enterprise response.

Automated report generation: The automated report generation function has significantly shortened report generation time. According to statistics, the time required to generate a single report has been reduced from two hours to within 10 minutes, while report quality remains stable. Reports are also richer and more detailed, providing strong support for decision-making.

(IV) Application Case

Midu and the Law Enforcement Corps of the Shanghai Municipal Emergency Management Bureau jointly developed the “Dual Assistants” intelligent emergency response solution. By combining the “Emergency Law Enforcement Learning Assistant” agent and the “Emergency Law Enforcement Intelligence Assistant” agent, the solution addresses practical problems such as the growing complexity of safety production accidents, delayed updates of emergency law enforcement knowledge, inefficient cross-departmental collaboration and insufficient risk early warning capabilities. Through the integration of AI and big data technologies, it builds an auxiliary system with intelligent question-answering and real-time perception capabilities. Its knowledge base covers more than 10 key fields, including industry and trade and hazardous chemicals, providing law enforcement personnel with instant answers to regulatory inquiries and rapid location of legal and regulatory basis. The accuracy rate of original text citation exceeds 98 percent, reducing tedious transactional work. The system processes more than two billion pieces of data per day, enables second-level monitoring of emergencies and provides strong digital and intelligent support for emergency management modernization.

Case 8

Automatic Generation of Enterprise Research Briefings for Data-Driven Decision-Making

Through Collaboration Between Two AI Models

Zhongshulian Data (Shanghai) Co., Ltd.

I. Basic Enterprise Information

Zhongshulian Data (Shanghai) Co., Ltd. is a provider of AI vertical applications and data product services focused on the industrial economy. The company provides data resources with compliance certification and is committed to building a trustworthy system for knowledge value co-creation. It provides one-stop digital solutions for government and enterprise clients, serving industrial upgrading and digital transformation.

The company has Level III protection under the Classified Protection of Cybersecurity scheme, multiple software copyrights and Shanghai technology-based small and medium-sized enterprise certification, and has completed the registration of generative AI services in Shanghai. The company is committed to enhancing industrial insight with data intelligence and comprehensively supporting high-quality development of the industrial economy. It addresses business needs related to industrial planning, industrial investment promotion, market monitoring and digital transformation of industrial parks.

Relying on the organizational foundation of its own team and shareholder units, Zhongshulian has built a multidisciplinary team covering product design, data analysis, large language models and technology development. Team members are outstanding R&D personnel with complete professional configuration, a reasonable age structure, rich industry experience and a strong sense of innovation. Core project members all hold postgraduate degrees or above.

II. Case Introduction

This case takes enterprise visits by key leaders as the data entry point. It automatically extracts enterprise entity information, matches names and industrial labels, and integrates the vertical insights of local multidimensional enterprise small models with the global analytical capabilities of general-purpose large models. It generates in-depth briefings covering dimensions such as enterprise status, operation, layout and executives. The system not only enables accurate assessment of enterprise status and potential, but also efficiently identifies investment and development opportunities. It provides data-driven decision support for targeted investment promotion, precise enterprise retention services and scientific empowerment of enterprise growth, and has important practical value for promoting high-quality industrial development and optimizing the business environment.

(I) Background and Needs

In the current wave of digital transformation, enterprise briefing generation has become a key link for governments, enterprises and investment institutions to obtain effective information and support decision-making. Traditional methods of enterprise information acquisition face challenges such as scattered data, fragmented information and untimely updates. AI-based enterprise briefing generation systems integrate multisource data, extract key information and analyze development trends, achieving a value leap from information acquisition to decision support.

Enterprise briefing generation refers to using automation technologies to extract enterprise entity information from various information sources, such as news reports, government announcements, corporate financial reports and social media. It matches industrial labels and uses AI models to generate structured and multidimensional enterprise panoramic reports. Such reports comprehensively reflect the development status, competitive advantages, potential risks and future prospects of enterprises, providing strong support for government investment promotion, enterprise investment and industrial chain collaboration.

Information acquisition scenario requirements

1) Demand for efficient all-channel collection: The system needs to cover core information sources for key leader visits to enterprises, including government websites, local news and meeting minutes. It should support automatic capture of

multichannel data, replace manual article-by-article screening and reduce the time needed to extract a single research item from several hours to minutes, avoiding efficiency losses caused by manual work.

2) Demand for standardized processing of enterprise information through automatic extraction of key information: The system needs enterprise name normalization capabilities, automatically associating short names and former names. It should establish a standardized industrial label system to resolve mixed use of vague expressions such as high-end manufacturing and precision machinery. It should also automatically extract key information such as executive changes and new industrial layouts, ensuring that the enterprise basic information database is complete and reusable and preventing omissions caused by human oversight.

Intelligent enterprise analysis scenario requirements

1) Demand for multidimensional correlation analysis: The system needs to break through the limitations of fragmented information, automatically integrate financial reports, industrial and commercial data, research dynamics and regional industrial policies, and build a full-dimensional analytical framework covering status, operation, layout and executives. It should support policy-linked assessment of enterprise potential and avoid one-sided analysis caused by relying on a single data source.

2) Demand for collaborative model judgment: The system needs to deploy local small models adapted to regional industrial characteristics to accurately identify the fit between enterprises and local needs, such as filling gaps in the new energy industrial chain. It should also integrate the global perspective of general-purpose large models to support comparison of competitive advantages among enterprises in the same industry. Professional models replace subjective manual judgment and provide an objective basis for identifying investment opportunities.

Scenario requirements for precise decision support and compliance assurance

1) Scenario-based rapid output: The system needs customized information templates based on client priorities. Investment promotion emphasizes industrial fit, cooperation emphasizes industrial chains and exchange dynamics, and service emphasizes development potential. It should reduce briefing production cycles from days to hours, solve the problem of delayed decision-making information and avoid excessive reliance on experience by decision-makers.

2) Compliance and timeliness management: The system needs to establish a compliance verification mechanism for enterprise sensitive information, automatically filter privacy risks, achieve real-time data updates and validity checks, and automatically correct outdated information such as enterprise relocation and changes in business scope. This ensures decision security and reliability and avoids problems such as deviations in investment promotion direction and delayed service response.

(II) Implementation Plan

1. System architecture design

The technical architecture of this system adopts a layered design and the principle of collaborative construction by multiple agent modules, ensuring system flexibility, scalability and maintainability. The overall architecture covers data processing, model computation and application display from data collection to final service output.

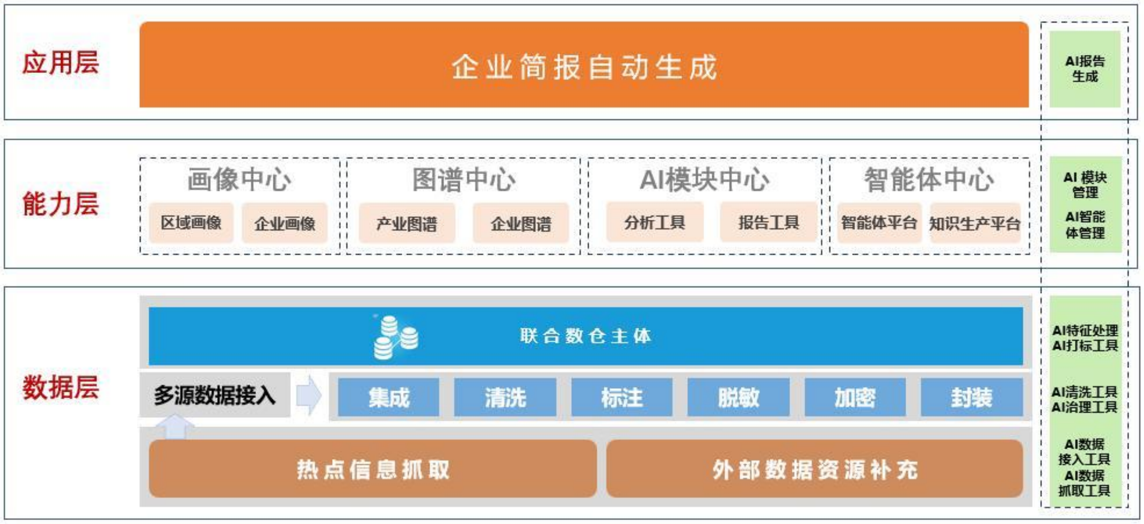


Figure 1 System Architecture

The core features of this architecture are event-driven operation and a hybrid model. The event-driven mechanism ensures that the system can automatically respond to external changes and start processing workflows in real time. The hybrid model architecture skillfully combines the strong general capabilities of general-purpose large models with the precise professional capabilities of domain-specific small models, ensuring both analytical breadth and depth while taking processing efficiency and cost into account.

2. Technical implementation of key agent modules

This system depends on the precise design and collaborative work of several key agent modules. Each agent module is optimized for a specific task, jointly ensuring the automated transformation from raw events to briefing generation.

1) Event capture and extraction agent module: This module serves as the system trigger. It continuously monitors various information sources and automatically captures relevant information when it identifies enterprise visit events involving leaders. The module deploys multiple web crawlers or uses API interfaces to scan news websites, government affairs disclosure platforms, social media such as WeChat official accounts and enterprise announcements in real time. Once it detects information containing keywords such as leader visit, inspection visit or enterprise reception, together with names of government or enterprise leaders, it triggers the collection process. After information capture, the module uses a deep-learning-based event extraction model to accurately extract key information from unstructured text, including the organization leading the visit, participating leaders, the enterprise visited, and the time, location and main content of the visit.

2) Enterprise information matching and enhancement agent module: After an enterprise name is obtained, this module confirms the identity of the enterprise and enriches its basic information. It first matches the extracted enterprise name with a local or external enterprise knowledge graph database. For example, it verifies enterprise identity through optical character recognition (OCR) for business licenses and obtains key identifiers such as the unified social credit code. After successful matching, the module retrieves multidimensional enterprise data from multisource connectors, including but not limited to basic industrial and commercial information, industrial labels, operating status and public opinion data. This information is stored in a structured manner, laying the data foundation for subsequent in-depth analysis.

Basic industrial and commercial information includes registered capital, establishment date, legal representative and business scope. Industrial labels are automatically assigned through an industry classification model. Operating status covers bidding information, patent data and recruitment information. Public opinion data includes news exposure and social media evaluation.

3) Model calling and computation agent module: This is the brain of the system, adopting a hybrid architecture of large model + multiple small models. The general-purpose large model handles tasks requiring strong language understanding and generation capabilities, such as interpreting implicit meanings in research events, generating fluent

natural language descriptions and making comprehensive judgments on enterprises. A series of small models trained for specific fields are responsible for accurate and efficient professional analysis. The enterprise status analysis model analyzes the market share, technological strength and influence of an enterprise within its industry. The operating health analysis model evaluates enterprise operation based on operating data and financial and tax data. The industrial layout analysis model identifies industrial layout priorities from enterprise investment relationships, patent applications and business scope. The intelligent scheduler of this module dynamically calls the most appropriate combination of large and small models according to the briefing template requirements, ensuring that the analysis is both comprehensive and professional.

4) Briefing generation and output agent module: This module integrates the outputs of previous modules into a structured and content-rich enterprise briefing. It usually includes a configurable briefing template that defines the perspectives to be covered, such as enterprise status, operation, layout and executives, as well as the corresponding content format. The module receives analytical results from the model computation layer and uses natural language generation technology to transform structured data into coherent narrative text. For example, data points such as a 25 percent profit margin and 15 percent annual growth are transformed into descriptions such as the enterprise shows strong profitability and maintains rapid growth momentum. Finally, the briefing can be output in various forms, such as a PDF document, a web dashboard or direct delivery to decision-makers' office systems through an API interface.

III. Project Highlights

(I) Overall Highlights

With the core idea of empowering government digital transformation with AI and driving high-quality regional economic development, the project builds an enterprise information service system characterized by intelligent efficiency improvement, precise empowerment and secure compliance. Its highlights focus on three dimensions.

1. AI agents lead precise collaboration between two models and break analytical limitations.

AI agents act as the scheduling hub, linking a general-purpose large model with domain-specific small models. The general-purpose large model is responsible for semantic understanding and global analysis, while domain-specific small models focus on vertical scenarios. Their complementary strengths avoid the limitation of a single model that understands either the region but not the global picture or the global picture but not the region, improving both briefing accuracy and response speed.

2. Precise assessment empowers industrial collaboration and activates the momentum of coordinated industrial development.

By combining local industrial policy models with general-purpose large models, the system intelligently evaluates enterprises from three dimensions: industrial matching degree, development potential and cross-regional cooperation value. It accurately screens investment promotion and cooperation targets that fit the leading industries of the city, marks enterprises that require focused cultivation, helps connect weak links in the industrial chain and activates contact points for regional cooperation, promoting optimized industrial layout and economic restructuring.

3. AI agents anchor government scenarios and realize a value closed loop.

AI agents precisely connect with the government entry point of leader research visits, extract decision-related information from the research process and directly transform it into support for cooperation and exchange services. At the same time, they link with compliance agents to automatically desensitize sensitive data and verify information timeliness. This both fills the gap in transforming research visits into analysis and decision-making and ensures government data security, supporting precise investment promotion, cooperation and regional industrial collaboration.

(II) Key Technological Breakthroughs

1. Hybrid model collaborative computing architecture

This project innovatively builds a three-layer collaborative architecture of general-purpose large model + domain-specific small models + agents, realizing optimized allocation of AI capabilities. The general-purpose large model provides strong semantic understanding and content generation capabilities, supporting long-context modeling and complex logical reasoning. Domain-specific small models trained for enterprise analysis vertical scenarios focus on professional tasks such as industrial label classification and enterprise risk assessment, ensuring the accuracy of analytical results. Agent modules are responsible for workflow scheduling and collaboration, allowing large and small models to perform their respective roles and maximize effectiveness. This architecture overcomes the limitations of a single model that either understands the region but not the global picture or understands the global picture but not the region, while balancing computing efficiency and cost.

2. Event-driven real-time information processing mechanism

The system adopts an event-driven design and can automatically monitor, capture and respond to relevant events such as leader visits to enterprises. Through intelligent crawlers and API interfaces, the system scans multichannel information sources in real time. Once a relevant event is identified, it automatically triggers the full process from information extraction and enterprise matching to briefing generation and push, achieving real-time response capability in which an event is discovered and immediately analyzed. Compared with traditional periodic report production, the enterprise briefing generation cycle is shortened from days to hours, greatly improving the timeliness of decision-making.

3. Multisource heterogeneous data fusion and knowledge graph construction

The project overcomes technical difficulties in integrating multisource heterogeneous enterprise data and builds a multidimensional knowledge graph covering enterprises, executives, operations, industries and policies. Through named entity recognition (NER) and entity linking technologies, the system automatically resolves ambiguity in enterprise names and transforms unstructured text into structured data. This technological breakthrough enables the system to automatically associate enterprise information with industrial policies and provide data support for precise industrial-chain investment promotion.

(III) Scenario Uniqueness

1. Precisely anchoring the government entry point of leader research visits and filling the transformation gap from research dynamics to information analysis and decision support

Unlike general-purpose products on the market that generate briefings from broad-source information such as financial reports and news, this project uses leader research visit events as a dedicated data entry point and precisely matches government decision-making scenarios.

Research visit events themselves are high-value. Enterprises that receive leader attention are often regional key enterprises or potential enterprises, so there is no need to screen core targets from massive information.

The research content contains recent enterprise dynamics and directions of leader concern. AI can specifically extract decision-related information, such as supply chain support needs raised by enterprises, and directly connect it with enterprise retention services. This realizes rapid transformation from research dynamics to information analysis and decision support, solving the traditional problems of information being difficult to accumulate and reuse after research visits.

2. Balanced design that emphasizes both automation and compliance security

While pursuing automation, the system innovatively introduces AI-driven data compliance and quality control mechanisms. The sensitive information identification model can automatically desensitize sensitive data such as revenue and patents and output blurred results according to regulations. The data quality verification model uses authoritative data sources, such as industrial and commercial data and government platform data, as benchmarks, and combines professional knowledge such as government industrial policies and investment promotion standards to verify the timeliness and accuracy of enterprise information in real time. This solves privacy protection and data update issues, and also ensures that outputs are based on objective data and aligned with government needs. Rigor and quality are significantly improved, providing dual assurance for reliable government application.

IV. Implementation Results

(I) Application Scenarios

1. Regional industrial collaboration scenario

When Shanghai conducts industrial cooperation with other regions, such as cities in the Yangtze River Delta, it is difficult to quickly identify complementary enterprises, resulting in low collaboration efficiency. Through the project's industrial label matching algorithm, Caohejing Industrial Park screened upstream and downstream enterprises in the industrial chains of two regions and realized rapid output of enterprise briefings. In the scenario of upstream and downstream industrial chain matching, the working time was improved from a weekly scale to an hourly scale, providing a solid foundation for improving the industrial coordination capabilities of industrial parks.

2. Investment promotion scenario

Traditional investment promotion relies on manual enterprise screening and offline research, and faces problems such as fragmented enterprise information, low precision in industrial matching and long project landing cycles.

Using the enterprise information processing results produced by this project, Shanghai Sanxi transformed its previous model of finding enterprise leads mainly through individual experience into a model that precisely matches local industrial status with potential enterprises through core information such as industrial labels, operating scale, strategic layout, investment needs and technological innovation. This realizes targeted investment promotion. The time for finding enterprise leads was improved from three to seven days to seconds, greatly improving efficiency in investment promotion scenarios.

3. Enterprise development planning scenario

SMEs often lack professional teams to analyze their industry positioning and competitor dynamics, making it difficult to formulate precise development strategies.

Through the industry comparison briefings generated by the project, enterprises can clarify their strengths and weaknesses and optimize product R&D and market expansion directions.

(II) Economic Benefits: From Cost Reduction and Efficiency Improvement to Value Creation

For government departments, the system can reduce labor costs for investment promotion and shorten project landing cycles. Under the traditional model, investment promotion personnel need five working days to manually screen public information of more than 200 enterprises. With this system's automatic matching agent model, the time is shortened to three hours, improving efficiency by 92.5 percent.

For enterprises themselves, information acquisition costs are substantially reduced, strategic decision-making cycles are significantly shortened and the probability of SMEs obtaining cooperation opportunities is improved. Traditional cross-regional enterprise cooperation matching requires manual sorting. For example, connecting enterprise resources between Shanghai and cities such as Suzhou and Hangzhou takes five working days for a single batch. After the system automatically identifies industrial chain needs of Shanghai enterprises, only 1.5 working days are required, improving efficiency by 70 percent.

(III) Technological Innovation Points: Building Core Capabilities of AI + Professional Knowledge + Objective Data

Innovation point 1: intelligent extraction technology for multimodal information

The project breaks through the challenge of text ambiguity recognition. It can accurately extract enterprise names, industrial labels and operating data from unstructured texts such as press releases and research reports, with extraction accuracy above 90 percent, higher than the industry average.

Innovation point 2: collaborative mechanism between local enterprise small models and general-purpose large models

Local small models are responsible for domain knowledge matching, such as industrial label classification, while general-purpose large models are responsible for complex reasoning, such as enterprise potential prediction. Their collaboration improves briefing generation accuracy and shortens response time to 30 seconds per briefing.

Innovation point 3: dynamic knowledge graph construction and real-time updating technology

The project integrates multisource enterprise information in real time, including research information, industrial and commercial data and patent data, to build a dynamic enterprise knowledge graph. Through an event-driven mechanism, the technology continuously monitors changes in various information sources. Once an enterprise information update is detected, such as an address change or executive change, it automatically triggers incremental updating of the knowledge graph, effectively avoiding the lag of traditional enterprise information databases.

(IV) Social Value: Supporting Regional Collaboration and Industrial Upgrading

Value 1: Promoting industrial collaboration in the Yangtze River Delta. The project helps break down information barriers between Shanghai and surrounding cities, promotes cross-regional industrial cooperation projects, drives upstream and downstream collaboration in industrial chains and supports implementation of the Yangtze River Delta integration strategy.

Value 2: Supporting SME development. The project provides industry analysis briefings for SMEs, helping them clarify development positioning and supporting the cultivation of specialized, refined, distinctive and innovative enterprises.

Value 3: Improving government service efficiency. Government departments can shift toward proactive services, such as pushing matched enterprise information in advance. Through the project, they can help some enterprises solve issues such as policy connection and resource matching.

Case 9

Pre-Sales AI Marketing and Customer Acquisition Agent Solution Based on Knowledge Distillation and Few-Shot Learning

Shanghai Guanglin Network Technology Co., Ltd.

I. Basic Enterprise Information

Shanghai Guanglin Network Technology Co., Ltd., operating under the brand Custouch, is a leading business-to-business (B2B) marketing technology service provider in China. Since its establishment, the company has been committed to promoting enterprise digital transformation and upgrading through innovative technologies. One of its core products, the AI Marketing and Customer Acquisition Agent, is an enterprise-level intelligent solution built on the latest AI technologies. It focuses on providing enterprises across industries with all-round services including intelligent marketing customer acquisition, customer service and decision support.

The company's technical team consists of members from well-known internet companies and senior experts, with deep technological accumulation in large language models, NLP, knowledge graphs and related fields. To date, the company has successfully provided digital transformation services for more than 300 leading international brands in sectors such as chemicals, manufacturing, electronics and healthcare, including several Fortune Global 500 enterprises. It has accumulated extensive industry knowledge and project implementation experience.

In R&D, the company continues to invest and has obtained a number of independent intellectual property rights and technical patents. The Custouch AI Marketing and Customer Acquisition Agent not only supports access to multiple mainstream large language models (LLMs), but also innovatively develops core technologies such as vertical-domain knowledge distillation and few-shot learning, providing enterprises with more accurate and efficient AI service solutions.

II. Case Introduction

(I) Background and Needs

1. Project background

The client in this case is a global leader in specialty chemicals with more than a century of technological accumulation and market experience in the chemical industry. Its business covers three core units: care chemicals, catalysts, adsorbents and additives. Its product portfolio contains more than 13,000 highly specialized stock keeping units (SKUs), and its service network spans many countries and regions around the world. As a technology-driven enterprise, the company has always been committed to creating value for customers through innovative solutions. Its products are widely used in dozens of industries, including plastics, coatings, textiles, energy and pharmaceuticals, serving customers ranging from large industrial enterprises to innovative SMEs.

With the rapid development of the global digital economy, traditional manufacturing is facing unprecedented pressure for transformation. The specialty chemicals industry is a typical knowledge-intensive industry, and its service model has long relied on manual services provided by professional technical personnel. This traditional model has obvious bottlenecks. First, the cultivation cycle for professional talent is long and human resource costs are high. Second, knowledge transfer is inefficient, and customers often need to wait a long time for professional answers. Third, service coverage is limited, making it difficult to provide 24/7 uninterrupted service.

The company's product system is extremely complex, containing more than 13,000 SKUs, each with unique technical parameters, application scenarios and usage requirements. Non-professional customers often feel overwhelmed when facing such a large product system. Even through official websites, product manuals and other channels, it is still difficult for them to quickly find solutions that match their needs. This information asymmetry not only affects customer experience, but also leads to the loss of many potential business opportunities. In addition, the company's customer service also faces the challenge of multi-channel collaboration. Customers may make inquiries through official websites, WeChat official accounts, emails, phone calls and other channels, while traditional service models struggle to achieve seamless connection and information sharing across channels, resulting in fragmented customer experience and low service efficiency.

2. Scenario needs analysis

In product selection scenarios, customers often provide only vague requirement descriptions, such as needing "a coating material with corrosion resistance in a high-temperature environment". Under the traditional model, customer service staff need to communicate with customers multiple times to gradually clarify specific technical parameters, usage environments and performance requirements. This process usually takes several days. Moreover, due to limitations in the professional knowledge of customer service staff, technical experts often need to be brought in, further extending response time.

In technical consulting scenarios, customer questions often involve in-depth product knowledge and technical details. For example, a customer may need to understand the conversion rate and selectivity of a catalyst under specific reaction conditions, or compare the performance differences of different products in a given application scenario. These questions require service staff not only to have solid professional knowledge, but also to quickly retrieve and integrate technical materials scattered across various systems.

In sales conversion scenarios, timely identification of high-intent customers and rapid response have become key challenges. Under the traditional model, sales lead screening and assignment mainly rely on manual judgment, resulting in delayed response, subjective assessment and incomplete information transfer. Statistics show that more than 60 percent of sales leads are lost because of delayed responses.

(II) Implementation Plan

1. System construction and integration

Knowledge base construction and structured processing: The project team systematically sorted more than ten types of data sources, including the company's official websites, product databases, microsites, WeChat official accounts, technical white papers, case libraries and safety data sheet (SDS) files. Invalid and outdated data were cleaned, and a structured knowledge graph covering product parameters, application scenarios, solutions and frequently asked questions was built, containing more than 100,000 knowledge entries.

AI model deployment and tuning: The project adopts a "hybrid model + RAG" technical architecture. It connects to LLMs such as DeepSeek and GPT to ensure conversational generation capability. At the same time, RAG technology ensures that every response can be traced back to specific paragraphs in the knowledge base, effectively preventing model hallucinations and improving answer accuracy to over 95 percent.

Omni-channel deployment and consistent experience assurance: The AI agent is seamlessly embedded into the company's WeChat official account, official Chinese website, product microsite in H5 format and on-site interactive screens at exhibitions. This ensures that customers receive a consistent and coherent service experience no matter which channel they use to initiate inquiries.

Backend system integration: Through API connections, the solution is deeply integrated with the company's existing customer relationship management (CRM) system, Salesforce, product information management (PIM) system and marketing automation (MA) platform. It realizes "conversation-to-ticket" conversion: customer information and interaction records are automatically synchronized to the CRM system; high-intent leads automatically generate tickets and are assigned to the corresponding sales staff; product information and pricing strategies are connected in real time to ensure response accuracy.

2. Functional module implementation

Natural language conversation engine: The engine supports Chinese natural language interaction and has capabilities in intent recognition, entity extraction and multi-turn dialogue management. It can understand complex and vague needs such as "my equipment requires a corrosion-resistant coating in a high-temperature environment".

Intelligent product selection and recommendation: The system adopts a combination of "option-guided interaction + open question answering". In the early stage, options are used to quickly narrow the scope of customer needs; then specific parameters are refined through multi-turn dialogue; finally, precise product models or solutions, and even combined solutions, are output.

Embedded conversion touchpoints: In the dialogue flow, the system intelligently inserts call-to-action (CTA) buttons such as sample application, technical document download, engineer appointment and livestream event registration, greatly improving information collection efficiency and conversion rates.

Sales collaboration and lead escalation: Compound rules are set based on keywords, such as "budget", "tender" and "testing", together with behaviors such as repeated comparison of parameters and downloading technical documents. These rules automatically trigger lead scoring escalation, notify sales staff immediately through Enterprise WeChat, email and short message service (SMS), and attach complete dialogue summaries and customer profiles.

3. Testing and optimization

Internal testing and knowledge calibration: Before launch, the sales, technical and customer service teams carried out multiple rounds of real-scenario testing, with more than 2,000 accumulated test cases. The coverage of the knowledge base and the accuracy of answers were continuously optimized.

A/B testing and continuous iteration: After launch, A/B testing was carried out on dialogue flows, scripts and recommendation strategies. Based on real user interaction data and conversion rate data, the AI model and the knowledge base tagging system were continuously iterated and optimized.

4. Core value of the AI solution

In response to the above challenges, the AI agent solution deployed by Custouch creates significant value at three levels.

At the service efficiency level, the agent provides uninterrupted 24/7 service and reduces the response time from more than 24 hours on average to seconds. Through NLP technology, the agent can understand customers' vague needs and rapidly clarify specific technical requirements through multi-turn dialogue, greatly improving service efficiency.

At the knowledge management level, the solution builds a unified knowledge hub, integrating more than ten types of knowledge resources such as product technical documents, application cases and SDS files, and forming a structured knowledge graph with more than 100,000 knowledge points. This not only ensures the accuracy and consistency of knowledge transfer, but also enables continuous knowledge accumulation and sharing.

At the business conversion level, through intelligent intent recognition and lead scoring mechanisms, the system can automatically identify high-intent customers and promptly push them to the sales team. It also automatically generates tickets containing dialogue summaries and customer profiles, helping sales staff quickly understand customer needs and improve conversion efficiency.

III. Project Highlights

(I) Key Technological Innovations and Breakthroughs

This project has achieved several important innovations in the industrial application of AI technologies. These innovations not only address current business pain points, but also provide a replicable technical path for digital transformation in the industry.

The first innovation is the application of vertical-domain knowledge distillation. The specialty chemicals industry is highly specialized and has a complex terminology system, making it difficult for general-purpose LLMs to directly meet business needs. Custouch innovatively uses knowledge distillation to extract and compress chemical-domain expertise from large general-purpose language models into lightweight specialized models. This breakthrough reduces model inference costs by 70 percent and improves response speed by three times while maintaining an answer accuracy rate above 95 percent, making it possible for SMEs to deploy professional-grade AI customer service at low cost and high efficiency.

The second innovation is the successful practice of few-shot learning. Chemical products are updated rapidly, and new products emerge continuously. Custouch has introduced few-shot learning capabilities so that the system needs only three to five sample question-and-answer pairs or technical parameter examples for a new product to quickly understand and master the new knowledge and integrate it into the dialogue recommendation logic. Compared with traditional methods that require large amounts of training data, few-shot learning shortens the launch cycle of new knowledge from weeks to hours, greatly improving system agility and practicality.

In knowledge management, Custouch adopts an advanced architecture combining RAG and multimodal knowledge integration. This architecture ensures that every AI response comes from the enterprise's official and real-time knowledge base, fundamentally eliminating the problem of AI hallucinations. At the same time, the system supports parsing and vector indexing of PDF technical manuals, PowerPoint (PPT) proposals, Excel parameter tables and even chart data in images, building a truly multimodal enterprise knowledge hub. This innovation enables AI to conduct deeper technical interpretation and solution recommendation, rather than merely performing simple question answering.

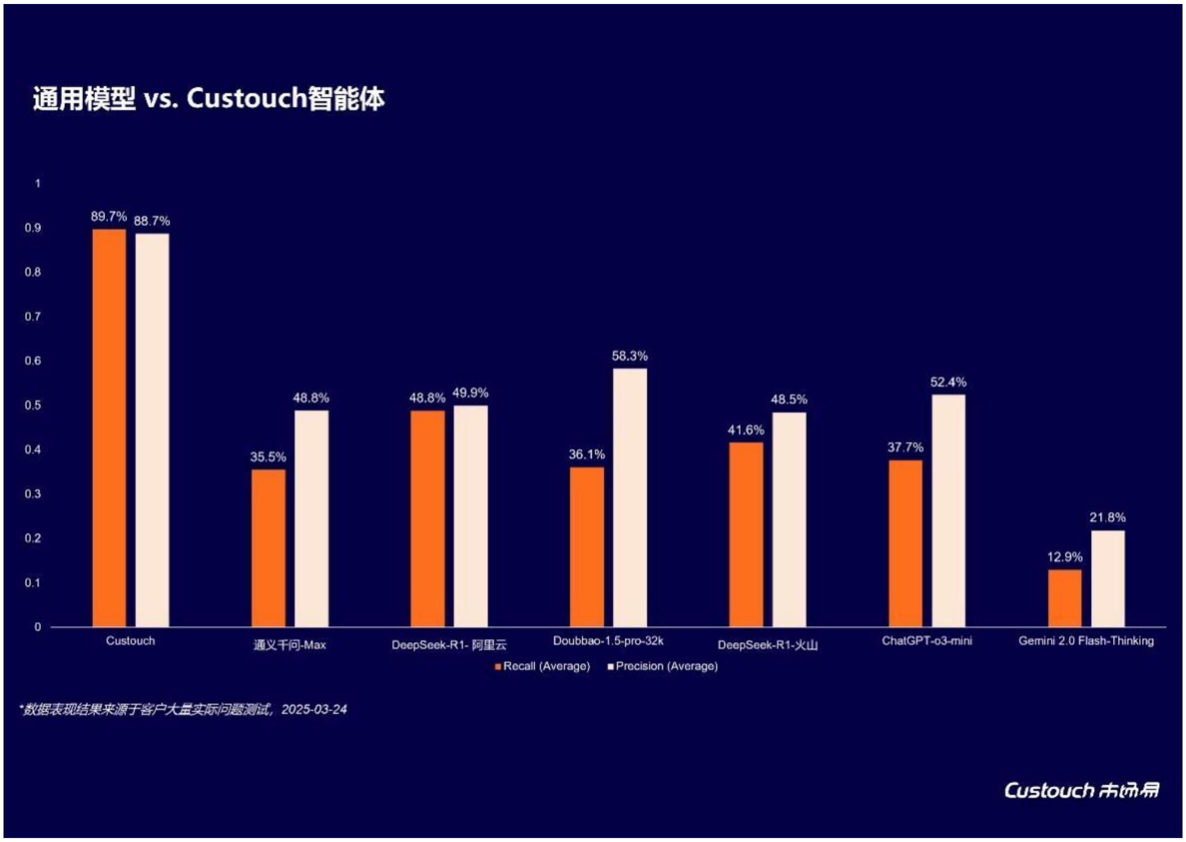


Figure 1 Custouch agent outperforms general models in recall and precision

(II) Innovative Application of Business Scenarios

This project not only achieves technological breakthroughs, but also demonstrates unique innovative value in business scenario applications.

The establishment of an intelligent sales collaboration mechanism is a major highlight. Custouch designed a compound rule engine based on keywords and behavioral analysis. When high-intent keywords such as "budget", "tender" and "testing" appear in a conversation, or when customers repeatedly compare parameters and download technical documents, the system automatically triggers lead scoring escalation. At the same time, through integration with Enterprise WeChat, email and SMS systems, it realizes instant multi-channel notifications for sales follow-up and attaches complete dialogue summaries and customer profiles. This mechanism shortens the average response time for sales leads from 42 hours to within one hour, significantly improving business opportunity conversion rates.

Another innovation is the realization of an omni-channel consistent service experience. Through One ID technology, Custouch connects data across multiple customer touchpoints, including the official website, WeChat official account and product microsite, building a unified user profile system. Regardless of the channel through which customers initiate inquiries, they receive a coherent and consistent service experience. The system also provides personalized product recommendations and services based on customers' historical interaction records, greatly improving customer satisfaction.

In decision support, the project innovatively develops an intelligent product selection engine. The engine combines "option-guided interaction + open question answering" and clarifies customer needs step by step through intelligent dialogue, eventually outputting precise product models or solutions. For complex needs, it can also recommend optimal product combination plans to help customers obtain the best technical solutions. This function enables non-professional customers to complete complex product selection with ease, lowering the professional threshold.

IV. Implementation Results

After comprehensive practical verification in a global specialty chemicals giant, this project has achieved significant returns far beyond expectations. Its results are reflected not only in direct economic benefits, but also in operational efficiency, technological innovation and social value, setting a new benchmark for the industry.

(I) Economic Benefits: Efficient Investment Return and Business Growth in the Case Enterprise

Remarkable ROI and short payback period: In the case enterprise's implementation, the project investment was approximately RMB 100,000. Within six months after launch, by improving lead conversion efficiency, shortening the sales cycle and reducing service costs, the solution directly or indirectly helped create approximately RMB 350,000 in revenue. The ROI reached 250 percent, far exceeding the industry average, and the payback period was shortened to within six months.

Customer acquisition and sales efficiency improved significantly: Practical data from the case enterprise show that the AI agent increased the number of high-quality sales leads by 150 percent year on year and increased sample applications by 100 percent. More importantly, through intelligent product selection and accurate question answering, the average customer cycle from product selection and quotation to signing a sales contract was shortened from 14 days to 72 hours, resulting in a qualitative leap in sales efficiency.

Operating costs were significantly optimized: The AI agent automatically handled about 70 percent of the case enterprise's repetitive pre-sales consulting questions. It is expected to save customers 70 percent of related customer service labor costs each year, allowing human customer service staff to focus on more complex and higher-value customer issues.

(II) Operational Benefits: Building a Measurable Service Upgrade

Service experience improved by leaps and bounds: The solution provides uninterrupted 24/7 second-level response service. Customer consultation satisfaction scores rose from below 70 percent at the initial stage of the project to over 85 percent in a short period and have remained stable above that level.

Knowledge management and application efficiency were renewed: A structured knowledge graph covering more than 100,000 knowledge entries was built for Clariant. Based on this knowledge graph, the AI agent achieves an answer accuracy rate above 95 percent, ensuring the accuracy and authority of professional information delivery.

(III) Outstanding Value of Technological Innovation

This project has important reference value for technological innovation in the industry. First, it successfully verifies the feasibility of applying the technology combination of "knowledge distillation + few-shot learning + RAG" in vertical industries, providing a replicable technical solution for the digital transformation of knowledge-intensive industries. Second, it creates an implementation model for lightweight models in enterprise-level scenarios. Through model optimization and architectural innovation, deployment and operation and maintenance costs are greatly reduced while performance is ensured. In knowledge management, the project explores an effective path for multimodal knowledge integration and solves the pain points of scattered enterprise knowledge assets and inconsistent formats. The established knowledge governance system ensures knowledge accuracy, timeliness and consistency, providing reliable support for enterprise knowledge accumulation and inheritance.

(IV) Far-Reaching Social and Industry Impact

The successful implementation of this project has produced broad social and industry impact. First, it provides an excellent example for the digital transformation of traditional manufacturing. The project won the "AI Scenario Marketing Gold Award" at the 2024 DMSM Golden Marketing Awards, receiving authoritative industry recognition. Second, through lightweight models and low-code platforms, it lowers the threshold for enterprises to apply AI technologies, enabling more SMEs to share the dividends of technological innovation.

In terms of industry promotion, the project experience has been successfully replicated in several related industries. Similar projects have now been successfully implemented in new materials, fine chemicals, pharmaceuticals and other fields, all achieving good results. The technical standards and implementation specifications established by the project are becoming reference standards for the industry and promoting the digital transformation of the entire sector.

The project also generates positive social benefits. By improving service efficiency and quality, it helps customers obtain professional solutions faster and better, indirectly improving the overall efficiency of the industrial chain. Meanwhile, the project’s environmental philosophy and practices reduce carbon footprints by optimizing logistics and reducing business travel, contributing to sustainable development in the industry. Most importantly, this project reshapes the service model and value creation method of B2B enterprises. Through the deep application of AI technologies, enterprises not only improve operational efficiency, but also create new customer value and competitive advantages, opening new possibilities for industry development. This innovative model is being learned from and adopted by more enterprises, driving the entire industry toward greater intelligence and digitalization.

Case 10

Intelligent Analysis and Operational Application of the LLM-Based Tiki Assistant in the Pan-Entertainment Industry

ThinkingData (Shanghai) Information Technology Co., Ltd.

I. Basic Enterprise Information

ThinkingData (Shanghai) Information Technology Co., Ltd. was founded in 2015 and has offices in multiple cities around the world, including Shanghai, Tokyo, Seoul and San Francisco. With the mission of "making data value accessible", ThinkingData is committed to helping global enterprises fully unlock data value from three dimensions: data awareness, data methodology and data tools. The company builds data intelligence infrastructure for global enterprises. Its self-developed data intelligence infrastructure, ThinkingEngine, referred to as the "TE system", provides global enterprises with integrated data-intelligence-driven solutions, meeting their data needs across all categories, dimensions, scenarios and life cycles. The TE system has multiple capabilities, including data analysis, refined operations, data development and community insight. Through deep AI empowerment, it integrates data production and consumption and helps enterprises drive business growth through data intelligence.

II. Case Introduction

(I) Background and Needs

In the early stage of the mobile internet industry, most enterprises focused their data needs on basic data statistics. At that time, data tools were mostly single report systems and could not support deeper business analysis. As market competition intensified and user needs became more diverse, most enterprises gradually realized that basic data alone could no longer accurately locate user pain points. A data insight system capable of responding quickly to business questions has become essential for refined operations in the future.

ThinkingData launched its product in 2018 and released version 5.0 of the TE system in 2025, gradually improving data processing and analysis capabilities. However, in the process of serving customers, it found that even with a powerful system, many customers still faced problems such as low data utilization efficiency, difficulty in reusing accumulated knowledge and long data analysis cycles.

To solve this "last-mile" problem, ThinkingData built the Tiki Assistant based on the AI capabilities of TE System 5.0 and the core capabilities of LLMs, transforming the core capabilities of the TE system into a lightweight tool for natural language interaction.

(II) Full-Link Data Closed Loop

The "full-link data closed-loop system" built by TE System 5.0 forms a complete value chain from data collection to intelligent application through deep collaboration among the open data warehouse, AI platform, AI Agent, Model Context Protocol (MCP) gateway and LLM gateway. This provides a solid foundation for the scenario-based implementation of the Tiki Assistant.

Figure 1　ThinkingEngine product architecture

The open data warehouse serves as the data hub. Through capabilities such as stream-batch processing, ad hoc query and multi-warehouse integration, it realizes real-time integration and processing of multiple data sources such as MySQL and Object Storage Service (OSS). This module synchronizes user behavior logs and third-party platform advertising data to a unified storage layer in real time, increasing the response speed of reach tasks from hours to seconds. The AI platform uses the MCP gateway to solve the platform dependency problem of traditional Function Call, standardizing the interaction protocol between AI models and business systems. Like a "universal socket" for appliances of different brands, it enables the LLM gateway to flexibly connect to enterprise private models and third-party APIs, supporting the dynamic model query capability of the Tiki Assistant.

Therefore, in efficient data query scenarios, the Tiki Assistant can directly index metric data in the open data warehouse. For example, when an e-commerce customer gives the voice command "analyze the payment rate of new users in April", the Tiki Assistant can automatically associate user segment data with payment event tables and generate real-time visualized results.

In addition, the Tiki Assistant can automatically optimize recommendation logic according to product operation data, transforming the traditional process of "data - decision - action - feedback" into a quantifiable and iterative intelligent loop. Through organic coordination among modules, the Tiki Assistant efficiently transforms technological value into business growth.

The multi-cloud elastic infrastructure of the TE system can also pool computing resources and automatically scale during traffic peaks, ensuring the smooth operation of the Tiki Assistant. Deep AI embedding further optimizes resource efficiency. When the Tiki Assistant handles routine requests such as "query the ROI in March", it prioritizes certified report assets to avoid repeated calculations; complex analysis tasks trigger dynamic LLM queries, realizing intelligent allocation of computing resources.

(III) Implementation Plan

The Tiki Assistant is deeply rooted in the TE system. It empowers the entire data production and consumption chain and helps enterprises drive business growth through data intelligence.

1. Real-time user behavior analysis

As a collaborative empowerment tool within the TE system, the Tiki Assistant focuses on releasing data value and improving work efficiency. It builds a full-scenario service system covering data query, tool application, usage support and practical functions. Its core capabilities focus on efficient data acquisition and analysis and provide two precise data query modes. In existing report query scenarios, it can quickly retrieve target reports, such as daily active user (DAU) and ROI reports, flexibly adjust the time range according to analysis needs, and output data with unified definitions and reliable results, such as monthly ROI calculation for March and April. In AI dynamic model query scenarios, the Tiki Assistant can intelligently identify needs, match relevant analysis models and initiate queries, making data-driven decision-making more efficient and targeted.

At the same time, the Tiki Assistant connects to the MCP protocol and can implement core operations such as metadata query and metric query, providing comprehensive support for data development and analysis. In addition, it supports quick query and creation of special date labels in this project, further simplifying workflows and improving overall collaboration efficiency, providing TE system users with a full-process and professional intelligent analysis experience.

Figure 2 Intelligent generation of data charts

2. One-stop refined operations

In operational scenarios, the Tiki Assistant builds a dual empowerment system of "intelligent copy processing + data insight interpretation". It has capabilities in intelligent copy generation, expansion, summarization, polishing and translation, and can efficiently produce push notification copy required for operational tasks. For metric data in the product data analysis module, the Tiki Assistant clearly presents trends, distribution characteristics and correlations of core metrics in a structured form. It automatically identifies data fluctuations and outliers without cumbersome manual analysis, deeply extracts the business logic and potential problems behind the data, and outputs actionable optimization suggestions. This enables data insights to be quickly transformed into actual operational actions, comprehensively improving the precision and execution efficiency of operational strategies.

3. Data development and governance platform

The core value of the AI log interpretation function of the Tiki Assistant lies in locating and interpreting error information in logs through correlation analysis between code and log content. It also provides directly referable correct code examples or implementable solutions, greatly reducing troubleshooting costs and making development, operation and maintenance processes more efficient and smooth.

This function has achieved deep integration and adaptation across multiple scenarios and the entire chain. At the integration level, it supports execution log interpretation for offline integration solutions. During the development stage, it covers analysis needs for product debugging logs and runtime logs. In operation and maintenance scenarios, it can interpret execution logs for a single task or an entire task-flow instance. It also supports query log interpretation in integrated development environment (IDE) scenarios and execution log analysis during the synchronization of space tables to the application layer. It comprehensively covers log interpretation needs from integration deployment to development debugging and daily operation and maintenance, providing technical teams with professional full-process intelligent troubleshooting support.

In addition, the Tiki Assistant can use Text-to-SQL functions to implement the TE product capability based on Amazon Web Services (AWS) GenBI, enabling SQL code generation through natural language.

Figure 3 Rapid generation of SQL statements

4. Self-service business intelligence data visualization

In addition to data queries, the Tiki Assistant can further help product teams complete data insights through rapid report generation and intelligent interpretation. Based on dashboard data, it outputs multidimensional analysis results, presents trends, distribution characteristics and correlations of core metrics, automatically identifies data fluctuations and outliers, deeply extracts data value and outputs actionable optimization suggestions. It also supports follow-up questions on interpretation results, helping users further explore data value. After interpretation, results can be directly generated as notes and displayed at the top of the dashboard for quick access. Later, the Tiki Assistant can continue optimizing dashboard or report content through conversational instructions, including adding, modifying or deleting reports, adjusting key metrics and analysis periods, and renaming dashboards. The entire process requires no complex operations and efficiently meets product teams' data visualization needs.

5. User community analysis platform

The Tiki Assistant can efficiently and accurately interpret player public opinion information in various communities, providing a more comprehensive basis for understanding current product status and optimization direction. It extracts keywords from player-generated content and scores sentiment, accurately identifying player sentiment tendencies and comprehensively perceiving the player community ecosystem. It also displays hot topic trends in real time, enabling the team to obtain current popular information and public opinion trends at the earliest opportunity.

Figure 4 Omni-channel community insight

III. Project Highlights

As a one-stop Agent assistant built into the TE system, the Tiki Assistant relies on the core capabilities of LLMs and covers the full process of data analysis and product operations.

It builds a core capability framework based on LLMs. Through prompt engineering optimization and domain knowledge base customization, it achieves accurate understanding of TE system functions. It supports "existing report query", such as retrieving ROI reports and adjusting time dimensions, and "AI dynamic model query", which automatically matches TE analysis models to generate results. This compresses the data acquisition chain from multi-step operations into natural language interaction.

For the pan-entertainment field, the assistant uses a large number of accumulated industry experience templates and improves the efficiency and reduces the cost of data analysis through a "template center". For example, in report generation scenarios, users do not need to write SQL. By selecting preset templates, they can generate standardized reports, shortening traditional operations from hours to minutes.

The Tiki Assistant also breaks through the functional boundary of traditional customer service chatbots. Through a global navigation entrance, which remains in the top navigation bar of the system, and context awareness, it provides unobtrusive assistance. For example, when users encounter difficulties while using TE event analysis, they can directly call Tiki to obtain guidance on dimension drill-down operations. When editing dashboards, they can call AI to generate outlines, forming a closed loop of "operation - question - solution". Its adaptability in weak network environments and support for localized deployment further strengthen the reliability of enterprise-level applications.

Through dual innovation in technology and scenarios, the Tiki Assistant lowers the threshold of data analysis from professional tool operation to natural language interaction, promoting the TE system's upgrade from a functional platform to an intelligent partner.

IV. Implementation Results

(I) Economic Benefits

Enterprises can significantly reduce operating costs with the Tiki Assistant. Using the Tiki Assistant, Jiujiu Interactive seamlessly accesses both internal and external data across all domains. Single-table queries achieve 100 percent stable success, respond in milliseconds and return accurate and unbiased results. The assistant can also support zero-threshold analysis according to the needs of business personnel, quickly generate visualized reports and help business operations proceed efficiently.

In addition to cost reduction, efficient and rapid data feedback also directly drives product growth by enabling faster implementation of operational decisions. A social platform targeting overseas markets connected to Tiki and optimized its user reach strategy through AI dynamic model queries. As a result, the next-day retention rate of high-value users increased by 8.1 percent, driving average monthly platform revenue growth of 5.3 percent. In addition, the intelligent report generation function helped customers shorten the strategy iteration cycle from 30 days to 14 days, greatly extending product vitality.

The intelligent question-answering module of the Tiki Assistant maintains a product usage issue resolution rate of 80 percent, greatly facilitating the daily work of product teams and further lowering the operational threshold.

(II) Social Value

Starting from "Data + AI", the Tiki Assistant generates reports and answers business questions through natural language. Users can complete data queries without SQL skills. This lowers the threshold for data analysis, allows more people to participate in data analysis and promotes the democratization of data analysis.

Case 11

Digitalized Perioperative Anesthesia Risk Precision Prevention and Control System

Shanghai Lanjia Medical Technology Co., Ltd.

I. Basic Enterprise Information

Founded in 2021 and headquartered in Shanghai, Shanghai Lanjia Medical Technology Co., Ltd. is an enterprise integrating research and development, production and sales, with a focus on integrated solutions for intelligent airway management. Lanjia Medical focuses on high-risk airway scenarios such as anesthesia, emergency care and intensive care. Its business covers core products including video laryngoscopes, disposable electronic airway tools and intelligent stylets, as well as AI-based airway recognition, risk prediction and intraoperative navigation systems. The company has established an R&D and AI laboratory in Shanghai and production and manufacturing bases in Hunan and Suzhou, forming an integrated layout from R&D design and intelligent manufacturing to market sales. In 2023, it completed a Series A financing round of tens of millions of yuan, accelerating the domestic substitution of imported high-end airway management devices. Upholding the philosophy of "technology-driven, clinically oriented and safety-assured", the company continues to promote the R&D of multimodal airway data integration and intelligent decision-support systems, with the aim of improving the safety, standardization and efficiency of clinical airway management and building an independently innovative Chinese technology system and brand for intelligent airway management.

The co-implementing institution, Eye & ENT Hospital of Fudan University, is a Class A tertiary specialized hospital integrating medical care, teaching and research in ophthalmology and otorhinolaryngology. It has leading specialty capabilities in China, houses the National Health Commission Key Laboratory of Hearing Medicine and the Key Laboratory of Myopia, and serves as a standardized resident training base for ophthalmology and otorhinolaryngology under the National Health Commission, as well as a Shanghai standardized specialist training base for ophthalmology, otorhinolaryngology and anesthesiology. The hospital's Department of Anesthesiology features laryngeal mask anesthesia, inhalation anesthesia and difficult airway management as clinical strengths, has nationwide influence in airway management skills training, and has made important contributions to the modernization of hospital medical services.

II. Case Introduction

(I) Project Introduction

Focusing on the core needs of anesthesia safety and preoperative risk management, Shanghai Lanjia Medical Technology Co., Ltd. built a "Digitalized Perioperative Anesthesia Risk Precision Prevention and Control System", forming an integrated platform for preoperative anesthesia risk assessment and intelligent decision support. Relying on the internet hospital system and digital transformation achievements of Eye & ENT Hospital of Fudan University, and closely aligning with the needs of the Department of Anesthesiology, the system carries out technology R&D and system construction. It focuses on high-frequency clinical scenarios such as anesthesia clinics and day surgeries, and builds an integrated online-offline model for preoperative risk assessment.

The system platform is built around the patient mobile terminal, doctor terminal and management terminal. It covers the whole process of online appointment registration, medical history filling, remote consultation, risk stratification, report generation and electronic signature, and realizes interconnection with hospital information system (HIS), picture archiving and communication system (PACS), electronic medical record (EMR), laboratory information system (LIS) and other in-hospital systems. By introducing AIGC and large-model algorithms, the system can automatically parse and extract features from multisource patient data, generate individualized anesthesia risk reports and decision prompts, and significantly improve the accuracy and efficiency of preoperative assessment.

The system has been operating stably across the hospital, providing intelligent preoperative assessment services for thousands of patients nationwide. More than 80 percent of mild cases have completed online assessment, effectively reducing patients' medical costs and improving the quality of medical services. It is an important demonstration achievement in the development of smart anesthesia.

(II) Background and Needs

Against the background of high-load operation in modern hospitals, preoperative anesthesia assessment is an important link in ensuring perioperative safety. It is also a key prerequisite for surgical scheduling, risk control and patient management. However, in long-term clinical practice, preoperative assessment has generally relied on offline consultation, paper-based records and manual judgment, leading to fragmented information, delayed processes and low efficiency. With the rapid growth of surgical volume and the popularization of day surgery, this traditional working model has become increasingly unable to support high-intensity and fast-paced medical needs.

Taking Eye & ENT Hospital of Fudan University as an example, its surgical volume has increased by an average of 6 percent to 8 percent annually in recent years, and the proportion of day surgeries has risen significantly. A large number of patients need to complete preoperative preparation and undergo surgery on the day of admission, while anesthesiologists must complete assessment, communication, risk confirmation and anesthesia plan formulation within a very short period of time, resulting in a sharp increase in workload. Problems such as repeated patient visits, incomplete materials and fragmented information have become major bottlenecks affecting perioperative safety and efficiency. Clinical statistics show that surgical postponements, rescheduling and cancellations caused by insufficient preoperative preparation, information asymmetry or delayed assessment still account for more than 5 percent in some departments.

Traditional anesthesia assessment has several common problems. First, information is fragmented and standards are lacking. There are no unified standards or data interfaces among preoperative visit information, surgical scheduling systems, EMRs and laboratory or examination data. Information transmission relies on manual aggregation and switching among multiple systems, which can easily cause omissions and delays. Meanwhile, risk assessment lacks a unified quantitative standard. Doctors mainly rely on experience-based judgment, leading to significant differences among physicians, and the assessment process lacks objectivity and traceability. Second, efficiency is low and manpower burden is heavy. Traditional assessment is highly dependent on offline consultation and manual records, with each patient taking 10 to 15 minutes on average. For incomplete histories or complex cases, repeated supplementation and visits are often required, consuming a large amount of non-operative time for doctors and affecting overall surgical scheduling efficiency. Third, patient experience and information utilization are insufficient. Some patients receive anesthesia assessment for the first time on the day of surgery due to inadequate preoperative education and risk communication. Incomplete information or anxiety can hinder communication and affect assessment quality. Assessment results cannot be called upon in a timely manner during or after surgery, and data are not incorporated into the hospital's overall quality management system, leaving their value insufficiently explored and used.

With the advancement of hospital digitalization and smart healthcare systems, clinical managers and anesthesiology experts generally believe that preoperative assessment urgently needs to transform from an "experience-based manual model" to a "data-driven model". It must not only improve the efficiency of information collection and circulation, but also realize objective and standardized risk identification to support clinical safety decisions. In particular, as day surgery volume continues to increase and patient conditions become more complex, continuing to use traditional assessment methods will inevitably lead to bottlenecks in assessment work, delayed risk warnings, limited resource scheduling and a series of related problems.

From the patient perspective, as expectations for medical services rise, patients want a more convenient, safer and more understandable medical process. Under the traditional assessment model, patients need to arrive at the hospital in advance, queue, fill out paper medical histories and wait for doctor assessment, which is time-consuming, repetitive and inconvenient. This is especially burdensome for non-local and elderly patients. Clinical feedback shows that insufficient preoperative information communication and long waiting times for assessment have become important factors reducing patient satisfaction. From the doctor perspective, anesthesiology is a department with high work intensity and high time density in most hospitals. Preoperative assessment occupies a large amount of non-operative time, and the assessment data lack structure and reuse value. Doctors must switch among different information systems and manually enter data, which can easily lead to omissions or repeated work and affect assessment quality and decision-making efficiency. How to use technological means to reduce manual workload and improve assessment accuracy and standardization has become an urgent practical issue.

In summary, there is an urgent clinical need to build a preoperative anesthesia risk management system covering patient self-reporting, remote doctor review, automatic risk identification and real-time information sharing. Its core objectives are to standardize and connect assessment data, improve the intelligence and foresight of risk identification, optimize doctors' workflows, improve patient experience, and provide technical support for the hospital to establish a traceable and quantifiable anesthesia safety management system.

(III) Implementation Plan and Problems Solved by AI

With an AIGC large model as its core, the Digitalized Perioperative Anesthesia Risk Precision Prevention and Control System realizes full-process intelligent preoperative anesthesia risk assessment through natural language processing, knowledge graph construction and multimodal data fusion.

First, at the data collection and standardization stage, the system automatically identifies patient-filled medical history information and performs semantic word segmentation, professional terminology matching and semantic disambiguation to ensure that assessment data are structured and consistent.

Second, in intelligent assessment and risk prediction, the model integrates patients' basic information, past medical history, imaging data and examination results, extracts key features and automatically matches assessment scales, enabling individualized risk stratification and early warning for high-risk patients. Third, in report generation and clinical decision support, the system automatically generates structured anesthesia assessment reports based on AIGC text generation algorithms, covering risk levels, precautions and anesthesia recommendations. Doctors can complete confirmation and signature on the terminal, forming a collaborative model of AI initial assessment followed by doctor verification. The system also has long-term memory and continuous learning capabilities, allowing it to maintain information consistency across multiple consultations and update model parameters in real time based on newly added cases, so that assessment results are continuously optimized.

With AI intervention, the average assessment time per case has been reduced from 15 minutes to 6 minutes, and assessment accuracy has improved by about 30 percent. This significantly reduces doctors' workload, improves risk identification, and truly realizes the goal of allowing data to diagnose first and enabling doctors to make more precise decisions.

III. Project Highlights

(I) Key Technological Breakthroughs

The key technologies of the Digitalized Perioperative Anesthesia Risk Precision Prevention and Control System are summarized as follows:

1. Anesthesiology-specific large language model and multimodal fusion algorithm

Targeting the characteristics of anesthesiology, the system independently trains a Chinese medical pre-trained model and combines instruction fine-tuning with reinforcement learning, enabling the model to identify professional terminology, understand anesthesia pathways and analyze drug associations. Through the Transformer architecture and deep feature mapping mechanism, the system jointly analyzes text, images and structured data, automatically extracts key variables related to anesthesia risk, completes multidimensional information fusion and precise modeling, and provides a data foundation for individualized risk prediction.

2. AIGC intelligent content generation and prompt memory mechanism

Based on the anesthesiology knowledge graph and guideline consensus, the platform builds an AIGC intelligent content generation module that can automatically generate standardized assessment reports and conduct medical standardization checks, ensuring that the content is traceable and verifiable. Through prompt engineering, the platform optimizes the model's understanding of complex conditions. Combined with a long-term memory module, it realizes contextual association and information continuity in multi-round consultations, significantly improving the accuracy and consistency of assessment reports.

3. Intelligent assessment system and clinical explainability innovation

The system introduces the modified Mini-Clinical Evaluation Exercise (Mini-CEX) standard and builds an automated assessment system for anesthesia intelligent models. It quantitatively scores medical history collection, clinical reasoning, treatment planning and patient communication, systematically verifying the clinical applicability of the model. Meanwhile, it builds a visual explanation interface and tracking mechanism, enabling autonomous control of the entire chain from information collection and feature extraction to decision generation, and promoting the transition of preoperative anesthesia assessment from experience dependence to intelligent drive.

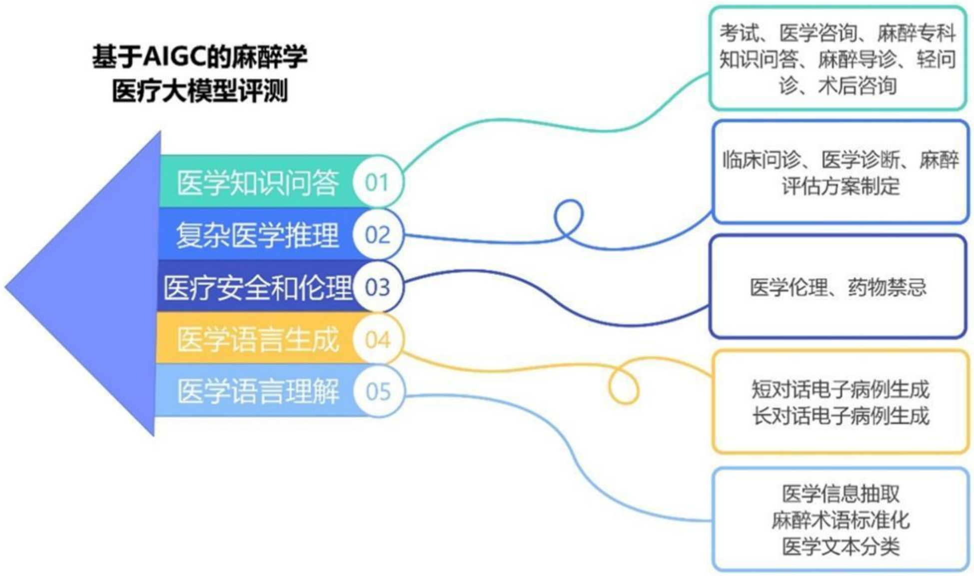


Figure 1 Construction of an AIGC-based knowledge base for a medical large model in anesthesiology

(II) Description of Scenario Uniqueness

The Digitalized Perioperative Anesthesia Risk Precision Prevention and Control System takes the high-frequency preoperative anesthesia assessment process in hospitals as its entry point and builds an intelligent decision-making scenario that can be directly implemented. It has strong clinical orientation and originality. Its scenario features are reflected in the following aspects:

1. Deeply embedded and closed-loop interconnected business system

The system is fully integrated into the hospital information architecture and is seamlessly connected with surgical management, EMRs and anesthesia documentation systems. Assessment results can be automatically returned to preoperative preparation and surgical scheduling modules, forming data sharing and process closure. Online patient information is synchronized in real time to in-hospital terminals, and doctors can check and sign reports immediately at any stage, making the entire assessment process traceable and interconnected.

2. Integrated online-offline model and intelligent triage

The platform adopts a collaborative design of out-of-hospital pre-assessment and in-hospital reassessment. Patients can fill in their medical history and complete remote consultation through the internet hospital, while the doctor terminal automatically receives, supplements and confirms the information, enabling early stratification and precise visits. Based on large-model analysis results, the system automatically recommends reassessment pathways, directing high-risk cases to offline assessment and enabling fast-track access for low-risk patients, thereby improving the utilization efficiency of medical resources.

3. Modular architecture and replicability

The system adopts a modular and standardized design and can flexibly adapt to the information system standards of different hospitals, enabling rapid deployment and smooth upgrades in multiple scenarios. As one of the earliest anesthesia risk assessment systems in China to apply AIGC technology, the project has formed typical digitalized and intelligent scenarios on the clinical, management and patient sides, providing a replicable paradigm for the full-process development of smart anesthesia.

Figure 2 Online process of the Digitalized Perioperative Anesthesia Risk Precision Prevention and Control System

IV. Implementation Results

(I) Economic Benefits

After the Digitalized Perioperative Anesthesia Risk Precision Prevention and Control System was put into operation, the hospital's preoperative assessment process became fully electronic and intelligent, significantly improving operational efficiency and resource utilization. On one hand, anesthesia clinic reception efficiency increased by about 40 percent, offline clinic registration volume decreased by 20 percent to 30 percent, and doctors saved an average of three hours of manual visit time per day. Through online and automated processes, about RMB 1.5 million in labor costs can be saved each year. On the other hand, patients' average waiting time for visits has decreased by about 40 percent, the cancellation rate of day surgeries has fallen by 15 percent, and bed turnover has significantly improved. At the same time, the system reduces paper material use and repeated data entry, improving management transparency and statistical accuracy. After being promoted in the hospital, the platform has driven the application of digital medical equipment and AI tools, formed a basic support system for smart anesthesia management, and saved about 20 percent of integration costs for the hospital's overall information construction, producing significant economic and management benefits.

Figure 3 Entrance of the Digitalized Perioperative Anesthesia Risk Precision Prevention and Control System

(II) Technological Innovations

With intelligent algorithms and the deep integration of medical knowledge as its core, the Digitalized Perioperative Anesthesia Risk Precision Prevention and Control System achieves technological breakthroughs in three directions: preoperative risk identification, model optimization and data security. It builds an explainable, verifiable and replicable intelligent anesthesia assessment system.

1. AIGC-driven individualized risk generation and knowledge fusion mechanism

Based on natural language generation algorithms and an anesthesiology knowledge graph, the system can automatically generate standardized anesthesia risk reports according to patient history and examination data, realizing full-process intelligent generation from data collection to result output. By integrating multiple scale standards such as American Society of Anesthesiologists (ASA) classification, Mallampati score and STOP-BANG, together with rule engine logic, the model realizes hybrid decision-making through deep learning and knowledge reasoning, greatly improving the accuracy and clinical matching of risk identification.

2. Multi-task learning and dynamic optimization model system

The system introduces reinforcement learning and continuous learning mechanisms, enabling it to dynamically adjust parameter weights according to real clinical feedback. Prediction results are continuously iterated and optimized as case data accumulate, gradually forming an adaptive learning model. This system can automatically correct abnormal outputs, maintain high generalization ability across different populations and surgical types, and ensure that assessment results are accurate, stable and updatable.

3. Integrated design of explainability, traceability and security protection

The platform establishes a visual reasoning display and path tracing mechanism, allowing doctors to clearly view the basis of model decisions and key variables, thereby enhancing the transparency and credibility of AI outputs. At the same time, the system adopts Advanced Encryption Standard 256-bit (AES-256) encryption, role-based access control and dynamic desensitization technology, forming a full-life-cycle data protection mechanism.

Together, these innovations build a clinically trustworthy, technically controllable and path-traceable intelligent assessment system, filling a domestic gap in the intelligent and standardized assessment of preoperative anesthesia risk.

(III) Social Value

The implementation of the Digitalized Perioperative Anesthesia Risk Precision Prevention and Control System marks a key breakthrough in the digital transformation of anesthesiology in hospitals. It provides a new paradigm for anesthesia safety management, medical quality improvement and the development of smart healthcare systems.

The system effectively alleviates the shortage of anesthesiologists by automating and standardizing assessment work, greatly improving preoperative assessment efficiency and service coverage, and promoting the improvement of the tiered diagnosis and treatment system and the perioperative risk prevention and control system. At the same time, patients can complete anesthesia assessment and consultation online, greatly reducing the number of hospital visits and waiting time, and significantly improving medical experience and patient satisfaction. The intelligent process design and visual report generation make doctor-patient communication more efficient and transparent, promote the establishment of a patient safety culture, and provide solid support for improving hospital service quality. In addition, the system promotes the structured accumulation of clinical data, provides high-quality sample sets for research, teaching and quality control, and supports the development of evidence-based anesthesiology research. The successful practice of the project promotes the deep integration of AI, knowledge engineering and clinical informatization, and provides a replicable path for building smart healthcare systems and creating a demonstration model for intelligent anesthesia management.

The wider promotion of this system will help more medical institutions realize intelligent preoperative risk assessment, standardized processes and visualized management, and has significant social, academic and industrial value.

Case 12

AIACCLBIO® Protein Engineering Large Model

Shanghai Matwings Technology Co., Ltd.

I. Basic Enterprise Information

Founded in 2021 by an elite team from Shanghai Jiao Tong University, Shanghai Matwings Technology Co., Ltd. is a global leader in AI protein design. The company is committed to using general AI technology to provide integrated AI protein solutions across the full process of protein design, experimental validation, pilot testing, scale-up and industrialization for biomedicine, synthetic biology and many other fields.

R&D personnel account for about 70 percent of the company's staff. The team has interdisciplinary backgrounds in AI, physics, biology and chemistry. Matwings Technology has successfully delivered more than 30 protein projects and has over 40 projects under development. It has gained recognition from many leading industry customers and established close long-term strategic partnerships with numerous upstream and downstream enterprises. At present, the company's technical services have expanded to multiple fields, including innovative drugs, in vitro diagnostics, nutrition and health, food and beverages, beauty and skincare, detergents and textiles, bioenergy, bioagriculture and environmental engineering, demonstrating broad influence and technological strength. Matwings Technology has received many honors, including National High-Tech Enterprise status, Shanghai "specialized, refined, distinctive and innovative" SME status, and the national first prize in the AIGC special competition of the "Maker China" contest organized by the Ministry of Industry and Information Technology. It was selected as a national typical case of AI biomanufacturing by the Ministry of Industry and Information Technology and was recommended as one of five outstanding cases. The company has completed three financing rounds, with investors including Qiming Venture Partners and Bencao Capital. Its cumulative financing has approached RMB 200 million, and its current valuation exceeds RMB 1 billion.

II. Case Introduction

(I) Background and Needs

1. Market background

The global markets for biopharmaceuticals, in vitro diagnostics, industrial enzymes and synthetic biology are all growing rapidly. The biological protein drug market is expected to grow from about USD 550 billion in 2024 to USD 1.2 trillion in 2032. Blockbuster biologics represented by semaglutide generated nearly USD 29.3 billion in revenue in 2024, demonstrating the key value of proteins across the entire biopharmaceutical chain. China's biomedical market reached about RMB 2.14 trillion in 2024, up 8.12 percent year on year, and is expected to grow to RMB 2.24 trillion in 2025. The in vitro diagnostics market was about RMB 137 billion in 2024 and is expected to exceed RMB 140 billion in 2025.

However, domestic enterprises remain highly dependent on imports in areas such as key enzyme preparations and high-end protein drugs. They have long been constrained by the technological and product monopolies of international giants such as Thermo Fisher, Merck, Roche, Novozymes and DSM, and the process of localization has been slow. Facing increasingly fierce international competition and technological barriers, promoting an independently controllable protein technology system has become an urgent need for national strategy and industrial development.

2. Demand for AI technology application

Traditional protein development relies heavily on researchers' experience and trial-and-error experiments, and lacks systematic modeling capabilities. This results in long R&D cycles, usually 2 to 4 years, low success rates and difficulties in multi-objective optimization. In metabolic pathway optimization and protein expression, expression efficiency is generally low due to the lack of system-level analytical tools, further constraining product performance and process development.

Domestic protein products still lag behind international advanced levels in key indicators such as catalytic efficiency, stability and environmental tolerance. AI technology is urgently needed to achieve performance leaps and promote domestic substitution. At the same time, enterprises generally face high R&D investment and long cycles. AI technology can precisely guide sequence design, process optimization and experimental validation, significantly reducing experimental rounds, shortening R&D cycles and lowering overall costs, thereby strengthening enterprises' core competitiveness in the global market.

(II) Implementation Plan

1. Enterprise positioning and core objectives

As an AI for Science (AI4S) enterprise, Matwings Technology aims to integrate "AI + life sciences" to overcome global technical bottlenecks in protein design, including data scarcity, modeling difficulty, slow validation and low success rates. It seeks to build an original technology system spanning underlying data elements, core algorithms and industrial applications, reduce costs and improve efficiency in R&D for biomedicine and synthetic biology, and help China's high-end biomanufacturing industry break through international competitive barriers.

2. Core technology system: the AIACCLBIO® large model

The company's independently developed AIACCLBIO® general AI large model for protein design has two core functions: "AI-directed evolution" and "AI enzyme mining". It adopts the Transformer architecture and a masked language model to learn the relationships among protein sequences, structures and functions. With zero-shot prediction capability enabled by multi-task learning, and reinforcement learning based on a small number of experimental results, the model can transcend structure and directly predict protein function. It creates a new protein design paradigm of "AIACCLBIO® automatic design + small-scale experimental validation", significantly addressing problems in traditional methods, such as reliance on expert experience, high trial-and-error costs, long cycles, low positive rates and difficulty in multi-indicator optimization. It helps customers improve product performance, shorten R&D cycles, reduce production costs and build high-barrier patent portfolios.

3. System architecture and business logic

The platform targets China's urgent demand for high-performance proteins in high-end biomanufacturing. Addressing long-standing global technical bottlenecks in protein engineering design, including data scarcity, modeling difficulty, slow validation and low success rates, it has carried out a series of original technical breakthroughs from the construction of underlying data elements and the innovation of core algorithmic models to the reconstruction of industrial application paradigms. The overall product technology has reached an internationally leading level, achieving multiple global firsts and domestic breakthroughs in data scale, technical route and system integration, and establishing a solid core competitive advantage for China in the strategic frontier of AI4S.

Data foundation: an internationally pioneering ultra-large-scale "public-private fusion" data system

The company has built the internationally leading VenusPod database, with a total of 15 billion sequences, far exceeding similar international datasets. Its core competitiveness lies in the integration of 8.3 billion public data entries and 6.7 billion highly strategic proprietary data entries. The proprietary data come from China's unique deep-sea and salt-lake extreme-environment microorganisms, which contain a large amount of protein "dark matter" naturally endowed with excellent characteristics such as high-pressure resistance, high-salt tolerance and extreme pH tolerance. This provides irreplaceable scarce resources for designing proteins suitable for harsh industrial conditions.

Model fuel: a high-value-density functional annotation system

To address the pain point of "large quantity but low quality" data, the company generated up to 6.5 billion environmental functional labels, such as optimal temperature, pH and pressure, a scale nearly 10,000 times larger than the total of public databases. This allows the training of the AIACCLBIO® model to be conducted on a "super textbook" with dense labels and clear semantics, fundamentally improving the prediction accuracy for protein functions, especially environmental adaptability.

Technical support: governance capabilities for massive multisource heterogeneous data

To activate data value, the company overcame three key governance technologies. First, automated cleaning and standardization: it established an efficient parsing pipeline that supports the processing of terabyte-to-petabyte-level raw data in multiple formats, such as FASTA and Protein Data Bank (PDB). Second, intelligent deduplication and quality control: it uses optimized clustering algorithms to remove redundancy while ensuring data diversity, thereby improving model training efficiency and generalization capability. Third, precise metadata extraction and association: it developed specialized algorithms to accurately extract labels from complex unstructured environmental records and reliably associate them with protein sequences, completing the key transformation from data to knowledge.

III. Project Highlights

(I) Disruptive Innovation in Core Algorithms and Model Architecture: From “Structure Prediction” to “Function Creation”

At the model level, AIACCLBIO® has realized a leap from following the mainstream international technical route to creating an independent technical path. Its core breakthrough lies in abandoning the traditional R&D paradigm that relies on structure prediction and establishing a function-oriented new pathway for protein design.

1. Dual breakthroughs in training efficiency and model performance

Based on the VenusPod database, Matwings Technology successfully trained the AIACCLBIO® large model. Its leading position is reflected not only in the scale of model parameters and training data, but also in an efficiency revolution in the traditional training paradigm. Through its self-developed attention mask bias technology, the company optimized the underlying mechanism of Transformer, enabling the model to automatically ignore sequence padding and minimize computational waste during training. This breakthrough increased training throughput by four times under the same hardware conditions, turning “training a ten-billion-parameter model on 15 billion sequences” from computationally infeasible to engineeringly achievable. The efficiency advantage directly translated into performance leadership. In the ProteinGym zero-shot protein function prediction benchmark established by Harvard Medical School, which brings together nearly 100 top models worldwide, this team’s model series has consistently ranked among the global top three, as shown in Figure 1.

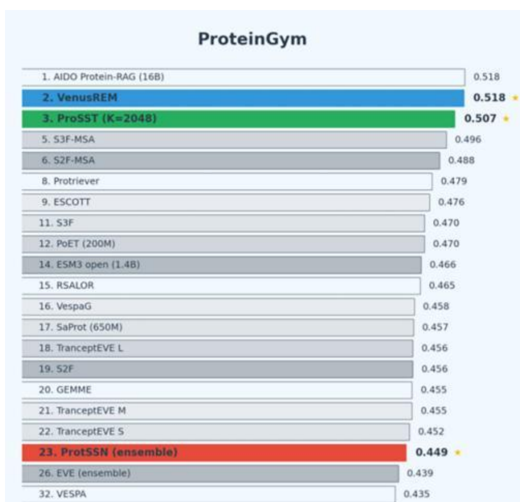


Figure 1 ProteinGym ranking of zero-shot prediction accuracy for protein function

2. Pioneering a “function-oriented protein generation” paradigm

Current mainstream routes represented by AlphaFold follow an indirect pathway of “sequence - structure - function”, which has bottlenecks such as high computational cost and unclear functional mapping. AIACCLBIO® is among the first to realize end-to-end direct generation from “functional requirement to protein sequence”, bypassing reliance on precise three-dimensional structures.

Through deep pretraining on VenusPod’s sequence-environment label pairs, the model establishes strong associations between sequences and macroscopic functions in latent space. By building a multi-task collaborative pretraining framework, the model not only masters the “grammar” of proteins, meaning sequence patterns, but also understands their “semantics”, meaning functional implications. Users can directly input functional intentions such as "design a lipase with the highest catalytic efficiency at 75 degrees Celsius under alkaline conditions", and the model can directly generate candidate sequences that meet the requirements.

(II) Integrated Innovation of System Platform and Industrial Closed Loop: Building a Self-Evolving Protein Design Agent

1. Building an AI-empowered protein data governance platform

The platform creates a data tool chain covering the entire process of "collection - governance - enhancement - service - iteration". It supports complex data query and annotation tasks driven by natural language instructions, realizing a paradigm upgrade from instruction-driven to intent-driven and greatly reducing governance thresholds and labor costs. By exploring a hybrid architecture of sparse Transformer and state-space models, the platform solves unified vector representation and semantic alignment problems for multimodal data, including sequences, structures and text, laying a foundation for cross-modal retrieval and knowledge discovery.

2. Realizing a "dry-wet experiment closed loop" and continuous model evolution

The company independently developed the Venus-Auto high-throughput automated experimental platform, enabling fully automated conversion from AI-designed sequences to purified proteins and performance data. It established a real-time data feedback and online learning mechanism, continuously feeding more than 300,000 precise functional data entries generated by wet experiments back to the AI platform for model fine-tuning. The resulting "design - validation - learning" closed-loop system increases the positive rate of protein R&D from less than 10 percent to 30 percent to 50 percent, and shortens the R&D cycle from years to months or even weeks, solving the industry problem of disconnection between AI models and the continuous knowledge updating of biological experiments.

In summary, the AIACCLBIO® protein engineering large model is a systematic innovation spanning three layers: data, algorithms and strategy platforms. Data innovation: it builds VenusPod, the world's largest protein sequence database, covering 15 billion protein sequences, including 6.7 billion high-value proprietary data entries, of which 700 million are from deep-sea exploration and 6 billion from salt-lake extreme environments; 6.5 billion sequences carry environmental functional labels such as temperature, pH and pressure. Both data scale and label richness are internationally leading, greatly enriching data diversity. Algorithm innovation: it designs a new multimodal deep-learning pre-trained large model for proteins and DNA, enabling end-to-end prediction, learning interaction relationships among biological elements, and designing pretraining tasks for specific scenarios. Strategy innovation: it creates a dry-wet iterative closed-loop pathway, shortening the R&D cycle to 2 to 6 months and increasing the positive rate from 1 percent to more than 30 percent, achieving a dual improvement in protein design efficiency and success rate.

IV. Implementation Results

The AIACCLBIO® protein engineering large model has achieved major success in both academia and industry. At present, the team's AI protein directed evolution module and AI enzyme mining module are largely mature. It has successfully delivered more than 30 commercial projects, with over 40 additional projects under development, and has been validated in multiple real industrial cases.

(I) The World's First Industrially Produced AI-Designed Protein: Alkaline Tolerance Modification of Growth Hormone Purification Filler Protein

In cooperation with Changchun GeneScience Pharmaceutical Co., Ltd., a leading enterprise in long-acting growth hormone, the project optimized the alkaline tolerance of a single-domain antibody used as an affinity chromatography filler in the growth hormone production process. Through sequence optimization by the AI large model, the alkaline tolerance of the target protein increased by about four times, greatly improving protein stability and helping the enterprise reduce costs and improve efficiency in production. The protein has now achieved 5,000-liter industrial production, making it the world's first AI-designed protein to realize industrialized production, as shown in Figure 2.

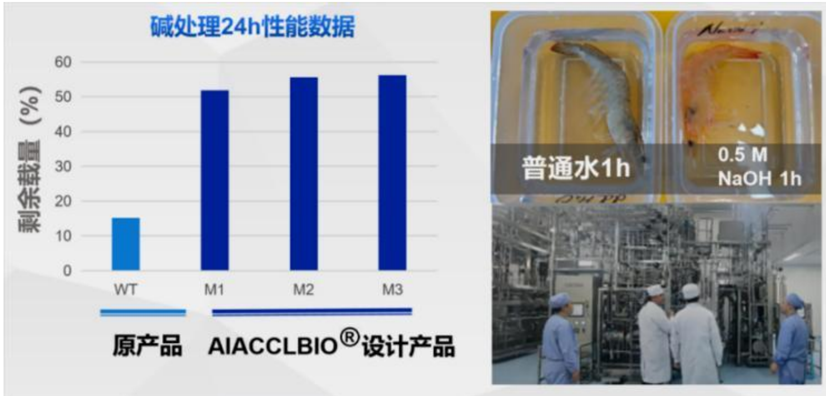


Figure 2 Alkaline tolerance modification of growth hormone purification filler protein

(II) Catalytic Activity Modification of Alkaline Phosphatase

In cooperation with a leading in vitro diagnostics enterprise, the project optimized and modified the activity of alkaline phosphatase (ALP). As a common labeling enzyme, ALP is often used in chemiluminescent enzyme immunoassay (CLEIA) systems. By combining highly sensitive alkaline phosphatase chemiluminescent detection technology with highly specific enzyme-linked immunoassay technology, high-sensitivity Alzheimer’s disease (AD) detection kits can be developed to detect early biomarkers of multiple major diseases. The suppliers of the most active commercial products on the market are Roche and Toyobo. Through optimization by the Matwings Technology AI large model, protein molecules with activity twice that of commercial products have been obtained, significantly improving the detection sensitivity of downstream kit development, as shown in Figure 3.

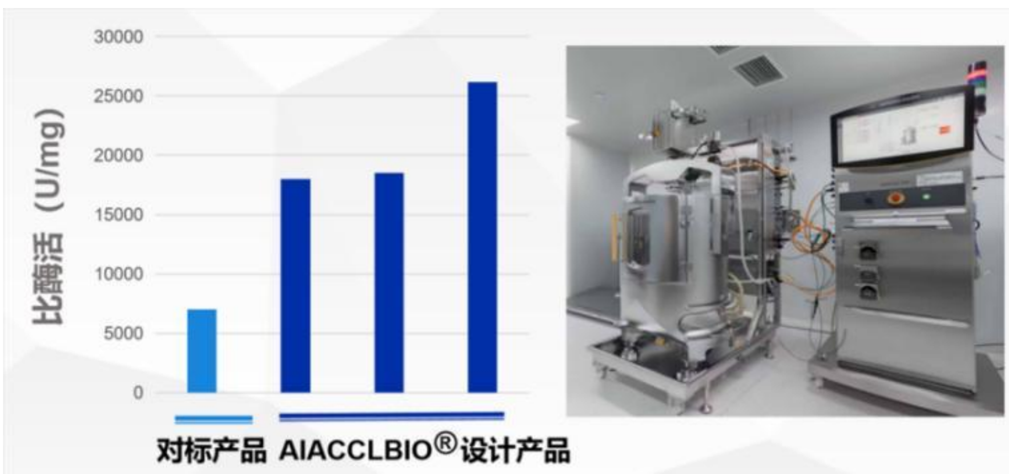


Figure 3 Catalytic activity modification of alkaline phosphatase

(III) Important Domestic Industrialization Breakthrough in Vitamin Derivatives: Enzymatic Production of Vitamin Derivatives

In cooperation with Changxing Pharmaceutical Co., Ltd., the project optimized the performance of a key enzyme used in the enzymatic synthesis of a vitamin derivative, focusing on improving the enzyme’s specific activity, stability and expression level. Under the same substrate loading, glycosyl donor ratio and reaction conditions, the project achieved a 50 percent increase in substrate conversion rate and a significant reduction in enzyme dosage within a short cycle, adding more than RMB 50 million in annual output value for the enterprise. At present, this AI-designed enzyme has achieved stable industrial production at the 20,000-liter scale and has become an important industrialization breakthrough in the field of vitamin derivatives in China.

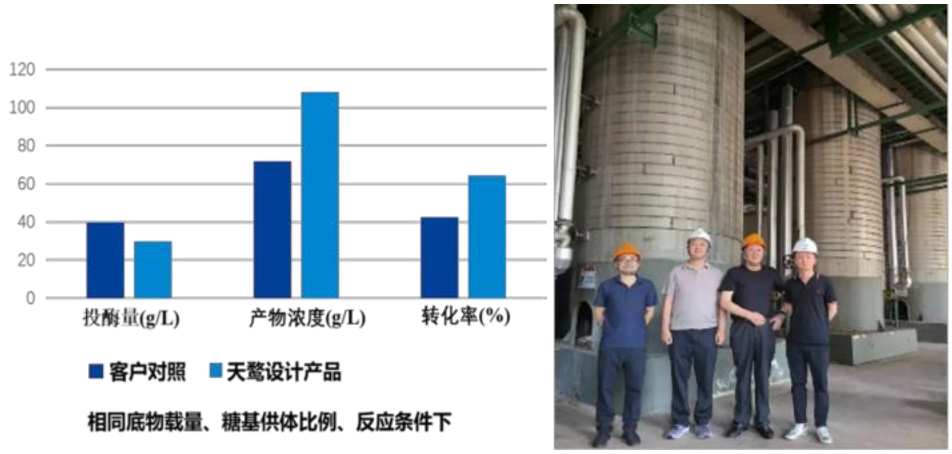


Figure 4 Enzymatic production of vitamin derivatives

Case 13

Mechanism- and Data-Driven Research on Enantioselectivity in a New Asymmetric Catalytic Reaction Based on the Wanhua Large Model

Yijia Zhisheng (Shanghai) Technology Co., Ltd.

I. Basic Enterprise Information

Founded in 2025 and located in Model Speed Space, Yijia Zhisheng (Shanghai) Technology Co., Ltd. is an intelligent-native enterprise focusing on the vertical field of “AI + chemical engineering”. With the mission of “enabling the global chemical industry to use a unified agent”, Yijia Zhisheng is committed to promoting the comprehensive transformation of the chemical industry from “experience-driven” to “intelligence-driven”. It provides customers with a series of “AI + chemical engineering” products and solutions covering the full lifecycle of chemical business forms, helping the chemical industry achieve green, high-quality and sustainable development.

The company has two core competitive barriers. The first is its data flywheel. Relying on nearly 30 years of accumulated chemical industry knowledge, it has built a professional chemical engineering data corpus. Through its self-developed intelligent chemical data corpus management platform and intelligent model training platform, the company continuously aggregates and activates full-lifecycle chemical data, forming a self-reinforcing data loop. The second is its technology flywheel. Based on the Wanhua multimodal chemical large model, the company realizes deep integration of industry knowledge, equipment data and AI algorithms, and builds a continuously evolving technical system. This system systematically addresses core pain points of chemical enterprises, such as fragmented systems, long R&D cycles, multiple safety hazards and high costs, while promoting intelligent upgrading across the entire process.

As an innovation-driven technology enterprise, the company has jointly established an “AI + Chemical Engineering Joint Laboratory” with the Shanghai AI Laboratory for Science, and has reached strategic cooperation with enterprises including Copilot Technology and DataCanvas. Together, they are building an open and collaborative intelligent chemical industry ecosystem. With “connecting the full value chain of chemical AI” as its strategic core, the company is determined to become the world’s first comprehensive service provider covering all business forms of AI + chemical engineering, and to contribute to the intelligent transformation of the chemical industry.

II. Case Introduction

(I) Background and Needs

In the field of modern fine chemical synthesis, the efficient and highly selective preparation of chiral molecules is both a core challenge and a key technical bottleneck. Asymmetric catalytic reaction, a synthetic strategy that achieves stereospecific conversion through chiral induction, is one of the core technologies currently used. However, the industrial production of chiral molecules is constrained by the limitations of existing catalytic systems. Low enantioselectivity leads to the formation of large quantities of by-products during the process, while traditional separation technologies often cannot achieve effective purification. This greatly increases downstream treatment costs and leads to an increase in three wastes. Therefore, developing a new catalytic system capable of highly enantioselective conversion for asymmetric catalytic reactions is of great scientific significance. The catalytic process involves complex multivariable interactions. The relationship between different components in the reaction system and reaction performance is difficult to quantitatively describe, resulting in a high degree of dependence on experience in catalyst screening and condition optimization, as well as a high cost of trial and error.

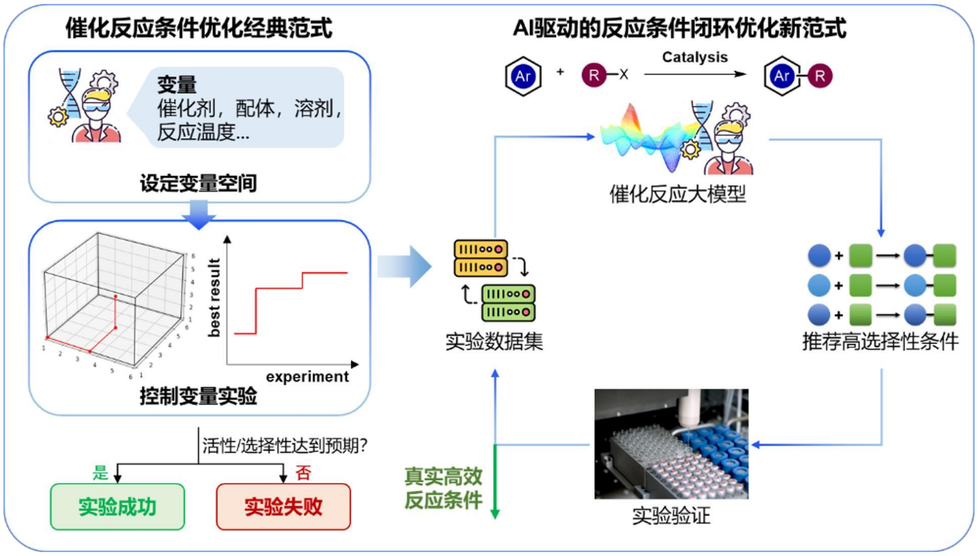


Figure 1　Differences in the catalytic reaction condition-screening paradigm before and after AI application

(II) Implementation Plan

To address the above challenges, more efficient and intelligent research methods are urgently needed to overcome the limitations of traditional optimization models in efficiency and accuracy. Specifically, the project implementation plan is built by introducing AI technology.

First, a rapid search scheme for high-dimensional condition space is introduced. The numerous condition variables in asymmetric catalytic reactions form an extremely high-dimensional parameter space, which is difficult to cover effectively through manual experiments. AI models can learn from historical data and build mapping relationships between variables and reaction outcomes. As a result, they can efficiently screen potential high-performance condition combinations in virtual space and greatly reduce the number of experimental validations required. Second, a precise prediction scheme for enantioselectivity is developed. Because the reaction mechanism is complex, experience alone cannot accurately judge the enantioselectivity performance of a certain combination of conditions. Machine learning methods, especially models with strong nonlinear fitting capability, can accurately predict the enantiomeric excess (ee) value and provide a reliable theoretical basis for experimental design. Third, a knowledge transfer and iterative optimization scheme is adopted. In the early stage, the target reaction may have relatively few data records. By transferring knowledge from structurally similar reactions, the model can be pretrained or initialized. After each round of experiments, the new results are fed back into the training set and used to continuously update the model, thereby forming a closed-loop optimization mechanism and progressively improving prediction accuracy and experimental success rates.

III. Project Highlights

(I) Key Technological Breakthroughs

To achieve the above objectives, this case has made four technical breakthroughs. First, in data collection and standardization, unified data recording specifications were developed to make data machine-readable and efficiently usable. The required field information for each experimental entry was clarified, including reactant structure, catalyst type and dosage, solvent type and volume, temperature, time, additive information, conversion rate and ee value. Data cleaning and validation processes were established to remove outliers and duplicate records and to ensure dataset quality and consistency. Second, in reaction characterization and feature engineering, the reaction system was digitally represented in a reasonable manner. A descriptor system was constructed from multiple dimensions, including reactant structure, catalyst properties and solvent parameters. Quantum chemical calculation or molecular dynamics simulation methods were introduced to theoretically calculate key transition-state structures and extract electronic structure parameters such as bond order, charge distribution and orbital energy levels. Environmental parameters such as steric hindrance and solvation effects were also used to achieve deep encoding of the reaction mechanism. Through feature selection and dimensionality reduction, a high-information feature set that fully reflects the essence of the reaction was eventually formed. Third, in model construction and algorithm optimization, the company's Wanhua model was used to fit the relationship between reaction inputs and enantioselectivity. Through cross-validation and hyperparameter tuning of different models, the prediction model with the best performance on the test set was selected. This ultimately produced a robust model capable of accurately predicting enantioselectivity under different conditions. This case fully relies on the company's independently developed Wanhua large model, as shown in Figure 2, and the molecular generation model, as shown in Figure 3. Fourth, in condition screening and iterative experimental validation, the predictive model was used to carry out large-scale virtual screening of candidate condition combinations. Top-ranked conditions with predicted ee values higher than the preset threshold of 90 percent were selected for experimental validation. If the current experimental results did not reach the final target of an ee value higher than 95 percent, the newly obtained data were added to the training set, and the model was updated and retrained. In this way, the model's predictive capability and generalization ability were gradually improved. Through multiple rounds of "prediction - experiment - feedback" iteration, the optimal reaction conditions were approached until the target of stably exceeding 95 percent was achieved. This not only solves specific problems in current reaction optimization, but also explores a new data-driven research pathway and provides technical accumulation and methodological reference for similar future topics.

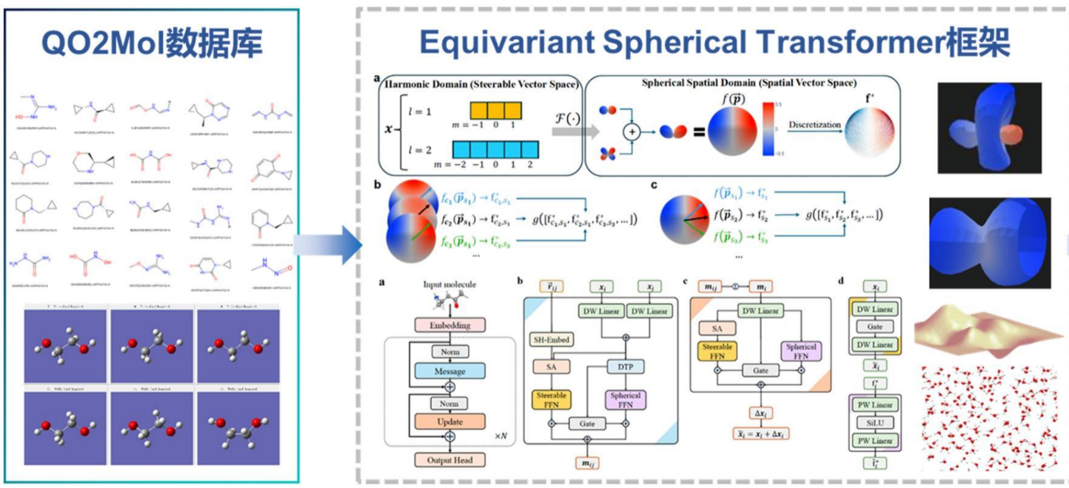


Figure 2 Wanhua large model for the chemical industry

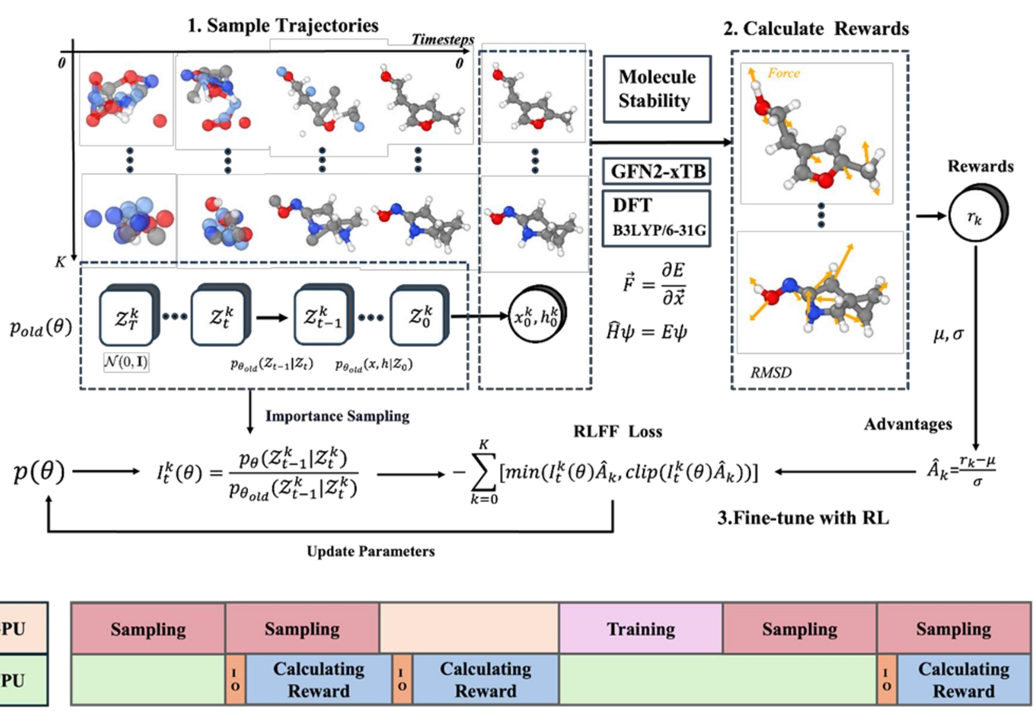


Figure 3 Molecular generation model

(II) Uniqueness of the Scenario

This case is unique in its integration of technologies. It pioneers a deeply integrated model of "AI large model + quantum chemical calculation + molecular dynamics simulation", breaks through the limitations of traditional research and realizes mechanism- and data-driven reaction analysis and optimization. This forms differentiated technical advantages in the research of asymmetric catalytic reactions. It is also unique in data application. To address the pain point of scarce target-reaction data in the early stage of the project, a closed-loop iterative mechanism of "prediction - experiment - feedback" was built, enabling efficient R&D breakthroughs in data-scarce scenarios. It is unique in goal orientation, focusing on the core target of an ee value greater than 95 percent. AI is used to precisely predict ee-value performance across a high-dimensional reaction condition space, filling a gap that traditional research has difficulty meeting for industrialization needs. Its value output is also distinctive: it substantially reduces experimental trial-and-error costs, shortens R&D cycles, improves product purity and provides strong support for the industrial production of high-value-added chiral molecules.

IV. Implementation Results

(I) Economic Benefits

Through an AI-driven optimization strategy, this case has significantly reduced the R&D cost of asymmetric catalytic reactions. It has already been implemented at Shanghai Huayi New Materials Co., Ltd. under Huayi Group. The Top 10 accuracy of retrosynthesis prediction for the reaction exceeds 0.6. The project empowers human efficiency and saves labor costs, directly saving more than RMB 500,000 in reagents, equipment and labor inputs based on a medium-scale case estimate. The R&D cycle has been substantially shortened, with potential benefits ranging from several million to tens of millions of yuan. The application of this case has significantly improved product purity and reduced by-product separation costs. For high-value-added chiral drugs or fine chemicals, annual incremental revenue may reach the ten-million-yuan level, forming a virtuous cycle of cost reduction, cycle shortening and benefit improvement.

(II) Technological Innovation

The case builds a hybrid-architecture catalytic reaction model that integrates data from quantum chemical calculations and molecular dynamics simulations. It has strong high-dimensional nonlinear fitting capability and breaks through the technical limitation that traditional empirical judgment cannot precisely predict ee values. To address the pain point of scarce target-reaction data in the initial stage of the project, transfer learning is used to reuse historical data from related reaction systems for model pretraining, solving the problems of low data utilization efficiency and delayed optimization in traditional research. Based on historical experiments and data from similar reactions, the AI model is trained to build mapping relationships between variables and reaction results, thereby efficiently locating potential high-performance condition combinations. The project innovatively integrates mechanism-level data from quantum chemistry and molecular dynamics, builds a mechanism- and data-driven feature set, reveals key factors influencing enantioselectivity, and provides a new method for understanding and optimizing asymmetric catalytic reactions.

(III) Social Value

This case empowers the upgrading of the pharmaceutical and fine chemical industries, safeguards the supply of key products, reduces dependence on imported high-end chemicals, helps the pharmaceutical and chemical industries transform toward efficiency and high-end development, and supports the security of industrial and supply chains. Its multi-technology integrated research model overturns the inefficient research path of traditional manual trial and error, enabling small and medium-sized research teams to conduct high-level research with the help of AI technology. It promotes overall innovation efficiency in catalysis and realizes inclusive access to advanced technology. By reducing resource waste and pollution, it also supports China's dual-carbon goals. In addition, it cultivates interdisciplinary talent, strengthens the talent foundation for industrial innovation and development, and helps the country build a high-level innovative talent system.

(IV) Expansion of Application Scenarios

This case can be applied in multiple practical scenarios. In innovative drug R&D, the research results of this case can be used to build an AI free energy perturbation (FEP) computing platform and construct key affinity prediction models to test specific target conformations. In experimental design for other reaction condition optimization tasks, the research results of this case can be fully used to accurately analyze historical research data, establish an automated process for generating reaction mechanism information, and provide reliable technical methods for reaction condition optimization in industrial catalysis and other production practices. In scenarios such as sustainable polymer material design, new electrolyte design, perovskite photovoltaic cell stabilizer design, optimization of highly enantioselective carbon-hydrogen bond activation reaction conditions, and formulation design models for daily chemical products, this case can also provide transferable technical support and methodological reference.

Case 14

Application of AI in an Intelligent Drug Development Platform

XImaging Intelligent Technology (Shanghai) Co., Ltd.

I. Basic Enterprise Information

XImaging is a company committed to developing AI-empowered unmanned intelligent operating systems and AI unmanned work systems. With AI, cloud computing and robotics as its core technologies, and unmanned laboratories as its implementation scenario, the company provides intelligent solutions for life sciences, medical and health care, scientific research and other fields, helping industries achieve AI-driven transformation and upgrading.

The company's founder and chief executive officer (CEO), Dr. Liu Jiapeng, graduated from Shanghai Jiao Tong University. He has spent 15 years in machine vision and AI big data application research, and has led the development of intelligent systems such as super intelligent laboratories and robotic workstations. The company brings together outstanding employees from well-known universities in China and abroad, including Shanghai Jiao Tong University, ETH Zurich and the University of Science and Technology of China, as well as leading domestic and international enterprises such as WuXi AppTec, Hengrui, Saint-Gobain of France and bioMerieux, laying a solid foundation for technological innovation and product R&D.

The company has decomposed and refined more than 100 standardized operation modules, covering over 90 percent of common experimental steps. The iMagicOS driver library has connected with more than 1,000 instruments and devices, enabling the drag-and-drop construction of arbitrary workstations, functional islands and whole-laboratory solutions through modular standardized products. The company has accumulated extensive experience in smart laboratory projects and established multiple industry benchmark cases, including projects for the National Institutes for Food and Drug Control, SGS, Insilico Medicine, WuXi AppTec, Porton Bio, State Grid and the Chinese Academy of Sciences.

II. Case Introduction

(I) Project Introduction

With breakthroughs in technologies such as AI reinforcement learning, drug R&D is undergoing a profound transformation from "experience-based trial and error" to "data-driven + AI decision-making". Traditional laboratories rely on manual operations and face problems such as low efficiency, with a single experimental cycle taking as long as 2 to 3 weeks, data silos, with data utilization below 30 percent, and high operational error rates. These issues severely constrain the iteration efficiency of AI models. Industry data show that traditional drug R&D requires an average of 10 to 15 years, costs USD 2.6 billion and has a success rate of less than 10 percent.

As a global leading AI pharmaceutical company, Insilico Medicine urgently needed to optimize its laboratory R&D process for validating new targets through automation and data management technologies, addressing pain points such as low efficiency of traditional manual operations, with daily sample processing capacity below 100 samples, high error rates of 5 percent to 8 percent in manual operations, and decentralized data management. Against this background, Insilico Medicine and XImaging reached a strategic cooperation and jointly launched the construction of a fully automated AI intelligent drug development platform, with a total project investment of RMB 50 million and a construction period of 12 months.

The platform is positioned as the world's first fully automated robotic laboratory assisted by AI decision-making. It deeply integrates core technologies such as AI, robotic automation and multimodal perception. It covers key links including target discovery, compound generation, high-throughput screening and personalized drug development, and aims to shorten the drug discovery process from the traditional 4 to 6 years to less than 18 months, significantly improving the success rate of drug R&D.

(II) Background and Needs

Improving efficiency across the entire drug R&D chain is an urgent need in the industry.

At the automation level, automated guided vehicles (AGVs), robotic arms and other automated equipment are needed to connect various experimental links and realize automated sample processing, with daily processing capacity required to reach more than 10,000 samples, and precise compound library management, supporting real-time tracking of more than one million compounds. This requires the platform to be highly integrated and collaborative, with equipment linkage response time below 100 milliseconds.

At the AI technology level, advanced algorithms such as AI and reinforcement learning need to be deeply integrated. In the screening stage, AI models are required to predict drug activity, with prediction accuracy above 85 percent. The platform also needs to be capable of processing petabyte-level multi-omics data and supporting real-time data analysis and model iteration.

At the system-performance level, the platform needs to greatly shorten the R&D cycle, with a target reduction of more than 60 percent, reduce trial-and-error costs, with a target reduction of 50 percent, and shorten the validation cycle for a single target from the traditional 6 months to within 2 weeks.

(III) Implementation Plan

1. System architecture design

The platform adopts a three-layer distributed architecture of “functional islands + AI middle platform + robot scheduling system”.

Modular design of functional islands: The 1,000-square-meter laboratory is divided into 8 standardized functional areas, including compound management, cell culture, high-content imaging and sequencing library construction. Standardized interfaces are used between modules to support hot swapping and rapid reconfiguration.

AI middle-platform integration: The platform deeply integrates the three core engines of Insilico Medicine’s Pharma.AI platform: the target discovery engine PandaOmics, which integrates 21 biological algorithms; the molecule generation engine Chemistry42, which integrates 45 AI models; and the clinical prediction system inClinico, which is based on the Transformer architecture. The platform processes an average of 2 terabytes (TB) of data per day.

Robot scheduling system: Through XImaging’s independently developed iMagicOS system, version 3.2, the platform uniformly schedules AGVs, robotic arms and humanoid robots, enabling 24/7 unmanned operation. The task allocation optimization algorithm increases equipment utilization to 92 percent, far higher than the industry level of 70 percent.

2. Key technologies and business logic

AI-driven target discovery and molecular design

PandaOmics target discovery engine: This engine integrates 12 types of multi-omics data, including genomics, proteomics and metabolomics, to achieve causal inference for targets and pathway reconstruction. Its causal inference model reduces the false-positive rate of target screening by 40 percent, and the accuracy of pathway prediction reaches 89 percent.

Chemistry42 molecule generation engine: Based on structure-based drug design and deep reinforcement learning, Chemistry42 can generate small-molecule compounds with specified properties such as high affinity and low toxicity from scratch. It generates more than 100,000 new molecules per day. Virtual screening is accelerated through quantum mechanical simulation and molecular docking, improving screening efficiency by 100 times and completing 5 million molecular docking calculations per day.

inClinico clinical prediction system: The system uses an ensemble learning framework integrating 8 machine learning models to predict the success rate of clinical trials. Its prediction accuracy for the success rate from Phase II to Phase III clinical trials reaches 81 percent, 35 percent higher than traditional methods.

Fully automated experimental execution

Sixth-generation LifeStar laboratory: The laboratory uses 18 AGVs and an AI decision-making system to realize unmanned operation across 10 core links from target validation to compound screening. Overall equipment effectiveness (OEE) reaches 90 percent. It generates 15 TB of multidimensional experimental data per day and improves optimization efficiency by 5 times.

Introduction of the humanoid robot “Supervisor”: As a typical application of embodied intelligence, reinforcement learning enables the robot to achieve pipetting precision of plus or minus 0.5 microliters, far exceeding the plus or minus 5 microliters level of manual operation.

Data closed loop and digital twin

Experiment-model iteration loop: Wet-experiment data generated by the laboratory, such as cell activity and metabolite concentration, are fed back in real time to the Pharma.AI platform through a high-speed data bus with a transmission rate of 10 gigabits per second (Gbps), and are used to optimize the prediction capability of AI models.

Multisource data fusion and standardization: The platform integrates internal experimental data, 56 public databases such as ChEMBL and The Cancer Genome Atlas (TCGA), and partner data, building a standardized dataset containing 8.5 billion records.

Application of digital twin technology: Physical simulation and machine learning are used to build a digital twin model of the experimental process, predict results under different experimental conditions and reduce the number of actual experiments by 67 percent.

III. Project Highlights

(I) Full-Stack AI Platform Driving an Experimental Closed Loop

The fully automated AI intelligent drug development platform jointly developed by XImaging and Insilico Medicine uses the Pharma.AI platform as its core and integrates the three major modules of target discovery, PandaOmics, molecule generation, Chemistry42, and clinical trial prediction, inClinico, forming a complete closed loop from data to decision-making.

AI brain for autonomous decision-making: The PandaOmics platform has 21 biological algorithms and 85 pretrained models. Based on multi-omics data, it autonomously identifies and ranks targets, increasing the accuracy of target discovery by 50 percent compared with traditional methods. The platform supports non-specialists in completing complex R&D tasks through natural-language interaction. In a case test, a high school student with a biology background used the platform to successfully discover a new target and publish an international paper.

Seamless connection between dry and wet experiments: The AI system controls 46 robotic arms in real time to complete sample processing, high-throughput screening and other operations. In the R&D of the anti-fibrosis drug ISM001-055, the platform completed the entire process of target discovery and validation within 14 days, shortening the process by 88 percent compared with traditional methods. The platform optimizes experimental paths through real-time feedback. In fibrosis drug R&D, it dynamically adjusted the compound generation strategy based on the results of the first three rounds of experiments, and the activity of compounds generated in the fourth round increased eightfold.

High-performance computing support: The platform uses the Intel Xeon CPU platform to optimize the inference efficiency of complex models such as AlphaFold2. Through high-bandwidth memory (HBM) and Advanced Matrix Extensions (AMX) technology, model inference speed is increased by 33 times and inference cost is reduced by 60 percent. Amazon Web Services provides elastic computing support, with peak use of 5,120 virtual central processing unit (vCPU) cores and an average daily processing volume of 45 TB of experimental data.

(II) Multi-Module Collaboration and Flexible Extension Architecture

Standardized functional unit design: The platform is divided into 8 functional islands, each using standardized interfaces and protocols. The compound management island realizes full-lifecycle tracking of more than one million samples through an automated storage system, with access accuracy of 99.99 percent. The cell culture island integrates environmental monitoring and automated passaging equipment, controlling culture environment parameters with a precision of plus or minus 0.1 degrees Celsius. The high-content imaging island is equipped with 12 imaging devices and has a daily processing capacity of more than 5,000 samples.

Deep scenario adaptation of humanoid robots: The deployed bipedal humanoid robot “Supervisor” has mastered 126 experimental operations through motion-capture learning, and can perform delicate tasks such as pipetting, with precision of plus or minus 0.5 microliters, and device operation. The robot achieves a task completion rate of 98.5

percent in complex scenarios, solving the problem of poor adaptability of traditional AGVs in unstructured environments.

Scalable technical middle-platform architecture: The Pharma.AI platform supports plug-in development. By integrating the large-model Copilot system, it enables non-specialist users to use professional tools through natural-language interaction. The platform provides 256 standard API interfaces and supports rapid integration with mainstream laboratory systems. The development cycle for new functions has been shortened from 3 months to 2 weeks.

(III) Core Algorithms and Technological Innovation Breakthroughs

Multimodal molecular design: Chemistry42 integrates 45 AI algorithms, including 28 deep-learning-based generative models and 17 optimization algorithms. In the R&D of fibroblast growth factor receptor 2/3 (FGFR2/3) inhibitors, the platform generated a compound library containing 150,000 molecules. Through absorption, distribution, metabolism, excretion and toxicity (ADMET) prediction and free-energy calculation, 23 candidate molecules were screened, of which 7 showed nanomolar-level activity in experimental validation. A highly selective inhibitor with a novel scaffold structure was successfully developed.

High-throughput precision screening technology: Through combinatorial optimization algorithms, the platform increases screening throughput to an average of 100,000 compounds per day, 50 times higher than the traditional method of 2,000 compounds per day. The deep-learning-based activity prediction model reaches an accuracy of 87 percent, and the hit rate of virtual screening increases from the traditional 0.1 percent to 3.5 percent.

Intelligent experimental planning and optimization: Bayesian optimization and multi-objective reinforcement learning algorithms are adopted, improving the efficiency of experimental design by 8 times. In anti-tumor drug screening, intelligent experimental design discovered a lead compound in only 3 rounds of experiments, compared with 15 rounds required by traditional methods, saving RMB 12 million in R&D costs.

IV. Implementation Results

(I) Economic Benefits

The fully automated AI intelligent drug development platform was built with a system value of RMB 50 million and took 12 months from launch to delivery. After the platform was put into operation, it brought significant economic benefits to Insilico Medicine.

Operational efficiency achieved a leapfrog improvement: the number of experimental personnel was reduced by 80 percent, sample and reagent consumption was reduced by 50 percent, and annual operating costs decreased by RMB 28 million. Time utilization increased by 3 times, throughput increased by 4 times, and R&D output per unit time increased by 5 times. The platform supports continuous 24/7 operation, and OEE reaches 92 percent, far higher than the industry average of 70 percent.

The R&D cycle was shortened in a revolutionary way: Through this platform, Insilico Medicine advanced its idiopathic pulmonary fibrosis pipeline from target discovery to the preclinical candidate compound stage in only 18 months, with R&D cost of USD 2.6 million. Compared with traditional methods requiring 3 to 4 years and USD 20 million to USD 30 million, R&D efficiency increased by 5 times and cost was reduced by 87 percent. At present, the platform has advanced 8 pipelines to the preclinical stage and is expected to save the company more than USD 200 million in R&D investment.

Standardization and industry-empowerment value: The 128 standardized operation modules abstracted from the platform have formed industry solutions and have been successfully promoted to 12 AI pharmaceutical enterprises, generating cumulative technology licensing revenue of RMB 85 million. Covering more than 1,000 square meters and integrating 8 functional islands and 102 analytical instruments, the platform realizes full-process intelligence across 16 core links from sample pretreatment to sequencing library construction. It is currently the largest and most complex automated laboratory in Asia.

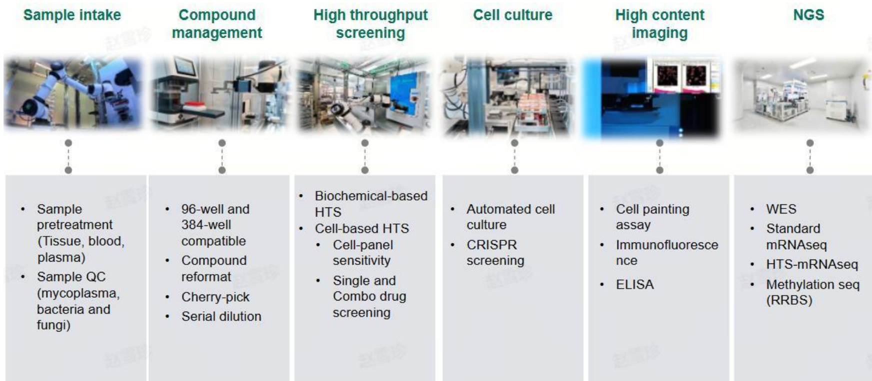


Figure 1 Six functional modules of the robotic laboratory

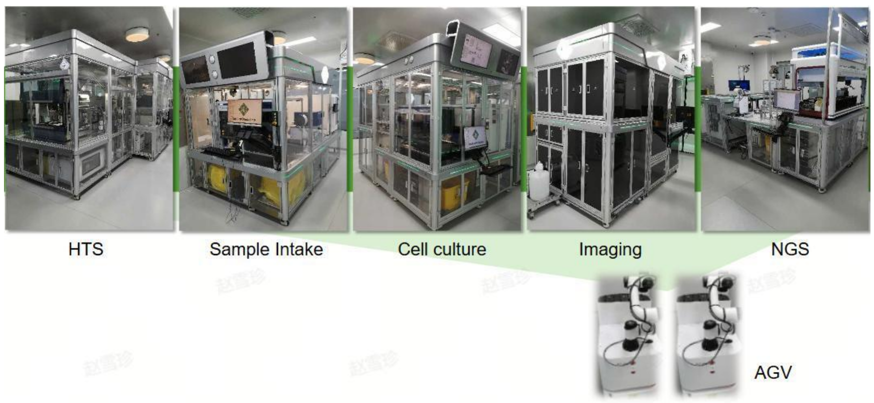


Figure 2 On-site images of functional islands

(II) Social Value

Accelerating breakthroughs in the treatment of major diseases: The platform focuses on difficult diseases such as fibrosis and tumors, and its AI-driven target discovery capability has broken through bottlenecks in traditional R&D. The first anti-fibrosis drug entirely discovered by AI, ISM001-055, has entered Phase II clinical trials, bringing new therapeutic hope to patients with idiopathic pulmonary fibrosis. The platform is expected to shorten the average time for new drug R&D from 10 years to 6 years, benefiting tens of millions of patients worldwide.

Technology inclusiveness and industry-standard innovation: By accurately predicting drug safety at an early stage through AI models, the platform reduces unnecessary animal experiments by 65 percent and advances the ethics of drug R&D. Recognition of the platform's AI prediction models by the United States Food and Drug Administration (FDA), based on the FDA Modernization Act 2.0, has promoted this technical route as a new industry standard. Its pipeline ISM001-055 has received FDA orphan drug designation and breakthrough therapy designation from China's Center for Drug Evaluation (CDE), providing a practical case for the upgrading of regulatory paradigms in the industry.

Industrial ecosystem aggregation and upgrading: The platform integrates technologies such as robotics, AI and synthetic chemistry, driving the collaborative development of 32 upstream hardware suppliers and 28 downstream contract research organization (CRO) enterprises. Insilico Medicine's Shanghai Zhangjiang R&D center has attracted 46 upstream and downstream enterprises to gather, forming an "AI + biomedicine" industrial cluster, increasing the scale of the Zhangjiang biopharmaceutical industry by 35 percent and helping Shanghai build a world-class biopharmaceutical industry hub.

Talent development and employment-structure optimization: The platform has created new jobs in the interdisciplinary field of AI + biology, trained more than 280 compound talents who understand both biotechnology and AI algorithms, and promoted the transformation of employment structures toward higher-end positions. It has established joint laboratories with 8 universities and trains 150 interdisciplinary graduate students each year, continuously supplying innovative talent to the industry.

Case 15

AIP Intelligent Modeling Platform and RIO Intelligent Strategy Generation System

Dual Systems Empowering Enterprise AI Applications

Shanghai AugereTech Information Technology Co., Ltd.

I. Basic Enterprise Information

Shanghai AugereTech Information Technology Co., Ltd. (AugereTech) is a professional modeling company in the field of AI. Led by data scientists and professors, the company has a full-stack algorithm engineering team as well as a professional marketing service team specializing in Internet user operations, campaign planning and operations. Its self-developed AIP efficient modeling platform provides AI solutions integrating AI model consulting, intelligent application development and digital-intelligence operations services, helping enterprises embrace AI-driven new quality productive forces and enter the era of intelligent operations 3.0 with ease. AugereTech owns dozens of authoritative certifications, patented technologies and software copyrights. Its chief scientist, Ouyang Yujun, is a renowned expert in the fields of financial and retail modeling and intelligent operations.

II. Case Introduction

(I) Background and Needs

As global digital transformation deepens, enterprises have an increasing demand for AI-driven scientific decision-making. According to data from International Data Corporation (IDC), the global AI solution market is expected to grow from USD 150 billion to USD 500 billion from 2023 to 2027, with a compound annual growth rate of 28 percent. The Asia-Pacific region is expected to grow by more than 35 percent, while AI platforms in vertical industries such as finance and manufacturing in China are becoming core growth drivers at a growth rate of 40 percent. In 2024, enterprise demand for AI models increased by 65 percent year on year. Seventy percent of enterprises believe that if they fail to achieve large-scale AI application within one year, they will face pressure on market share. Meanwhile, the required frequency of model iteration has become higher, making traditional modeling approaches difficult to adapt.

Against this background, AugereTech's AIP automated modeling platform offers multilayered value. It helps SMEs overcome bottlenecks in AI talent and technology, enables low-cost scientific decision-making and improves ROI. It also promotes digital-intelligence upgrading across multiple fields, supports the formation of new quality productive forces, lowers the threshold for AI application, and maximizes the inclusive and industrial value of AI capabilities.

At present, pain points in traditional modeling constrain the entire industry chain. First, AI decision-making talent is scarce. Professional data scientists are highly paid, team maintenance costs are high, and the whole modeling process remains heavily dependent on manual work. Second, most enterprises remain at the stage of data collection and visualization, lacking goal-driven AI decision logic, while existing solutions find it difficult to take responsibility for business outcomes. Third, most models are developed in a customized way, with limited reuse across scenarios, making it difficult to convert data resources into growth momentum. These factors highlight the urgent need for automated modeling platforms.

(II) Overall Project Objectives

In this context, AugereTech independently developed the AIP base platform and the RIO intelligent application deployment platform. Centered on "efficiency upgrading and value implementation", the AIP automated modeling platform uses coordinated design across four dimensions - efficiency improvement, quality optimization, cost control and platform stability - to build an end-to-end target system covering "process, performance, cost and operations and maintenance". Each dimension focuses independently on optimization in its corresponding area, while also forming a logical linkage in which efficiency improvement supports cost control, quality optimization safeguards platform value, and stability underpins full-process implementation, thereby achieving overall performance beyond traditional models.

The four dimensions form a closed-loop logic of “process efficiency improvement - cost optimization - quality implementation - stability support”. Efficiency improvement is the foundation. By shortening modeling cycles, accelerating iteration and improving data processing timeliness, it reduces resource occupation and supports cost control. Quality optimization is the core of value. By improving model accuracy and deployment success rates, it builds competitiveness while avoiding repeated investment and reducing hidden costs. Cost control is the direct output. It saves labor and tool costs and serves as a quantified result of efficiency improvement, reflecting commercial value. Platform stability is the guarantee for implementation. Reliable operation ensures the continuous achievement of the first three objectives.

Specifically, the efficiency improvement dimension focuses on the systematic compression of the time required for the entire modeling process. Taking the traditional model as the benchmark, it sets key performance indicators (KPIs) such as modeling cycles and iteration frequency, achieving large-scale improvements in process efficiency and providing a foundation for subsequent cost optimization. The quality optimization dimension is oriented toward performance compliance and practical usability. It benchmarks model accuracy and other indicators against core performance parameters of traditional modeling, such as AUC and mean absolute error (MAE), ensuring the value of the models output by the platform while reducing resource waste caused by quality defects. The cost control dimension refers to traditional resource allocation and sets KPIs such as labor cost savings, converting process efficiency into measurable cost benefits. The platform stability dimension focuses on system availability, fault recovery and data security, ensuring long-term and stable support for the implementation of enterprise AI applications.

Through the logical linkage of “process efficiency to reduce costs - quality assurance to safeguard value - stability to support implementation”, all dimensions form a target closed loop and jointly promote the comprehensive efficiency upgrade of the AIP automated modeling platform. The platform packages the complex decision modeling process into low-threshold and reusable applications, enabling enterprises to build expert-level AI decision models rapidly and at low cost and promoting the popularization of intelligent productivity.

(III) Implementation Plan

AugereTech’s AIP intelligent modeling platform and RIO intelligent strategy generation system are dual systems built around the core objectives of full-process empowerment, low-threshold implementation and high-value conversion. Through modular architecture design, layered technical support and scenario-based technological innovation, they build a full-chain AI application solution covering data processing, model construction, strategy generation, deployment and iteration, providing efficient and reliable support for intelligent transformation in multiple industries.

1. Core Contents of System Construction

The dual systems focus on the full-process needs of enterprise AI applications and form a collaborative support system of “modeling empowerment + strategy implementation”. With modular design at its core, the AIP platform builds four major functional modules. The rapid modeling module uses visual drag-and-drop operations and presets operators covering data preparation, model training and evaluation, enabling non-technical personnel to quickly build business processes and shorten project startup cycles. The feature generation module supports multi-source data import and custom operator development, balancing standardization in data processing with personalized needs of different industries. The backtesting engine module relies on step-by-step execution, quantitative evaluation and data lineage tracing to ensure model reliability and adapt to high-precision scenarios. The iterative optimization module supports automatic model updating and performance monitoring, helping enterprises respond quickly to changes in business scenarios.

System Architecture Design

The dual systems adopt a layered collaborative architecture to ensure functional adaptability and operational stability. The architecture is divided into three layers. The user interaction layer provides a web-based visual interface and integrated development tools, meeting both the drag-and-drop operation needs of non-technical personnel and the code editing needs of data scientists. The functional module layer includes core modules such as project management and data management. Through standardized API interfaces, it enables data flow, function invocation and collaborative linkage between AIP and RIO, ensuring the smooth advancement of business. The data and resource layer relies on multi-source data storage technologies and containerized resource scheduling to realize secure storage of massive data and elastic allocation of computing resources, providing underlying support for stable system operation. The sandbox management module uses environment isolation technology to support safe verification and controlled release of models and strategies.

Core Technologies and Business Logic

Based on scenario-based technological innovation, the dual systems build a core logic of "technology adapting to scenarios and strategies matching business needs". At the technical application level, for supply chain and manufacturing scenarios, the company has developed a "dual-engine" algorithm. It combines a self-developed unsupervised algorithm that captures naturally occurring business patterns with an open-source supervised multi-algorithm selection mechanism. Through innovative technologies such as dynamic model matrices, multi-model chain decision-making and algorithm optimization, it breaks through the bottlenecks of long implementation cycles and high implementation costs in traditional modeling, improving production efficiency, supply chain turnover and market response speed.

For financial and retail scenarios, AugereTech has developed a multi-model operations decision system and an improved uplift model. It uses technologies such as mixed re-division of experimental and control groups, analysis of variance (ANOVA), and threshold variable screening to build concise and stable predictive models, improving the effectiveness and accuracy of strategy matching.

At the level of business logic, the systems follow an end-to-end logic of "data - modeling - verification - strategy - implementation - iteration". First, the AIP platform completes multi-source data processing and high-quality model construction. Then the RIO system converts models into business strategies, which are verified through sandbox testing and deployed into the production environment. At the same time, the iterative mechanisms of the dual systems continuously optimize models and strategies based on business dynamics and indicator feedback, ensuring that AI applications always fit actual enterprise needs and that technical value is efficiently converted into business value.

III. Project Highlights

(I) The Intelligent Modeling Platform Lowers the Technical Threshold for AI Modeling

By integrating core technologies, the AIP automated modeling platform packages professional AI modeling and optimization processes into visualized and automated workflows. Ordinary business personnel can complete modeling without mastering underlying principles, greatly lowering the threshold for AI application. Its core support capabilities are reflected in two aspects. First, automatic feature engineering simplifies the process through preset and custom feature processing operators. Users can automatically generate and screen features by configuring parameters without writing complex code, and can also use the Structured Query Language (SQL) editor to quickly query and process data to extract effective features, significantly improving efficiency. Second, the platform provides automated machine learning (AutoML) capabilities. It integrates multiple classical algorithm models, supports automatic parameter optimization and can automatically try different algorithm and parameter combinations after users define optimization objectives, lowering the professional threshold for model tuning and ensuring that non-algorithm personnel can also build high-quality models.

(II) Introduction of a "Dual-Engine" Algorithm for Supply Chain and Manufacturing

The self-developed unsupervised algorithm captures naturally occurring industry patterns and is combined with open-source supervised multi-algorithm selection. The self-developed unsupervised algorithm does not require manual data labeling. Through clustering, such as K-means and hierarchical clustering, as well as anomaly detection and other technologies, it automatically mines hidden inherent patterns in supply chain data. It breaks through the limitations of human experience and accurately identifies natural patterns such as seasonal fluctuations in supply chain demand, potential correlations in inventory turnover and early signals of transportation anomalies. The open-source supervised multi-algorithm selection mechanism dynamically evaluates and selects among multiple open-source supervised algorithms, such as regression analysis and decision trees, based on historical labeled data, including previous production plans and delivery results, selecting the optimal algorithm for the current business scenario.

(III) Improved Uplift Model for Finance and Retail

AugereTech's self-developed RIO multi-model operations decision system has made several breakthrough improvements to the uplift modeling algorithm, mainly including multi-model coverage and refined training, gain-oriented modeling, quadrant probability judgment and input reconstruction, and mixed analysis of experimental and control groups.

The system refines customer-group classification and modeling objectives and builds hundreds of intelligent submodels. Multi-model coverage accommodates different business scenarios and customer-group features, effectively addressing confounding bias and inductive bias. By directly using "gain" as the modeling objective, it improves the ability of the model to predict strategy effects and avoids deviations caused by traditional dependent-variable modeling. Quadrant probability judgment and input reconstruction reduce interference caused by data precision fluctuations by calculating the probability distance of customers in each quadrant and reconstructing input features. Mixed analysis of experimental and control groups, combined with ANOVA, identifies inter-group differences, optimizes variable selection and builds a more concise and overfitting-resistant predictive model. In typical financial-industry scenarios, the model has achieved stable and significant incremental effects.

The following table presents uplift statistics from a marketing campaign implemented at a provincial branch of Agricultural Bank of China. The area under the uplift curve (AUUC) of AugereTech's model remains in the high range of 12‰ to 20‰, significantly outperforming traditional single-model and dual-model approaches and steadily surpassing the commonly used tree-based uplift model in the industry.

Algorithm Type	Campaign 1 AUUC	Campaign 2 AUUC	Campaign 3 AUUC	Campaign 4 AUUC
S-Learner algorithm (single model)	0.33‰	5.92‰	1.80‰	3.80‰
T-Learner algorithm (dual model)	11.26‰	15.57‰	1.59‰	1.58‰
Tree-based uplift model	10.50‰	14.20‰	7.10‰	12.80‰
AugereTech improved incremental algorithm	11.42‰	16.88‰	8.81‰	15.45‰
AugereTech improved incremental joint algorithm	16.08‰	20.23‰	12.66‰	17.96‰

Note: The AUUC value of the tree-based uplift model is an estimated range based on public industry literature and typical application scenarios.

With its self-developed improved uplift model, AugereTech not only achieves a leap beyond traditional methods, but also maintains clear advantages when benchmarked against advanced industry algorithms. This strongly demonstrates the advancement and effectiveness of its self-developed algorithmic approach. The system is applicable not only to precise marketing, customer segmentation and strategy optimization in the financial and retail sectors, but also has strong scalability and business adaptability, making it an ideal choice for promoting the digital-intelligence upgrading of enterprise operations.

(IV) Summary of Technical Advantages

1. Generational advantage in intelligence: Most mainstream systems in the market solve problems at a single or average static level, while AugereTech uses its dual engine of "discovery + selection" to realize a leap from process automation to decision intelligence. What it provides to customers is not a fixed tool, but a decision-making partner that can evolve independently and become smarter with use.

2. Agile advantage of technical architecture: Compared with closed commercial suites, the self-developed AIP + RIO dual-system empowerment is based on an open technology stack of self-development plus open source, avoiding vendor lock-in. It can quickly integrate the latest global algorithmic achievements and bring customers faster deployment, lower customization costs and future-oriented technological extensibility.

3. Efficiency advantage in total cost of ownership: Through high algorithmic adaptability and automation, the project fundamentally compresses the most time-consuming and costly links in the traditional model, namely implementation and continuous maintenance. Implementation results show that the full modeling process has been shortened from an average of 30 days per cycle in the traditional model to 8 days per cycle, a time reduction of 73.3 percent. Annual labor costs have been reduced by 50 percent, and modeling labor demand has decreased by more than 70 percent, substantially shortening the AI implementation cycle for enterprises. This means that customers can not only obtain better business results, but also enjoy lower total cost of ownership and faster ROI throughout the entire project life cycle.

4. High degree of industry focus: Unlike the generalized solutions and globalized focus of mainstream system suppliers, AugereTech's technical route accurately targets industry pain points. Its improvement effects on core industry indicators such as model accuracy, production efficiency and turnover speed far exceed those of traditional models.

IV. Implementation Results

Driven by the AIP intelligent modeling platform and the RIO intelligent strategy generation system, AugereTech has achieved deep implementation in complex business scenarios such as supply chain and manufacturing. Its self-developed "dual-engine" algorithm, which combines an unsupervised algorithm capturing naturally occurring business patterns with open-source supervised multi-algorithm selection, uses dynamic model matrices, multi-model chain decision-making and algorithm optimization to form comprehensive advantages beyond industry benchmarks in efficiency, turnover, response speed and risk control. The specific results are as follows:

(I) Implementation Results in Leading Enterprises: Quantitative Outcomes

In manufacturing and supply chain scenarios, the project has successfully empowered well-known enterprises such as Yuyuan Jewelry, 999 Pharmaceutical and GAC Honda, solving efficiency bottlenecks and supply-demand mismatches under the traditional single-algorithm decision-making model. Actual operation has verified that the dual systems helped customers achieve significant growth in business indicators.

On the sales side, overall sales volume increased by 22 percent, effectively capturing latent market demand.

On the cost side, monthly capital costs decreased by RMB 8 million, greatly reducing inventory backlog and capital occupation. On the operations side, the stockout rate decreased by 25 percent and the turnover rate increased by 18 percent, enabling precise and rapid supply chain response.

(II) Comparative Advantages in Key Performance Indicators

Based on AugereTech's multi-algorithm decision-making model, the project comprehensively outperforms single-algorithm decision-making and industry benchmarks across core evaluation dimensions, as shown below:

Production efficiency leap: The production efficiency improvement rate reached 28 percent to 35 percent, while the industry benchmark is only 8 percent to 15 percent, enabling flexible adaptation to complex production needs for different vehicle models.

Accelerated turnover: Supply chain turnover days were reduced by 18 to 22 days, far exceeding the 10 to 15 days achieved by a single algorithm, and precisely matching the supply and demand rhythm of components.

Rapid market response: Market response time was shortened to 2 to 4 hours, compared with the industry benchmark of 8 to 12 hours, providing the capability to respond rapidly to order fluctuations.

Fast implementation and high accuracy: The implementation cycle was shortened to 4 to 6 weeks. At the same time, unsupervised algorithms were used to identify hidden risks, increasing anomaly detection accuracy to 92 percent.

Comparative Analysis of the Effectiveness of AugereTech's Multi-Algorithm Decision-Making Model and Industry Benchmarks

Evaluation Dimension	AugereTech Multi-Algorithm Decision-Making	Single-Algorithm Decision-Making	Industry Benchmark	Advantage

Production efficiency improvement rate	28%-35%	12%-18%	8%-15%	Multiple algorithms adapt to production needs of different vehicle models
Reduction in supply chain turnover days	18-22 days	10-15 days	5-10 days	Precise matching of component supply and demand rhythm
Market response speed (hours)	2-4 hours	6-8 hours	8-12 hours	Rapid response to order fluctuations
Implementation cycle (weeks)	4-6 weeks	8-10 weeks	12-16 weeks	Modular algorithms reduce implementation costs
Anomaly detection accuracy	92%	75%	68%	Unsupervised algorithms identify hidden risks

Note: The results are based on a pilot project in the automotive parts supply chain and compare the project with a single ant colony algorithm and the industry average.

(III) Overall Value

The self-developed AIP + RIO dual-system empowerment differs from similar products that focus only on the modeling layer. AIP is not only a modeling base, but can also connect with the RIO strategy generation platform. It not only realizes the automated closed-loop capability from model output to business strategy implementation and completes the goal of process efficiency improvement and cost optimization, but also builds an industry-competitive intelligent decision-making brain for enterprises through high-precision multi-algorithm collaboration, further strengthening the conversion of technical value into business value. Its strong scalability and business adaptability promote the popularization of AI-empowered digital-intelligence upgrading of enterprise operations.

Case 16

Design and Simulation Agent: A Human-Efficiency Multiplication Engine for SMEs

Based on Multi-Agent Collaboration

China Mobile (Shanghai) Information and Communication Technology Co., Ltd.

I. Basic Enterprise Information

China Mobile (Shanghai) Information and Communication Technology Co., Ltd., referred to as China Mobile Shanghai Industrial Research Institute, is a wholly owned subsidiary of China Mobile. Based in Shanghai and serving the whole country, it focuses on three major fields: new industrialization, transportation to vehicle (ToV) large-scale transportation, and BeiDou positioning and navigation. It carries out research and development of technologies, platforms, products and solutions, with an emphasis on improving industrial innovation capability and deepening intelligent integrated applications. It brings together high-level national, provincial and municipal talent, focuses on the research and development of common technologies, key technologies and forward-looking technologies, and builds an industry-university-research-application platform centered in Shanghai.

China Mobile Shanghai Industrial Research Institute uses AI empowerment of new industrialization as its starting point and promotes the implementation of AI + manufacturing. Based on the capabilities of "connection + data + intelligence", it upgrades 5G + Industrial Internet, develops industrial large models and agents, builds a full-scenario capability system of computing power, data, models, platforms and applications, and creates an industrial intelligent hub. It deeply cultivates scenarios such as equipment intelligence, data elements and digital transformation of SMEs, providing intelligent connectivity, data mining and visual decision-making services for industrial, energy and financial sectors. OnePower has been selected for two consecutive years as a national cross-industry and cross-domain Industrial Internet platform by the Ministry of Industry and Information Technology. It has implemented more than 5,000 5G + industrial projects across 31 provinces, providing support for the intelligent upgrading of enterprises.

II. Case Introduction

(I) Background and Needs

In the equipment manufacturing industry, the research and development design process often faces challenges such as complex product structures, extensive design standards and variable process flows. Traditional methods that rely on human experience are inefficient and prone to misunderstandings of design standards and excessively long simulation calculation cycles. Meanwhile, as industry competition intensifies and delivery cycles shorten, enterprises urgently need to strike a balance among design compliance, simulation accuracy and research and development efficiency.

(II) Difficulties Before Implementation and AI Application Needs

1. Complex design standards and low efficiency in retrieval and verification: Customers have large volumes of technical standards documents with multiple languages and complex terminology. Manual interpretation is time-consuming and labor-intensive and is prone to missing key requirements. The amount of design data is large, and comparing drawings with specifications requires extensive verification work. Deviations are often discovered late, affecting project progress and delivery quality.

2. Dispersed compliance requirements and difficulty in quality risk control: In multi-disciplinary collaborative design, there is a lack of a unified compliance checking mechanism. The scope of impact caused by design changes is difficult to evaluate in a timely manner, and quality risks are identified late, which may lead to customer acceptance failure and rework losses.

3. Long simulation calculation time and insufficient prediction accuracy: In complex welding scenarios, traditional simulation software may take hours or even days to solve, seriously restricting research and development efficiency. At the same time, it is difficult to accurately reflect key factors such as weld microstructure changes and residual stress, affecting the reliability of process evaluation.

4. Complicated parameter settings and closed systems that are difficult to integrate: Simulation scenarios involve a large number of parameter configurations. Operations are complicated and error-prone, model debugging is difficult, and intelligent guidance is lacking. Mainstream software products are mostly foreign, with closed interfaces, and are difficult to deeply integrate with enterprise process databases and manufacturing execution system/product lifecycle management (MES/PLM) systems. The need for domestic substitution is urgent.

Based on these needs, AI technology is required to realize full-process intelligence covering design standard parsing, design basis document generation, simulation acceleration and parameter optimization. It addresses core pain points such as manual interpretation, low simulation efficiency and insufficient accuracy, and provides strong support for the digital transformation of the equipment manufacturing industry.

(III) Overall Project Objectives

In response to the complex needs of research and development design and process simulation in the equipment manufacturing industry, the project builds design and simulation agents covering core capabilities such as design standard parsing, knowledge retrieval, design generation, simulation proxy models and parameter optimization. It breaks through the traditional design quality control and inefficient simulation calculation models that rely on human experience. While improving design standardization, accuracy and compliance, it accelerates the digital verification and optimization process of welding processes, fully helping enterprises achieve efficient collaboration, quality improvement and shorter delivery cycles in research and development.

(IV) Implementation Plan

1. Product Solution for Design Agents

This solution centers on the design basis document generation agent. It drives the overall business process through semantic question answering, integrates the experience of senior engineers with the intelligent capabilities of large models, and uses an agent scheduling framework to realize knowledge retrieval, parsing and template output.

In terms of specific functions, the system first identifies user needs through intent parsing and completes project information, standard library and directory selection based on template configuration. Then, based on multi-format parsing and hybrid vector/keyword retrieval, it conducts full retrieval of the knowledge base title by title, supports back-translation of multilingual entries, and allows users to select required content and generate index files. The large model further combines paragraph content with template requirements to generate the body text of the design basis document, with citations to the original standard text. It supports secondary editing and version backtracking, and finally outputs and archives the design basis document. At the same time, the system builds a private knowledge base that supports standard version management, customer isolation and historical document management. In design knowledge question answering, it supports natural-language retrieval and traceable answers, helping engineers quickly obtain professional support.

Figure 1 Design basis document generation assistant

2. Product Solution for Welding Simulation Agents

The industrial welding simulation software built under this solution is supported by core technologies such as finite element analysis and AI large models, forming a domestic simulation platform with independent intellectual property rights. Its overall architecture covers basic functional modules such as model import, material library, welding path, mesh division, boundary conditions, measuring point definition, simulation analysis and simulation reports. It is also equipped with operational management capabilities such as license authorization, tenant management and log management to ensure controllability and reliability in enterprise-level applications. At the same time, an AI assistant function is introduced to provide modeling assistance, intelligent analysis and intelligent question answering services, significantly improving the efficiency and accuracy of users in complex process modeling and result analysis.

The software has a wide range of applications, covering industries such as shipbuilding, equipment manufacturing, automobiles and aerospace. It can serve multiple welding process scenarios, including arc welding, laser welding and electron beam welding, and supports simulation calculations for various weld structures such as T-joints, butt joints, corner joints and lap joints. Its simulation capabilities cover not only thermal simulation and structural simulation, but also multi-physics coupling analysis, enabling comprehensive prediction of temperature fields, deformation fields and residual stress. Overall, the architecture balances accuracy, scalability and usability, providing solid support for digital verification and optimization of complex welding processes.

Figure 2 Field welding process

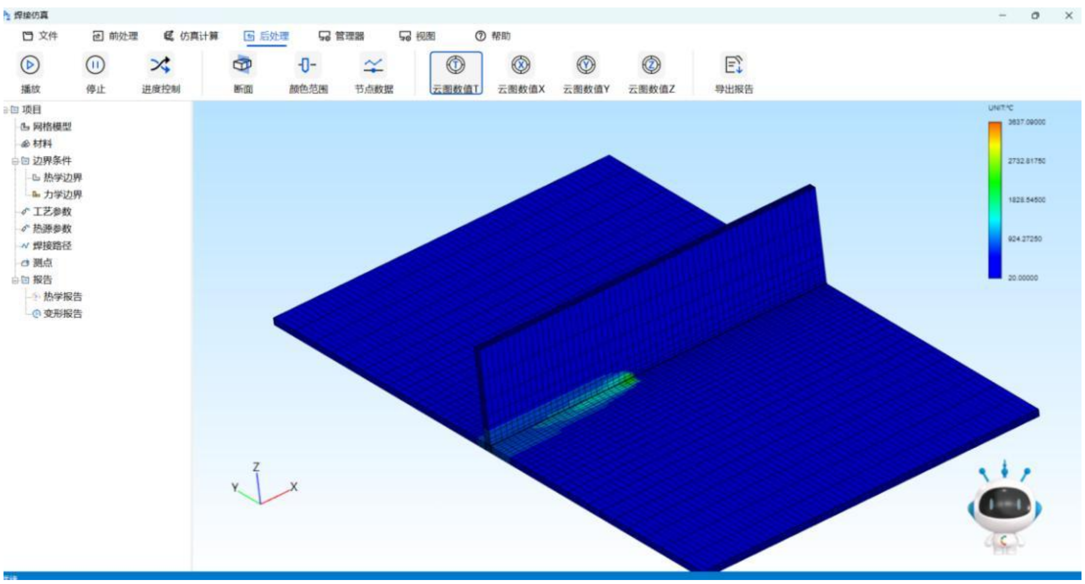


Figure 3 Simulation software

III. Project Highlights

(I) Case Highlights

Precise focus on industry pain points: The project focuses on key issues in the design and simulation links of the equipment manufacturing industry. It has successfully addressed core pain points such as complex design standards, low document generation efficiency and long welding simulation calculation time. The design agent can automatically parse standards and generate design basis documents, while the simulation agent greatly shortens simulation time by optimizing the calculation process, improving overall research and development efficiency.

Technological breakthroughs and innovative applications: The design agent uses semantic parsing and cross-modal retrieval technologies to realize automated conversion of design standards and document generation, breaking through the limitations of traditional manual operations. The simulation agent introduces transient heat conduction solving and sparse matrix deformation simulation technologies, combined with an AI proxy model, enabling a balance between high accuracy and high efficiency in welding simulation.

Improved research and development efficiency: Through intelligent generation of design basis documents and accelerated welding simulation, the project successfully shortened the generation time of design documents from 7 days to 10 minutes, and reduced the simulation time for a 5-minute welding physical process from 2 hours to 35 minutes. The accuracy of automatic design basis document generation reached more than 85 percent, and welding simulation prediction accuracy also reached more than 85 percent. This significantly improved the overall efficiency of design and simulation and saved substantial time for enterprise research, development and production.

AI assistants leading digital-intelligence transformation: After the introduction of AI assistant functions, the project provides engineers with intelligent question answering, parameter recommendations and anomaly warnings during design and simulation. Through natural-language interaction and intelligent analysis, the agent greatly lowers operational difficulty and improves work efficiency and accuracy.

(II) Replicable and Scalable Experience, Tools and Methods

1. Technological Innovation Points

In the design direction, the project integrates cross-modal retrieval, semantic understanding and large-model generation technologies to realize automated conversion from complex design standards to structured design documents, reducing manual intervention and improving design accuracy. In the simulation direction, the project uses transient heat conduction solving, large sparse matrix deformation simulation and thermo-elasto-plastic finite element algorithms, and introduces simulation proxy models to build an efficient parametric modeling system. It creates a domestic welding simulation tool that balances high precision and high efficiency. Overall, through the deep integration of design standard parsing, intelligent text chunking and simulation proxy models, the project realizes the practical implementation of intelligent design and autonomous simulation.

2. Key Technologies and Data Support

(1) Industrial-grade deep parsing technology for complex multi-format documents: This technology realizes precise parsing and structured information extraction from highly complex and highly specialized industrial documents, such as engineering design drawings, multidimensional process procedures, heterogeneous equipment manuals and cross-type technical standards.

(2) Scenario-aware intelligent text chunking technology for industrial semantics: The project develops algorithms that can deeply combine the logical structure of industrial documents, such as design reports, bill of materials (BOM) lists and process flow diagrams, with contextual semantics for intelligent chunking. It builds efficient, accurate knowledge indexes that support multi-granularity queries and meet the needs of rapid retrieval and knowledge association for massive engineering and technical information.

(3) Source attribution technology for industrial knowledge generation and application: This technology builds reliable and precise association chains between model-generated content, such as design suggestions, fault diagnosis conclusions and optimization schemes, and underlying original knowledge sources, such as specific specification clauses, historical case records and equipment parameter documents. It enhances the transparency of industrial AI systems, the traceability of decision-making bases and the trustworthiness of final outputs.

(4) Self-developed simulation solving engine: Combining finite element algorithms with simulation proxy models, the engine achieves process simulation capabilities that balance high accuracy and high efficiency.

(5) Private knowledge base and version management: The system supports rapid import and updating of customer standards, improving flexibility in adapting to different enterprises and industries.

3. Product Applications and Solutions

(1) Design agent: Based on advanced AI technologies, the design agent can understand complex design standards and automatically generate design documents. By combining a knowledge base and multilingual parsing capabilities, it provides enterprises with an integrated solution that helps them quickly complete design document generation and verification and ensure design compliance and accuracy.

(2) Simulation agent: Through the self-developed simulation solving engine, the simulation agent can automatically complete various calculation tasks in welding processes. It covers the entire process from parameter setting and model construction to multi-physics coupling simulation, and can track simulation results, helping enterprises quickly obtain high-precision simulation analysis results in complex processes.

(3) AI assistant function: The system provides natural-language interaction, intelligent question answering and process guidance, lowering the usage threshold and improving the engineer experience.

4. Replicable Promotion Model

The methodology and tools of the system in design and simulation scenarios have strong cross-industry adaptability and can be applied to automobiles, shipbuilding, aerospace, equipment manufacturing and other industries. The project adopts a promotion model of “benchmark case creation, standardized application promotion and frontline point-to-point support”. Specifically, it will first be implemented in leading enterprises to form demonstration samples, then copied and expanded to their upstream and downstream enterprises as well as external industries, and finally achieve large-scale technological application and industrial ecosystem construction.

IV. Implementation Results

(I) Project Results

1. Case 1, design category: Guangzhou MINO Equipment Design Basis Document Generation Agent Project.

Guangzhou MINO Equipment Co., Ltd., referred to as MINO Equipment, was established in 2008 and is headquartered in Guangzhou. It is the largest and most comprehensive supplier of high-end automotive manufacturing equipment in South China. Its business focuses on automated welding production lines for automotive body-in-white, powertrain and new-energy manufacturing equipment, robot integration services, and intelligent manufacturing solutions for general industries such as high-speed rail and aviation.

This project is an important part of the digital transformation and intelligent manufacturing initiative of MINO Equipment. Targeting mechanical design engineers in automotive equipment manufacturing enterprises and focusing on compliance-oriented design scenarios, it addresses pain points such as low efficiency in customer standard interpretation, difficulty in compliance verification and high manual error rates. The project built a design basis document generation agent. Based on general industry standards and standards of several major original equipment manufacturers (OEMs), the agent provides intelligent parsing and structured processing capabilities for design documents. It can automatically generate compliant design basis documents, provide intelligent question answering and deliver advanced capabilities such as design compliance checking. The project effectively reduces manual workload and design cycles, reduces human errors and improves the efficiency and standardization of design delivery.

2. Case 2, simulation category: Shanghai Waigaoqiao Shipbuilding Industrial Welding Simulation Project.

Shanghai Waigaoqiao Shipbuilding Co., Ltd. is located on the south bank of the Nangang section of the Yangtze River estuary in Pudong New Area, Shanghai. It was jointly established by China State Shipbuilding Corporation, Baosteel Group Co., Ltd. and Shanghai Electric (Group) Corporation, with China State Shipbuilding Corporation as the controlling shareholder. It is one of the most modern large-scale ship assembly yards in China. Its main business includes the production of bulk carriers, oil tankers, ultra-large container ships, offshore drilling platforms, drillships, floating production storage and offloading units, and offshore auxiliary vessels. The company also provides ship repair services for customers.

Shanghai Waigaoqiao has numerous small and medium-sized suppliers in its industrial chain, such as Suzhou Wuzhong Wucheng Stainless Steel Products Co., Ltd. and Shanghai Feijin Environmental Protection Technology Co., Ltd. These enterprises generally face shortcomings in welding simulation, including insufficient technology and lack of tools, making it difficult to meet the high requirements of large shipyards for accuracy and efficiency. Through AI empowerment, the project not only fills Waigaoqiao’s own gaps in digital welding processes, but also effectively improves the simulation capabilities of its upstream and downstream SMEs, driving and empowering the entire industrial chain.

This project is an important construction project for Shanghai Waigaoqiao Shipbuilding to promote the digitalization and intelligentization of ship welding processes. Through the development and application of industrial welding simulation software, the project can accurately predict and optimize complex welding processes such as arc welding and laser welding. The software integrates technologies such as finite element analysis, AI and cloud computing, and provides multi-physics coupling simulation capabilities for temperature fields, deformation fields and residual stress, enabling high-precision simulation of welding structure deformation. The system supports parametric modeling, dynamic boundary condition setting and online simulation, and can be remotely debugged and connected with enterprise process databases. Through this project, the shipyard can significantly improve efficiency and reliability in welding process research, development and verification.

(II) Economic Benefits

By automating design standard parsing, document generation and welding simulation, the agent breaks through the limitations of traditional manual operations. In the design process, the accuracy of automatic design basis document generation reaches more than 85 percent, significantly improving document preparation efficiency. The time required is reduced from 7 days of manual preparation to 10 minutes, with expected annual labor cost savings of about RMB 800,000. In the simulation process, welding process simulation speed is improved to one quarter of that of traditional commercial software. The simulation of a 5-minute welding physical process is reduced from 2 hours to 35 minutes. Prediction accuracy reaches more than 85 percent, helping shipyards reduce production trial-and-error costs and rework losses by RMB 5 million and comprehensively optimize the research and development process.

(III) Social Value

The project helps equipment manufacturing enterprises accelerate the digital transformation of research and development design and process verification, improve product quality and delivery efficiency, and enhance overall industry competitiveness. At the same time, it breaks the high dependence on foreign simulation software, promotes the implementation of domestic industrial software in major equipment manufacturing scenarios, and forms a replicable industry demonstration effect. Through the dual implementation of design and simulation agents, the project provides solid support for independent innovation and intelligent upgrading in China’s equipment manufacturing industry, with important social value and strategic significance.

Case 17

AI STORE on Zhiyun Shanghai: Enhancing Enterprise AI Capabilities and Shanghai's Urban AI Capacity

China Telecom Corporation Limited Shanghai Branch

I. Basic Enterprise Information

China Telecom has deployed more than 62,000 P (FP16) of intelligent computing power nationwide, including over 40,000 P (FP16) of self-owned intelligent computing power and 15,600 P (FP16) of independently controllable intelligent computing power. It has long collaborated with large-model organizations in Shanghai, including the Shanghai Artificial Intelligence Laboratory, the Shanghai Artificial Intelligence Research Institute, SenseTime, MiniMax, StepFun, Kupas, Infini-AI, GeFlow Technology, Fudan University and Shanghai Jiao Tong University, supporting the training and application of large AI models in the city.

China Telecom has actively implemented strategic layouts such as cloud-network convergence, computing power networks and new computing infrastructure. Relying on a "2+4+31+X" cloud resource-pool architecture, it has built China's largest Internet data center (IDC) cluster and computing power network. It has independently developed core technologies such as cloud operating systems and distributed databases, and obtained more than 2,000 cloud-computing patents. In new infrastructure fields such as 5G standalone (5G SA) networks and IPv6 upgrades, it maintains an industry-leading position and has built the world's largest 5G co-construction and sharing network. The intelligent computing centers in Shanghai should comply with the relevant indicators of the current Guidelines for Shanghai Intelligent Computing Centers (2025 Edition). Their machine rooms meet the IDC Class A standard, their measured power usage effectiveness (PUE) does not exceed 1.25, and they are connected to the city's data-center energy-efficiency monitoring and management platform.

In terms of industry position, China Telecom has long remained in the first tier of China's communications industry. As a listed company controlled by a central state-owned enterprise, it has been selected for the Fortune Global 500 for many consecutive years. In cloud computing, eSurfing Cloud ranks first in China's hybrid cloud market; in the IDC sector, its rack scale leads the country. As an important participant in the national "Eastern Data and Western Computing" project, it undertakes the construction of several national hub nodes. Looking ahead, China Telecom will focus on intelligent computing power services, accelerate the construction of new "cloud + AI + computing power" infrastructure, double its intelligent computing scale within three years, and prioritize hub nodes such as Shanghai to build a low-latency computing service circle. It will also promote innovative services such as AI large-model training and industry intelligent solutions, empower digital transformation across industries, and support Shanghai in building itself into an international digital city. China Telecom has the ability to connect to the municipal intelligent computing power resource coordination and scheduling service platform, and commits to giving priority supply to Shanghai's AI-related SMEs.

II. Case Introduction

(I) Background and Needs

Against the increasingly intense global competition in AI, Shanghai, as an important highland in China's national AI development strategy, is accelerating the construction of an AI public service platform with local characteristics and strong ecosystem collaboration capabilities. The aim is to meet the urgent needs of regional industries for intelligent upgrading. In recent years, Shanghai has made it clear that AI should serve as a core driver for the deep integration of new-generation information technology with the real economy and for building an internationally influential AI innovation center. Guided by this policy orientation, the AI STORE on Zhiyun Shanghai actively responds to Shanghai's AI development requirements, integrates itself into the overall layout of urban digital transformation and industrial upgrading, and fully leverages its advantages in resource integration and technological empowerment to improve Shanghai's urban AI capacity. By providing high-quality AI application services, the platform helps enterprises lower technical barriers, improve production efficiency and strengthen intelligent decision-making, thereby promoting the intelligent transformation and high-quality development of industries.

At the initial stage of the project, the challenges faced by the AI STORE on Zhiyun Shanghai were particularly complex and representative. The AI market was booming, and many enterprises were asking how to enter this field and how to use AI. These questions were precisely the solutions that the AI STORE on Zhiyun Shanghai aimed to provide.

Many SMEs encounter serious difficulties when trying to introduce AI technologies because they lack sufficient technical reserves and financial support. For example, when some manufacturing enterprises carry out intelligent upgrading, they need not only to purchase expensive computing resources, but also to assemble professional technical teams for model training and system deployment, creating heavy short-term burdens. At the same time, enterprises in retail, logistics and other industries find that existing AI solutions are often too general to fit their specific business needs, resulting in poor practical outcomes and low input-output ratios. In addition, the rapid development of AI has continuously increased enterprise demand for digital tools. Some enterprises lack independent R&D capabilities, but still hope to use this technology to improve productivity and urgently need “digital employees” that can be directly deployed and put into use to meet their practical needs in digital transformation.

(II) Implementation Plan

1. System Construction

The AI STORE on Zhiyun Shanghai was officially launched in September 2024. At the beginning, it was committed to providing public computing power services, supporting large-model enterprises in training and inference. In early 2025, DeepSeek ignited the domestic AIGC sector. China Telecom Shanghai closely followed this wave, leveraged its significant advantages in computing power scheduling, model aggregation and ecosystem collaboration, and combined them with the Zhiyun Shanghai brand concept. Focusing on the deep integration and innovative development of AI technologies in industrial applications, the platform was officially upgraded into a comprehensive public service platform integrating computing power, models and applications. Through systematic platform-architecture upgrading and the construction of a multi-level product and service system, it continues to promote the efficient implementation and practical application of AI technologies in key industries such as government affairs, finance and education, helping enterprises move from “usable” to “easy to use”.

System Architecture

Centered on the construction of the AI STORE portal, the platform has formed an overall capability framework of “three supermarkets, one portal and two gateways”: the computing power supermarket, model supermarket and application supermarket form the AI product matrix; the AI gateway and the Model Context Protocol (MCP) gateway build unified underlying scheduling and distribution capabilities.

Computing power supermarket: A self-developed process connects user authentication, product subscription and one-click deployment, enabling self-service delivery of public computing power.

Model supermarket: The AI gateway is developed to build unified access and aggregation of multi-source models. Twelve mainstream models, including DeepSeek and Qwen, have been launched, supporting online API calls, token-based metering and billing, and genuinely realizing “model as a service”.

Application supermarket: The “Xiao” series of AI applications has been launched to serve scenarios such as government affairs, legal services, office work and knowledge bases.

Portal construction realizes unified scheduling, authentication, metering and billing, and service governance.

The AI gateway aggregates core capabilities such as unified authentication, token rate limiting, quota control, multimodal support and observability, realizing unified encapsulation of models such as DeepSeek and Qwen.

The MCP gateway provides full-link MCP registration, routing and billing capabilities, and gradually supports automatic OpenAPI-to-MCP conversion. It not only supports model services, but also provides underlying support for the AI capabilities of the “Xiao” series, allowing the application layer and model layer to be truly decoupled and flexibly expanded.

Business Logic

The AI STORE on Zhiyun Shanghai innovatively builds three core modules - the computing power supermarket, model supermarket and application supermarket - corresponding to the three key links of resource acquisition, technical adaptation and application implementation. The computing power supermarket provides flexible, on-demand computing resources, enabling AI enterprises to quickly deploy and test their applications without one-time heavy investment in infrastructure, thereby significantly improving development efficiency and resource utilization. The model supermarket focuses on differences in enterprise demand for AI models, integrating multiple mainstream models and MCP tools to support flexible invocation or secondary development according to business scenarios, so as to enable efficient model adaptation and optimization. The application supermarket further lowers the threshold for AI use by packaging mature AI applications into plug-and-play toolkits. Enterprise users do not need deep technical backgrounds to directly integrate these applications into production processes and start intelligent transformation with “zero threshold”. Through the coordinated operation of the three modules, the AI STORE not only effectively solves enterprise pain points at the levels of computing power, models and applications, but also builds an open, flexible and sustainable AI ecosystem for enterprises, helping more SMEs achieve intelligent transformation.

III. Project Highlights

During the construction of the “Zhiyun Shanghai AI Portal”, the AI gateway and MCP gateway formed the infrastructure core of the entire AI STORE. They undertake model access, capability governance and agent orchestration, and are the key to enabling the portal to deliver a unified service experience and stable external intelligent capabilities. To allow AI to truly enter government and enterprise business flows, the team overcame multiple long-standing industry technical problems in less than one year, from unified model protocols and external capability orchestration to the implementation of multi-step agent processes, forming an AI engineering system with strong professionalism and technical depth.

The most representative value of the AI gateway is that it makes all models “become one type of model”. In the early stage of the project, model API compatibility varied greatly across vendors, making it almost impossible to directly connect them to the same agent system. To allow the “Xiao” series agents to flexibly switch models, compare costs and route on demand, the team reconstructed the underlying model-access method and designed a model-protocol abstraction layer, uniformly encapsulating the inputs and outputs of all models as standard capabilities. This abstraction layer may look simple, but in practice it had to solve dozens of inconsistencies in token metering rules, function-call fields, dialogue contexts and content formats. The team ultimately built a sustainably evolving model-adaptation system, allowing a new model to be connected within only a few hours. More importantly, this system has been successfully applied to more than 30 models from different sources and inference frameworks, setting up a unified multi-model entrance for AI STORE.

Unified model protocols were only the first step. The real technological breakthrough lies in model scheduling. In government and enterprise scenarios, users may initiate numerous high-token sessions simultaneously, such as policy interpretation, contract analysis and long-form text generation. Traditional API gateways cannot meet the scheduling requirements of LLMs. Therefore, the AI gateway built a model-scheduling and token-billing engine that can identify the occupancy status of each model inference channel in real time, dynamically allocate inference channels for new requests, control resource use by high-cost tasks and keep billing accuracy within 1%. This capability is a leading exploration among operators nationwide and provides an accurate metering foundation for the commercialization of AI STORE. These breakthroughs make AI STORE a truly operable platform for computing power and model resources.

Compared with model-side unification and scheduling, the MCP gateway focuses on capability governance and agent orchestration. Government and enterprise businesses often need access to numerous internal tools, such as approval-system APIs, database query interfaces, document management services, file systems and workflow services. These tools differ in format, permission settings and access paths. To allow agents to invoke enterprise capabilities just as they invoke large models, the team built a capability registration center based on the MCP protocol and automatically completed API-to-MCP tool conversion, enabling every enterprise API to become a “capability atom” that agents can directly invoke. In a typical scenario such as data inquiry for a cultural center, an agent needs to access the file system, knowledge-base retrieval service and structured table interface simultaneously. Traditionally, this required substantial customized development; after the MCP gateway was implemented, only capability registration was required for the agent to call them naturally, greatly reducing the threshold for integration.

After capability access was established, another challenge was security and authorization. Government and enterprise systems have extremely high requirements for permission management, while agents have automated invocation characteristics. Once unauthorized access occurs, serious risks may result. For this reason, the MCP gateway designed a fine-grained permission system at the capability and field levels. For each MCP tool, it defines visible fields, executable actions and callable scopes, and prevents out-of-scope invocation through dynamic token control. Every capability access by an agent is executed in a sandbox environment, ensuring that the agent cannot break through the established permissions.

In terms of implementation results, this system shows significant engineering value. The agent development cycle has been shortened from several weeks to several days; the model invocation error rate has decreased by more than 70%; and in some high-load scenarios, overall operating costs have been reduced by 30% to 40% through multi-model routing and API scheduling. These breakthroughs enable AI STORE not only to provide large-model services, but also to enter government and enterprise business flows in the form of agents, forming a genuine "AI middle platform".

The construction experience of the AI gateway and MCP gateway has become a core competitive advantage of the project and has replicable and demonstrative value for the industry. It solves long-standing problems such as model unification, capability access and permission governance, and fully builds the underlying capability system that allows large models to execute business tasks. This lays a foundation for future scaled applications in government services, enterprise knowledge bases, production and manufacturing, cultural services and other fields.

IV. Implementation Results

Focusing on the strategic layout of "Zhiyun Shanghai", China Telecom Shanghai follows the guidance of the Shanghai Municipal Commission of Economy and Informatization and leverages the elastic computing resource advantages of the Zhiyun Shanghai public computing power service platform. Based on the AI STORE on Zhiyun Shanghai, it promotes the deployment of foundation models, corpus products, vertical applications and agent applications, providing enterprises with one-stop ordering, use, subsidies and services. It inclusively empowers industries and supports enterprises in scaled development. At the same time, it leverages the operator's network advantages to provide strong computing-network capabilities and responsive public intelligent services. Relying on the metropolitan intranet architecture, it is committed to building and operating an open "agent store" platform.

The AI STORE on Zhiyun Shanghai has built three major modules: the computing power supermarket, model supermarket and application supermarket, which respectively address resource acquisition, model adaptation and application implementation. The computing power supermarket provides on-demand computing resources to help enterprises efficiently deploy and test applications; the model supermarket integrates mainstream models and MCP tools, supporting flexible invocation and secondary development; the application supermarket packages mature AI tools to realize "zero-threshold" integration. Through the collaboration of the three modules, an open and flexible AI ecosystem is formed to help SMEs achieve intelligent transformation.

(I) Computing Power Supermarket

Relying on the Xirang computing power scheduling platform, the computing power supermarket has successfully integrated self-owned and partner multi-source computing resources, forming a rich resource pool covering Ascend 910B, high-performance NVIDIA graphics cards and other types of intelligent computing cards. This effectively meets diverse and personalized customer needs for computing power. Through flexible resource scheduling and efficient management mechanisms, the platform not only improves the utilization of computing power resources, but also substantially lowers the threshold and cost for enterprises to obtain computing power, thereby enhancing the accessibility and operability of AI applications. In addition, Shanghai Telecom's self-developed single-card and multi-card intelligent computing products listed in the computing power supermarket provide enterprises with a new computing power service model beyond traditional physical servers. This "per-card, per-hour" billing method, with the lowest price reaching only RMB 1 per card-hour depending on the computing card type, greatly reduces the cost pressure on enterprises in computing power procurement and use.

(II) Model Supermarket

In the model supermarket, the platform actively builds multi-model aggregation capabilities and establishes deep cooperation with high-quality model service providers such as eSurfing, Huawei and Alibaba to enrich the types of available model services. At present, the platform has launched shared service mode and dedicated service mode, covering online model API call services and dedicated MCP services. Among them, the online model API call service has launched the DeepSeek series, Qwen series and other models; 12 dedicated MCP services and 43 dedicated model services are available. This multi-model aggregation model not only improves the flexibility and adaptability of model services, but also provides enterprises with broader model options and customized service capabilities. In the financial industry, enterprises can choose different model combinations according to their business needs to intelligently upgrade risk assessment, credit approval and other processes. In government affairs, model services are used in scenarios such as intelligent Q&A and policy interpretation, improving service efficiency and user experience.

(III) Application Supermarket

As the third core section of the platform, the application supermarket has launched the “Xiao” series digital employee zone, including Xiao Director, Xiao Lawyer and Xiao Assistant, providing enterprises with intelligent service solutions. It supports lightweight delivery for immediate use and can also provide private deployment services according to specific project needs, offering a higher level of security assurance. Relying on a rich AI application ecosystem and self-developed capabilities, the application supermarket provides standardized AI applications for multiple scenarios, including intelligent office work, legal assistance, knowledge-base Q&A, digital humans, publicity and copywriting, and is committed to simplifying complex workflows for SMEs. By building a standardized and modular application system, the platform lowers the threshold for enterprises to deploy applications on their own, enabling them to quickly access and use AI capabilities.

(IV) Ecosystem Cooperation

In ecosystem construction and external cooperation, the AI STORE on Zhiyun Shanghai is continuously deepening collaboration with the Shanghai Municipal Commission of Economy and Informatization, while simultaneously connecting with standardized MCP interfaces of government departments such as the meteorological bureau, promoting the empowerment of enterprises with government resources through an open platform. This initiative not only breaks down data barriers between government affairs and industries, but also realizes the integration and sharing of core capabilities, effectively addressing common problems in AI applications such as closed systems and data silos. By providing modular and component-based “building-block” development support, enterprises can quickly build personalized AI solutions, greatly improving innovation efficiency and the speed of technological iteration. This will help form a clustering effect in the AI industry, further strengthen Shanghai’s innovation advantages in AI and promote high-quality development of the regional digital economy.

In terms of implementation results, the AI STORE on Zhiyun Shanghai has achieved certain economic results. As of September 2025, the platform had accumulated 1,369 registered users, served nearly 400 enterprises and received more than 70,000 page views. Single-card and multi-card intelligent computing products, as inclusive computing power services, had delivered more than 350,000 card-hours of computing power, serving 97 customers in industries such as finance, government affairs and education. Three months after the launch of the “Xiao” series digital employees, they had served more than 300 customers covering streets, commissions and offices, medical institutions, human resources, law firms, newspapers and other sectors, with converted contract revenue reaching the million-yuan level.

Looking ahead, the AI STORE on Zhiyun Shanghai will continue to focus on industry scenario-based needs, deepen the iteration of standardized agent capabilities, and expand scaled applications in high-value scenarios such as government affairs, finance, and central and state-owned enterprises. It will also improve the model supermarket service system, leverage ecosystem collaboration and self-developed strengths, continue to enhance MCP and dedicated model service capabilities, provide selected MCP services based on Shanghai’s local characteristics, explore agent development and Model as a Service (MaaS) platform capability iteration, and build an end-to-end AI service closed loop. It will deepen ecological cooperation with leading model and application vendors, jointly build a delivery system covering API calls, model deployment, knowledge-base tuning, prompt engineering and other full-scenario services, promote AI technology from “usable” to “easy to use”, substantiate customer-side AI scenario delivery, and cultivate a team of 100 agent delivery engineers. In the future, China Telecom Shanghai will continue to consolidate the intelligent computing foundation, jointly promote industrial upgrading and urban governance, and support the construction of Shanghai as an AI highland.

Case 18

Solving Cross-Border Logistics Pain Points for Chinese E-Commerce Enterprises Based on Intelligent Data Analysis

Shanghai JUSHUITAN Network Technology Co., Ltd.

I. Basic Enterprise Information

Shanghai JUSHUITAN Network Technology Co., Ltd., hereinafter referred to as JUSHUITAN, was founded in 2014. It is a “technology little giant” enterprise specializing in the research and development of software as a service (SaaS) enterprise resource planning (ERP) application services. The e-commerce ERP software system developed by JUSHUITAN based on the SaaS model provides e-commerce enterprises with advanced application scenarios such as business intelligence (BI) data mining and analysis, intelligent warehouse management and omnichannel product management, increasing the average efficiency of participating enterprises by 35%.

JUSHUITAN has successively received honors including the National Public Service Demonstration Platform for SMEs of the Ministry of Industry and Information Technology, Shanghai Technology Little Giant Enterprise, Shanghai Enterprise Technology Center, Shanghai Key Service Unicorn Enterprise, Shanghai E-Commerce Demonstration Enterprise, Shanghai Industrial Internet Demonstration Platform, Shanghai Quality Benchmark and Shanghai Brand Leadership Demonstration Enterprise. It has been selected into the Ministry of Commerce’s Excellent Cases of Business Technology Innovation Application, Yangtze River Delta Regional Innovation Product Application Demonstration Cases, Shanghai Recommended Catalogue of Innovative Products and Shanghai Excellent Cases of “AI Plus” Application in the Commercial Sector. On October 21, 2025, JUSHUITAN was listed on the Main Board of the Hong Kong Stock Exchange. As a co-builder of the e-commerce ecosystem, JUSHUITAN strives to play the role of a digital hub, supports SMEs in participating in the new development pattern of domestic and international circulation, and provides strong technical support for merchants.

II. Case Introduction

(I) Background and Needs

As a new-generation leading enterprise in China’s e-commerce SaaS ERP industry, JUSHUITAN has, after years of development, upgraded its ERP from a simple management tool into an enterprise decision-making center. Through technologies such as cloud computing, AI and big-data analysis, it provides data support for enterprise managers’ decisions. As the domestic e-commerce industry becomes increasingly mature and stable, the cross-border e-commerce industry is rising rapidly and developing vigorously. In recent years, e-commerce has entered a stage of global layout, with both domestic and overseas markets expanding and competition intensifying. Cross-border e-commerce ERP systems, as key technical tools supporting cross-border operations, have gradually attracted industry attention.

In cross-border business, logistics receives the most feedback related to disputed orders. In cross-border e-commerce logistics, it is common that after parcels leave the country, timely parcel logistics information cannot be obtained. This situation is somewhat better in developed logistics markets such as Europe and the United States, where users can query relevant English websites after obtaining tracking numbers. However, for some countries using minority languages and countries with less-developed logistics sectors, even when tracking numbers are available, opening various minority-language websites may still not allow merchants to quickly and conveniently obtain delivery information.

JUSHUITAN hopes that its logistics system will not merely summarize data for query purposes. On this basis, it should realize monitoring and analysis, monitor and provide real-time feedback on abnormal logistics status, support not only order logistics trajectories but also first-leg document systems, customer-defined abnormality monitoring and other functions, and use logistics reports to provide accurate big-data analysis, including route analysis, country analysis, delivery rate, pickup rate and abnormality analysis.

Under these premises, JUSHUITAN established the cross-border logistics trajectory analysis system project, planning to use AI + robotic process automation (RPA), BI big-data analysis and other technologies to solve problems faced by Chinese e-commerce enterprises in cross-border logistics real-time monitoring, abnormality handling and after-sales service.

(II) Implementation Plan

The JUSHUITAN cross-border logistics trajectory analysis system works together with JUSHUITAN ERP and the data platform. It can directly query information in the ERP system by logistics provider, international tracking number and other dimensions. It also helps merchants quickly locate order logistics information, analyze logistics timeliness through big data, improve delivery and pickup efficiency, reduce parcel loss rates, control operational risks and costs, and thereby improve the overall management level and sustainable operation capability of enterprises.

1. Technical Solution

The technical solution mainly consists of three core modules and combines AI for big-data analysis.

The tracking module includes manual tracking, which tracks international orders by manually adding them and is suitable for merchants who only need to track key orders; automatic tracking, which presets automatic tracking rules for large-batch order tracking, reducing manual operations and eliminating the need for repeated manual order subscription; and batch import tracking, which is often used when there are fixed order files requiring tracking, the quantity is large and manual addition is inconvenient.

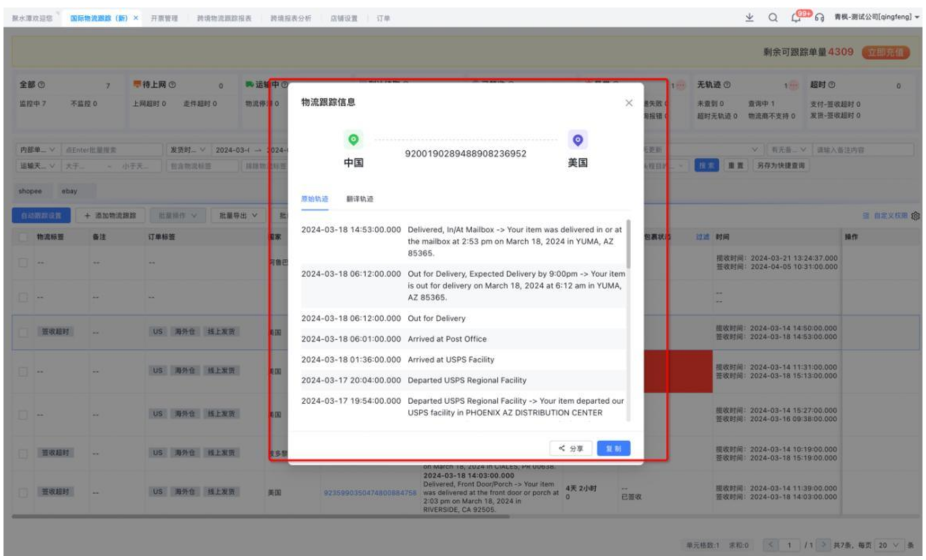


Figure 1 Cross-border logistics tracking information

The monitoring module includes timeout monitoring rule settings, which preset orders to enter timeout status after a certain stage exceeds the time limit; abnormality monitoring rule settings, which preset the binding between corresponding keywords in logistics trajectories and abnormal statuses; quick monitoring, through which the system automatically classifies documents in corresponding statuses and users can click the corresponding tags for quick monitoring and query; and a recycle-bin function, under which trajectory monitoring counts all document trajectories by default and allows document recycling through the recycle bin.

The query and modification module includes a query-pool function, which presets query conditions and enables quick queries based on preconfigured conditions; batch modification of logistics tags, which allows special orders requiring dedicated attention and not suitable for order tags to be quickly filtered through logistics tags; and combined query, through which quick status filters can be combined with search-box conditions.

2. Innovations

Big data provides the storage medium and query capability.

MySQL trajectory storage solution: MySQL retains data for one month, StarRocks stores data for six months, disk usage is reduced by 60%, and no disk expansion will be required in the next three years. Database queries per second (QPS) are reduced, and subscription tables use queue-based batch writing.

Memory queue + Message Notification Service (MNS) backup queue: The memory queue is used by default to receive and process logistics trajectory data; when an exception occurs, the data is pushed to the MNS queue for processing. There is only one application deployed in containers across six nodes. According to online operation, each node uses 200 to 300 MB of memory and one to two central processing unit (CPU) cores.

Monitoring dimensions: Logs for each module can be enabled or disabled separately; only exception logs can be recorded to reduce log storage costs; logs taking more than 500 ms can be recorded for performance troubleshooting; logs are stored in Alibaba Cloud Simple Log Service (SLS), which is low-cost and supports log alerts; container deployment supports resource monitoring and real-time monitoring of application resource usage.

Stability: Container deployment supports graceful launch and shutdown and convenient scaling. It is isolated from other applications and does not affect them. The daily number of request timeouts is 2,000 to 5,000. Log alerts allow exceptions during operation to be identified, investigated and repaired in a timely manner.

Release cycle: The system can be released independently. It is deployed on Alibaba Cloud containers and can be published online at any time after fixes or new features are added. It is independently developed and has a small application footprint.

Technology stack: The .NET 5 framework is cross-platform. Deployment on Linux servers is cost-effective, and performance in all aspects has improved. Independent development and deployment allow flexible use of suitable lightweight components to improve development efficiency and performance. WebAPI uses an asynchronous non-blocking request mode to improve throughput.

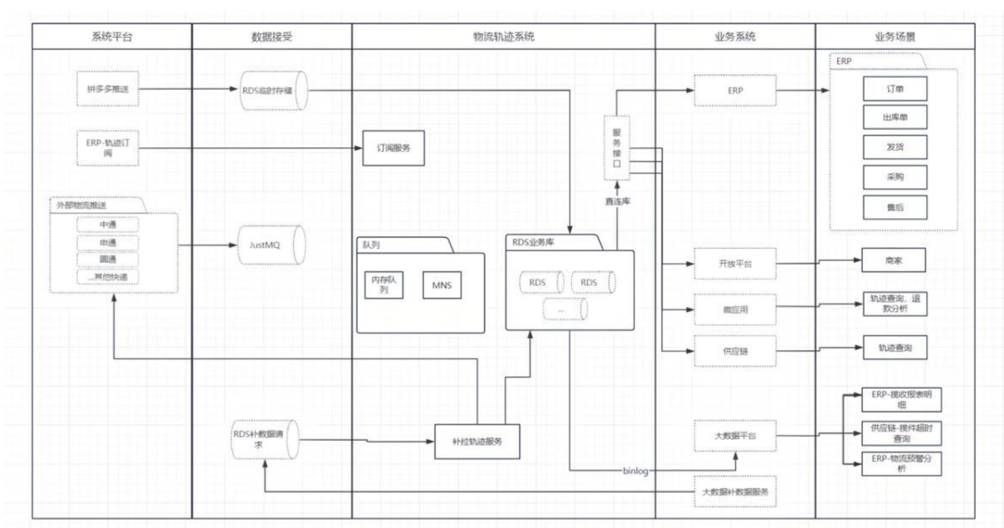


Figure 2 System technical architecture

3. Implementation Process

The JUSHUITAN cross-border logistics trajectory analysis system is a fully automated and intelligent analysis system. Merchants independently configure international logistics monitoring through menu-based settings in the system and adjust their service strategies according to the implementation process.

For example, logistics timeout can be fully customized by platform, status and time, enabling customization for all merchants. According to the different assessment standards of different platforms and customers' requirements for fulfillment timeliness, the system can configure the monitoring dimensions and time required.

For abnormal English words appearing in logistics trajectories, the system can identify and parse them and include them in abnormality monitoring. The powerful data parsing capability processes detailed logistics trajectory content in multiple languages. Once any configured word or sentence is hit, the system can analyze it in time and identify customer-defined abnormalities.

4. Core Technologies

JUSHUITAN has completed the construction of its data platform and built its own online analytical processing (OLAP) engine, StarRocks. By using technologies such as vectorization, cost-based optimization (CBO), intelligent materialized views and a columnar storage engine that can be updated in real time, the platform realizes multidimensional, real-time and highly concurrent data analysis. At the same time, AI algorithms including exponential smoothing, linear programming, heuristic algorithms, the Christofides algorithm, ant colony algorithms and operations optimization are used for intelligent data analysis.

In system deployment, cluster deployment is adopted. This method ensures that the system remains stable in high-concurrency scenarios and effectively distributes requests through load balancing, improving overall system throughput. To better manage the services and calls of each module, the project introduces the distributed framework Nacos. Nacos not only provides service discovery, configuration management and dynamic service management, but also supports advanced deployment strategies such as gray release and traffic scheduling, greatly improving operational efficiency and response speed.

In data processing, Alibaba Cloud MaxCompute is used for computing capability. MaxCompute, as a distributed computing platform, has powerful data processing capabilities, can handle large-scale data and supports multiple computing models. Through MaxCompute, data cleaning, integration and analysis can be completed quickly, providing accurate data support for upper-level businesses. Meanwhile, to meet the storage needs of different information types, the project uses high-performance storage solutions such as Hologres and Relational Database Service (RDS). Hologres is a high-performance distributed OLAP database suitable for real-time big-data analysis scenarios and can provide low-latency and high-concurrency data access. Alibaba Cloud RDS supports multiple relational databases and provides stable, secure and convenient database hosting services for platform operation.

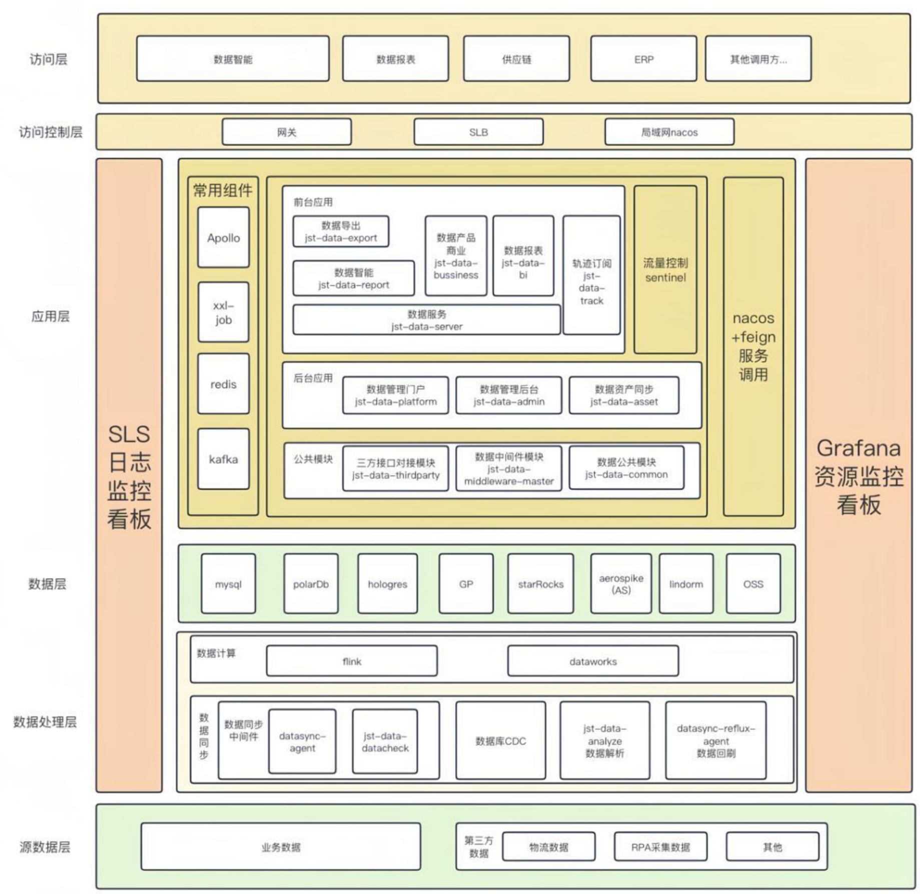


Figure 3 Data platform technical architecture

With the rise of AI technology, JUSHUITAN has also built its own AI agent architecture, JuXie Cloud. It includes an AI base, namely the basic capability layer; AI HUB, the middle layer and model-service hub; the AI agent platform, namely the upper-level development and runtime environment; AI applications, namely the business-scenario layer; and its own fine-tuned models. It is believed that in the future this architecture will provide stronger multidimensional intelligent processing for overseas business.

III. Project Highlights

The JUSHUITAN cross-border logistics trajectory analysis system helps cross-border e-commerce enterprises track and intelligently analyze international logistics, bringing great convenience to merchants.

First, when merchants need to check the logistics status of product parcels, traditional logistics tracking software still requires them to enter tracking numbers for queries. In the JUSHUITAN system, merchants only need to click the tracking number to immediately know the logistics details of the parcel. After a sales order is shipped, international logistics tracking is performed automatically without manual addition by customers, ensuring timeliness. Logistics trajectories are translated into Chinese across languages, greatly reducing merchants' learning costs in cross-border logistics.

Second, the logistics information monitored and fed back by JUSHUITAN can confirm whether product parcels are delivered to customers within the required time. If delivery is not completed in time, the system can immediately detect it. The data monitoring function can promptly identify causes and find solutions, actively helping customers monitor abnormal logistics status. This allows feedback on the timeliness and accuracy of cargo delivery in advance and helps merchants maintain the shopping experience of customers placing sales orders.

Third, logistics reports tracked through the JUSHUITAN system can effectively analyze logistics routes from the origin country to the destination country, including but not limited to delivery rate, receipt rate and capability comparisons among different freight forwarders used by merchants, providing a comprehensive multidimensional data analysis system.

Against the background of cross-border logistics tracking difficulties, the core logic of the JUSHUITAN cross-border logistics trajectory analysis system is to monitor international logistics intelligently and conveniently. Compared with most similar systems currently on the market, which can only view and monitor logistics tracking numbers and cannot collaborate with ERP systems for multidimensional processing of internal orders, the JUSHUITAN system has fully broken through this preliminary application scenario. By effectively integrating with e-commerce ERP software, it helps merchants intelligently handle a series of logistics-related problems in the operation and management of international logistics, providing merchants with true decision-making data.

IV. Implementation Results

The JUSHUITAN cross-border logistics trajectory analysis system has currently connected to more than one hundred mainstream e-commerce platforms, including AliExpress, Shopee, Lazada, TikTok, Amazon and eBay; logistics trajectory platforms such as Track123 and 51Tracking; and more than 1,300 international logistics freight forwarders. It aggregates e-commerce platform order status and logistics trajectories, helping cross-border e-commerce enterprises efficiently handle pre-sales and after-sales issues.

(I) Economic Benefits

JUSHUITAN ERP advocates one-stop services. After logging into a JUSHUITAN ERP account, users can enter different systems or functional interfaces through different menus and complete various customer needs in a one-stop manner. Although this analysis system is not currently purchased separately, it has played an important role in attracting traffic for JUSHUITAN ERP customers, especially cross-border merchants.

After the cross-border logistics trajectory analysis system went online, the number of subscribed orders rapidly increased from more than 10,000 orders per month to around 500,000 orders per month and has continued to remain at a high level.

Relying on years of project experience in domestic e-commerce ERP, JUSHUITAN provides cross-border e-commerce enterprises with integrated business-finance solutions through seamless connection with global e-commerce platforms, logistics and warehousing institutions, efficiently empowering the sustainable development of cross-border businesses. In the past two years, cross-border business has maintained stable and rapid growth. A subsidiary has been established in Thailand, and cooperation with platforms such as Shopee, Lazada and Temu aims to radiate across Southeast Asia. The business has developed healthy self-generating capabilities and has initially achieved profitability.

(II) Social Benefits

First, the system effectively breaks information silos and addresses important pain points in cross-border logistics. Within only six months after the system went online in 2024, merchants had prepaid more than ten million yuan into their ERP accounts in order to use this function. The system also successfully helped cross-border e-commerce customers monitor more than five million international logistics orders, greatly improving merchants' efficiency in international logistics tracking, problem-handling speed and ability to solve omitted issues, thereby improving cross-border order fulfillment capability and customer satisfaction.

Second, it effectively empowers all kinds of merchants in the cross-border industry and helps merchants avoid international logistics losses. In 2024, the system worked with cross-border e-commerce merchants to effectively monitor more than 100,000 abnormal international logistics orders, providing early warning and data support to help customers recover losses from orders and international logistics fees.

The system is easy to operate, provides a very good customer experience, requires no training personnel to intervene, has a highly user-friendly process and covers various user scenarios. It is expected to become a benchmark application scenario in the industry.

At present, this case has been selected as an excellent "AI Plus" application case in Shanghai's commercial sector.

(III) Technological Breakthroughs and Innovation Capability

JUSHUITAN is mainly engaged in the development of e-commerce application software and provides the most advanced and effective application scenarios for e-commerce enterprises. In relative terms, it is a leader in the e-commerce ERP industry in making chain innovation, paradigm innovation and model innovation breakthroughs with existing technologies. It currently has 25 authorized invention patents, far exceeding the total number of authorized patents held by other competitors in the industry. This project has applied for one software copyright and one invention patent. While continuously improving its product matrix and business model, the company is also increasing investment in AI and big data, focusing on research into AI + RPA application scenarios while also developing the electronic platform as a service (EPaaS) platform to achieve open-source empowerment for the industry's rapid development.

(IV) Market Adaptability

After the JUSHUITAN cross-border logistics trajectory analysis system went online, it showed the following advantages over other cross-border ERP tools in the logistics dimension by enabling users to easily grasp the real-time logistics status of cross-border parcels.

For example, a new lifestyle digital brand, Shenzhen Baseus Technology Co., Ltd., used the JUSHUITAN cross-border logistics trajectory analysis system to track more than three million orders, discovered nearly 50,000 abnormal orders and recovered losses of hundreds of thousands of yuan. These losses were originally inevitable costs that customers had to bear in cross-border logistics, but now they may be transformed into order revenue.

In addition, the JUSHUITAN system provides customers with multidimensional logistics trajectory analysis reports, namely analysis of all logistics used within a time range, including summary and detailed reports. Examples include online timeliness, which platforms usually assess and which may lead to order cancellation if timeliness standards are not met; receipt timeliness, which analyzes the logistics receipt timeliness of different channel providers and provides effective data to logistics providers for assessment standards; country-level online timeliness reports covering global countries; and country-level receipt timeliness reports for different countries. These reports are well recognized by customers and the market, and the number of system users is growing rapidly.

(V) User Feedback and Market Recognition

Merchants who have used the system report that after completing automatic subscription settings in the logistics trajectory tracking module, shipments are automatically tracked and no longer need to be added manually. According to different configurations, after an international logistics tracking number is generated for an order, it automatically enters the analysis system, and abnormal information is marked and promptly reminded. In addition, the system covers translation of almost all languages, allows switching between original and translated languages at any time, and enables quick sharing and copying.

Case 19

Application of AI in International Maritime Logistics

ONEAIX Technologies (Shanghai) Co., Ltd.

I. Basic Enterprise Information

ONEAIX Technologies is a world-leading AI technology company focusing on the broader supply chain sector, including production and manufacturing, logistics and distribution, and new retail. The company has independently developed core technologies in AI, automation and big data mining. Based on large-model technology, ONEAIX has built the world-leading enterprise-level digital employee platform - ONEAIX Platform - and has comprehensively developed an innovative office model of "white-collar workers + AI Agents". It has successfully empowered nearly 2,000 enterprise customers to achieve efficient and intelligent human-machine collaboration. The company is a national-level "specialized, refined, distinctive and innovative" Little Giant enterprise and has received a range of recognitions for its innovation in digital employees and enterprise AI applications.

II. Case Introduction

(I) Background and Needs

In international maritime logistics services, the traditional operating model faces many challenges that seriously affect industry efficiency and service quality:

1. Efficiency bottlenecks and high labor costs: International maritime logistics involves a large volume of high-frequency and repetitive tasks, including data entry, verification, information inquiry and information transmission. Typical examples include customer inquiries, order intake, shipping instruction recognition and system order creation, booking, space release monitoring, shipment tracking and monitoring, exception alerts, and bill of lading comparison. These tasks have a low level of automation and depend heavily on manual work, resulting in low efficiency. A leading logistics enterprise processes tens of thousands of orders each month, and fully manual data entry creates a huge workload and obvious efficiency bottlenecks.

2. Fragmented information and low collaboration efficiency: Freight rates, vessel schedules, cargo status and customer requirements are scattered across carrier websites, agency systems, instant messaging tools, emails and offline paper documents. Information acquisition and integration are extremely time-consuming. Communication costs among sales, route management, operations, customer service and finance departments are high, information circulation is inefficient, and an effective cross-departmental collaboration system is difficult to establish.

3. Lagging customer experience and unstable service quality: Customers increasingly expect real-time response, proactive service and personalized solutions. Under the traditional model, however, business personnel often cannot provide 24/7 responses, and services are mostly reactive. Customers face long waiting times and poor experience. Service quality depends heavily on the experience and professional competence of individual employees, making stable and high-standard service difficult to provide.

4. Insufficient decision support and lack of risk management: Massive business data has not been effectively accumulated, cleaned or analyzed. As a result, management lacks unified data dashboards and in-depth insights, making it difficult to accurately predict market trends, optimize resource allocation or support scientific decision-making. Supply chain processes are complex and vulnerable to external disruptions, but real-time monitoring and exception alert mechanisms remain insufficient, making it difficult to respond effectively to delays, shortages, accidents and other emergencies.

(II) Implementation Plan

1. This project integrates generative AI large models with AI Agent technology and deploys a team of professional AI digital employee Agents to provide an end-to-end intelligent solution from inquiry and quotation to service fulfillment. The core technical support includes the self-developed international logistics vertical large model

CubeLLM, a multimodal intelligent document processing engine, a freight-rate big data analysis platform, and an industry knowledge graph-based intelligent question-answering system.

2. The project is implemented through five core Agents that cover the entire process. The intelligent inquiry and quotation Agent accepts inquiries 24/7, automatically searches prices and generates standardized quotations. The intelligent order intake Agent responds to instructions in real time and completes shipping instruction parsing and order creation within 30 seconds. The booking and space release automation Agent completes booking with one click, monitors space release in real time and reports exceptions. The bill of lading automation Agent integrates multi-source data, quickly prepares bills and verifies carrier requirements. The customer service automation Agent proactively pushes logistics milestone information, conducts shipment tracking every 30 minutes and issues exception alerts.

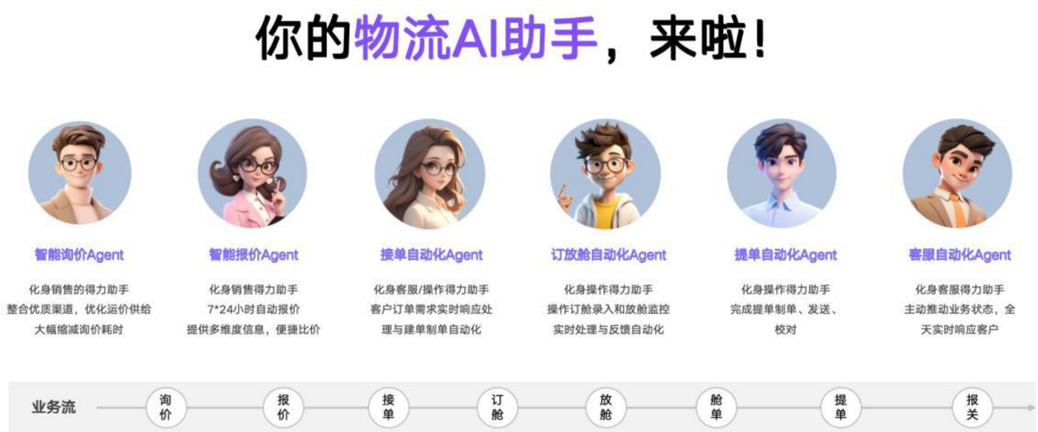


Figure 1 The professional AI digital employee Agent team independently developed by ONEAIX enables intelligent services across the full logistics process.

III. Project Highlights

(I) Key Technological Breakthroughs

1. Industry-specific AI digital employee Agent framework and executor:

Agents as core productivity elements: The project innovatively takes “AI digital employee Agents” as its core value carrier and productivity element. They are not merely chatbots, but intelligent executors with full-chain capabilities in perception, cognition, decision-making, execution and feedback. They realize a leap from “you ask, I answer” to “you say, I do”, which is the fundamental difference between this project and traditional RPA or single AI tools.

Task-chain executor and contextual memory: To address the complexity and multi-step nature of international logistics processes, the project designs an efficient task-chain executor that intelligently decomposes complex business instructions into subtasks and executes them in logical order. Its core innovation lies in the task-context memory module, which stores and manages all relevant information from multi-round interactions and multi-step execution. This ensures coherence and consistency of the Agent during long-duration and complex tasks, effectively solving the limited memory-window problem of general large models.

Business process definer and low-code/no-code orchestration: The project provides visual tools that allow business personnel or developers to define and modify complex international logistics processes with an extremely low threshold. It seamlessly connects large-model intelligent modules with other automation tools, greatly accelerates AI application development and iteration, lowers the technical threshold and enables business agility.

2. Deep fine-tuning and application of the international logistics vertical large model CubeLLM:

Highly professional knowledge accumulation: The self-developed international logistics vertical large model CubeLLM is fine-tuned with a dataset of up to 20 GB, covering a massive volume of professional terms, industry books, actual documents in multiple languages, and millions of quotation inquiry conversations. As a result, CubeLLM reaches an unprecedented level of precision in understanding, perceiving and reasoning about industry knowledge.

Proactive mitigation of large-model hallucination: To address the factual and stochastic limitations of general large models, CubeLLM uses a unique prompt assembly structure, enhanced data verification mechanisms and continuous model quality control. Combined with RAG, it greatly reduces hallucination. Through deep integration with the international logistics knowledge graph, RAG enables the large model to dynamically retrieve and reference the most accurate and up-to-date industry data when answering questions or executing tasks, ensuring reliable information output.

Multilingual and complex-context understanding: International logistics naturally involves multilingual environments as well as complex colloquial inquiries and instructions, which impose high requirements on AI language understanding. Through training and fine-tuning on massive multilingual corpora and the integration of common industry abbreviations and specific expression habits, CubeLLM significantly improves Agent understanding and interaction in cross-language and complex contexts.

3. Breakthrough innovation in the multimodal intelligent document processing engine:

Core technology of the vision-language model (VLM): Moving beyond the limitations of traditional OCR, which relies only on vision or character recognition, this project innovatively integrates multimodal AI technology enhanced by large-model vision capabilities. This means that the Agent can not only “read” text, but also “understand” the layout, structure and visual elements of images, enabling deeper document content understanding.

Complex layouts and cross-page table processing: International logistics documents, such as commercial invoices and customs declarations, often contain complex table structures, multi-row and cross-page content, and different unit systems and alignment patterns across countries and regions. This project creates an intelligent layout understanding approach combined with AI-assisted template configuration technology, greatly improving OCR recognition accuracy and information extraction for such complex documents. It solves a long-standing difficulty that traditional engines struggle to handle and provides high-quality structured data input for subsequent automated processes.

(II) Scenario Uniqueness

Pioneering practice of supply-chain-specific “professional digital employees”: The project is among the first globally to propose and implement AI digital employee products for the supply chain sector. It builds a diversified team of “professional assistants”, covering customer and sales inquiry and quotation Agents, order intake automation Agents, booking and space release automation Agents, bill of lading automation Agents and customer service automation Agents. This realizes end-to-end coverage of the full international logistics business process.

Operational model transformation from passive response to proactive management: Through intelligent customer service Agents, big data visualization and intelligent decision support, the project shifts the service model from passive inquiry response to proactive notification. Agents can monitor business status in real time and proactively push logistics milestone information and exception alerts to customers, greatly improving customer experience and operational efficiency.

First-mover advantage and vertical specialization in business-to-business (B2B) international logistics: The project precisely targets and deeply specializes in the complexity, multilingual nature and multiple rules of B2B international logistics. Through the high-level integration of Agent technology and industry large models, it establishes a unique first-mover advantage and strong competitive barriers in this specific high-value market.

Figure 2 Platform architecture.

IV. Implementation Results

(I) Economic Benefits

The platform has achieved significant economic benefits and business growth at Shanghai Longsun, a national 5A international logistics enterprise:

Significant improvement in economic benefits: AI digital employees replace a large amount of repetitive manual labor, substantially reducing labor costs, operational error costs and time costs, and directly bringing considerable cost savings to the enterprise. By providing real-time, accurate quotations and efficient services, they directly help improve order conversion rates, optimize supply chain links and enhance overall profit margins. In core business processes, the order intake automation Agent automatically integrates shipping instructions, emails and historical order information, creating orders within 30 seconds. This lowers the personnel threshold, simplifies operating steps and doubles labor efficiency. The booking and space release automation Agent enables one-click booking, automatic space release monitoring, exception identification and automatic customer forwarding, increasing single-ticket operation efficiency tenfold. The bill of lading automation Agent integrates multi-source data, including emails, online booking confirmations and systems, for rapid document preparation, automatic verification of carrier order requirements, and faster employee onboarding. Human efficiency increases threefold, while document fees are reduced by 50 percent.

Substantial enhancement of business capacity: Intelligent automation enables the enterprise to handle a larger business volume with fewer personnel. This means that the enterprise can seize more market opportunities and cope with business peaks without significantly increasing labor costs. Through this platform, business carrying capacity increased by 2.5 times. During the peak maritime season, when order volume surged by 30 percent, the enterprise added no front-line operations staff while still maintaining efficient business processing. At the same time, the employee training cycle was shortened by half, and the time for new employees to move from onboarding to independent work was reduced from three months to 1.5 months, lowering the cost of talent development.

Continuous optimization of quotation conversion rate and overall profit margin: The intelligent inquiry and quotation Agent can handle 2,000 inquiries per day, respond to customer needs in real time 24/7, provide multi-dimensional freight-rate information and convenient price comparison services, and shorten customer inquiry response time from an average of four hours to within one minute. The quotation conversion rate increases by 15 to 20 percent. Through visualization and controllability of the inquiry and quotation process, the platform helps the enterprise upgrade refined management of its marketing system, optimize resource allocation mechanisms, build sustainable digital management capabilities, ensure efficient use of resources and directly drive profit growth.

(II) Social Value

Supporting high-quality development of the Shanghai International Shipping Center: The project injects advanced AI technology into Shanghai's international logistics service industry. Through the hyperautomation of core business links such as inquiry and quotation, order intake, booking and space release, and document processing by AI digital employees, it significantly improves the operational efficiency, service precision and intelligent level of Shanghai Port in the global market.

Optimizing the human resource structure and promoting worker development: AI digital employees take over a large number of high-frequency, repetitive, rule-based, tedious and low-value tasks in international logistics. Human employees are therefore freed from heavy and mechanical work and can be redeployed to positions with greater creative, strategic and interpersonal value.

Improving customer experience and service quality: AI digital employees provide uninterrupted real-time response services, eliminating the time limitations of traditional human services and ensuring that global customer needs can be answered immediately. This creates an excellent "zero-waiting" service experience. By proactively pushing logistics milestone information and exception alerts, the platform enables customers to track cargo transportation status in real time, enhances business transparency and trust, and improves the global competitiveness of Chinese international logistics enterprises.

Promoting sustainable development and green low-carbon operations: Through AI-based intelligent scheduling, capacity optimization and space management, the platform effectively reduces unnecessary transport mileage, vessel empty-load rates and resource waste. Logistics enterprises applying this technology have seen average vessel empty-load rates fall by 15 percent and transport mileage optimized by 8 to 10 percent, directly or indirectly reducing fuel consumption and carbon emissions. This contributes technological strength to the green and low-carbon development of the global logistics industry and aligns with China's dual-carbon strategic goals.

Case 20

R&D and Application of an Intelligent Optimization Control System for Waste Incineration

Shanghai SUS Environment Co., Ltd.

I. Basic Enterprise Information

Shanghai SUS Environment Co., Ltd. is a world-leading provider of waste incineration equipment and technologies, and one of the world's top three investors and operators of waste-to-energy projects. Centered on waste-to-energy projects, the company's business covers the investment, construction and operation of low-carbon circular industrial parks, core equipment supply, sludge treatment, environmental remediation and comprehensive energy utilization. SUS Environment attaches great importance to independent R&D and technological innovation. It has a high-level R&D team, has achieved more than 30 scientific and technological outcomes, has been granted more than 270 patents, and has participated in the preparation of over 20 national standards and industry specifications. The company has led three national key R&D programs and has received major awards including the National Science and Technology Progress Award.

II. Case Introduction

(I) Project Introduction

The "intelligent optimization control system for waste incineration" developed by Shanghai SUS Environment Co., Ltd. addresses problems in waste incineration such as unstable operation and difficult pollutant emission control caused by complex fuel composition and large fluctuations in calorific value. By applying AI technology, the system realizes intelligent control of the incineration process. It integrates four AI algorithms - reinforcement learning, operating parameter prediction, anomaly identification and visual recognition - to analyze operating data and flame images in real time, automatically optimize key parameters such as feeding, grates and air volume, and improve control accuracy and response speed. After implementation, the system operation rate exceeds 96 percent, manual operation frequency is reduced by 80 percent, power generation per ton of waste increases by 2.04 percent, and economic benefits from power generation and operation and maintenance savings reach more than RMB 1 million per year.

(II) Background and Needs

Municipal solid waste comes from wide-ranging sources, has complex composition and fluctuating calorific value, and can easily cause unstable combustion equipment operation and excessive pollutant emissions. Mechanical grate incineration adapts well to municipal waste incineration and has become the mainstream technical option. However, due to fuel characteristics, it still faces problems such as large fluctuations in combustion and original pollutant concentration, low automation and low efficiency. Municipal waste combustion and flue gas pollutant purification form a typical nonlinear, strongly coupled and lagging system. Traditional automatic combustion control (ACC) and flue gas purification control systems based on proportional-integral-derivative (PID) control logic have poor adaptability to such systems, and it is difficult for them to meet the needs of long-cycle, high-stability and efficient operation.

At present, domestic and international applications of AI technology in waste incineration mainly focus on parameter prediction and combustion diagnosis. Relatively little work has been done on intelligent control, and most applications are limited to a single link in the overall waste treatment facility. Intelligent facilities that reflect coupling among process systems are still under development, and engineering applications that realize long-cycle intelligent management and control are even more limited. Combining big data analytics with AI technology to realize data-driven intelligent management and control of waste-to-energy plants is an important trend for the sustainable development of the industry and an important path toward forming new quality productive forces.

(III) Implementation Plan

1. Accurate identification of combustion conditions

In the actual in-furnace combustion process, the combustion state is affected by the coupling of multidimensional factors such as dynamic flame features and process parameters. This complex interaction makes the characteristics of inefficient operating conditions highly concealed and difficult to identify accurately. From the perspective of technical implementation, how to use advanced image processing algorithms and machine learning technology to accurately extract multidimensional dynamic flame features and realize accurate identification of combustion conditions has become a key technical issue to be solved in this project. Therefore, it is necessary to design an operating-condition identification framework that integrates multidimensional features, providing strong technical support for precise regulation of the waste incineration process.

2. Real-time prediction of key indicators driven by both mechanism and data

Due to nonlinear factors such as fluctuations in fuel properties, uneven furnace temperature distribution, changes in flue gas composition and load disturbances, the coupled interference among different signals makes it difficult to accurately capture the changing patterns of key indicators in the combustion system. Therefore, how to effectively extract process states and abnormal signs from multi-source signal interference and accurately predict key indicators under constraints has become a core difficulty that urgently needs to be solved for control system optimization. From the technical route, traditional modeling methods cannot accurately describe the impact of complex operating conditions. It is necessary to build a prediction framework driven by both mechanism and data, so as to balance the logical consistency of the mechanism with the adaptive capability of data.

3. Multi-objective optimization and adaptive composite control

Since the calorific value of waste fluctuates by season and region, relying solely on expert rules or AI models cannot simultaneously achieve stable combustion and excellent production capacity. The project must therefore solve the problems of rigid rules in traditional expert control when facing sudden changes in waste calorific value and load disturbances, as well as model mismatch and the lack of a dynamic weight-switching mechanism in pure deep learning control under small-sample extreme conditions. It aims to realize adaptive regulation under different waste calorific values and load disturbances, ensure stability and adaptability, overcome the technical obstacles of “model mismatch” and “rule rigidity”, and form an implementable low-carbon adaptive feedforward-feedback composite control technology. This provides technical support for adaptive low-carbon operation and stable operation of incineration systems under complex conditions.

III. Project Highlights

(I) Key Technological Breakthroughs

1. Multivariable reinforcement learning

The multivariable reinforcement learning algorithm includes three parts: feature input, decision algorithm engine and decision output. Its principle is as follows: main operating parameters such as steam flow, primary air volume, pressure in each air chamber, main steam pressure, furnace flue gas temperature and pusher stroke are selected as variables representing waste incineration conditions. Through multidimensional extraction of historical data, the corresponding relationships between relevant equipment instructions at different operating points are summarized, and benchmark operating parameter guidance values are provided.

Through feature extraction from historical data, the corresponding relationships between changes in main steam flow and feeder speed commands, grate cycle, and primary and secondary fan speed commands can be obtained, forming benchmark values for control instructions. This avoids the conventional control process of obtaining benchmark data through a large number of tests and quickly obtains the benchmark feature data of each furnace.

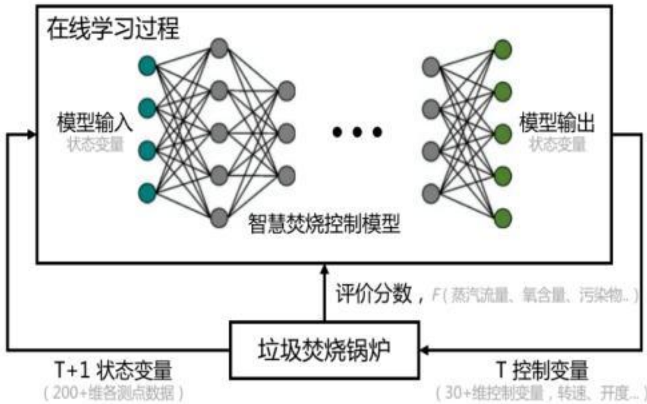


Figure 1 Schematic diagram of the reinforcement learning model for waste incineration control.

2. Operating parameter prediction

Waste incineration boilers involve large fluctuations in waste calorific value and strong delay and inertia characteristics. Traditional industrial control methods have difficulty effectively solving the control lag caused by such long delay and strong inertia. In addition, large and uncertain fluctuations in waste calorific value further increase the difficulty of control. By leveraging data-driven advantages and combining industrial process principles, the project develops an industrial prediction engine. Based on prediction results and recommendation algorithms, it guides the control system or operators to act in advance, alleviating poor control performance and large fluctuations in operating parameters caused by long delay, strong inertia and model uncertainty.

Industrial prediction is based on production process principles and historical operating data. By analyzing input-output correlations, constructing feature variables and combining the characteristics of different systems, the system identifies feature variables, including original and constructed features, and target variables, namely the parameters to be predicted. Industrial prediction is the algorithmic core. It is built through machine learning and deep learning algorithms. The industrial prediction engine periodically trains and updates models according to operating results, maintaining high prediction accuracy.

The predicted parameters include main steam parameters and pollutant emission parameters. Based on the prediction model, control parameters are optimized to further improve boiler stability. Taking steam flow prediction as an example, the steam flow prediction model relies on deep learning algorithms and is built using dozens of historical operating parameters, including pusher speed, pusher action, primary air volume, primary air pressure, secondary air volume, furnace temperature, flue gas oxygen content, feedwater flow, drum liquid level and main steam pressure. It can basically and accurately predict steam flow 180 seconds into the future, providing an effective basis for subsequent steam flow trend estimation and feeding-time decisions, and mitigating the adverse impact of waste calorific value fluctuations on stable combustion.

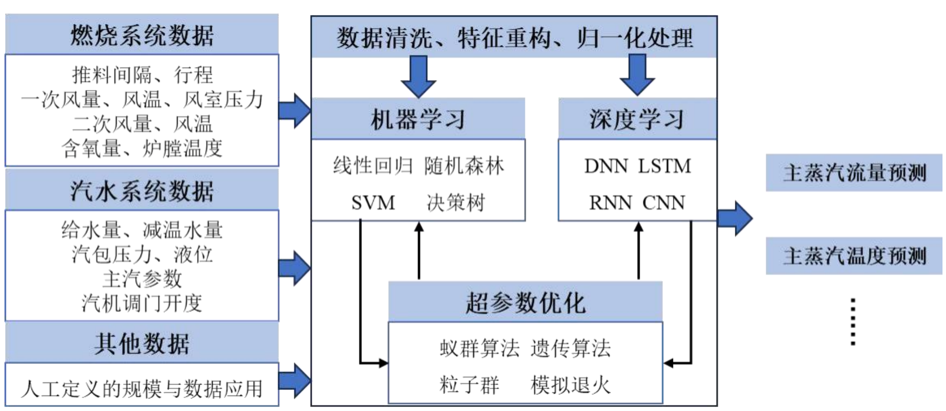


Figure 2 Schematic diagram of operating parameter prediction.

3. Anomaly identification algorithm

After obtaining historical operating data, the anomaly identification algorithm uses unsupervised clustering algorithms such as K-means and K-MEDOIDS to divide operating conditions, remove clearly abnormal operating-condition data and obtain data within the normal generalized operating-condition range. At the same time, it uses a modeling engine to model normal operating-condition data and obtain a big-data black-box model of the target variable. Through the model, the theoretical normal value can be calculated in real time and compared with the actual value to determine the degree to which the current state deviates from the normal value.

In this project, in order to effectively judge the amount of waste entering the furnace, grate differential pressure is selected as the basis for determining the amount of waste in the furnace according to the structure of the control device. However, grate differential pressure is also affected by primary air volume, drying-section damper opening, combustion-section damper opening and other section damper openings. Its instantaneous value cannot directly reflect the amount of waste. Therefore, an anomaly identification algorithm is introduced when using air pressure to judge waste feeding volume, and big-data modeling is carried out for grate differential pressure. Historical operating data for three to five months is selected. Abnormal conditions are removed through unsupervised learning to obtain normal operating-condition data. In normal conditions, variables such as primary air frequency feedback, primary air pressure feedback and secondary air pressure feedback are selected as model input features. A data modeling engine is used to establish a grate differential pressure model. Based on the deviation between the model output predicted value and the actual air pressure value, the current amount of waste entering the furnace is determined, and the feeding and grate speed are dynamically adjusted to ensure feeding uniformity.

4. Flame image recognition

The visual recognition algorithm mainly identifies flame video images. Through visual recognition, it timely perceives the combustion condition and combustion position of waste in the furnace, and then rapidly adjusts the grate and primary fan to ensure a stable combustion process. The visual recognition algorithm in this project mainly uses image classification and image segmentation to identify the fire line, flame area, smoke and dust, and brightness in flame images. It then combines post-processing to analyze the real-time combustion state of grate flames as input for subsequent control optimization algorithms. Based on on-site flame images, the recognition algorithm divides images according to the identified flame color, fire-line area and corresponding position relationships, thereby conducting more detailed analysis of flame combustion status. It can also use the flame color and brightness inside the image to evaluate the combustion condition of waste in the furnace and the generation of carbon monoxide, thus conducting targeted fuel optimization adjustment.

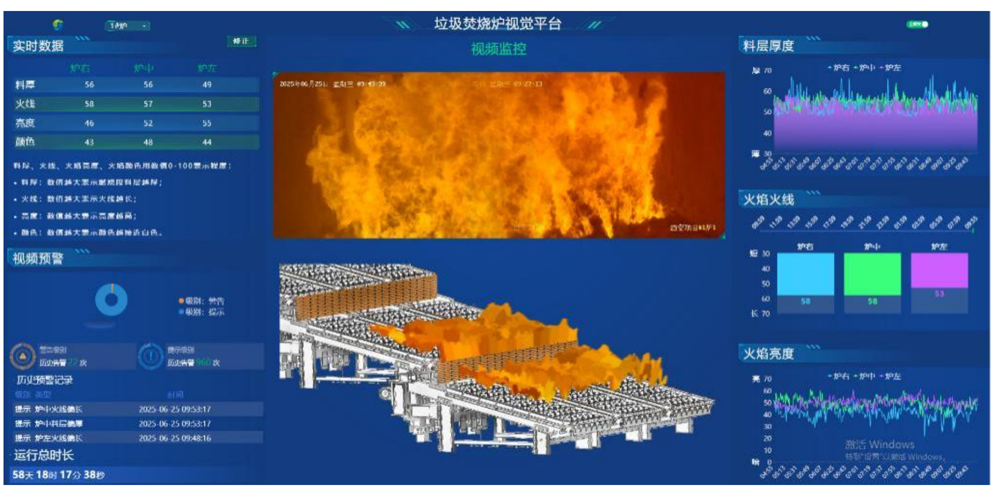


Figure 3 Visual platform for flame image recognition.

(II) Scenario Uniqueness

The scenario uniqueness of the intelligent optimization control system for waste incineration lies in its successful deep integration of cutting-edge AI technology into a highly complex, variable and traditional industrial process control scenario. It solves two long-standing core pain points in the waste incineration industry: "experience dependence" and "control lag". Municipal solid waste has complex composition and highly fluctuating calorific value, making the incineration process a typical nonlinear, strongly coupled and long-lag complex system. Traditional automatic control systems, such as PID or conventional ACC, have poor adaptability to this process, resulting in very low operation rates. Production operation has long relied heavily on frequent manual intervention by experienced operators. This not only involves high labor intensity, but also has strong subjectivity and delayed feedback, easily leading to unstable combustion, low efficiency and fluctuations in pollutant emissions.

The uniqueness of this system is that it is not a simple optimization of a single link. Instead, it builds an intelligent control brain covering the full chain of “perception - prediction - decision-making - control”. Its core breakthroughs are reflected in four aspects:

Transformation of decision-making from “experience-driven” to “data-driven”: Through reinforcement learning algorithms, the system mines the control logic of excellent operators from massive historical data and forms data-based control benchmarks. It transforms tacit “master operator experience” into replicable and optimizable algorithmic models, realizing knowledge accumulation and automation.

Advanced prediction and intervention for processes with “large inertia and long delay”: In response to the slow response characteristics of the incineration process, the system uses deep learning models to accurately predict key parameters, such as steam flow, 180 seconds into the future. This enables control instructions to “foresee the future” and achieve advanced regulation, fundamentally overcoming the lag of traditional feedback control.

Multimodal perception fusion breaks through bottlenecks in key parameter measurement: The system creatively integrates distributed control system (DCS) operating data with flame video images. Through visual recognition algorithms, it analyzes in real time in-furnace combustion states that cannot be directly measured, such as fire-line position and flame morphology, and quantifies them as control parameters. Combined with the “soft measurement of material layer thickness” realized by the anomaly identification algorithm, it builds a comprehensive perception capability far beyond traditional DCS.

Closed-loop intelligent control with multi-objective collaborative optimization: The system integrates the outputs of four major AI engines - reinforcement learning, parameter prediction, anomaly identification and visual recognition - and collaboratively optimizes dozens of control loops such as feeding, grates and air volume. Ultimately, it achieves a dynamic optimal balance among improving combustion stability, increasing power generation efficiency and reducing pollutant emissions, forming an adaptive closed-loop intelligent control system that can cope with complex and variable operating conditions.

In summary, the uniqueness of this system lies in the fact that, for the first time in a typical waste incineration scenario, it realizes a leap from reliance on manual experience to data-driven intelligent decision-making. It builds an “industrial AI brain” with perception, prediction and collaborative optimization capabilities, providing a high-value model for the intelligent upgrading of process industries.

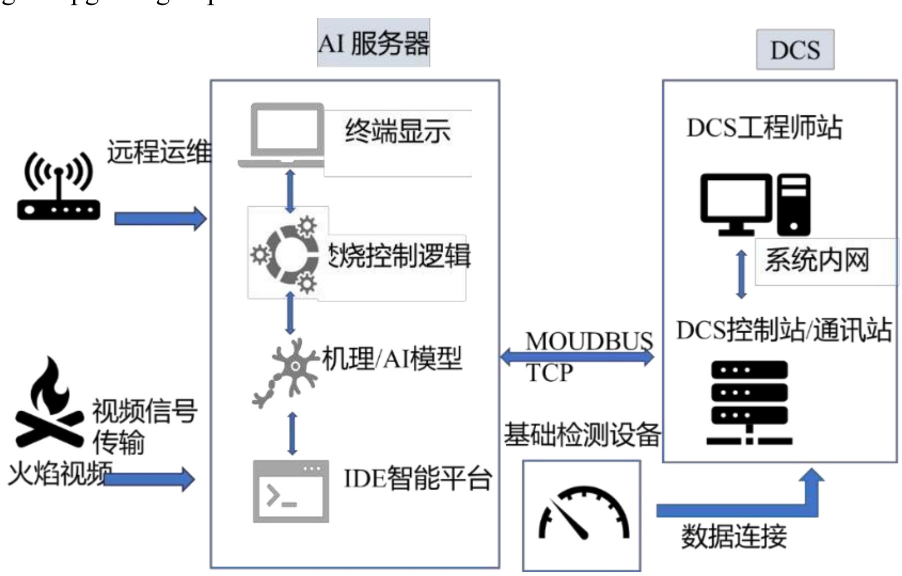


Figure 4 Architecture of the intelligent optimization control system for waste incineration.

IV. Implementation Results

(I) Economic Benefits

By stabilizing control of furnace temperature, boiler outlet oxygen content and main steam flow, the intelligent optimization control system makes combustion stable and the material layer uniform. The system has been applied in Xi’an, Shaanxi, and Langfang, Hebei.

Application at Xi’an Jingwei SUS Environment Energy Co., Ltd.: Compared with purely manual operation, main steam flow stability improved by more than 20 percent, and power generation per ton of waste increased by about 2 percent. Based on a single incineration line with a daily treatment capacity of 750 tons, annual power generation increased by 2 million kilowatt-hours, adding approximately RMB 800,000 in power generation revenue. Furnace temperature stability improved by more than 10 percent, significantly improving furnace slagging and providing strong support and assurance for equipment risk management. The maintenance cycle can be effectively extended by six months, reducing unplanned shutdown time and saving RMB 300,000 in operation and maintenance costs each year. Combining power generation revenue and saved operation and maintenance costs, annual economic benefits increased by RMB 1.1 million.

Application at Sanhe SUS Renewable Energy Co., Ltd.: Compared with purely manual operation, main steam flow stability improved by more than 35 percent, and power generation per ton of waste increased by 1.5 percent. With a single furnace treatment capacity of 1,000 tons per day, annual power generation increased by 2.48 million kilowatt-hours, adding RMB 920,000 in power generation revenue. Furnace temperature stability improved by more than 15 percent, saving RMB 400,000 in equipment operation and maintenance costs. Combining power generation revenue and saved operation and maintenance costs, annual economic benefits increased by RMB 1.32 million.

(II) Social Value

The project results will promote the transformation and upgrading of China’s municipal waste incineration control technology from traditional manual control and PID-based control to AI-driven control. The intelligent optimization control system developed in this project achieves an operation rate of more than 96 percent, significantly improves the intelligent operation level of municipal waste incineration plants, and reduces operator operation frequency by more than 80 percent. It effectively reduces dependence on skilled operators, improves the working environment and helps cultivate a group of technical personnel with AI application capabilities for the industry. At the same time, the successful implementation of the project forms a replicable and scalable intelligent control model, injecting new technological content into the waste incineration equipment manufacturing and technical service industries, and supporting green, low-carbon and high-quality development of the environmental protection sector.

Case 21

Edge AI Vision Empowering Defect Detection for Tablets in Pharmaceutical Enterprises

Shanghai Yuben Intelligent Technology Co., Ltd.

I. Basic Enterprise Information

Shanghai Yuben Intelligent Technology Co., Ltd., founded in September 2020, is headquartered in Model Community, Zhangjiang AI Innovation Town, Pudong, Shanghai, with branches in South China and Huzhou. It is recognized as a national high-tech enterprise, a national technology-based SME, and a Shanghai innovative SME. The company owns more than 60 intellectual property assets, including invention patents, utility model patents, design patents and software copyrights. Its core team has more than ten years of experience in the development, application and implementation of computer vision, cloud computing and edge computing. Research and development (R&D) personnel account for 76 percent of its workforce, and employees with master's degrees or above account for 65 percent. Since 2023, the company's annual revenue growth has continued to exceed tenfold. With the mission of "making factories smarter", Yuben Intelligent has developed two product series: "YobiAI Industrial Safety Eye", which focuses on personnel safety protection in factories, and "YobiAI Industrial Inspection Eye", which focuses on finished-product quality inspection. Yuben Intelligent ranks among the leading providers by market share among top domestic and overseas enterprise customers. Its products have been sold to benchmark factories including CATL, Huawei Southern Factory, Colgate, Goertek, Orion, Emerson, GlaxoSmithKline, TCL, TGOOD, Karrie Group, Kostal, Yaskawa, Mahle and Baoti, and it continues to receive intended orders from Fortune Global 500 enterprises and other customers.

I. Background and Needs: Detection Bottlenecks in the Pharmaceutical Industry and the Limitations of Manual Inspection

Against the backdrop of increasingly strict drug-quality regulation and growing public health awareness, drug safety is not only the lifeline of enterprises, but also a social responsibility closely related to national welfare and people's livelihood. At present, defect detection in the pharmaceutical industry is still dominated by traditional manual inspection or statistical sampling. These methods are highly subjective, random and inefficient, and have become key weaknesses and potential risk sources that restrict the modernization of pharmaceutical quality-control systems.

In October 2025, at the third-quarter drug quality and safety risk consultation meeting, the Shanghai Medical Products Administration further emphasized the importance of full-life-cycle quality management for pharmaceuticals and required close attention to the implementation progress of the 2025 edition of the Chinese Pharmacopoeia and other relevant new regulations. The new regulations clearly require solid preparations such as oral tablets and capsules to strictly meet content-uniformity standards and prohibit non-conforming products from entering the market. Any slight deviation may lead to substandard active-ingredient content. In mild cases, this weakens therapeutic effects; in severe cases, it threatens patients' lives.

(I) Major Challenges Facing Pharmaceutical Enterprises

1. Low Inspection Efficiency

Most pharmaceutical enterprises still rely on traditional manual visual inspection. Four workers usually operate beside one tablet-inspection machine at the same time. In a high-temperature and noisy production environment, workers must wear protective equipment for long periods while staring closely at tablets for inspection. This easily causes visual, physical and mental fatigue, leading to a significant decline in inspection efficiency as working time increases.

2. Insufficient Accuracy

On high-speed production lines, manual inspection is easily affected by attention fluctuations and subjective judgment, causing frequent missed detections and false detections. In particular, broken pieces and fragments close to the size of intact tablets, such as pieces around four-fifths of a full tablet, are extremely difficult for the human eye to identify, making stable and consistent quality control hard to achieve.

3. Interference from Complex Environments

Factory lighting conditions are complex, and the white belt conveyor on the tablet-inspection machine is prone to reflection in the workshop, further increasing the difficulty of manual judgment. Even when ordinary camera systems are introduced to assist inspection, complex backgrounds and changing light conditions still interfere with imaging quality and affect detection accuracy and reliability.

Figure 1 Limitations and challenges of traditional inspection methods

II. Implementation Plan: Innovative Practice of Machine Vision and Edge AI

Machine vision covers not only traditional image-processing technologies, but also AI-based visual analysis enabled by deep learning. Edge AI moves computing power down to edge devices, such as controllers deployed inside factories, and offers advantages including low latency, low dependence on bandwidth and high data security. It is especially suitable for industrial scenarios that require strong real-time performance and involve sensitive data.

As an AI enterprise centered on independent algorithm and software R&D, Yuben Intelligent deeply integrates machine vision with edge AI technologies in response to the pain points and needs of the pharmaceutical inspection sector. By fully leveraging deep learning in pattern recognition and decision-making, the company developed “intelligent detection of broken tablets and tablet fragments on belt conveyors” within its YobiAI Industrial Inspection Eye series, achieving both high precision and high efficiency in tablet-defect detection.

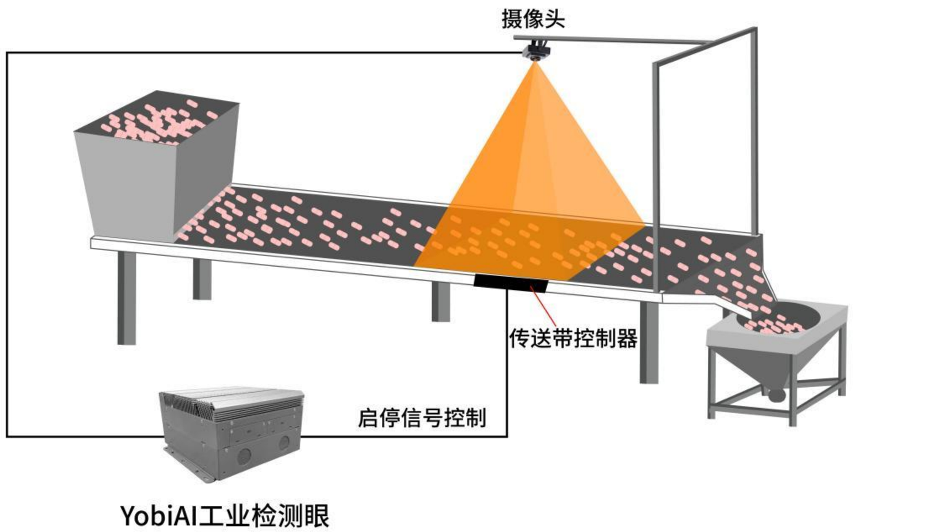


Figure 2 Application scenario of intelligent detection for broken tablets and tablet fragments on belt conveyors

(I) System Composition

As a split-type, high-precision AI inspection product, the intelligent detection system for broken tablets and tablet fragments on belt conveyors supports the synchronous operation of multiple cameras, requires no additional light source and delivers excellent inspection performance. The system is delivered to customers as a kit. Its core components include Yoben Intelligent's self-developed YobiAI Industrial Inspection Eye with independent intellectual property rights, integrating algorithms, software and hardware, as well as industrial cameras and industrial relays. It can also be connected to auxiliary devices such as all-in-one monitors and audible and visual alarms.

(II) System Implementation Plan

1. High-Speed and High-Frame-Rate Visual Inspection

The system is equipped with a 120 frames per second (FPS) high-frame-rate industrial camera and achieves a tablet inspection speed of 5,000 pieces per minute (pcs/min). It can easily match the pace of high-speed production and ensure that every tablet is captured clearly and without omission, thereby providing high-quality raw data for subsequent analysis.

2. Multi-Frame Voting Mechanism

Traditional visual inspection usually uses an industrial camera system that takes one image and produces one result; whether the result is correct depends entirely on the accuracy of that single recognition. Yuben Intelligent breaks through the "one-shot decision" limitation of traditional single-frame recognition by innovatively introducing a multi-frame voting mechanism based on video streams. For example, for a single tablet, the system continuously captures multiple frames and votes on the collective result. Only when six out of ten frames determine the target to be a fragment will the result be output as not good (NG). This effectively resists single-time misjudgment, systematically improves recognition accuracy by 5 percent to 10 percent, and greatly reduces wrongful judgments.

3. Real-Time Detection and Conveyor Start-Stop Control

The system can complete the entire process from image acquisition and AI defect judgment to control-signal output within 20 milliseconds, instantly linking with the conveyor for start-stop control. In the real-time display image, it marks the positions of non-conforming tablets to guide quality inspectors in timely tablet removal. It can also work with non-standard components or other linked equipment to automatically reject NG tablets.

4. Standardized and Flexible Adaptation

To meet the trend toward multi-variety and small-batch production in pharmaceutical enterprises, Yuben Intelligent's system has a built-in intelligent recipe-management program. It can preset and switch with one click among hundreds of inspection parameters for tablets of different colors and shapes, realizing "one-time configuration and long-term reuse”. This significantly reduces the time and cost losses caused by production-line changeovers, empowers flexible manufacturing and meets diverse needs.

5. Fully Functional Software Platform

The system is equipped with a fully functional web-based software platform. It requires no professional training, lowers the threshold for use by pharmaceutical workers, and provides a simple and intuitive whole-process operation experience. It supports multi-screen view modes for seamless monitoring of the full-line inspection status and instant abnormality alarms. It also provides multi-level user-permission management to ensure system security, and offers backup and restore functions to prevent parameter loss caused by misoperation. In addition, the system has fault self-diagnosis capability, ensuring continuous and stable 24/7 operation and protecting production rhythm.

Figure 3 Partial types of non-conforming tablet defects detectable by the system

III. Project Highlights and Core Technologies for Building Detection Barriers

(I) Parallel-Light Detection Technology

The system uses a non-attenuating parallel light source generated by laser or light-emitting diode (LED) arrays, fundamentally eliminating light scattering and reflection interference caused by different tablet positions and uneven tablet surfaces. As a result, the captured tablet images have clear contours and stable dimensions, providing high-quality “raw data” for subsequent precise analysis.

(II) Sub-Pixel Processing Technology

Because tablet surfaces have curvature and reflective characteristics, acquired images may have uneven grayscale distribution. Yuben Intelligent therefore uses sub-pixel processing technology to identify subtle edges on the tablet surface. Sub-pixel processing improves the accuracy of feature localization by analyzing the distribution changes of grayscale values around edge pixels. It divides one physical pixel into 1/N, where N is usually 5, 10 or 20, before edge detection. This means each pixel is subdivided into smaller units, with support for up to 1/100 segmentation precision for detailed processing, improving measurement accuracy by dozens of times.

This means that on a mainstream 12-megapixel industrial camera, the system equivalently creates a digital ultra-high-definition discrete dot matrix at the level of hundreds of millions of points, around 40,000 by 30,000. It achieves 0.02 mm-level measurement precision and can capture micro-defects invisible to the human eye and even to conventional machine vision, thereby determining more accurately whether a tablet is broken or fragmented.

(III) Automatic Dimensional Measurement Technology

Based on Yuben Intelligent's self-developed algorithm using deep convolutional neural networks, the system can autonomously learn and extract deep features of various tablets, completing classification, screening and judgment within seconds and improving efficiency by dozens of times.

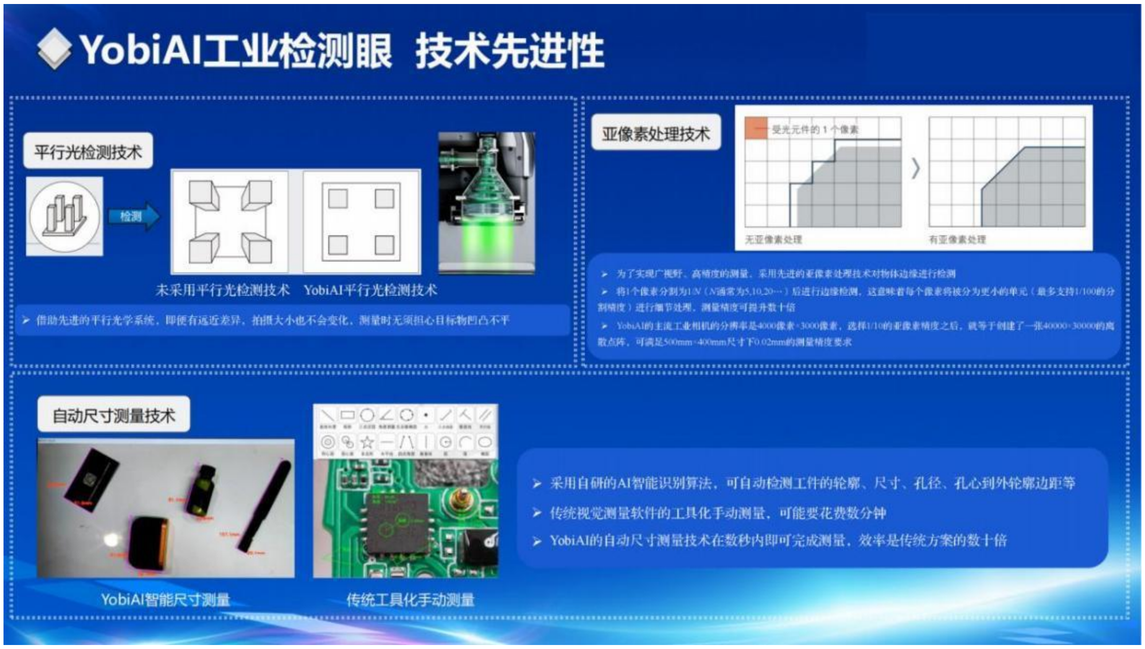


Figure 4 Technological advancement

IV. Implementation Results

At present, the intelligent detection product for broken tablets and tablet fragments on belt conveyors has been successfully deployed in the factory of Yuben Intelligent's partner Haleon (China) Co., Ltd., and has been seamlessly connected to the existing production line. Yuben Intelligent also provided system operation and maintenance training for the pharmaceutical company's operators and technical team, successfully meeting the company's two core demands for intelligent inspection equipment: replacing manual inspection and significantly improving inspection efficiency.

(I) Automated Inspection Replacing Manual Inspection

With high precision and high efficiency, the YobiAI intelligent detection system for broken tablets and tablet fragments on belt conveyors has successfully replaced traditional manual inspection. Each production line previously required at least four quality inspectors; after deployment, only one quality inspector is needed for tablet removal, system monitoring and maintenance. Labor costs have been reduced by 75 percent, and problems caused by human factors, such as hygiene risks, inconsistent standards and efficiency fluctuations, have been effectively eliminated.

(II) Significant Improvement in Inspection Efficiency and Accuracy

The system supports a conveyor-line speed of 5,000 pcs/min, with accuracy of at least 99.9 percent. For every 100,000 tablets, the number of false detections is no more than five pieces, and missed detections no more than two pieces. Meanwhile, the system automatically stores up to 100,000 abnormal records, including images, and supports multi-dimensional filtering and statistical analysis. This provides a solid data foundation for production-process optimization and quality traceability, ensuring transparency and traceability throughout the production process.

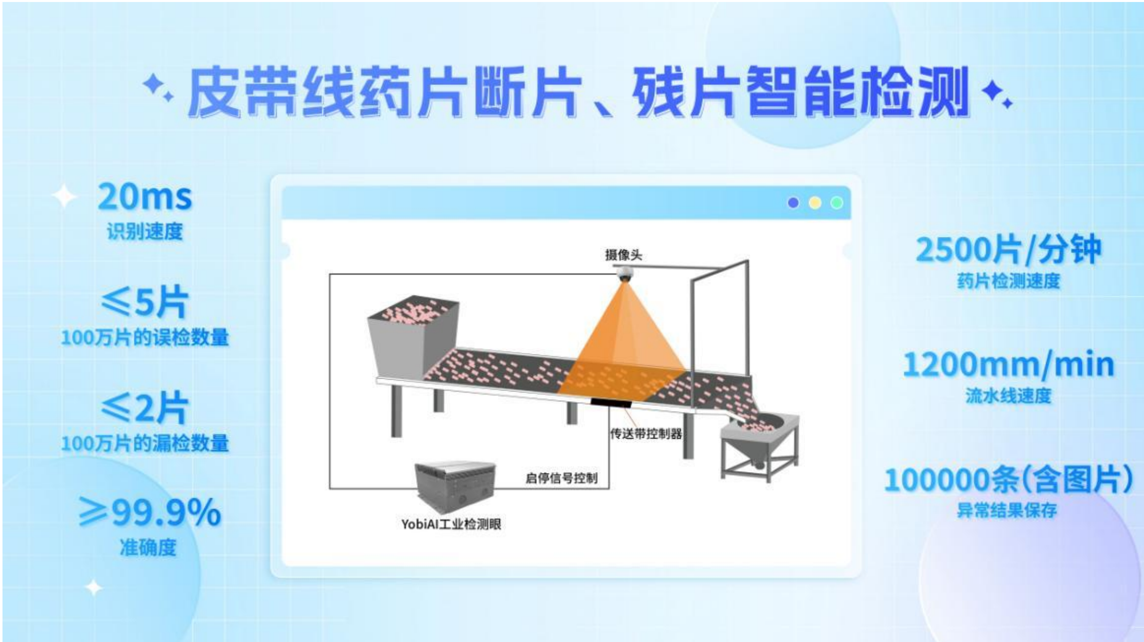


Figure 5 System data chart

The successful application of the YobiAI intelligent detection system for broken tablets and tablet fragments on belt conveyors at Haleon, together with the continuous accumulation of operational data and algorithm optimization, verifies its robustness and adaptability in complex industrial environments. Recognition and repeat purchases by Haleon, an industry benchmark enterprise, have further stimulated procurement interest from many pharmaceutical enterprises, providing a replicable example for the digital and intelligent transformation of inspection processes in the industry.

With its efficient inspection capability, precise defect recognition and strict compliance with Good Manufacturing Practice (GMP), the system is gradually becoming a preferred solution for intelligent inspection in the pharmaceutical industry. Yuben Intelligent has also demonstrated the great potential of edge AI vision controllers in the pharmaceutical industry, and is confident that the concept and technology of YobiAI Industrial Inspection Eye are equally applicable to manufacturing fields such as food, electronics and automobiles. By making factories smarter, it is becoming a key force driving intelligent transformation in manufacturing.

Case 22

Application of AI in Manufacturing Quality Inspection

Shanghai Youyun Information Technology Co., Ltd.

I. Basic Enterprise Information

Shanghai Youyun Information Technology Co., Ltd. has long been deeply engaged in industrial vision and edge intelligence. Centered on "AI recognition without training", the company has developed quality-inspection solutions for both discrete and process manufacturing.

The system has a built-in cross-industry visual knowledge base and multi-task embedding model, with zero-shot recognition and adaptive threshold updating capabilities. It can be used out of the box without sample collection or manual annotation. During operation, it automatically performs steady-state correction through feature-distribution drift detection to ensure long-term accuracy and robustness. The solution supports multimodal feature-level fusion of visible light and infrared imaging, millisecond-level inference on an edge AI-BOX, and connection to programmable logic controllers (PLCs) through Open Platform Communications Unified Architecture (OPC UA) for a closed loop of rejection and sorting, meeting the real-time and stability requirements of high-throughput production lines. The multimodal and OPC UA capabilities come from existing implementation practices. It also provides batch-level full-chain image and event traces, supporting root cause analysis (RCA), statistical process control (SPC) and compliance audits. It has been replicated and applied in industries such as chemicals, new energy, automotive parts and 3C. The company positions itself as a provider of plug-and-play solutions that remove quality-inspection bottlenecks, accumulate "digital quality" assets, and help customers achieve scalable yield improvement without modifying processes or increasing training costs.

II. Case Introduction

(I) Background and Needs

In high-end manufacturing fields such as key materials for new energy vehicles, production-line automation has increased capacity to thousands of pieces per hour. If quality inspection still relies on manual visual inspection, enterprises inevitably fall into a triple dilemma of bottlenecks, blind spots and broken links. The core contradiction lies in the sharp conflict between the efficiency ceiling, vague standards and non-inheritability of experience in traditional manual inspection, and high-end manufacturing's rigid demand for efficiency, precision and quality stability. Facing this dilemma, enterprises hope to build a truly AI-based quality-inspection system that requires no manual training - neither sample annotation nor algorithm experts - and can independently identify defects and automatically learn process changes.

(II) Implementation Plan

1. Design Concept: Enabling AI to Operate Without Human Dependence

Youyun Information proposed a solution combining "zero-shot recognition + feature-adaptive evolution". Its core objective is to enable AI to be plug-and-play without samples, training or parameter tuning. The system has a built-in cross-industry general defect knowledge base and semantic representation model. Through "semantic-feature" mapping, it can directly identify unknown defects and feed back judgment results in real time, achieving true zero-shot inspection.

2. Technical Architecture

The platform adopts a three-layer edge-intelligence architecture.

Front-end perception layer: A multimodal acquisition system composed of industrial cameras, line light sources and infrared thermal imagers serves as the "compound eyes" of AI, stably capturing texture and temperature differences under different lighting and motion conditions.

Edge computing layer: The AI-BOX is deployed to run a training-free feature-matching engine. Based on preset feature embeddings and anomalous-feature clustering, the algorithm performs defect localization and discrimination, with a response time of less than 20 milliseconds.

Back-end management layer: A web console and quality cockpit are provided for real-time monitoring of inspection speed, defect distribution and system-drift status, enabling remote visualized operation and maintenance.

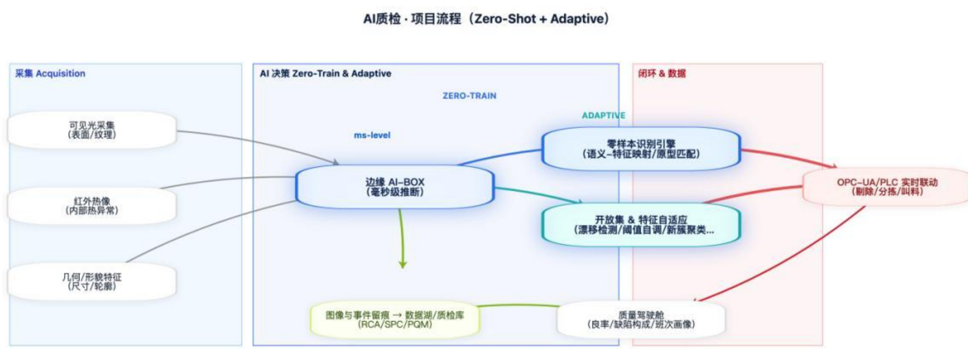


Figure 1 Overall process of the intelligent quality-inspection system

3. Zero-Training Online Deployment Process

When the system goes online, no historical samples need to be imported. A baseline model can be generated through one-time automatic calibration. Afterwards, AI automatically discovers abnormal areas and outputs classified results through the feature-distribution drift detection mechanism. If an unknown defect type appears, the system automatically clusters it and generates a “new concept label”, which can be put into use without manual annotation or retraining.

For example, on a rubber-block production line, AI captured a new defect called “material X bubble” on the third day. The system identified it and incorporated it into the inspection logic only by comparing feature differences, with zero manual intervention and zero model retraining throughout the process.

4. Platform Capability Expansion

The system supports seamless connection with manufacturing execution systems (MES), PLCs and quality databases. Inspection results can reversely drive process-parameter adjustment, enabling closed-loop optimization. Through traceable data flows, enterprises can carry out RCA, SPC and predictive maintenance, moving quality management from “discovering problems” to “preventing problems”.

III. Project Highlights

(I) Zero-Shot Recognition: Enabling AI to “See at Birth”

The most breakthrough technology of this system is “Zero-Shot Visual Inspection”. Traditional machine-vision systems must collect samples, annotate defects and train models manually before they can operate, while Youyun Information’s visual engine completely skips this step.

Based on a multimodal semantic embedding model, the system integrates general texture, contour, shape and thermal-distribution features from dozens of industrial sub-industries. By establishing a semantic-feature mapping matrix, AI can understand abstract concepts such as “crack”, “bubble” and “contamination” in a human-like manner, and find the closest feature clusters in image space, thereby recognizing unknown defects without training.

This means enterprises no longer need to wait for hundreds of samples to “feed” the model, nor do they need to depend on algorithm experts. The system can run stably immediately after startup, reaching over 98 percent inspection accuracy at first deployment and truly achieving “zero-threshold AI onboarding”.

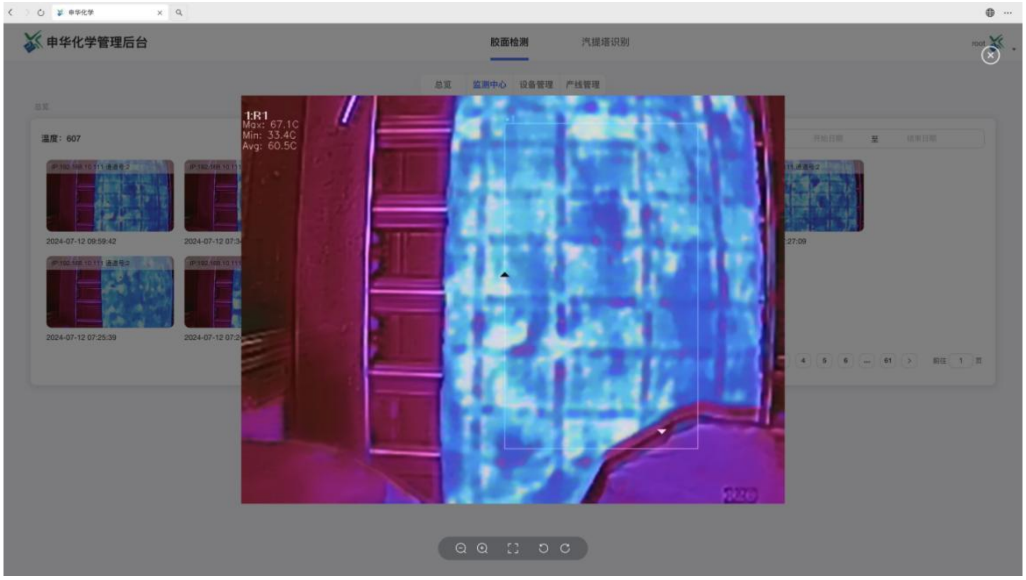


Figure 2 Infrared temperature recognition for rubber blocks

(II) Feature-Adaptive Evolution: AI Self-Repair Capability

Different from the static weight structure of traditional models, the AI engine in this project adopts an adaptive feature evolution mechanism.

When production-line environments change, such as lighting shifts, slight raw-material variations or equipment vibration, causing image-distribution drift, the system automatically performs distribution drift detection. It dynamically updates decision thresholds through the Kolmogorov-Smirnov (K-S) test and energy-distance calculation. When a completely new feature cluster is detected, open-set recognition is triggered, and the cluster is incorporated into a new category with updated prototype vectors without model retraining.

This mechanism gives AI “self-healing and evolution capabilities”: it does not rely on human intervention and will not become “blind” due to changes in working conditions. It can continuously maintain accuracy and recall above 99 percent.

(III) Multimodal Feature Fusion: From “Seeing the Surface” to “Understanding the Essence”

Youyun Information innovatively introduces a three-modal fusion inspection architecture into industrial quality inspection, combining visible light, infrared thermal imaging and geometric texture.

1. Visible-light channel: used to identify surface defects such as scratches, contamination and bubbles.

2. Infrared channel: captures thermal-distribution abnormalities to identify potential internal issues such as material sintering and uneven curing.

3. Geometric-texture channel: extracts geometric morphology features to achieve stable recognition of deep texture.

These types of information are mutually constrained and confidence-weighted in the feature-level fusion layer, ensuring cross-modal decision consistency. The final inspection results can not only identify whether a surface has a problem, but also infer where the problem originates, providing diagnostic evidence for process optimization. This multimodal fusion approach extends system inspection capability from two-dimensional surfaces to three-dimensional material layers, realizing a transition “from inspection to cognition”.

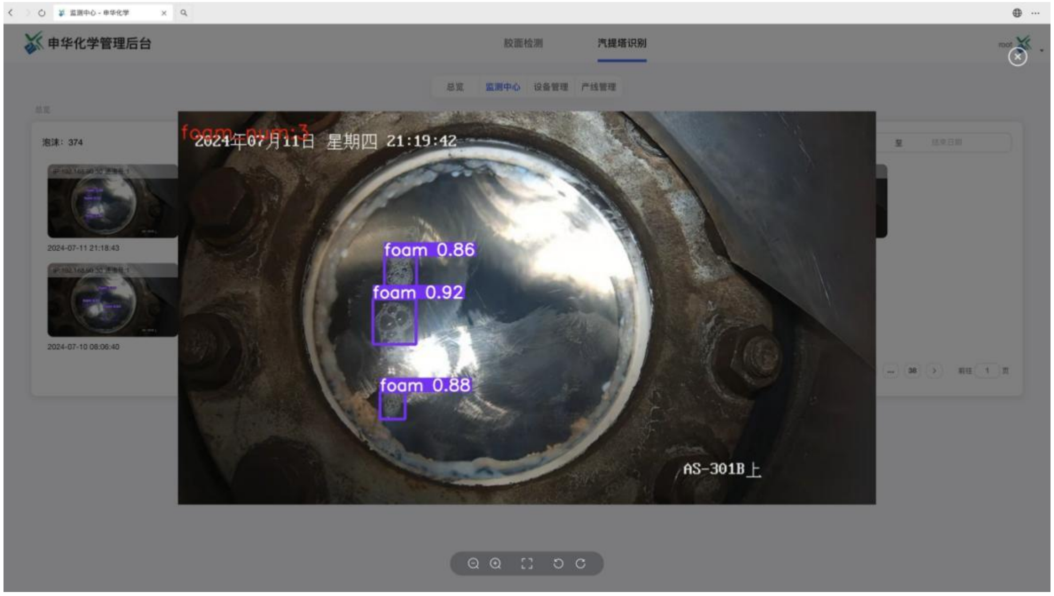


Figure 3 Recognition example in a bubble-detection scenario

(IV) Edge-Intelligence Closed Loop: Millisecond-Level Response and Industrial-Grade Stability

The system architecture adopts an "edge computing + cloud collaboration" model. The edge-side AI-BOX integrates a high-performance neural processing unit (NPU), completing image inference within 20 milliseconds and realizing hard real-time signal interaction with PLCs through the OPC UA standard interface to drive synchronized action of rejection mechanisms. The cloud side is responsible for model version management and remote monitoring, ensuring algorithm stability throughout the full life cycle.

This architecture not only solves latency and bandwidth bottlenecks on industrial sites, but also ensures long-term stable operation under 24/7 high-frequency inspection conditions. Even when the network is disconnected, the edge side still has full autonomous capability, truly realizing "cloud disconnection without AI downtime".

(V) From AI Algorithms to Digital Quality: Systematic Innovation

The innovation of this project is not limited to algorithms. It is also reflected in the reconstruction of a "Digital Quality System". With "data as knowledge" at its core, every inspection, abnormality and process change is transformed into structured data assets. Through the "quality cockpit", production managers can view yield, defect distribution and team performance in real time. Combined with MES and ERP systems, the solution forms closed-loop decision-making, shifting quality management from post-event correction to real-time prevention. This design makes AI not only a recognition tool, but also the "intelligent hub" of enterprise digital transformation.

IV. Implementation Results

(I) Typical Case Application: A Large Synthetic Rubber New-Materials Enterprise

This solution has been successfully implemented at a well-known synthetic rubber manufacturer in East China with annual production capacity exceeding 170,000 tons. The company has long supplied products to leading international tire brands and maintains extremely strict inspection standards for product appearance and internal impurities.

In this project, Youyun Information deployed an AI-based intelligent quality-inspection system for the enterprise's new factory production line. The system covered eight core inspection points in key rubber post-processing procedures, focusing on solving the pain points of traditional manual visual inspection on high-speed production lines: fatigue, inconsistent standards, and difficulty distinguishing small bubbles from impurities.

(II) Specific Quantified Results

By introducing zero-shot recognition and multimodal fusion technologies, the project achieved significant practical results for the enterprise. The specific data comparison is as follows.

1. Human-Resource Optimization and Cost Reduction

Workforce optimization: Before system deployment, the eight key quality-inspection points on the production line required three-shift manual rotation. After the system went online, it successfully replaced high-intensity repetitive human visual judgment and directly optimized around ten quality-inspection positions. These employees were redeployed to higher-value roles such as process optimization and equipment maintenance. This not only eased recruitment difficulties, but also saved the enterprise more than RMB 1 million per year in direct labor and related management costs according to estimates.

Return on investment (ROI) cycle: Based on hardware investment and labor savings, the overall ROI cycle of the project was shortened to within eight months.

2. Dual Improvement in Quality and Efficiency

Missed-detection rate and false-judgment rate: When handling complex defects such as surface bubbles and foreign-color impurities on rubber blocks, the AI system reduced the missed-detection rate from 2.5 percent under manual inspection to below 0.1 percent. Meanwhile, thanks to deep learning's intelligent filtering of "pseudo-defects", such as water-stain interference and surface gloss, the false-judgment rate dropped significantly, effectively avoiding waste of qualified products.

Capacity matching: The system's single-point inspection response speed is less than 20 milliseconds, perfectly matching the enterprise's high-speed production rhythm for an annual capacity of 170,000 tons. It completely eliminated the previous production-line bottleneck caused by the limited speed of manual visual inspection and supported 100 percent full-load operation of the production line.

3. Customer Satisfaction and Compliance

Zero customer complaints: Since the system went online, products delivered to end customers such as leading international tire manufacturers have achieved zero appearance-quality complaints, greatly consolidating the enterprise's A-level supplier position in the international supply chain.

Data assetization: Compared with past paper-based or fragmented records, every outgoing rubber block now has high-definition image and inspection-data traces, achieving 100 percent full-chain quality traceability. This fully meets the strict compliance-audit requirements of ISO 9001 and downstream key customers in high-end manufacturing.

Case 23

An Intelligent Predictive Maintenance Platform for Induced Draft Fans in Coal-Fired Power Plants Based on AI and Multi-Source Data Fusion

Shanghai Donghao Testing Technology Co., Ltd.

I. Basic Enterprise Information

Founded in 2002, Shanghai Donghao Testing Technology Co., Ltd. focuses on AI-based equipment fault early warning and intelligent operations and maintenance (O&M) systems, and provides customer-oriented AI + simulation digital twin platforms. Its parent company, DongHua Testing Technology Co., Ltd., founded in 1993, is a national high-tech enterprise and a leading company in structural mechanics performance research and intelligent testing and analysis systems. It was listed on the A-share market in September 2012 with the stock code 300354.

II. Background and Needs

As the proportion of power generated from new energy sources continues to increase, thermal power units are gradually shifting toward flexible peak-shaving operation. Frequent start-ups, shutdowns and load fluctuations have significantly increased the risk of unplanned equipment outages. As a key auxiliary machine of large boilers, the health status of the induced draft fan directly affects unit efficiency and safety. Once a fault occurs, the unit may be forced to shut down, causing economic losses ranging from hundreds of thousands to more than one million yuan.

Figure 1 Physical induced draft fan monitored by the project

Under the existing O&M model, the operation and maintenance of induced draft fans mainly face the following problems:

(I) Insufficient Equipment Fault Early-Warning Capability

Induced draft fans operate for long periods under harsh conditions such as high temperature and high dust concentration. Their faults are sudden and concealed. Traditional management depends on regular manual inspections and operating-parameter threshold alarms, making it difficult to identify early abnormal signs in time. Inspection personnel must work close to the equipment and face risks such as high temperature, noise and high-voltage electricity. Inspections are also limited by shifts and labor availability, usually only once or twice per shift, and cannot cover all-weather changes in equipment status. Manual judgment of parameters such as vibration, noise and temperature is highly subjective and has a high rate of false alarms and missed alarms, making early fault warning and accurate fault localization difficult.

(II) Prominent Data Silos and Insufficient Value Mining

Although power plants have collected large amounts of data such as vibration, temperature, pressure and current through distributed control systems (DCSs), supervisory information systems (SISs) and other systems, such data are scattered across different systems and lack unified fusion and correlation analysis. Vibration signals mostly remain at the level of simple threshold alarms, without sufficient spectrum analysis, trend analysis or multi-parameter coupling analysis, forming obvious information silos. As a result, it is difficult to build cross-parameter and cross-system fault reasoning chains, accurately determine fault types, locations and evolution trends, or fully release the value of data.

(III) Lack of Scientific and Forward-Looking Maintenance Decisions

The current maintenance strategy still relies heavily on empirical judgment and lacks reliable data support. The prevailing modes of time-based maintenance (TBM) and breakdown maintenance (BM) ignore the true health status of equipment. Decisions are mostly based on service age and operating hours rather than health indicators. This can lead either to over-maintenance, such as overhauling healthy equipment in advance and wasting spare parts and labor, or to under-maintenance, where hidden risks are not identified and minor faults evolve into unplanned shutdowns, thereby affecting life-cycle economic performance.

Against the broader background of smart power plant construction, it is necessary to introduce AI and prognostics and health management (PHM) technologies to solve the above pain points and build an intelligent predictive maintenance system for key equipment such as induced draft fans. By constructing an intelligent predictive maintenance platform based on AI and multi-source data fusion, the project aims to achieve the following objectives:

Adaptive alarms and dynamic threshold setting are introduced together with a scenario self-learning mechanism to reduce repeated alarms, false alarms and missed alarms, and to improve alarm accuracy and reliability.

Early fault features of key components such as bearings, rotors and impellers are mined to identify abnormal trends and accurately locate fault types and fault positions.

Heterogeneous data such as vibration, temperature, pressure and current are integrated to build a multi-source fusion diagnosis model, improving diagnostic accuracy and robustness through methods such as deep learning.

Historical maintenance data and expert experience are accumulated to build standardized feature libraries and case knowledge bases, providing a unified methodology and technical foundation for replication and promotion across more units and equipment types.

Through the above construction, the health management level of induced draft fans can be significantly improved, supporting the safe, stable and economical operation of units and providing strong support for thermal power units to undertake flexible peak-shaving and system-supporting roles in the new power system.

III. Implementation Plan

The platform developed in this project covers the whole process of operation and maintenance for boiler induced draft fans in coal-fired power plants. During operation, it reduces the risk of unplanned shutdowns through real-time monitoring and early warning for key components such as bearings, rotors and impellers. During maintenance, it provides maintenance strategies based on AI diagnostic results, optimizes maintenance plans, extends equipment life and reduces maintenance costs. At the management level, it combines knowledge bases and digital twins to provide data and model support for smart O&M decision-making, thereby improving overall management efficiency.

System construction centers on four layers: the sensing layer, the edge computing layer, the AI diagnosis layer, and the platform and business application layer. Together, they realize online monitoring, intelligent diagnosis and O&M decision support for key equipment such as induced draft fans. Implementation includes sensor deployment and data acquisition, edge computing and data processing, AI model construction and training, platform development and integration, and system commissioning and online operation.

(I) Sensor Deployment and Data Acquisition

Integrated temperature-vibration sensors and shock pulse sensors are installed at the drive-end and non-drive-end bearing positions of the motor and at the fan bearing housing to monitor three-axis vibration, temperature and bearing shock pulse characteristics. These are used to identify early faults such as bearing wear, rotor imbalance or looseness, and impeller imbalance. Eddy-current displacement sensors are deployed at the rotors on both sides of the fan bearing housing to monitor rotor vibration in the X and Y directions, enabling online prediction of rotor-related faults.

The selected sensors combine high precision with adaptability to high-temperature, high-dust and other working conditions, meeting the requirements for long-term stable operation and providing high-quality raw data for subsequent edge computing and AI diagnosis.

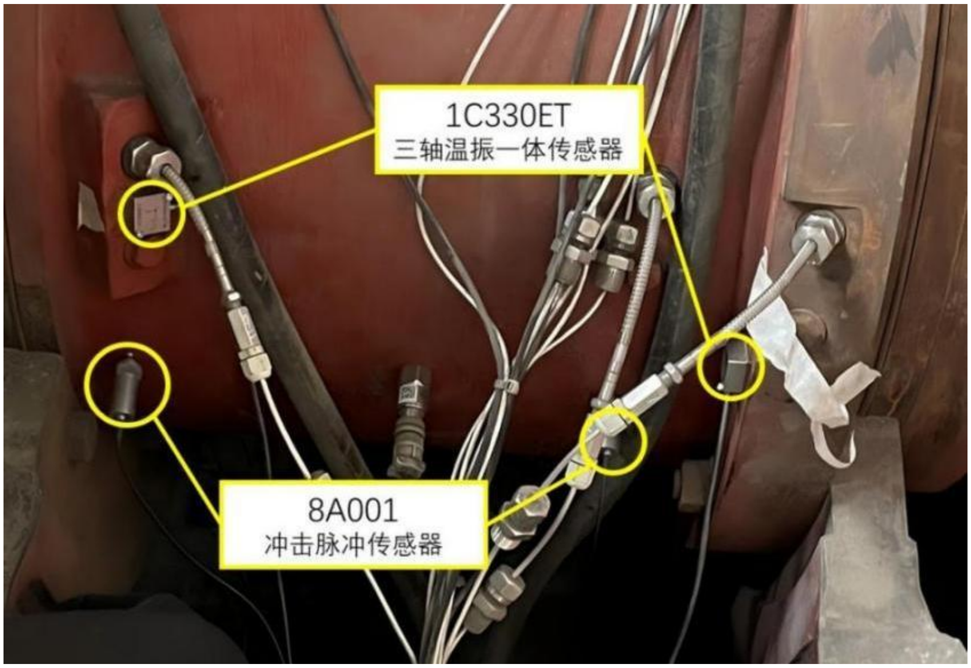


Figure 2 Sensors deployed in the project

(II) Edge Computing and Data Processing

The data acquisition layer integrates high-performance embedded digital signal processor (DSP) chips to build edge intelligent nodes. These nodes process and extract features from signals such as vibration, temperature and shock pulse locally in real time. Sampling rates are adaptively adjusted according to different fault diagnosis requirements, such as high-frequency sampling for shock pulse analysis and low-frequency sampling for temperature and vibration trend tracking.

On the edge side, the system completes time-domain features such as peak value, root mean square and kurtosis, frequency-domain features such as fast Fourier transform (FFT) spectra and order analysis, envelope demodulation, and shock pulse method (SPM) evaluation. More than 20 key fault features are extracted. This provides a low-latency and highly reliable data foundation for real-time diagnosis and warning, while significantly reducing the amount of raw data transmitted back, improving system response speed and reducing the load on the upper-level platform.

(III) AI Model Construction and Training

The system integrates operating data, historical maintenance records and expert experience to establish a standardized feature database covering multiple typical faults, including rotor imbalance, misalignment, bearing damage, rotor-stator rubbing and fluid-induced vibration.

Deep learning algorithms such as one-dimensional convolutional neural networks (1D-CNNs) and autoencoders are adopted to build a multi-label diagnosis model, enabling simultaneous identification and coupling analysis of multiple faults. Data augmentation and transfer learning are combined to improve the model's generalization capability across different units and operating conditions.

Through preprocessing, feature selection and model optimization validation, the model's diagnostic accuracy can stably reach a high level of more than 95 percent. It also supports an online self-learning mechanism, continuously updating parameters with real-time data to adapt to long-term operation and changes in working conditions.

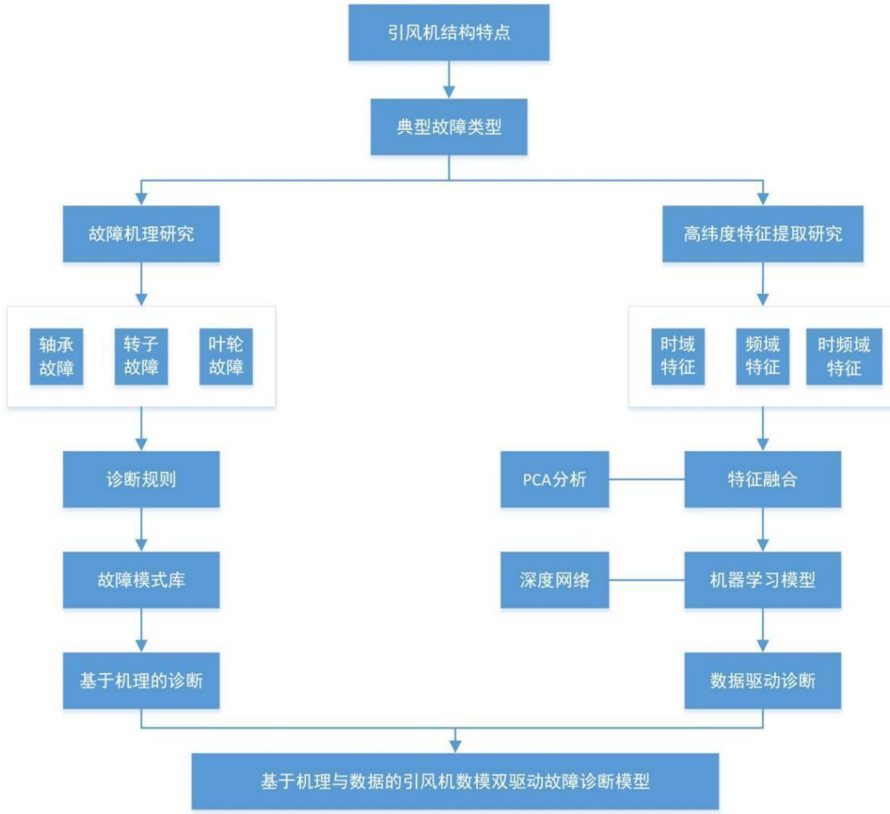


Figure 3 Mechanism- and data-driven digital-analog dual-drive fault diagnosis model for induced draft fans

(IV) Platform Development and Integration

The platform adopts a browser/server (B/S) architecture, featuring high configurability and multi-source data fusion capabilities. It supports cross-platform access from personal computers (PCs) and mobile terminals. Through standard interfaces such as WebAPI, Modbus TCP and OPC Unified Architecture (OPC UA), it can be flexibly integrated with existing DCS, SIS, management information system (MIS), manufacturing execution system (MES) and other systems in power plants, breaking down data silos and forming a unified equipment status monitoring and O&M decision-making platform.

Core functions include a visual dashboard with three-dimensional (3D) models, maps and videos; equipment health status and alarm statistics; trends, spectra and shaft centerline trajectories; and an alarm calendar. They also include equipment and measuring-point management, which supports two-dimensional (2D) and 3D model import and point deployment, a maximum sampling rate of about 1 MHz and million-level data points, with real-time features in the time domain, frequency domain, order domain and envelope domain extracted and analyzed by DSP chips. The platform also provides a knowledge base and fault model library, covering full life-cycle data retrieval, feature-maintenance solution association and automatic disposal recommendation after alarms; intelligent alarms and status monitoring, such as over-limit alarms, rate-of-change alarms and trend alarms with multi-channel push notifications; and data query and export, supporting .txt, .csv, .xls and .cmadat formats. The platform adopts an independently controllable software architecture and feeds features and alarm information back to the digital twin system through WebAPI, while balancing data security, system stability and future expansion compatibility.

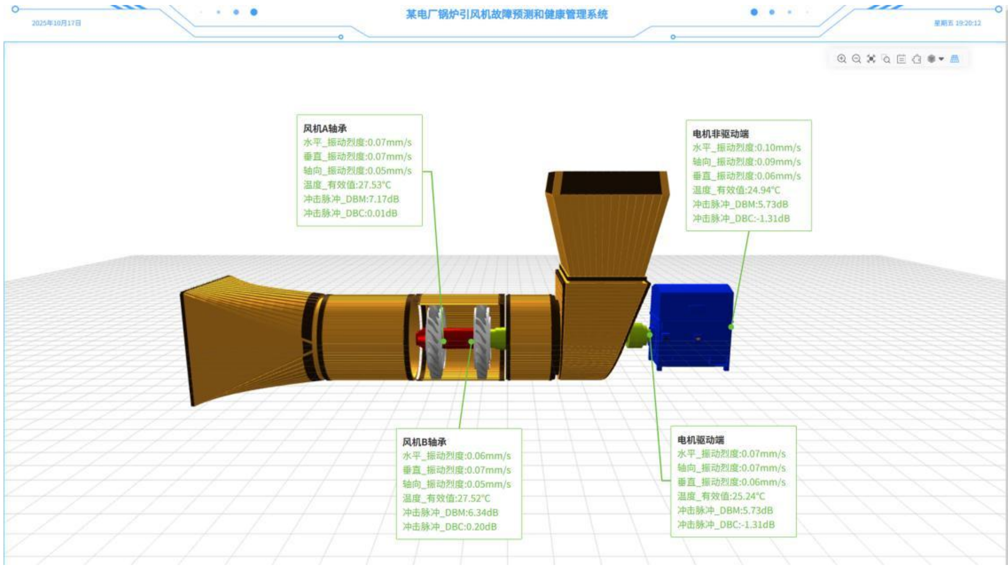


Figure 4 Status visualization dashboard

(V) System Commissioning and Online Operation

After sensor deployment and system integration are completed, on-site surveys and parameter tuning are carried out. Joint testing is organized under typical operating conditions to verify the entire process, including the acquisition link, edge computing, AI model and platform display. Through staged trial operation and iterative model optimization, the system gradually forms a stable and reliable operating state, realizing continuous online monitoring and intelligent warning for induced draft fans and related equipment, and providing a mature engineering solution for replication and promotion in different units and power plant scenarios.

IV. Project Highlights

(I) Key Technological Breakthroughs

1. Hybrid architecture combining edge computing and server-side deep learning

The project innovatively builds a hybrid architecture of edge feature extraction and server-side deep learning. The edge side uses high-performance DSPs for real-time processing and multi-rate adaptive feature extraction, reducing transmission burden and latency. The server side undertakes complex model training and inference, achieving both low latency and high precision. Overall real-time performance and reliability are significantly improved.

2. Fault diagnosis model combining multi-source information fusion and 1D-CNN

By integrating multi-source heterogeneous data such as vibration, temperature, pressure and current, the project builds a multi-label model based on 1D-CNNs, enabling accurate classification and early warning of multiple types of faults. The model captures spatiotemporal and coupling features, significantly improving diagnostic accuracy and robustness. The early fault recognition rate is more than 30 percent higher than that of traditional methods.

3. Online self-learning and adaptive optimization capability

The system has online self-learning capabilities and uses real-time data to continuously optimize the model so that it can adapt to variable working conditions. Combined with digital twins for status mapping and fault simulation, it can dynamically adjust alarm thresholds and maintenance strategies, significantly improving the level of intelligence.

4. Integrated PHM platform and standardized API interfaces

The platform integrates monitoring, diagnosis and decision-making, providing comprehensive O&M support functions. Through standardized application programming interfaces (APIs), it seamlessly connects with existing power plant systems, eliminates data silos, realizes multi-system collaborative analysis and decision-making, and offers good scalability and compatibility, laying a foundation for industry-wide promotion.

(II) Scenario Specificity

1. High reliability requirements in harsh environments

Induced draft fans in coal-fired power plants operate in harsh environments, including high temperature, high dust concentration and high noise, which place extremely high requirements on sensor and system reliability. The project uses highly protected and strongly interference-resistant sensors, and the edge computing nodes are designed with both high-temperature resistance and stability in mind, ensuring long-term stable operation under extreme conditions.

2. Complexity of multi-source data fusion

Power plant data come from diverse sources such as DCSs, SISs and vibration sensors, with different formats and sampling frequencies. Through unified data standards and multi-protocol integration, the project realizes deep fusion and correlation analysis of multi-source data, establishes cross-parameter fault reasoning chains, and effectively solves the problem of data silos.

3. Industry demonstration value and replicability

The project not only serves induced draft fans, but its technical framework and solution can also be replicated for pumps, compressors, gearboxes and other key equipment, as well as in petrochemical, metallurgical, manufacturing and other industries. It provides a practical example for the intelligent transformation of traditional industries. The platform’s knowledge base and model library support rapid adaptation and promotion, significantly lowering the application threshold.

V. Implementation Results

(I) Economic Benefits

After being put into operation on the No. 1 induced draft fan of the Jiangyin thermal power plant in Jiangsu Province, the project achieved significant economic benefits in several dimensions.

The number of unplanned shutdowns was greatly reduced. The annual average number of unplanned shutdowns of the induced draft fan decreased from three to one, a reduction of more than 65 percent. The unplanned shutdown rate fell by about 30 percent, and the frequency of emergency maintenance was significantly reduced.

Power generation losses were effectively controlled. Based on unit output and on-grid electricity price estimates, the system can reduce annual electricity losses caused by fault-related shutdowns by about 500,000 yuan.

The maintenance cycle was extended scientifically. The average maintenance interval increased from six months to 7.5 months, extending the maintenance cycle by about 25 percent. Maintenance resources are allocated more rationally, and planning flexibility has improved.

O&M costs decreased significantly. Thanks to the lower fault rate and improved status warning capability, spare-parts turnover increased and redundant procurement was reduced. It is expected that, over the next five years, the system will save about 350,000 yuan per year on spare parts and labor costs.

Comprehensive unit benefits improved. Enhanced operating stability of the induced draft fan increased the overall availability of the unit, enabling it to generate higher returns in peak-shaving and power generation tasks, with an average annual revenue increase of about 800,000 yuan.

The total investment of the project was 233,800 yuan. Based on the above benefit estimates, the payback period is expected to be less than one year, indicating significant long-term economic benefits.

(II) Technological Innovations

The project realizes the collaboration of edge computing and server-side deep learning, alleviating the problems of large data volume and high transmission latency in industrial sites and significantly improving real-time performance. The 1D-CNN fault diagnosis model based on multi-source data fusion can simultaneously identify and analyze multiple coupled faults. In actual operation at the Jiangyin thermal power plant in Jiangsu Province, its accuracy exceeded 95 percent. The system has multi-level adaptive alarms and online self-learning capabilities, dynamically adjusting strategies and parameters according to equipment status and reducing false alarms and missed alarms. Combined with digital twins and a fault knowledge base, it supports virtual simulation, maintenance decision-making and cross-scenario replication, offering strong technical innovation value.

(III) Social Value

The project provides replicable intelligent O&M solutions for multiple thermal power enterprises, including the Jiangyin thermal power plant in Jiangsu Province. It helps improve quality and efficiency, supports low-carbon transformation, and serves the dual carbon goals. At the above power plant, the project reduces the need for about 600 person-times of inspections in high-temperature and high-noise environments each year, significantly improving operational safety. The results are highly promotable and can be adapted to predictive maintenance for key equipment clusters in small and medium-sized manufacturing enterprises, promoting the deep integration of AI and the industrial internet and advancing intelligent manufacturing. At the same time, the project has established a fault model library and a knowledge system, cultivated professional technical talent, and created positive social value in terms of industrial safety, energy conservation, emission reduction and talent development.

Case 24

Innovative Application of AI in Laser Cleaning of Special Steel

Shanghai Shoukun Intelligent And Technology Co., Ltd.

I. Basic Enterprise Information

Founded in 2012 and headquartered in the Shanghai Robot Industrial Park, Shanghai Shoukun Intelligent And Technology Co., Ltd. is a national high-tech enterprise, a Shanghai specialized, refined, distinctive and innovative enterprise, and a recognized enterprise technology center. It has repeatedly been honored as a benchmark enterprise for intelligent robots in Shanghai. The company focuses on the R&D and manufacturing of high-end laser intelligent equipment, AI robot integrated applications and industrial internet of things systems.

Shoukun Intelligent attaches great importance to R&D innovation and has built a young, highly educated professional team covering multiple interdisciplinary fields such as laser technology, AI vision and robot control. The team is led by senior engineers and national manufacturing talents, and core members have deep industry experience and innovation capabilities. The company has accumulated more than 100 independent intellectual property rights, undertaken several key scientific and technological innovation projects in Shanghai, and established industry-university-research cooperation with multiple universities and research institutions to continuously promote technological breakthroughs and industrial application.

The company is equipped with advanced R&D laboratories and production equipment, and has built a full-scenario product matrix covering laser marking, welding, cleaning, cutting and cladding. It is committed to providing integrated intelligent solutions for food and beverage, energy and automotive parts, biomedicine, electronics and home appliances, aerospace and military industries. Shoukun Intelligent adheres to technology-driven development and innovation-led industry progress. It has served more than 1,000 customers and has passed certifications including quality management system certification, intellectual property management system certification, integration of industrialization and informatization management system certification, and CE certification. It has also won many national and provincial-level honors, demonstrating its strong technical strength and industry influence.

II. Case Introduction

(I) Project Background and Introduction

As an indispensable key basic material in fields such as new energy vehicles, aerospace, national defense and military industry, special steel has surface treatment quality that directly determines the performance, reliability and service life of end products. During the production process of special steel, a layer of dense, strongly bonded oxide scale with a thickness of about 10 to 50 micrometers forms on the steel surface after hot rolling. The oxide scale is mainly composed of FeO, Fe3O4 and Fe2O3. This layer must be thoroughly removed before cold rolling. Otherwise, it may cause surface defects, roll scratches, dimension deviations and other problems during cold rolling, ultimately affecting the deep-processing quality and service performance of steel products.

At present, the domestic steel industry generally uses chemical pickling and physical sandblasting to remove oxide scale. However, traditional methods have unavoidable disadvantages. In terms of the environment, pickling uses strong acids such as hydrochloric acid and sulfuric acid, producing harmful gases and waste acid liquid, with high treatment costs and environmental risks. In terms of product quality, acid liquid corrodes the steel substrate, reducing material fatigue strength by 10 to 15 percent, while the high-speed impact of sandblasting introduces residual stress of up to 200 to 300 MPa on the material surface, increasing the risk of deformation during subsequent processing. In terms of process performance, traditional methods have low flexibility, high energy consumption, many consumables and limited adaptability to high-end special steel production characterized by small batches, multiple varieties and high quality requirements.

Given the urgent requirements of China’s dual carbon strategy, the inherent need for manufacturing transformation and upgrading, and the market’s strict quality standards for high-end steel, the development of a green, efficient and intelligent new oxide-scale cleaning technology has become an urgent demand for the steel industry.

(II) Implementation Plan

1. Application scenario analysis

The team conducted in-depth investigations of several domestic steel enterprises and found that several key technical challenges must be overcome to realize large-scale industrial application of laser cleaning, especially in complex scenarios involving high-end special steel.

First, insufficient accuracy and stability in visual detection of oxide scale lead to recognition difficulty. Oxide scale is not a uniform film. Its inner layer is loose and differs from the outer layer in bonding strength. Under different steel grades and rolling processes, its composition, thickness, color and distribution vary greatly. The steel surface also has complex background interference such as oil stains, water stains, reflection and uneven color. Traditional machine vision algorithms based on fixed-threshold binarization or edge detection have extremely high false detection and missed detection rates in such scenarios, and cannot meet the industrial requirement of a removal rate above 99.5 percent. In highly difficult scenarios, such as black scale on dark steel surfaces and thin-film scale, traditional visual inspection almost becomes ineffective.

Second, a lack of adaptive adjustment capability for process parameters causes control difficulty. The effect of laser cleaning depends on more than ten parameters, including power, frequency, pulse width, scanning speed and defocus amount. Different steel grades, such as silicon steel, automotive high-strength steel DP980 and martensitic steel, require very different parameter combinations. Previously, parameter tuning relied entirely on process engineers' experience and trial-and-error debugging, which was inefficient and hard to reproduce. Existing equipment cannot dynamically adjust laser output according to the real-time detected state of oxide scale, such as thickness, coverage and coverage position, making it difficult to balance cleaning quality and production efficiency.

Third, precise processing under dynamic working conditions creates decision-making difficulty. Hot-rolled plates may warp or deform during conveying and storage, with fluctuations of up to plus or minus 5 mm. A fixed-focus laser head can cause focus deviation and drastic changes in energy density. For plates of different sizes and shapes, as well as regions with uneven oxide-scale distribution, robots need to use different cleaning paths to avoid repeated cleaning or omissions. Traditional preprogrammed methods cannot cope with this unstructured work scenario and are highly inflexible.

In summary, the core challenge of the project is how to enable the machine to have eyes that can accurately identify defects like an experienced master worker, a brain capable of intelligent decision-making, and hands that can drive the laser robot to perform flexible and precise operations. This is precisely the core value that AI, especially deep learning technology, can provide.

2. Solution overview

Against this background, the company integrated frontier AI technologies such as computer vision (CV) and deep learning (DL) with high-power laser technology and robotics, and initiated the R&D of an AI deep-learning-based laser intelligent processing system. The project aims to build an intelligent solution integrating intelligent perception, real-time decision-making, precise execution and autonomous learning. It fundamentally solves the pain points of traditional processes, empowers the steel industry to achieve leapfrog development toward green, flexible and intelligent manufacturing, and builds new core advantages for global competition.

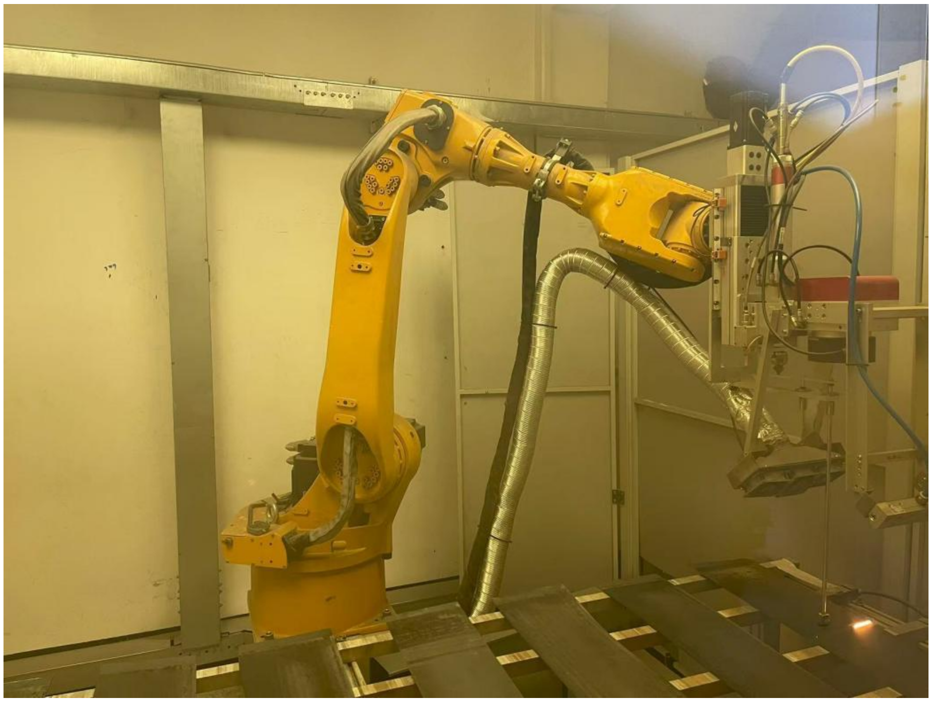


Figure 1 AI deep-learning-based laser intelligent processing system

3. Key AI applications

In this project, the AI vision system plays a key decision-guiding role. It solves the technical difficulties in the scenario and guides the hardware system to work efficiently and stably.

The system has high-precision and highly robust visual recognition capabilities. It can eliminate interference such as lighting changes, surface reflection and oil stains, accurately identify oxide-scale regions, and quantitatively determine their thickness and distribution. The recognition precision is required to reach plus or minus 0.05 mm, and the accuracy exceeds 99.8 percent. The system builds a deep-learning-based process parameter recommendation model for real-time adaptive process decisions. According to the real-time state information input by the vision system, it automatically matches and outputs optimal parameter combinations such as laser energy, frequency and scanning speed, ensuring that the laser cleaning process is both efficient and stable.

III. Project Highlights

(I) Key Technological Breakthroughs

This project represents profound technological integration and re-innovation, mainly reflected in the following aspects.

The first breakthrough is the multimodal fusion intelligent visual perception system. The project innovatively adopts an imaging solution in which dual-band industrial cameras work collaboratively. This multispectral strategy can effectively capture differences in reflection characteristics between oxide scale and substrate under different wavelengths, significantly improving imaging quality and information richness under harsh conditions such as strong reflection and oil contamination. At the algorithm level, transfer learning is carried out based on the ResNet-50 model pretrained on large image datasets. The model is then combined with an improved U-Net++ segmentation network to achieve pixel-level precise segmentation of oxide-scale areas. In addition, a self-built special steel oxide-scale image dataset containing more than 100,000 annotated images greatly enhances the model’s generalization capability and robustness.

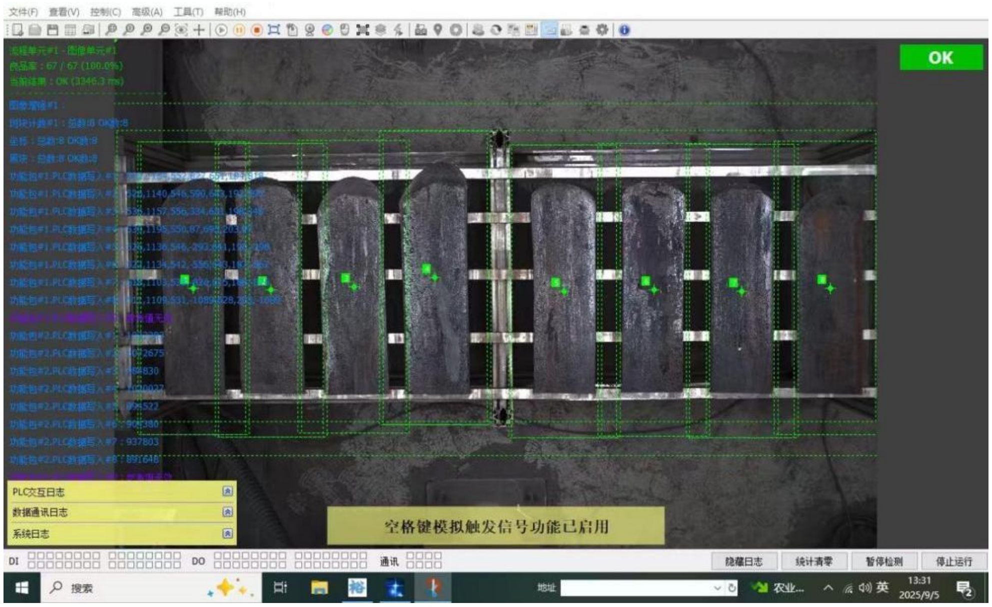


Figure 2 Multimodal fusion intelligent visual perception system

The second breakthrough is adaptive control decision-making based on deep reinforcement learning. The project is among the first to apply the deep Q-network (DQN) reinforcement learning algorithm to process control in laser cleaning. This transforms the system from a passive execution tool into an intelligent agent with autonomous decision-making and optimization capabilities.

Another breakthrough is the real-time closed loop integrating perception, decision-making and execution. The project connects the entire data flow and control flow from visual perception to AI decision-making and then to robot execution. It realizes high-speed and stable communication between the AI edge computing module, a 2,000 W pulsed fiber laser, a six-axis industrial robot, a high-precision linear-motor focusing mechanism and the production line programmable logic controller (PLC) control system. Through carefully designed software architecture and communication protocols, key technical problems such as data synchronization among multiple devices and real-time assurance are solved, forming a responsive, stable and reliable industrial closed loop.

Finally, the project builds an edge-edge-cloud collaborative ecosystem for continuous evolution. A unique three-level system architecture of device, edge and cloud is constructed. On the device side, a Jetson module deployed locally on the equipment is responsible for real-time inference and control. On the edge side, factory-level servers conduct preliminary data aggregation and model fine-tuning. On the cloud side, Shoukun Intelligent has built a central process data platform and model training center. The central server uses massive data from different production lines and steel grades to conduct regular federated learning. Without disclosing users' original data, the model is continuously optimized and upgraded, enabling the system to become increasingly intelligent as more applications are deployed.

(II) Scenario Specificity

The laser intelligent cleaning scenario for special steel selected in this case is highly representative of industrial applications and technically challenging. It is not an isolated laboratory application under restricted conditions, but directly confronts multiple challenges common in heavy industrial environments, including material diversity, complex working conditions, strict quality requirements and urgent production rhythm. The success of this solution proves that AI technology can not only play a role in asset-light industries such as the internet and finance, but also penetrate heavy-asset basic industries such as steel and solve pain points in their core production processes.

Its uniqueness lies in the first realization of full-process and full-element AI empowerment in the field of steel surface treatment. It builds an intelligent processing organism that can see accurately, think clearly, work precisely and learn continuously. This is not only a technical product, but also an evolvable and promotable intelligent manufacturing methodology, providing a highly valuable model for the intelligent transformation of other traditional industrial processes.

IV. Implementation Results

(I) Economic Benefits

Taking the actual application at Baowu Group, a large steel group, as an example, the group deployed Shoukun Intelligent's AI laser cleaning system on its special steel hot-rolling production line. According to actual operation calculations, a single production line can save about 3.2 million yuan annually in consumables such as acid liquid, neutralizers and steel grit. Because the demand for waste acid and exhaust gas treatment is completely eliminated, annual environmental treatment costs are reduced by 750,000 yuan. At the same time, because the energy utilization efficiency of laser cleaning is much higher than that of traditional processes, the comprehensive energy-saving rate reaches 38 percent. In total, the annual comprehensive cost reduction of a single production line exceeds 3.95 million yuan. The cleaning quality remains above 99.8 percent, and customer satisfaction has improved significantly.

In the application of Shanghai Baokuang Technology, a private small and medium-sized enterprise engaged in the R&D and production of mining machinery, the system demonstrated excellent flexible production capability. The enterprise mainly produces large structural components such as hydraulic supports, with many material types and small batch sizes, resulting in high switching costs for traditional cleaning methods. After introducing the AI laser cleaning system, it realized one-click switching for cleaning workpieces of different materials, with no need to retune the production line. A single device saves about 1.85 million yuan annually in outsourced cleaning costs and consumables, shortens the product cleaning cycle by 60 percent, and solves the surface treatment problem of small-batch, multi-variety orders.

By mid-2025, Shoukun Intelligent had successfully obtained multiple orders for this system from the above enterprises and other customers. Its products cover both large steel groups and small and medium-sized manufacturing enterprises, indicating that the technical solution has gained market recognition among enterprises of different sizes and needs. Based on current market feedback and industry demand, the product's market size is expected to exceed 100 million yuan within the next three years.

(II) Technological Innovation and Intellectual Property Achievements

The project has achieved multiple breakthroughs from zero to one. It has successfully developed China's first fully closed-loop intelligent laser cleaning system integrating AI visual recognition, deep-learning decision-making, precise robot execution and continuous process library evolution. It fills a domestic gap in this technological field and breaks the technological monopoly held by foreign manufacturers in certain high-end laser processing equipment.

The project team did not simply apply existing AI models, but carried out in-depth algorithm optimization and innovation for specific industrial scenarios. Frontier models such as ResNet-50, U-Net++ and DQN were creatively applied to solve oxide-scale recognition and laser parameter control problems. A series of engineering challenges, including small-sample learning and lightweight model deployment, were also solved. This represents industrial-grade innovative application of frontier algorithms.

The project realizes deep embedded integration between the AI edge computing module and high-power laser processing equipment, rather than a simple plug-in connection. This integrated design ensures the reliability, stability and real-time performance of the system in harsh industrial environments, meeting the requirement for continuous and stable 24/7 operation. It is an innovative integration of deep embedded systems.

Around the core technologies of this system, Shoukun Intelligent has applied for two invention patents, three utility model patents and two software copyrights, forming a complete intellectual property protection network and building a solid intellectual property system. This has laid a strong foundation for market competition and subsequent technological evolution of the product.

(III) Far-Reaching Social Value and Industry Influence

The technology practices green manufacturing and supports the dual carbon goals. It is a purely physical and non-contact processing method. No chemical additives are consumed throughout the entire process, and no toxic or harmful gases or liquid emissions are produced. According to calculations, a single production line can reduce annual acidic waste gas emissions by about 200 tons and heavy-metal-containing waste liquid and sludge emissions by about 300 tons, completely eliminating the source of heavy metal pollution in related links. At the same time, compared with traditional high-energy-consumption processes, its life-cycle carbon emissions are reduced by more than 75 percent. It is therefore an important technological path for the steel industry to respond to the national dual carbon strategy and achieve green and low-carbon transformation.

The system improves product quality and strengthens industry competitiveness. Through precise AI control, the system significantly improves the consistency and stability of surface treatment quality for high-end special steel products. This provides downstream industries such as new energy vehicles, aerospace and precision instruments with basic materials of more reliable quality and superior performance, enhancing the core competitiveness and brand image of China's manufacturing in the international market.

It also forms a replicable and promotable intelligent model. The deep integration model of AI technology explored in this project has a highly universal and adaptable technical framework, core algorithm modules and system architecture. Through targeted scenario adaptation and accumulation of process data, the solution can be rapidly extended to many scenarios and industries, including rust removal before welding in shipbuilding, surface cleaning and paint removal of rail transit vehicle bodies, surface treatment of aerospace composite components, and surface treatment of non-ferrous metals such as copper, aluminum and titanium alloys. It has broad market prospects and great social benefits, and is expected to become an important model for the intelligent and green transformation of traditional manufacturing.

Case 25

AI Model for Alloy Optimization in Converter Steelmaking

Luoxu Energy Technology (Shanghai) Co., Ltd.

I. Basic Enterprise Information

Luoxu Energy Technology (Shanghai) Co., Ltd. was established in January 2022. The company focuses on the deep application of AI in steel metallurgy and other high-energy-consumption industrial fields, and is committed to promoting intelligent, green and low-carbon transformation across the industry. Supported by the Institute of Energy Process Engineering at Technische Universität Bergakademie Freiberg in Germany, the company has developed a comprehensive R&D system. Its founder, Dr. An Fengbo, previously led large-scale Sino-German optimization projects for high-energy-consumption installations. The core team consists of doctors and engineering professionals with research backgrounds in China and abroad, specializing in the intersection of industrial process modeling and AI. The company has obtained multiple independent intellectual property rights and software copyrights, and has established partnerships with a number of steel enterprises. Its AI R&D outcomes have already been applied in intelligent alloy addition calculation and other scenarios.

II. Case Introduction

(I) Background and Needs

1. Industry background

(1) Policy dimension

China is promoting the green transformation of the steel industry. The Special Action Plan for Energy Conservation and Carbon Reduction in the Steel Industry released in 2024 explicitly encourages the application of AI in production optimization and requires the numerical-control rate of key processes to reach 80 percent by the end of 2025. After the steel industry is included in the carbon trading market, AI can accurately optimize key production links, including alloy addition, reduce material waste and carbon emissions, and become an important tool for enterprises to implement energy conservation and carbon reduction policies.

(2) Industrial dimension

Affected by fluctuations in raw material prices and declining demand, the steel industry increasingly emphasizes cost-accounting-based operations and demands extreme cost reduction and efficiency improvement. Traditional processes, including alloy addition and other key links, suffer from high consumption, low precision and bottlenecks in process improvement. AI can integrate real-time data, break through the limitations of linear calculation, and realize dynamic optimization of key processes, matching the needs of cost reduction, quality improvement and efficiency enhancement.

(3) Process dimension

Traditional alloy addition relies on manual experience and simple calculation. It faces problems such as delayed real-time response, which may cause thousands of yuan in waste for a single heat, and insufficient precision, which may extend the smelting cycle. Under market competition, enterprises urgently need solutions featuring high precision, low waste and fast response. AI, with its data processing and real-time optimization capabilities, has become a core means of breaking through these constraints.

2. Core project needs

(1) Efficiency and cost: replace manual calculation, automate alloy feeding, reduce errors, improve production efficiency, and optimize batching according to real-time market prices so as to reduce the cost per ton of steel.

(2) Quality stability: dynamically adjust the feeding strategy based on real-time data, narrow composition fluctuations, and improve the consistency of steel quality.

(3) Standardization and weak data foundation: establish unified calculation standards and a dedicated database to solve problems such as scattered data, poor data quality and reliance on operator experience.

(4) Insufficient process optimization and emergency response capability: use systematic data analysis and AI models to identify production bottlenecks and build mechanisms for process optimization and abnormal-situation response.

3. Problems to be solved

(1) Efficiency improvement and cost optimization: AI models are needed to replace manual calculation, automate alloy addition, reduce human intervention and errors, accelerate the feeding process, improve production smoothness, and meet the requirements of capacity expansion. At the same time, production data, target composition and alloy market price fluctuations must be integrated in real time to accurately calculate alloy demand, formulate optimal feeding schemes, reduce resource waste and consumption, and significantly lower production costs.

(2) Quality stability and narrow-band breakthrough: AI models are needed to accurately adjust alloy feeding quantities according to real-time molten steel composition and production targets, control composition fluctuation within a very small range, and ensure stable product quality and performance. With the dynamic adaptability and precise control capability of AI, the project also promotes narrow-band production, improves steel strength, toughness and other properties, meets stringent customer requirements for different steel grades, and strengthens competitiveness in the high-end market.

(3) Establishing standards and accurately predicting parameters: AI models are needed to establish unified calculation standards for alloy feeding, eliminate differences caused by teams and individual experience, improve standardization of production operations, and support cost-accounting-based operations. Meanwhile, AI algorithms accurately predict key parameters such as tapping quantity, alloy yield and residual manganese content, correct distorted calculation baselines, and improve alloy feeding precision and production stability.

(4) Data management and process optimization empowerment: A dedicated alloy feeding database should be built to integrate procurement data such as prices, quality inspection data such as raw material composition, production data such as heat conditions, and warehousing data such as inventory distributed across ERP and laboratory information management systems (LIMS). This enables rapid retrieval, precise transmission and traceability, supports AI in formulating feeding schemes that coordinate quality, cost and inventory, and provides accurate data support for cost statistics and management decisions. Through AI-based data analysis and deep mining, key production indicators can be quantitatively analyzed, bottlenecks and defects accurately identified, and a data-driven process optimization mechanism formed to promote continuous process iteration and support enterprise intelligent transformation.

4. Project construction objectives

(1) Management objectives

The project aims to improve feeding precision by analyzing molten steel parameters in real time, accurately predicting alloy usage, dynamically correcting schemes, reducing redundant addition, improving control precision and efficiency, and reducing manual intervention. It deepens cost management by optimizing feeding combinations based on alloy composition and prices, accurately calculating heat-level costs, increasing cost transparency, and supporting refined management. It optimizes production processes by analyzing data to identify process weaknesses and dynamically adjusting feeding strategies, thereby improving energy utilization, promoting process upgrading and meeting the requirements of high-end steel grades. It also improves the level of intelligence by introducing AI to build a real-time data analysis and decision-making platform, achieving automation across the alloy feeding process, breaking data barriers, and laying a foundation for intelligent manufacturing. Finally, it seeks to form a low-cost steel AI solution applicable to the steel industry and to provide a replicable model for industry progress.

(2) Technical objectives

序号	具体指标	预期目标
1	模型部署覆盖率	模型覆盖率 ≥ 85%，实现关键计算自动化
2	数据驱动决策采纳率	模型推荐方案被生产现场采纳比例 ≥ 80%
3	资源利用效率	合金消耗量或原料利用率提高 ≥ 3%
4	工艺稳定性	关键工艺参数波动幅度降低 ≥ 10%

(II) Implementation Plan

Project technical solution

(1) Key technologies

Data collection and preprocessing: data on alloys, production and prices are collected, and algorithms such as principal component analysis and clustering are used for cleaning and standardization to ensure data integrity and consistency. Model construction and optimization: machine learning and deep learning are combined to establish a mapping model for alloy addition. Linear regression and support vector machines (SVMs) are used to analyze relationships, deep neural networks (DNNs) improve prediction accuracy, and genetic algorithms optimize the ratio. Data storage: an alloy control database is built, and Hadoop distributed storage is used to ensure scalability and reliability. Visual interaction: a visual human-machine interaction interface displays heat process requirements, model inputs and recommended addition quantities. Data interaction management: ETL tools and API interfaces automatically process data, and stream processing frameworks support real-time analysis to ensure data accuracy and timeliness. Intelligent optimization decision-making: intelligent decision models are built, multiple algorithms are used to construct mapping models, and DNNs improve prediction accuracy so as to optimize feeding schemes. Collaborative optimization of production performance: methods such as principal component analysis are used to identify associated patterns, and genetic algorithms optimize composition control, feeding schemes and costs.

(2) System architecture

The system adopts a four-layer deeply collaborative architecture, building a full-process closed loop from real-time production data collection, multidimensional processing and intelligent algorithm analysis to accurate decision output. Its functional architecture is shown below:

Figure 1 System architecture of the AI calculation model for alloy feeding

The system architecture consists of four parts. The basic resource layer provides the system with a stable and reliable hardware and software support environment. The data collection layer realizes high-precision, high-frequency access and preprocessing of key data across the converter steelmaking process. The algorithm modeling layer completes the core transformation from data to decision-making through four AI model systems. The system application layer turns intelligent decision results into business functions that can directly guide production.

Through standardized data interfaces, real-time communication protocols and collaborative scheduling mechanisms, all layers are seamlessly connected to form an intelligent life-cycle control system covering the alloy feeding link of converter steelmaking. This effectively promotes the transformation of production from traditional experience-driven operation to data-driven operation and realizes precise, efficient and cost-optimal alloy feeding.

(3) Core system functions

Data management module: Process standard management builds a structured process procedure database and automatically synchronizes updates to relevant models. Material management realizes life-cycle management of alloys and other materials, covering basic information, composition data, warehousing, inventory changes and consumption statistics. Price management establishes a dynamic alloy price management mechanism, records price fluctuation trends, generates analysis reports and automatically synchronizes them to the economic calculation model, ensuring that cost optimization is based on the latest price data.

Alloy optimization center module: This module realizes real-time monitoring and operational integration of key data across the smelting process. It dynamically displays core information such as molten iron, composition and auxiliary material additions, supports operator interaction, and enables seamless connection between model output and production execution.

Smelting data statistical analysis module: Through multidimensional cost monitoring, production performance statistics and historical data review, this module builds a full-process analysis system. First, multidimensional cost monitoring takes alloy cost as the core and analyzes costs by heat, day, week, month and steel grade, generating cost trend curves and calculating deviations between actual and standard costs. When deviations exceed thresholds, the system automatically marks and prompts them. Second, production performance statistics generate periodic reports by team, including KPIs such as alloy unit consumption, record operator behavior data, and compare them with preset baselines to generate performance deviation reports. Third, historical data review builds a full production data archiving and retrieval system, enabling traceability of process parameters, model decision processes and other data.

Abnormality warning module: A multi-level abnormality detection and warning mechanism is built to compare process standards, model predictions and actual data in real time, monitor abnormalities across the entire process, and generate alarms for abnormal feeding, cost deviations and composition abnormalities.

III. Project Highlights

(I) Key Technological Breakthroughs

1. R&D technology based on mechanism-data fusion modeling

The project breaks through the limitations of traditional single models by integrating computational fluid dynamics (CFD) simulation, multiscale thermodynamic mechanism equations and deep learning algorithms to build an interpretable hybrid model. This addresses both the black-box problem of AI and the insufficient real-time performance of mechanism models. For example, the converter model controls temperature error within ±10 degrees Celsius, and the carbon content hit rate reaches at least 85 percent.

2. Closed-loop intelligent control

The project builds a full-chain closed-loop system of prediction, decision-making, execution and feedback. It not only predicts indicators such as temperature and composition, but also outputs control strategies through reinforcement learning, such as dynamically adjusting process parameters in converter steelmaking to avoid rework and shutdowns.

3. Lightweight deployment technology

The project requires no transformation of production equipment and no production interruption. Upgrading is achieved only through algorithm optimization, resulting in a short deployment cycle and suitability for older plant scenarios. For example, the alloy feeding system was rapidly implemented at Baowu Zhongnan Steel, reducing the cost per ton of steel by 1.6 yuan.

(II) System Structure Innovation

1. Architecture of core platform plus industry modules

The project builds an integrated Luoxu intelligent industrial modeling platform and pairs it with vertical industry model packages for steel, chemicals and other sectors. It supports modular expansion. For example, to add an adaptation module for the cement industry, only a dedicated module needs to be installed, without rebuilding the core system.

2. Dual-end deployment capability

The system supports edge-side real-time computing and cloud-side collaborative management. The edge side ensures low latency in industrial scenarios, such as flame recognition responses within five seconds, while the cloud side enables multi-plant data aggregation and model iteration, adapting to different enterprise information technology (IT) architectures.

3. Integrated tool chain

The project integrates data cleaning, feature extraction, model training, deployment and operation and maintenance tools into a complete tool chain, lowering the threshold for enterprises and enabling frontline operators to learn and use the system quickly.

(III) Uniqueness of Application Scenarios

The uniqueness of this project in the full process of converter steelmaking lies in the deep adaptation of AI technology to core scenarios, which fundamentally changes the pain points of traditional production models.

Full-process intelligent closed-loop scenario: The project breaks data silos and technical discontinuities across traditional links and builds a coherent data-model-decision-iteration scenario, realizing end-to-end integration from data collection to intelligent decision-making rather than optimization of a single link.

Dynamic alloy feeding scenario: Different from static ratios driven by manual experience, the project creates a scenario of multi-source data fusion in real time and dynamic optimization by steel grade. It accurately uncovers hidden correlations between parameters and yield, achieves an error reduction of more than 40 percent and second-level response, and solves the problems of delayed and inaccurate traditional feeding.

Cost prediction and control scenario: The project abandons passive after-the-fact accounting and forms a closed loop of real-time price fluctuation tracking, automatic optimization when deviation exceeds five percent, and accurate calculation of cost reduction benefits. It proactively responds to cost risks and provides a scientific basis for decision-making.

Strict quality assurance scenario: For production scenarios with strict composition requirements, such as rebar, the project uses time-series prediction of molten steel composition with an accuracy above 92 percent as the core to ensure composition hit rate and finished-product narrow-band hit rate, solving the pain point of unstable traditional quality.

Iterative production scenario: The project builds a dynamic scenario of collection, training, deployment, validation and iteration. By continuously optimizing parameters and identifying process bottlenecks, it promotes the upgrade from automated execution to intelligent decision-making rather than maintaining a static production model.

(IV) Project Innovations

1. Technological integration innovation: The project breaks the limitations of single algorithms by integrating machine learning and deep learning algorithms to mine complex nonlinear relationships. In complex scenarios, the prediction error of alloy feeding is no more than five percent, and the control deviation for composition narrow-banding is ±0.05 percent, solving technical pain points in steel enterprises and filling an industry precision gap.

2. Cost adaptation innovation: In response to cost control needs, genetic algorithms and other methods are linked with real-time cost data to build a cost-quality-efficiency model. When alloy prices fluctuate by more than five percent, the system automatically adjusts the ratio to minimize cost without degrading quality, reducing overall costs by another three to five percent.

3. Data application innovation: A lightweight dedicated database is built to realize real-time visualization and multidimensional statistics. Visual dashboards help quickly locate abnormalities, while multidimensional statistics support low-cost data-driven production optimization. For example, the unit consumption of niobium-nitrogen alloy for a certain steel grade decreased by eight percent.

4. Management mechanism innovation: The project develops an operation-level evaluation model. It uses three categories of indicators to realize quantitative skill scoring and file management, automatically identifies abnormalities and outputs improvement suggestions, raises the composition qualification rate by 12 percent, and builds a closed-loop mechanism to solve pain points in personnel management.

IV. Implementation Results

This project covers an AI application R&D project for a production line with an annual steelmaking capacity of four million tons at a steel plant. It focuses on alloy optimization in converter steelmaking and follows the principle of light assets and high adaptability. It is aligned with enterprise needs, has a high input-output ratio and meets economic requirements.

(I) Economic Benefits

After the AI calculation model for converter steelmaking alloy optimization was put into production, the system realized optimized control over alloy feeding in the converter steelmaking process. Actual operating data show that by optimizing the alloy batching structure, the use of high-value alloys, such as ferrosilicon, silicomanganese alloy (6818) and niobium-nitrogen alloy, was effectively controlled, reducing the cost per ton of steel by 3.9 yuan.

Based on actual operating results and the annual production scale of four million tons covered by the project, including 2.4 million tons of rebar, 800,000 tons of industrial wire rods and 800,000 tons of plates, the project’s economic benefits were comprehensively calculated. According to differences in niobium-containing characteristics among steel grades, the actual cost-reduction effect in the refining process calculated at 60 percent, and the actual refining ratios of different steel grades, a professional algorithm was used to calculate economic benefits:

Economic benefits = production capacity × cost reduction per ton of steel × (niobium-containing ratio × 100 percent + non-niobium ratio × 10 percent) × (refining ratio × 60 percent + M-station ratio × 100 percent)

After calculation, the annual economic benefits of the project are as follows: rebar achieved a cost reduction of 5.631 million yuan, industrial wire rods 187,200 yuan, and plates 646,500 yuan. The total annual economic benefit reached 6.4647 million yuan, equivalent to a reduction of 1.6 yuan per ton of steel. Through AI technology, the project realized precise control of alloy costs and provided a successful example for improving quality and efficiency in steel enterprises.

(II) Social Benefits

1. Improving product quality and industry competitiveness: Precise control of molten steel composition narrows the fluctuation range, reduces defective products and resource waste, helps enterprises steadily produce high-end products, enhances downstream supply reliability, improves the international reputation and competitiveness of China’s steel industry, and promotes high-quality development of the industrial chain.

2. Optimizing production processes and operating efficiency: The project improves the hit rate of arrival at the refining station, reduces supplementary alloy addition and refining time, accelerates the production rhythm, lowers material and time costs, supports large-scale continuous production, and improves overall industry operating efficiency.

3. Increasing production efficiency and ensuring supply capability: Automated feeding overcomes the drawbacks of manual operation, reduces process interruptions, improves production stability and efficiency, releases capacity, meets market demand, and provides stable raw materials for national infrastructure and industry.

4. Promoting energy conservation, carbon reduction and green transformation: The project precisely matches energy inputs, improves thermal efficiency, lowers unit energy consumption and carbon emissions, responds to the dual-carbon strategy, provides a reusable technical pathway for the green transformation of the industry, and sets a green benchmark.

5. Promoting enterprise transformation and intelligent upgrading of the industry: The project helps enterprises establish a standardized data management system, realize data integration and analysis, support scientific decision-making, promote transformation from experience-based management to data-driven management, and inject momentum into industry-wide intelligent upgrading.

Case 26

Intelligent Optimization Collaboration Platform Based on Model-Driven and High-Performance Responsive Computing

Gudou Technology (Shanghai) Co., Ltd.

I. Basic Enterprise Information

Gudou Technology (Shanghai) Co., Ltd. was founded on July 31, 2019. It focuses on providing AI-based decision optimization software for manufacturing customers, addressing the pain points of complex, inefficient and costly enterprise resource collaboration decisions. In fields such as operational decision-making, job scheduling, supply chain management, logistics distribution and intelligent replenishment, the company provides integrated intelligent collaboration solutions covering strategic consulting and software implementation.

Gudou Technology has completed the layout and implementation of group-level large-scale supply chain projects in areas including new energy materials, new energy batteries, complete vehicles and auto parts manufacturing, food and medicine, steel, chemicals and 3C manufacturing. Representative customers include Shanghai Electric, China Resources Sanjiu, Yanfeng Automotive Seating, Seres Auto, Contemporary Amperex Technology Co., Limited, Hang Lung Auto Parts, Weifu High-Technology Group, China Petrochemical Corporation, HBIS Group, Xiangtan Iron and Steel, Xingcheng Special Steel, Yongfeng Steel, Rizhao Steel, Jingye Steel, Zenith Steel, Hengrui Pharmaceuticals, Shineway Pharmaceutical, Kelun Pharmaceutical, Botanee Cosmetics, Carslan Cosmetics, Xiaomi and Shengtai.

Gudou Technology adheres to innovation. Through independent R&D, it has obtained multiple independent intellectual property rights and qualifications including high-tech enterprise, specialized, refined, distinctive and innovative enterprise, technology-based SME and district-level technology center. It has also passed TCMMI3 certification and, among industry peers, obtained ISO system certification and dual-software enterprise certification at an early stage, laying a solid foundation for high-quality, efficient and competitive software products and services.

II. Case Introduction

(I) Background and Needs

Enterprises whose production mode is mainly engineer-to-order (ETO) design and produce according to orders. Under this mode, products are mostly single-piece customized products. The production process is long, involves extensive collaboration and supporting activities, and includes many special processes. Therefore, an efficient planning and scheduling system is needed. However, the formulation of all workshop plans relies entirely on manual decisions, making it difficult to systematically and efficiently address numerous production challenges, such as frequent production abnormalities, urgent inserted orders, substantial inter-process transfer activities and unstable movement times. These problems prolong the overall project production cycle and make it difficult to quickly respond to downstream market demand. Such enterprises mainly face the following problems:

1. Low transparency and poor real-time performance of planning and scheduling information, low efficiency of manual scheduling and heavy workload.

2. Delayed arrival of raw materials and the need for manual tracking of production conditions, resulting in poor real-time performance and delayed feedback.

3. Frequent inserted orders, which easily disrupt existing production plans and lead to low resource utilization efficiency.

4. Delayed raw material arrival, machine tool failures that cannot be repaired in time, insufficient fixtures and incomplete technical preparation all require adjustment of pre-scheduled plans. Manual plan adjustment makes it difficult to coordinate all resources. Most urgent inserted orders have very tight delivery dates and leave very limited time for manual decision-making. Meeting inserted orders through shutdown and switching may also delay other orders.

5. Low inventory turnover: Production cannot proceed according to the original plan due to resource changes, and shutdowns and switching caused by inserted orders lead some blanks and semi-finished products to remain in

warehouses without moving downstream. As a result, turnover from arrival of blanks to delivery of finished products is low.

(II) Implementation Plan

The resource intelligent optimization collaboration platform based on “AI + business model” provided by Gudou Technology, hereinafter referred to as the Gudou Platform, can solve the above problems faced by ETO-dominant enterprises. The core idea of the technical solution is as follows: Gudou business models solve users’ actual production problems; the agent-based resource intelligent optimization collaboration platform built through AI orchestration provides convenient natural language interaction and query; by structuring enterprise business indicators and business rules and digitizing the planning experience of business personnel, the platform coordinates the overall situation, automatically optimizes the best decisions, senses execution feedback in real time, calculates the optimal scheme under the current situation according to changes, and provides natural language interaction through agents so that users can query and analyze current scheduling results.

The first step for the Gudou Platform to solve business problems is to build business optimization models based on digital twin technology. Through modeling, business problems are abstracted and key business indicators and rules are extracted. Business modeling is carried out in an object-oriented way. The general process is as follows:

1. Through business modeling, all process information and supply chain paths that may be involved in each product are parameterized to facilitate subsequent algorithmic solution of mathematical models.

2. Through code, business models are transformed into mathematical models for algorithmic solution.

3. Through operations research optimization algorithms, optimal planning decision schemes are obtained.

After the advanced planning and scheduling (APS) model is built, master data, business data such as orders and inventory, and bills of materials (BOMs) are required as inputs to perform finite-capacity calculation. Therefore, the second step is data integration and governance.

The Gudou Platform can seamlessly integrate with other key existing business systems of enterprises, including planning control systems (PCSs), manufacturing execution systems (MESs) and master data platforms. Such integration ensures data consistency and efficient process collaboration. The platform can receive and process data from different business systems, including production plans, production orders, BOMs and inventory information, so as to realize comprehensive production planning and scheduling. It obtains necessary data on planning, production, materials, inventory and sales from other systems, and feeds the planning and scheduling results of the intelligent scheduling system back to each business system. Data integrity and accessibility are key to the successful launch of the platform. Therefore, throughout the APS implementation process, data governance requirements such as order number generation specifications are defined to ensure that data issues do not affect successful project launch. Meanwhile, the powerful rule library provided by the Gudou Platform can manage all imported data and allow users to subsequently add, delete, modify, query, manage versions and perform other maintenance operations.

Finally, after business modeling and data input from other related systems are in place, the Gudou Platform can perform one-click finite-capacity scheduling based on established optimization strategies and constraint rules. It outputs planning results and process execution tracking functions including order management, workshop master plans, capacity load, outsourcing plans, delivery commitments, equipment scheduling, work order issuance and job reporting.

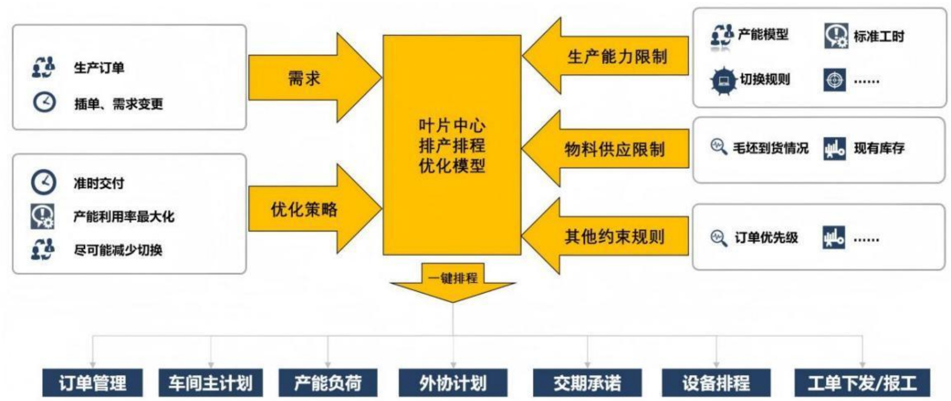


Figure 1 One-click scheduling diagram

In addition, the Gudou Platform can realize what-if multi-scenario simulation through its built-in scenario management function. It can conduct sand-table simulations by adjusting business rules, including simulations of new equipment, production lines and personnel resources or simulations of inserted orders. Through its built-in visual comparison function, the system can compare and display specific indicator parameters between the production environment and each simulation scenario, including on-time delivery rate, production capacity, equipment utilization, inventory-related indicators and even financial indicators. By setting different factories and workshops as different scenarios, rapid horizontal expansion across factories and workshops can also be achieved.

After the Gudou business model solves production scheduling problems through business modeling and algorithms, the AI-orchestrated agents on the platform combine the capabilities of large language models and business models. By using AI's natural language understanding capability, the platform allows users to query orders, equipment, inventory and other information in planning results through natural language dialogue. After fully understanding user intent through its built-in knowledge base and corpus, the agent calls the interface views of Gudou business models and displays the business data that users care about in a visualized manner, making it convenient for users to directly query planning results.

III. Project Highlights

(I) Technological Advancement and Practicality

1. High-performance responsive in-memory computing: The platform supports automatic calculation of hosted attribute dependencies and automatic real-time data refresh, enabling efficient linkage of information flows. In scheduling scenarios, when 2,000 task adjustments trigger automatic rescheduling, second-level or minute-level responses can be achieved, ensuring real-time business decisions. The response granularity covers any object and any attribute change of an object, while the response depth supports unlimited levels of object and attribute changes, enabling dynamic responses across all dimensions without blind spots.

2. Integrated model construction and management: The platform comprehensively supports model development for various planning and scheduling decision scenarios such as marketing plans, delivery commitments and production master plans, and supports expansion into sub-scenarios. It also provides full-process online model management, covering online editing, compilation, debugging, multi-person collaborative development, online release, import and export backup, start-stop monitoring and other functions.

3. Model hot-update technology: While the platform and models remain running and do not need to be stopped, model logic can be updated and iterated in real time, ensuring uninterrupted business continuity while supporting rapid model optimization and upgrading.

4. Efficient online model debugging technology: Relying on model hot-update capabilities, logic updates take effect directly. Compared with traditional model packaging updates and platform startup, this greatly shortens the time required and can improve model debugging efficiency by more than 80 percent, significantly reducing iteration waiting time and improving development efficiency.

5. Rule customization and management: The platform supports flexible rule customization and convenient maintenance to meet diverse business scenario needs. It allows any number of branches to be defined under the same rule group, supports refined splitting and expansion of complex logic, and enables rapid copying of rule branches to greatly improve configuration efficiency. For any data changes in the same branch, different versions can be defined to realize full-chain traceability of rule parameters and ensure controllable iteration. At the same time, rule assets can be imported and exported, enabling flexible transfer and management.

6. Scenario customization and scenario simulation: The platform supports the creation of multiple data-isolated scenarios for a single model, enabling multi-scenario applications. Scenarios can be quickly copied, and what-if scenario simulation is provided to support multidimensional business verification and decision-making.

7. Scenario persistence and recovery: Scenario data supports minute-level persistence to ensure secure data retention. It can also be fully restored from persisted data into memory within minutes, enabling efficient data continuity and rapid recovery.

8. Custom configuration engine for views and components: The platform supports efficient reuse of configured pages. Different views can be flexibly assembled through drag-and-drop operations, quickly generating functional interfaces that meet business needs and greatly improving interface construction efficiency. It also supports secondary personalized configuration of page components, such as using different icons to visually present planning status and completion conditions.

9. Visual interface configuration based on model object associations: Without code development, page configuration can be completed only through intuitive user interface (UI) drag-and-drop operations and assignment binding with business objects. Users do not need to understand underlying data structures such as background database tables. For example, configuring a page with three query search conditions and one right-click row menu containing five form operation details takes only 0.5 to one hour, significantly lowering the threshold and time cost of page construction.

10. Flexible and powerful system integration: The platform fully supports protocols such as Hypertext Transfer Protocol (HTTP), SOAP Webservice and REST Webservice, message queues such as RabbitMQ and RocketMQ, database access, Excel file import and export, and multiple other integration methods, covering diverse data interaction scenarios. It also supports flexible protocol expansion to accommodate future integration needs of new businesses.

(II) Scalability

The model layer and the basic platform layer are separated, realizing decoupling between business and technology. A unified technical platform supports intelligent decision-making in multiple domains. Based on a single system architecture, customized models are built for different business scenarios and practical needs to handle diverse business problems. Users can continuously build their factory digital system without purchasing additional systems.

At the agent application layer, the platform continuously enhances and expands business models in various domains by integrating knowledge bases, corpora and large-model orchestration technology, and continuously implements knowledge graphs for different business domains. It can build agent applications adapted to different business scenarios at low cost, significantly improve user interaction experience and reduce the cost of human-machine interaction.

IV. Implementation Results

Implementation of such projects can significantly improve collaboration efficiency and automation across the full cycle of manufacturing. For example, the advanced intelligent scheduling project at Shanghai Turbine Plant, accepted in April 2025, achieved multiple economic and social benefits.

(I) Economic Benefits

In terms of economic benefits, the advanced intelligent scheduling project at Shanghai Turbine Plant relies on real-time data updated 40 times per day and the APS system to transform manual scheduling experience into standardized logic. It generates plans with one click and supports multi-scenario simulation, connecting the full chain of raw material procurement, technical preparation and tooling preparation. The planning horizon extends to the monthly level, and the granularity is refined to the minute level. Planning preparation time was shortened from more than two hours to within half an hour. At the same time, the project provides customized functions for the technical department, including hierarchical prompts, technical preparation and real-time information synchronization, and for workshops, including rapid generation of global plans and real-time display of scheduling results. Closed-loop issue management handled 334 issues within two months, and all-staff collaboration supported more than 1,500 operations in the launch month, ensuring implementation and improving the efficiency and reliability of production planning.

(II) Social Benefits

In terms of social benefits, the digital platform improves information transparency and business collaboration efficiency through intelligent technologies, reducing reliance on manual operations and the potential errors arising from them. It not only promotes green production and sustainable development, but also greatly reduces carbon footprints and energy consumption, contributing to environmental protection.

(III) Social Value

The successful practice of such projects provides the industry with referenceable experience in digital transformation and in the application of AI agents in manufacturing. It drives related enterprises toward intelligent manufacturing and helps improve the technical level and market competitiveness of China’s manufacturing industry as a whole.

Part III AI Applications

Case 27

WeMeet Multi-agent AI Business Interconnection Platform

Shanghai Huishen Digital Technology NetworkCo.,Ltd.

I. Basic Enterprise Information

Shanghai Huishen Digital Technology Network Co., Ltd. is an innovative technology enterprise specializing in the deep integration of AI with business scenarios. Its core offering, the WeMeet Multi-Agent AI Business Interconnection Platform, is the world's first AI business ecosystem built on a multi-agent architecture. The platform has made a breakthrough by moving from relying on large language models to intelligent agents, which integrates diverse business scenarios including conferences, exhibitions, marketing, business travel, and communities to create an intelligent, interconnected global business network.

Leveraging agent technology, WeMeet can fully understand users' intentions and offer active guidance. It supports core functions including multi-agent collaboration, text-to-application generation, and intelligent matching and recommendation. These capabilities significantly enhance efficiency and experience of enterprises across confertence planning, exhibition operations, marketing propagation, and business travel. The platform has successfully enabled some notabe projects such as the GDPS Global Developer Pioneers Conference CIIE Shanghai Events and Conferences, Chinese Business Global Industry Service Alliance, AIPC 2025 International Conference Center Association, and MeetSG Multi-Agent Destination AI Platform, demonstrating it's robust technological innovation strength and sound global service capacity. In compliance with *the Interim Measures for the Management of Generative Artificial Intelligence Services* issued by the Shanghai Cyberspace Administration, the corporate has completed the registration for generative AI services. The generative synthesis algorithm of WeMeet Huishen AI Business Interconnection Plarform has also obtained official filing with the Cyberspace Administration of China.

Adhering to the philosophy of "Intelligence Leads the Future, Interconnection Creates mutual benefits", the company is collaborating with global partners to build an open, trustworthy, and sustainable interconnected ecosystem of AI and business so that the users can achieve achieve efficient, cross-language, cross-platform intelligent collaboration and resource integration.

II. Case Introduction

(I) Background and Needs

Today, the transformation of AI applications in the business sector is accelerating. Gartner predicts that by 2026, 75% of World' s top 500 companies will adopt decision intelligence practices, including recording decisions for subsequent analysis. By 2027, 50% of business decisions will be improved or automated through AI agents for decision intelligence. IDC forecasts that by 2026, 50% of data teams at China's top 500 enterprises will use AI agents to finish data preparation and analysis and these agents will become important organizers and coordinators; and 40% of China's 500 top companies will integrate data intelligence with AI model intelligence to unify governance policies, practices, and technologies for AI models and data.

In typical business activities such as conferenceas, exhibitions, marketing, business travel, communities, business associations, digital systems generally suffer from multiple access portals, fragmented subsystems, poor cross - ssubystem collaboration and information silos. The key pain points are as follows:

Dispersed information and fragmented processes: Information including agendas, exhibitor materials, product catalogs, marketing content, travel schedules, and community data is scattered across multiple systems, posing a challenge to cross-platform collaboration.

Poor intelligence and high operational costs: Operations including content generation, customer recommendations, product matching, travel services, and community operations are mainly manual, driving continuous cost increase.

Obvious cross-language communication barriers: There exists surging demand for cross-language content generation, bilateral communication and real-time translation across global business activities, which cannot be fully satisfied by conventional manual services.

Inefficient supply-demand matching: Without semantic understanding and personalized recommendation function, traditional search systems rely on keyword-based retrieval, leading to the loss of business opportunities.

Difficulty in scaling multi-role, multi-task collaboration: Enterprises need an end-to-end capability covering content generation, data query matching and recommendation, interactive communication and intelligent service, yet lack a unified intelligent framework to support it.

To address the above problems, the WeMeet Multi-Agent AI Business Interconnection Platform was built. Centered on core technologies including agents, semantic vector retrieval and multimodal AIGC, the platform can reshape collaborative workflows for business activities and delivers intelligent linkage across diverse roles, systems and scenarios.

Implementation Plan

The Implementation Plan for the WeMeet Huishen Multi-Agent Business Interconnection Platform is not limited to a single function, but is an intelligent infrastructure system designed for organization-wide business activities. The core development logic can be summarized as follows: leveraging a platform-based architecture to support multi-agent collaboration, using a semantic vector knowledge base to break data silos acros different business scenarios, employing multimodal AIGC to enable business content generation, and integrating a unified API and tool orchestration system to achieve cross-system execution. This builds an intelligent business service system capable of understanding, reasoning and autonomous execution.

1. Overall System Architecture

Centered on six typical business scenarios — conferences, exhibitions, marketing, business travel, communities, and business associations, the platform establishes a unified AI foundation, structured into three layers:

Fundamental Technical Layer: It provides the multi-agent collaboration framework, semantic vector retrieval engine, knowledge middle platform, multimodal content generation engine, and unified tool scheduling system.

Service Engine Layer: It offers general intelligent service modules including intelligent agenda generation, exhibition item analysis, customer recommendation, business travel planning, and community operation.

Scenario Application Layer: It deploys pluggable agents to form specific business modules, which include conference agents, exhibition agents, marketing agents, travel agents, community agents, and business association agents.

This architecture ensures both technical consistency and scenario adaptability, enabling end-to-end intelligent support covering content understanding, knowledge matching, task planning and action execution.

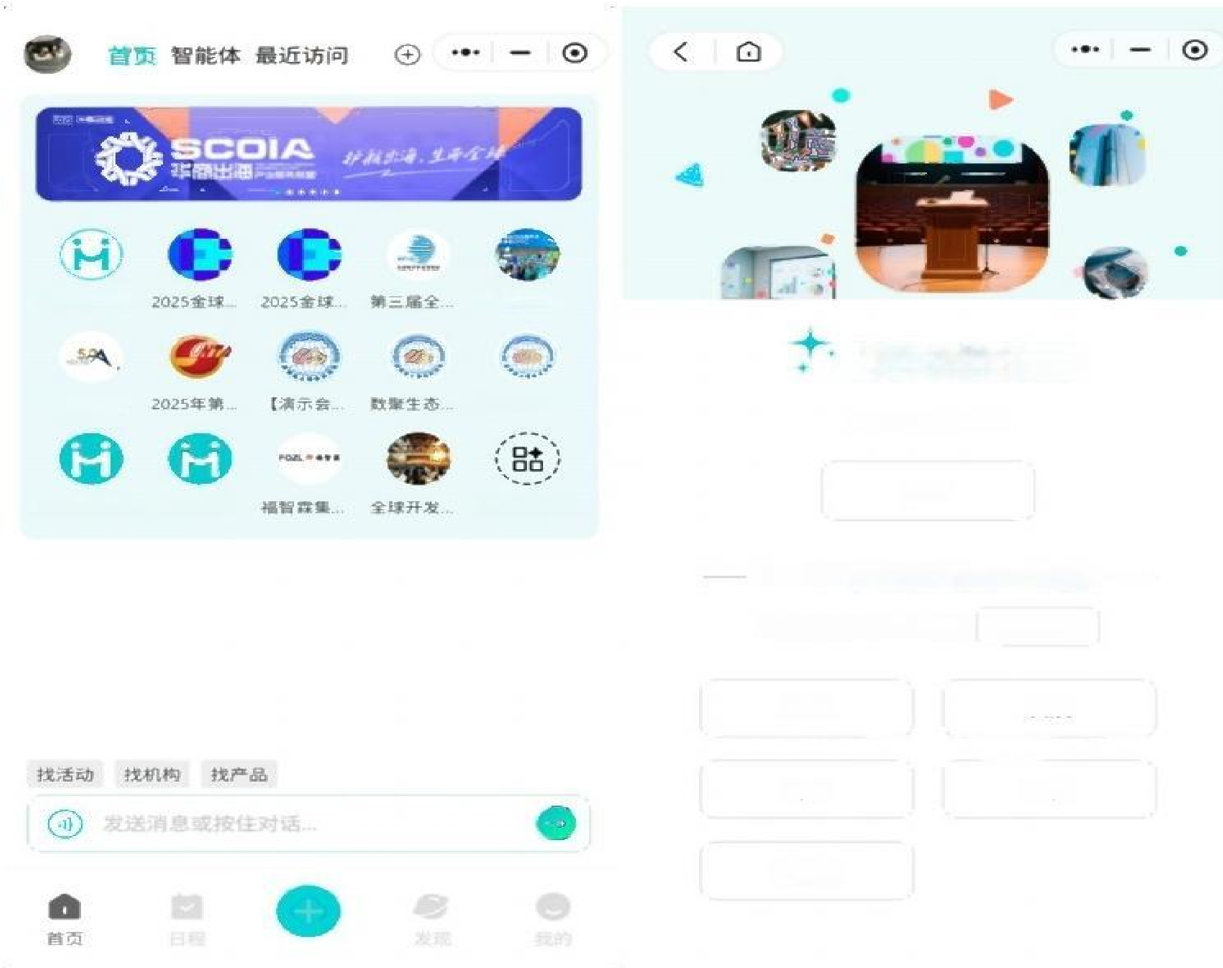


Figure 1 WeMeet Multi-Agent AI Business Interconnection Platform

2. AI Technical Architecture and Operating Mechanism

The platform has a three-tier AI architecture consisting of Perception & understanding, Knowledge Matching and Decision & Execution and operates driven by a self-developed multi-agent collaboration framework.

Tier 1: Perception & understanding Layer

By leveraging the WeMeet business large language model and multimodal recognition technologies, the system can identify text, speech, image and video content, extracting conference topics, product features, customer intentions, and behavioral signals, which helps convert unstructured business information into structured data and automatic identification of tasks, thereby providing semantic input for agent scheduling.

Tier 2: Knowledge & Matching Layer

Leveraging self-developed data vectorization middleware, the platform encodes enterprise information, exhibitor materials, industry resources, user behaviors and trade association knowledge into unified vector representations. This enables cross-language semantic retrieval, business knowledge correlation, and supply-demand identification. This layer not only supports human-initiated information retrieval but also offers core business intelligence functions such as active information push, automatic supply-demand matching and domain knowledge alignment.

Tier 3: Decision & Execution Layer

The self-developed Multi-Agent Framework serves as the core execution component of the platform, with capabilities covering task decomposition, intent analysis, agent collaboration, reflective self-correction, and tool invocation. Based on the task content, the system can automatically orchestrate different agents to work in collaboration. For example, for the task of “generating a marketing plan for Southeast Asia exhibition”, the system automatically coordinates the marketing agent, retrieval agent, AIGC agent, translation agent and business opportunity agent to form a closed-loop execution process that proceeds through five stages: understanding, planning, generation, distribution and business opportunity feedback.

3. Key Business Logic Process Design

In specific business activities, the semantic-driven platform relies on agent scheduling mechanisms to realize the following five typical business collaboration models:

Intelligent conference organization: Intelligent agenda generation, speaker-topic matching, multilingual interactive Q&A and automatic post-meeting summarization and archiving.

Intelligent exhibition matchmaking: Semantic recognition of exhibits, supply-demand intent matching, precommendation push before the exhibition, AI customer service during the exhibition and business opportunity tracking and conversion after the exhibition.

Intelligent marketing growth: Multi-channel content generation, semantic user behavior analysis, automatic recommendation and outreach and refinement and distribution of highly-converting content.

Intelligent business travel itinerary management: Integrated planning of flights, hotels, and conference agendas, intelligent schedule conflict detection, destination resource recommendations, and MICE service route generation.

4. Platform Integration and Construction Model

Based on a microservices architecture and API Gateway design, the platform can integrate with external systems such as CMS, CRM, ERP, and exhibition management systems rapidly. It also delivers single sign-on, multi-role permission permission, and data security governance. In addition, the system is equipped with multilingual, multi-time-zone, and multi-cultural adaptation capabilities, allowing quick deployment fr overseas exhibitions, cross-border marketing, international trade association operations and other scenarios.

Through a unified architecture, open interfaces, and a sustainable intelligent evolution mechanism, the platform can realize scalable application and continuous iteration across different organizations and industries.

III. Project Highlights

(I) Overall Innovation Logic

The core innovation of the project lies in its fundamental logic: agentification, data vectorization, scenario semanticization, and service personification. It reconstructs the intelligent infrastructure for business activities from the ground up. Relying on the self-developed multi-agent framework and multi-layer semantic models, the platform achieves a closed-loop, end-to-end capability covering data understanding and intelligent decision-making. It equips AI with the full spectrum of intelligent capabilities including understanding, reasoning, generation, interaction, and execution across scenarios like marketing, exhibitions, business travel, and community, making it an AI-native business interconnection platform for future needs.

(II) Core Technological Breakthroughs

1. Self-Developed Multi-Agent Collaboration Framework

A three-tier memory system consisting of NoSQL, vector database, and cache is constructed to support instant memory, semantic memory, and long-term knowledge accumulation. Agents that are capable of task planning, reflective optimization, and dynamic tool scheduling can break down complex business tasks and execute cross-system operations, breaking through the limitations of traditional AI that only offer Q&A and content generation.

2. Data Vectorization and Unified Semantic Retrieval Model

A unified data vectorization middleware has been developed to realize semantic indexing and cross-language retrieval of structured and unstructured business data. The project deploys an RRF (Reciprocal Rank Fusion) algorithm to improve recall accuracy, and HNSW (Hierarchical Navigable Small World) dynamic indexing to achieve millisecond-level search for ten-thousand-volume data. In addition, it supports semantic equivalence matching across over 100 languages, with a 42% improvement in matching accuracy and a 68% increase in response speed, greatly enhancing the capacity for international exhibition and cross-border marketing.

3. Personification and Context Awareness of AI Agents for Business Scenarios

Through identity configuration, tone settings, industry knowledge embedding, and memory management, AI assistants are personified, such as AI exhibition assistants, AI marketing consultants, and AI destination advisors and other roles that can act on behalf of enterprises to handle customer consultation, content generation, material push, and lead follow-up over a long term. These assistants evolve from Q&A-based AI into “digital employees” capable of executing tasks.

4. AIGC-Driven Intelligent Content Generation Engine

Based on the WeMeet Huishen large model and combined with related industry corpora, the engine can offer automatic multilingual generation and proofreading of exhibition press releases, investment brochures, conference summaries, product introductions, and other materials, with an accuracy rate of over 91%. Combined with content security detection capabilities, it realizes automated production of business content with high quality and consistency.

5. Full-Link Closed -Loop Data System and Intelligent Feedback Mechanism

The platform establishes a tracking system for behaviors, intentions, conversions, and business metrics, supporting recommendation accuracy validation, model parameter fine-tuning, and self-learning of prompt templates. This creates an adaptive AI system that the more it is used, the smarter it becomes, enabling continuous improvement in model capabilities and business outcomes.

(III) Scenario Uniqueness and Application Innovation

Multi-scenario integration capability: The platform connects marketing, exhibitions, business travel, and communities, forming a closed-loop business growth chain, which covers the whole process of customer acquisition, exhibition conversion, travel support, and community retention.

Global Cross-language Support: The platform supports interaction and content generation across more than 100 languages. With unified semantic vector encoding it also enables users to conduct global business smoothly by communicating in their native language.

Multi-role intelligent collaboration: Marketing assistants, exhibition assistants, and customer service agents can execute tasks collaboratively based on the task scheduling engine, creating an intelligent ecosystem featuring coordinated operation of multiple agents.

Openness and sustainability: The platform provides open SDKs and APIs, enabling enterprises, associations, and developers to build customized AI applications and industry-specific plug-ins with high scalability.

(IV) Comparative Advantages Over Traditional Models

Dimension	Traditional Model	WeMeet AI-Powered Model
Data understanding	Static keyword matching	Semantic understanding &contextual modeling
Customer Acquisition	Manual outreach and ad putting	AI-driven proactive marketing & semantic business opportunity recommendations
Exhibition Services	Human customer service and fixed processes	AI exhibition assistant &multilingual real-time response
Travel Promotion	Manual information integration	Intelligent destination recommendations & automatic itinerary generation
Cost Structure	Labor force-driven	Automated operation cutting costs by over 60%
User Experience	Fragmented information and communication barriers	Full-process multilingual AI companion service

IV. Implementation Results

(I) Overview of Overall Achievements

Since launch, the platform has established a replicable intelligent application system covering six core scenarios: conferences, exhibitions, marketing, business travel, communities, and trade associations, with large-scale deployment in multiple flagship projects. Centered on AI Hui Business Lounge, multi-agent framework, vector semantic retrieval, and AIGC content engine, the platform achieves four major objectives: cost reduction, efficiency enhancement, revenue growth, and quality improvement:

Cost reduction: Service and operation labor costs reduced by 40%-65%; multilingual outsourcing costs reduced by over 60%;

Efficiency enhancement: Cross-language response latency reduced by 70%; business opportunity response speed increased by 3 to 5 times; the hit rate of supply-demand matching improved by 30%-45%;

(II) Implementation Results in Representative Enterprises and Project

1. CIIE Shanghai Events and Conference

Serving as the core intelligent AI collaboration support system for the 2025 CIIE Shanghai Events and Conference, the platform provides services including conference content generation, multilingual interaction, agenda scheduling, intelligent retrieval, and documentation archiving. The system processed over 2,500 AI-powered conference-related conversation, covering scenarios such as conference inquiries, conference and guest recommendations, agenda Q&A, guest introduction, and cross-language communication.

It automatically generated and updated conference content over 320 times, including agenda scheduling, event posters, session banners, and multilingual materials, replacing manual design, copy-pasting, and translation that previously took 1to 3 hours per session, saving the time cost of each material by over 80% on average. It also completed AI-powered translation and updates for over 100 conferences with seamless compatibility between Chinese, English, and other languages, greatly improving pre-event preparation efficiency for international participants.

After deployment, the organizer's previously manual workflows, which relied on copy-editing, translation, and email confirmations, were greatly shortened, with content response time reduced from days to minutes , creating a multi-role collaborative intelligent conference management model.

2. GDPS Global Developer Pioneers Conference

The 2025 Global Developer Pioneer Conference & International Embodied AI Skills Competition was held in Shanghai in the form of an immersive AI event featuring multilingualism, diverse interactive sessions, and international developer community connectivity. Served as the conference’s intelligent interaction and content engine, the WeMeet platform completed over 1,250 AI assistant interactions, primarily for event inquiries, venue route, competition technical explanations, and recommendations for events, technologies, policies, training, and job opportunities.

The system of the platform generated more than 100 sets of conference materials, thematic content, posters and agenda displays, greatly reducing the time previously required for manual editing and coordination. The event also deployed embodied intelligence with AI-powered exhibition for the first time. The platform enabled AI assistants to deliver real-time recommendations on events, technologies, policies, training and jobs, helping participating developers achieve key interaction goals including technical research, scenario matching and partner sourcing.

Validated by this conference, the platform is proven viable not only for conference operations but also provides a new AI application model integrating scenario recommendations, exhibition interaction, and talent matchmaking. It further demonstrates the system's scalability in MICE destination services and industry events.

Figure 2 Typical Successful Deployment Projects

(III) Technical and System Benefit Performance

The hit rate of multilingual retrieval Top-3 has increased by 42%, and the response time of cross-semantic intelligent matching has decreased by 68%;

The three-tier memory system (NoSQL, vector database, and cache) has reduced the misjudgment rate of multi-turn dialogues by approximately 35%;

The automatic task orchestration coverage of SkillChain exceeds 70%, drastically reducing manual configuration costs;

The fine-tuned AIGC content generation model maintains a stable semantic accuracy rate of over 91%, delivering consistent performance for exhibition, investment promotion, product and conference scenarios.

These key metrics prove that the platform can not only replace manual workflows, but also possesses the capability to restructure intelligent businesses at the structural level.

(IV) Social and Industry Value

The platform demonstrates exemplary value in industry empowerment, business goes global, talent cultivation, and digital inclusion:

The platform breaks the limitation of operation only during exhibition period, enabling year-round online investment promotion, digital community maintenance, and international business opportunity matchmaking;

Multilingual AI assistants has become digital foreign trade specialists for Small and Medium-sized Enterprises, lowering cross-language communication barriers and cross-cultural costs;

It drives the transformation of the MICE industry from a human-service-oriented model to an intelligent-collaborative one, enhancing the service competitiveness of city-level destination;

It reduces redundant business travel and paper material usage, embodying the concept of green and low-carbon development;

It gives birth to new jobs such as AI operations specialists, AI content planners, and data intelligence analysts, promoting the structural upgrading of talented personnel.

(V) Conclusion and Prospects

The platform has reconstructed the entire chain of conferences, exhibitions, marketing, business travel, communities, and trade associations, forming a scalable, auditable, and evolvable intelligent foundation:

Short term: Continue to maximize the dividends of cost reduction and efficiency improvement and codify the feasible operational system into practical operation manuals and industry templates;

Medium term: Build a marketplace for industrial knowledge assets and an ecosystem of AI agent plugins to create a low-barrier access for more partners to use the platform;

Long term: Establish a four-tier interconnected AI business network linking cities, industries, enterprises, and individuals, driving high-quality, low-threshold business collaboration worldwide.

Case 28

A Short Video Creation Service Platform Based on Multimodal Understanding and Multi-Agent Collaboration

Shanghai Weiling Information Technology Co., Ltd.

I. Basic Enterprise Information

Founded in December 2020 and headquartered in Zhangjiang, Shanghai, Shanghai Weiling Technology Co., Ltd. is an internet company dedicated to providing leading digital marketing services for enterprises. The company has received strategic investments from Tencent Group, Sequoia Capital, and joint investments from Sky9 Capital, Yunqi Partners, and Lightspeed China Partners. Focusing on the innovative track of AI-empowered marketing, the company integrates intelligent technologies with professional teams to help enterprises achieve both efficiency and quality improvements in their digital marketing transformation. Its flagship product, Weiling · Jisi is an innovative video creation service platform driven by both AI and a professional video production teams, designed to make vedio content generation a daily productivity tool for enterprises. Powered by AI-driven capabilities including hit-content analysis, automatic script generation, highly realistic digital humans, multimodal material understanding, intelligent editing and post-production, and matrix distribution and data feedback, Weiling · Jisi provides enterprises with one-stop video content generation solutions, helping enterprises turn piles of raw footage into high-quality cinematic videos.

II. Case Introduction

(I) Background and Needs

Actually every communication-related scenarios including brand promotion, event documentation, ad putting, and entrepreneur personal IP building, are shifting toward video formats. However, most enterprises have not yet established the production capabilities that match this trend. For instance, video materials often accumulate to hundreds of gigabytes, yet traditional editing workflows take days to process such volumes, as a result, by the time the final vedio is ready, the optimal opportunity for distribution has often been missed. While enterprises are fully aware of the importance of video, they struggle to find a sustainable balance among time, cost, and human resources.

The cost burden of traditional vedio production methods is another critical challenge. Building an in-house team requires long-term investment in manpower, equipment, and management; outsourcing is costly and time-consuming, with extra charges for every revision; and while AI tools can boost efficiency, most rely on rigid templates and often fails to meet brand standards.

	速度	性价比	视频质量	灵活性	可规模性
集思（AI+服务）	★★★★★	★★★★★	★★★★	★★★★	★★★★★
内部团队 视频质量高但人力成本巨大，产能有限且难以规模复制。	★★	★	★★★★	★★★★★	★
广告/创意公司 出品专业但报价极高、流程冗长，灵活性与扩展性差。	★	★	★★★★★	★★	★
自由职业者 单次成本较低且灵活，但产能与质量不稳定，项目管理成本高。	★★	★★★	★★★	★★★	★
AI剪辑工具 成本低且速度快，但效果粗糙同质化，需专业人力调试。	★★★★★	★★★★	★★	★★	★★★

Figure 1 Jisi vs. Other Video Production Methods

Therefore, whether enterprises choose in-house production, outsourcing, or basic AI tools, none can resolve the core pain points of the video era: original materials cannot be systematically understood by the system, content cannot be produced at scale, and multi-platform adaptation costs remain excessively high. The massive volume of footage accumulated by enterprises are difficult to be turned into structured assets, while the efforts invested by teams fail to yield reusable methodologies, resulting in a lack of systematic workflows, standardized processes, and continuous optimization capabilities.

(II) Implementation Plan

The implementation plan of Weiling · Jisi is not a mere aggregation of functional modules, but rather an end-to-end production system centered on AI and built on an engineered workflow, shifting video content from manual production to AI-enhanced scalable production

1. System Development Content

The system development begins with underlying material governance, employing multimodal understanding technology to perform semantic segmentation and tagging on digital assets including historical video, image, audio, document. In this way, information including subject, movements application, emotions, and scenarios of every shot is made retrievable. The establishment of such a material database transforms what was once a pile of accumulated footage into a true video asset library, providing foundational data for hit-vedio structure analysis, script generation, and shot matching.

Figure 2 Jisi Interface for Cutting Long Vedios into Short Clips

On this foundation, the system adopts a Multi-Agent architecture, encapsulating capabilities including script analysis, digital human driving, rhythm inference, style identification, and post-production generation into “atomic capabilities” that are automatically invoked by the scheduling center based on task requirements. These agents can not only process content independently but also collaborate on cross-stage tasks. For example, generating multiple versions of a script, instantly returning candidate footage, synchronously matching BPM, and guiding the post-production workflow to complete beat cuts.

Through no-code workflow orchestration, the entire workflow is managed to enable modular visual management. This transforms the process consisting of hit-content analysis, script generation, digital human dubbing, intelligent automatic editing, matrix distribution, and data feedback into a replicable, reusable enterprise production line.

2. System Architecture Design

The overall system architecture consists of five layers: Input Layer, Understanding Layer, Generation Layer, Distribution Layer, and Feedback Layer.

Input Layer: It receives multi-source data including live streams, events, courses, and advertising materials, and automatically performing preprocessing tasks such as authorization verification, security scanning, video structure segmentation, and audio-video synchronization.

Understanding Layer: It is the core of the system. Based on technologies including knowledge graphs, semantic parsing, sentiment analysis, and shot grammar analysis, it transform raw materials into structured content. For example, when processing a speech vedio, the system can identify logical segments, viewpoint and key quotes, tonal variations, audience interaction highlights, and visual highlights.

Generation Layer: It covers capability chains including script generation, digital human dubbing, automatic editing production, and special effects composition. Script generation is driven by a library of 3-second hook factor for hit vedios, content commitment structures, and CTA frameworks. The digital human system maintains expressive consistency through voice cloning and emotional curve modeling. The automatic editing system automatically completes beat cuts, subtitles, scene transitions, and cover generation based on rhythm curves, shot grammar, and style libraries.

Distribution Layer: It handles multi-platform adaptation, title recommendation, cover version generation, and content scheduling. It generates tailored content based on the differentiated rules of platforms such as Douyin, WeChat Channels, and rednote, ensuring consistency and efficiency across different accounts.

Feedback Layer: It aggregates metrics including view count, bounce rates, completion rates, engagement rate, clicks, and review status into a unified dashboard. Through continuous optimization of script structures of the model, shot selection logic, and rhythm templates, it drives the data flywheel for ongoing iteration.

The entire architecture supports model routing, abnormal bypass switching, gated launch, and hot-swappable capabilities to guarantee system stability and scalability. Industry-grade SLAs cover key indicators including availability, accuracy, latency, and compliance.

3. Business Logic

The business workflow of Weiling · Jisi follows the core pipeline of Understanding, Generation, Reuse, Optimization. The system first captures high-growth content in real time through a hit-vedio analysis mechanism and analyzes its rhythm structure, shot transition logic, and hook factors, and consolidates them into rules and templates. Subsequently, AI combines brand style and industry knowledge to generate multiple versions of script drafts.

Upon receiving the script, the digital human agent automatically synchronizes lip movements and facial expressions according to the emotional curve of the text, achieving human-like and credible delivery. Meanwhile, based on the logic of multi-scene storyboard script, the intelligent automatic editing system aligns different scenes from the footage in terms of content expression, visual mood, and musical intensity to ensure highly consistent brand presentation across all outputs.

Once the finished video is produced, the matrix distribution system generates tailored titles, covers, and subtitle according to the exposure mechanisms of different platforms to enable bulk distribution. After data feedback, the system automatically extracts effective factors through A/B testing and significance analysis and updates elements including opening hooks, shot structures, and rhythm beat sync points into the rule base to form a continuously evolving data flywheel.

III. Project Highlights

The core advantage of Weiling · Jisi lies in its ability to fully systematize the complex video creation chain. The synergy between AI and professional teams is realized by the integrated interplay of the Multi-Agent architecture, multimodal understanding system, rhythm and shot grammar models, digital human expression system, and data flywheel mechanism. This enables enterprises to rapidly produce popular and brand-aligned content across diverse scenarios.

(I) Efficient Clipping of Live Stream and Course Footage

Education enterprises and knowledge service organizations generally possess massive materials of live stream and course footage. However, manual clipping is time-consuming and labor-intensive. Moreover, the information is of high density and difficult to sift. As a result, these materials cannot be promptly transformed into shareable videos, so enterprises cannot fully unlock the commercial value of them.

Built on a multimodal understanding model, Jisi leverages techniques including semantic structure recognition, speech rhythm analysis, emotional fluctuation detection, and audience interaction peak positioning. These enable the system to automatically parses the logical framework of long-form videos and extract highlight clips. It also automatically aligns footage beats based on BPM curves and generates short videos with subtitles and background music.

The livestream footage can be structured on the same day the livestream concludes. By the next day, ready-to-release content can be produced in batches. This transforms livestreams from being a terminal point into the starting point of dissemination, enabling the continuous extension of knowledge-based content.

(II) Mass Production and Continuous Optimization of Advertising Materials

Competition in advertising and marketing has shifted from a contest of creativity to a contest of iteration speed. Traditional ad production suffers from long production cycle, heavy labor costs, and limited versions of output, making it difficult to support today’s advertising workflow of rapid trial-and-error and data feedback.

Leveraging hit content factor libraries, opening hook structures, visual grammar models, and rhythm template libraries, Jisi streamlines ad production into a standardized assembly-line workflow. It generates multiple draft versions based on prevailing hit content trends and automatically completes clip matching and rhythm fitting. The automatic editing system deals with subtitles, background music, and cover design.

Enterprises can thus obtain multiple versions of material for A/B testing in a short time. As data flows back continuously, script structures and shot selection logic are refined iteratively. Advertising creativity evolves into a reusable, continually optimizable content system, reducing advertisement putting costs and improving exposure efficiency.

(III) Digital Production for Entrepreneur and Expert Personal IP

The personal discourse of entrepreneurs and experts is becoming a valuable brand asset, yet in practice, they are often constrained by time, shooting scenarios, and manpower, making sustained output difficult.

Jisi offers highly realistic digital humans, voice cloning, emotional curve models, and lip-sync alignment technology to build ready-to-use digital avatars for entrepreneurs. Through one-time learning, the system can accurately replicates the subject’s tone, speech cadence, and facial expressions, and generates multiple script versions based on industry knowledge. The digital human agent delivers human-like and fluent narration.

Freed from recording schedules or venue constraints, entrepreneurs can maintain high-frequency expression. This enables personal IP production to achieve scalability, delivering steady content generation ability for the brand.

(IV) Content Collaboration Between Sales Matrix and Marketing Teams

Within the sales system, brands desire standardized sales pitches while sales teams need personalized expression. A traditional model often fail to find a balance between them.

Jisi resolves this longstanding conflict . The marketing department provides a standard script, and the AI generates multiple content variations based on each salesperson’s persona, tone, and customer profile. The digital human system enables every sales representative to maintain a distinct image, achieving a fusion of unified logic and personalized expression. Once headquarters uploads source materials, the entire team can produce a large volume of client-facing video materials.

Figure 3 Video Matrix Interface

Through data feedback, enterprises can also track frontline content performance and further refine their pitch system. This achieves full integration between content production and sales scenarios for the first time.

(V) Rapid Production of Event Highlights and Brand Communication

Events, exhibitions, and launch events often carry critical brand publicity missions, but the cycle of traditional highlight production is lengthy and slow, causing brands to miss the peak window of audience attention.

Jisi employs shot grammar models, sentiment analysis, people detection, and keyframe extraction technologies to automatically identify keynote speeches, emotional peaks, interactive highlights, and key brand footage before automatically compiles highlight reels. Professional teams then conduct visual unification based on the AI-generated cut to ensure the final output feature distinctive brand styling. Enterprises can complete full-platform distribution within a day after the event, transforming event communication from post-event summarization to real-time engagement, significantly extending the brand exposure lifecycle.

(VI) Lightweight Innovation in Knowledge Dissemination and Educational Content

Knowledge-based content is characterized by deep logic, slow pacing, and dense information, making it difficult for traditional editing methods to quickly extract key points.

Jisi can truly understand knowledge-based content: it identifies logical segments, extracts key quotes and viewpoints, analyzes emotional intensity, and generates lightweight version of short video based on rhythm models. A one-hour course can be split into a dozen short clips, making knowledge dissemination more inclusive and accessible. Educational institutions can deliver front-end teaching and sustain back-end continuous dissemination, turning knowledge production into a process of accumulating lasting value.

IV. Implementation Results

(I) Economic Benefits: Quantitative Results Centered on the Real-World Performance of Jiulongtu

In the practical application of Jiulongtu, the automatic editing capabilities of Jisi significantly improved the operational efficiency of the WeChat Channels. Previously, producing one short video took three to five hours but now the process only lasts 15 to 25 minutes, lifting overall output capacity by roughly 8 to 10 times. Daily output per employee rose from 1-2 videos to 12-15 videos, establishing a stable daily posting schedule. With the system automatically handling post-production tasks including subtitling, background music matching, and cover design, production costs dropped by about 65%, and labor input for operation decreased by over 70%.

Higher output directly drove business growth: within three months, Jiulongtu's WeChat Channels organic traffic increased by 120%, vedio completion rate rose by 30%, and customer acquisition costs from private domain traffic via short videos fell by about 40%. Jisi became the core driver of growth for the WeChat Channel.

(II) Overall Outcomes Across Multiple Clients: Statistical Results from Long-Form Video Slicing in Diverse Scenarios

According to the statistics, for multiple knowledge-based educational clients, the time required for long-form video clipping has been reduced from the original 2 to 3 days to just 30 to 60 minutes. For e-commerce businesses, conventional short-video montage production, which previously took 4 to 8 hours, has been shortened to roughly 20 minutes. Meanwhile, the content quality has improved simultaneously: the 3-second retention rate for short videos has risen from 52% to 68%, the completion rate from 18% to 28%, and the combined favorite-and-share ratio from 4.2% to 5.9%.

Cross-platform release efficiency has also been greatly enhanced. Platform adaptability rate rose from 88% to 98%, and the entire pipeline from final cut to official release has been compressed from 10 hours to just 3 hours. The feedback cycle for campaign data collection shrank from 24 hours to 1 hour, enabling a complete vedio creation, official launch， and iterative optimization cycle within a single day.

Overall, clients have achieved an average 60% reduction in production costs and a 70% decrease in marketing manpower input. For the first time, enterprises gain sustainable capabilities for video-driven business operations.

(III) Social Benefits Outcomes

Jisi transforms professional video production from a high-barrier, specialized skill into a universally accessible digital infrastructure. Small and medium-sized enterprises no longer need to invest in expensive dedicated teams to produce professional-level videos. Individual creators and sales teams can also deliver ordinary content through digital human technology.

In terms of inclusive knowledge dissemination, educational institutions and industry experts leverage Jisi to break lengthy courses into lightweight, information-dense short videos. This significantly lowers the threshold for public access to professional knowledge and improves the efficiency of knowledge dissemination.

In terms of equal right to expression, digital human technology gives users who are uncomfortable on camera the same expression experience as professional streamers. Voice cloning technology enables people with dyslexia or verbal communication difficulties to generate natural audio and visual content automatically from text input.

Moreover, the platform’s industrial-grade SLA and compliance framework guarantee safe and controllable AI-generated content. It provides an auditable and traceable AI-driven content production model, serving as a critical technical foundation for the sound development of the digital content industry.

Case 29

BOOKSGPT: A Professional Foundation Model Empowering the Transformation and Upgrading of the Entire Publishing Chain

Shuchuan (Shanghai) Enterprise Development Co., Ltd.

I. Basic Enterprise Information

Established on September 30, 2022, Shuchuan (Shanghai) Enterprise Development Co., Ltd. serves as the comprehensive management headquarters, operational headquarters, and innovation technology center of DCG. It oversees capital operations, technological innovation, and international business, with a commitment to driving the digital and intelligent transformation of the traditional publishing industry. Centered on its self-developed RAYS Publishing Convergence Service Cloud Platform, the company embeds intelligent QR codes with big data tagging capabilities to upgrade print books into smart interactive print books. Such books support two-way interaction, user services, data analysis, and online monetization. Its parent company, DCG, ranks among China's Top 30 High-Growth Cultural Enterprises, a "Little Giant" among National-level Specialized, Sophisticated SMEs and Forbes China Top 50 AI Technology Company. The Group has officially launched its Hong Kong IPO initiative. In 2024, the company designated AIGC as a key strategic priority. Leveraging its self-developed big data algorithms and natural language generation technologies, it successfully launched BOOKSGPT, a vertical large language model tailored for the publishing industry. Adopting the "One Book, One Model" strategy, the model empowers the transformation and upgrading of the entire publishing chain, and has become a benchmark case in Hongkou District for promoting the integrated development of culture and technology.

II. Case Introduction

(I) Background and Needs

1. Industry Transformation Trends

With breakthrough advances in generative artificial intelligence, large language model technologies are accelerating their penetration and integration across diverse vertical sectors. At present, publishing institutions in China face three core challenges in rolling out AI application: improving the quality and efficiency of content production, optimizing distribution and sales strategies, and advancing the digitalization of books. To systematically address these issues, it is essential to move beyond the technical constraints of general LLMs and build a specialized model system that truly understands the operational rules of the publishing industry.

2. Key Industry Pain Points

The application of AI in the publishing industry still suffers from notable shortcomings.

In the content production workflow, topic selection and planning rely heavily on editorial experience, with preliminary research cycles lasting 2 to 3 months. Manual oversight is necessary for image-text adaptation and terminology standardization during content creation. During proofreading and editing, lengthy texts suffer from information loss and missed errors.

In marketing, traditional distribution models struggle to adapt to the new media content e-commerce environment, facing high traffic costs and difficulties in reaching target users.

In reader services, most digital resources are merely simple text-scanning conversions, lacking interactivity and failing to meet readers' demand for in-depth services such as video explanations, personalized exercise sets, and intelligent Q&A.

3. Limitations of General Large Language Models

General-purpose LLMs exhibit significant limitations in professional publishing scenarios. First, insufficient domain expertise. Trained primarily on publicly available internet data, general-purpose LLMs lack deep understanding of publishing-specific terminology and standards. They are better suited for open-domain, low-specialization tasks, and cannot handle the highly specialized, in-depth tasks prevalent throughout publishing workflows. Second, excessive resource consumption. General-purpose LLMs rely on trillions of parameters to build general cognitive capabilities. Such resource-intensive model is incompatible with localized deployment within publishing institutions. Third, prominent data security risks. Publishing operations involve sensitive data such as unpublished manuscripts and materials for major topic selection filings. The public-web training model of general-purpose LLMs carries potential copyright infringement risks. Moreover, while they offer common data security designs, they lack industry-specific depth, leaving publishing institutions to bear additional costs for copyright clearance and privacy protection.

(II) Implementation Plan

To systematically address the above challenges and satisfy the publishing industry's demands for professionalization, security, and intelligence, DCG had the foresight to launch research and development of intelligent book session services back in 2019. In 2024, the Group's RAYS Digital Human Intelligent Conversation Algorithm was included in the fourth batch of domestic deep-synthesis algorithm filings published by the Cyberspace Administration of China (CAC). In the same year, the Group officially launched BOOKSGPT, China's first professional AI large language model exclusively built for the publishing industry. In April 2025, BOOKSGPT obtained the filing certificate for generative AI service issued by the CAC. With these achievements, DGC has become one of the few enterprises in China holding dual filings for both the generative AI service and the deep-synthesis algorithm.

BOOKSGPT provides a full spectrum of foundational capabilities covering publishing operations and reader interaction services. Its core mission is to help the publishing industry build high-quality intelligent publishing operation systems, online book content service systems, and new media book marketing systems.

1. Development of a Technology Foundation Platform Dedicated to the Publishing Industry

Through its self-developed BOOKSGPT, DGC delivers an AI capability infrastructure that is deeply coupled with publishing business logic. The system employs a model architecture optimized specifically for publishing scenarios, integrates industry data to build a comprehensive knowledge system, and incorporates robust security and compliance mechanisms to ensure both effectiveness and security in practical applications.

2. Intelligent Empowerment of Core Publishing Workflows

Intelligent publishing operations: Through an AI-powered toolchain, the platform reshapes the entire production workflow covering topic selection and planning, content creation, and three-round review and proofreading process, resolving the longstanding challenge of balancing production efficiency and content quality.

Intelligent distribution and marketing: Deeply embedded within the new media ecosystem, the platform leverages automated generation of marketing materials and precise optimization of ad putting strategies to break through the traffic bottlenecks inherent in traditional distribution models.

Digital-intelligent book services: Guided by the "One book, One model" philosophy, the platform creates tailored digital resources and interactive services for each physical book, transforming static printed volumes into dynamic knowledge service platforms.

2. Full-Process Security Assurance

It supports private deployment and zero-trust permission management to ensuring compliant use of core business data.

III. Project Highlights

Figure 1 Core Advantages of BOOKSGPT v.s. General-Purpose LLMs

Figure 2 Complete Publishing Process Empowered by BOOKSGPT

(I) Key Technological Breakthroughs

1. Efficient Model Architecture Technology for the Publishing Industry

Employing a Mixture of Experts architecture, the platform leverages a 120-billion-parameter scale, combined with LoRA model compression technology. This ensures professional-grade model performance while significantly reducing computational resource requirements, enabling efficient operation on publishing institutions' on-premises server, effectively overcoming the high resource consumption and deployment hurdles of general-purpose LLMs.

2. In-depth Integration of Publishing-Domain Knowledge Technology

A dedicated knowledge graph for publishing has been constructed, with training data integrating 1.6 million books, 2.63 million bibliographic records, 60 million book reviews and marketing copy materials, and 300 million reader profiles and behavioral records. This endows the model with deep domain expertise, delivering significantly superior accuracy and contextual understanding compared to general models when handling specialized publishing tasks.

3. Reinforcement Learning from Human Feedback (RLHF) Alignment Technology

The model training incorporates an RLHF mechanism and integrates specially the feedback from senior editors' review and proofreading decisions. This enables the model to deeply comprehend and internalize publishing standards, achieving high alignment with industry logic in tasks such as topic evaluation, content compliance assessment, and copyright risk early warning, ensuring both professionalism and safety of AI-generated content.

3. Full-Process Data Security and Compliance Assurance Technology

A security framework compatible with private deployment and zero-trust access control has been designed. Through a data desensitization pipeline, the system provides full-lifecycle protection for sensitive information including core business data and unpublished manuscripts, addressing publishing institutions' critical concerns regarding data sovereignty and regulatory compliance.

5. Dual-Model Collaboration Technology for Precise Proofreading of Long-Form Publishing Texts

Targeting the unique challenge of long-form text proofreading in publishing industry, the platform innovatively develops a collaborative mechanism connecting a proofreading model and a reasoning model. The proofreading model focuses on rapid and precise localization of surface-level errors such as character, word, and grammatical mistakes by suppressing invalid token generation rates. Focuses on complex issues like factual consistency and logical coherence, the reasoning model constructs a multi-stage reasoning chain for in-depth verification. Guided by dedicated reward functions built on PPO reinforcement learning algorithms, the dual-model collaboration effectively resolves the information dilution suffered by general-purpose models when processing long-form texts.

6. Multimodal Content Generation Technology

In text generation, a mechanism that combines pre-training and fine-tuning enables the model to accurately emulate dozens of professional writing styles.

In image generation, a technology stack combining Diffusion Models, LoRA fine-tuning, and ControlNet has been employed to build a publication-grade image library encompassing over 200 sub-styles. The key innovation lies in the fact that all generated images are trained on a copyright-cleared publishing image database, systematically avoiding copyright infringement risks at the technical source.

7. Platform-Based Intelligent Marketing Material Generation Technology

Through automated analysis of massive high-quality content from mainstream content platforms, the platform distills the content structures and success factors for each platform. Through targeted training, the large models can automatically understand and adapt to styles of different platforms, enabling batch generation of diverse marketing materials including copy, images, and video scripts. This realizes automated and platform-aligned production of marketing content.

8. Data-Driven Closed-Loop Precision Advertisement Putting Decision Technology

By processing user behavioral data to build refined profiles, the model generates initial advertisement putting strategies based on industry knowledge graphs and tracks core metrics in real time to automatically optimize delivery parameters, the platform establishes a data-driven intelligent marketing decision-making capability, forming a closed-loop workflow covering data collection, strategy formulation, performance monitoring and iteractive optimization.

9. Personalized Digital Resource Generation Technology

Adopting the "One Book, One Model" decomposition and analysis approach, the system performs deep textual analysis of each book to understand its style, themes, and logical structure. Through Retrieval-Augmented Generation technology, it dynamically calls upon external knowledge bases such as structured book catalogs and index data, correlates reliable information via knowledge graphs, and automatically generates accompanying digital resources including practice exercises, mind maps, significantly extending the value of printed books.

10. Interactive Book Retrieval and Q&A Technology

Adopting Retrieval-Augmented Generation technology, a system consisting of book knowledge base, semantic parsing, and multi-turn dialogue is built. Book content is transformed into a structured knowledge base, with catalog hierarchies, terminology systems, and logical frameworks parsed and indexed. Through semantic understanding technologies, the system captures the core intent of user queries and precisely locates relevant information within the knowledge base. It also supports multi-turn dialogue, enabling cross-chapter information reasoning based on context, and ensures traceability of answer by citing original text passages from the book.

(II) Uniqueness of Application Scenarios

1. Full-Process Auxiliary for Editorial Production

Editors can use AI tools on a unified platform to complete the entire workflow including topic selection report drafting, content-assisted creation, and intelligent proofreading, which significantly boosts both efficiency and standardization of content production.

Figure 3 Empowerment of Editorial Workflow

2. Data-Driven Intelligent Marketing for New Media Books

By inputting basic book information, marketers can leverage the system to automatically generate marketing material packages tailored for platforms such as Douyin and rednote. According to real-time data feedback, the system automatically optimizes advertisement putting strategies to deliver precise audience reach and higher conversion rates.

Figure 4 Empowerment of Distribution Business

3. Knowledge Service Enhancement Based on the “One Book, One Model” Strategy

By scanning the QR code embedded in the book, readers can have access to exclusive knowledge spaces, where they can engage in intelligent Q&A, access supplementary materials, and complete personalized exercises for an immersive and in-depth reading experience.

Figure 5 Empowering Reader Experience

IV. Implementation Results

Since its deployment, BOOKSGPT has achieved substantial progress across multiple key procedures of the publishing industry, delivering efficiency gains, sales growth, and service upgrades for partner publishing institutions.

In terms of editorial production efficiency, AI-powered editorial assistant tools have been adopted by nearly 400 publishing and distribution organizations, with approximately 15,000 editors registered for use. As of October 2025, cumulative usage of AI digital employees has exceeded 1.02 million interactions, with total usage time surpassing 400,000 hours. In specific business scenarios, AI topic selection improves the topic planning efficiency by over 150%, generating a complete topic selection in just 1–3 minutes; AI proofreading delivers a 15.8% performance improvement when processing 128K long-form texts, with an average accuracy of 40.2%, representing a 19.6% improvement over general-purpose models; AI illustrator supports over 200 artistic styles and generates high-resolution illustrations that meet printing standards; AI-optimized advertisement putting contributes to an overall 150% increase in book marketing efficiency.

In terms of book marketing, the AI-driven new media marketing solution has delivered outstanding results for partner organizations. For The Seniors University of China, the official Douyin account’s debut live stream achieved 4.595 million impressions and over 1.5 million views, ranking among the Top 3 on the live stream popularity chart; In collaboration with Changjiang Children’s Publishing House, GMV reached RMB 674,000 with an ROI conversion rate of 5.1, and a single-session sales record of RMB 18,000; In collaboration with Bookuu (Bookuu.com) , GMV exceeded RMB 850,000, with a single-session peak of RMB 72,000; For Tianjin Ancient Books Publishing House, a new account achieved GMV of RMB 145,000 starting from zero, successfully breaking through the industry bottleneck of low conversion rates for high-priced ancient books.

In terms of reader services and resource value enhancement, over 1.2 million book titles have been served cumulatively, demonstrating significant achievements of the “One Book, One Model” strategy. For example, in collaboration with Chutian Publishing House, the commemorative edition of *The World’s No. 1 Sword—The Sword of King Goujian of Yue* was upgraded into an interactive smart interactive print books empowered by BOOKSGPT. AI reading digital humans create deep connections among readers, authors, and editors, substantially enhancing the reading experience and service capabilities

Case 30

AI-driven SD-WAN Information Technology Application Innovation Networking and Intelligent Operations Practice

Shanghai Dienian Information Technology Co., Ltd.

I. Basic Enterprise Information

Shanghai Dienian Information Technology Co., Ltd., known as 7x Networks was established in July 2017, with its headquarters in Shanghai and branch offices in Hangzhou, Guangzhou, Wuhan, and Hong Kong. The company is a National High-Tech Enterprise and a Shanghai Specialized and Sophisticated SME. Specializing in the Software-Defined Networking domain, the company leverages its core product portfolio, "7x ONE", to build secure, agile, and intelligent global connectivity paradigms, driving enterprises' successful transformation into a software-defined world.

The company's core team consists of senior specialists from leading domestic and international enterprises including Cisco, Alibaba, Siemens, and China Telecom, possessing solid technical expertise and extensive industry service experience. In terms of technological innovation, 7x Networks has been named a Representative Vendor in the Gartner's *Market Guide for Zero-Trust Network Access, China, 2025*, and recognized as a Representative Vendor for AIOps in the 2025 Gartner Hype Cycle for Infrastructure and Communications Technology, China. Additionally, the company has received multiple awards from the China Academy of Information and Communications Technology, including the "2024 SD-WAN IT application innovation Pilot Program Outstanding Case Award" , the "SASE Best Practice Award", the "SD-WAN Service Quality Evaluation" Excellence-Level Certification, and the industry's first SD-WAN IT application innovation Security Solution Evaluation Certificate, fully demonstrating its technical strength and industry recognition in the convergence of SD-WAN IT application innovation.

Looking to the future, 7x Networks will continue to deepen its expertise in Software-Defined Networking technologies, driving the evolution of SD-WAN from a connectivity tool to an intelligent business infrastructure. This will support enterprises in business expansion and agile transformation amid the trends of globalization, multi-cloud collaboration, and remote work, while consistently contributing to national information security and the development of the IT application innovation industry.

II. Case Introduction

(I) Background and Needs

Driven by the national IT application innovation strategy and the digital upgrading of industries, enterprises are actively pursuing comprehensive localization of their IT infrastructure. However, this transition poses multifaceted challenges to existing WAN architectures, including insufficient controllability, performance bottlenecks, and security risks.

Specifically, many enterprises still rely on core network equipment and technology stacks from foreign vendors, which not only brings about supply chain and data security risks but also makes it difficult to meet the requirements of self-reliance and strength in science and technology. Meanwhile, with surging business volumes and the widespread adoption of cloud-based applications, traditional network transmission suffers from low efficiency, high latency, and significant packet loss, making it difficult to ensure the smooth operation of critical applications including ERP systems, remote diagnostics, and collaborative R&D. In terms of security, network management policies are fragmented, user access system lacks effective identity verification, and defensive measures lag behind dynamically evolving new threats such as zero-day vulnerabilities and stealthy attacks. Furthermore, operations remain heavily dependent on manual intervention, resulting in low efficiency and high costs, with a lack of intelligent early warning and automated control capabilities. Enterprises cannot deal with trouble in large-scale distributed networks, compromising business continuity.

To address these challenges, 7x Networks' SD-WAN has emerged as an ideal technology choice for enterprise Information Technology Application Innovation transformation. It integrates the flexibility of software-defined networking, self-developed WAN optimization technologies, integrated SASE architecture, and high degree of autonomy and control. In collaboration with 7x Networks, Shanghai Pharmaceutical Holding Co., Ltd. has jointly built a new-generation enterprise network that is secure, of high-performance, intelligent, and fully independently controllable through an AI-powered IT application innovation SD-WAN solution.

(II) AI-Driven 7x ONE

Traditional network management remains stuck in a stage of passive response, manual reliance, and low-efficiency and redundancy, failing to meet the demands for high reliability, intelligence, and security for digital transformation. These fundamental issues faced in practical applications stem from the profound contradiction between the static nature of traditional network architectures and dynamic business requirements. These pain points are not superficial phenomena but represent systemic challenges caused by network technology that fail to keep pace with growing business complexity.

In terms of path selection, the static routing protocols, which traditional SD-WAN relies on, such as OSPF and BGP, are essentially passive response mechanisms based on preconfigured rules. They are incapable of sensing micro-level changes in network state such as millisecond-level latency fluctuations or sudden congestion in real time, and can only make decisions based on periodically updated link-state databases, resulting in severely delayed path adjustments. For example, when there is a sudden surge in video conference traffic, traditional mechanisms must wait for link-state protocol convergence, typically taking dozens of seconds, making it impossible to avoid congestion in advance.

The limitations in security threat detection stem from inherent flaws in rule base. Traditional security tools rely on signature databases of known attack patterns. However, emerging APT attacks and zero-day vulnerabilities often disguise themselves by using legitimate protocols, making it impossible for rule base to identify such anomalous activities. For example, insider threats usually move laterally by using legitimate credentials, so traditional detection methods require manual log analysis to uncover such activities.

The inefficiency of network operation and maintenance fundamentally arises from an imbalance between data scale and processing capacity. Modern networks generate terabytes of telemetry data per second, including traffic metrics and device logs, making it impossible for manual inspections to extract meaningful information from the massive datasets. The sources of failure often involves multi-device interdependencies, for instance, link packet loss may trigger application timeouts, which in turn activate security policy blocks. Traditional approaches require cross-team collaboration for troubleshooting, typically taking hours to locate the root fault.

Technology research on these pain points has driven the convergence of SD-WAN and AI. This integration is not a simple superposition of functions. Instead, it reconstructs the SD-WAN control plane through data-driven decision-making, upgrading the network from a passive framework responses passively to an autonomous system capable of proactive identification, prediction and optimization. An SD-WAN without AI is like driving solely by looking in the rearview mirror. AI-enabled SD-WAN gains panoramic perception and predictive planning capabilities, representing a core technical breakthrough to cope with the dynamic nature of modern networks.

7x Networks stands as an active promoter and leader in this convergence trend. With its AI-powered IT application innovation SD-WAN solution, 7x Networks is driving the evolution of enterprise networks from simple connectivity pipes to intelligent business enablement platforms.

1. Intelligent Path Selection: 7x PATH Algorithm Delivers Dynamic Optimization

Through its self-developed 7x PATH intelligent path algorithm, the system collects multidimensional data including link latency, packet loss rate, and bandwidth utilization in real time, and dynamically models network states via reinforcement learning. The AI engine not only responds to real-time congestion but also predicts bottleneck risks based on historical traffic patterns, proactively allocating optimal paths for critical applications such as video conferences and remote collaboration. Predictive optimization, what was once impossible with traditional static policies, has become a normal operation, significantly enhancing business fluency and reliability.

2. Security Threat Detection: XDR Intelligent Computing Engine Reshapes Security Defense System

Equipped with an intelligent threat detection engine 7x XDR employs unsupervised learning to establish baselines for normal traffic behavior, enabling the detection of unknown threats such as emerging APT attacks and zero-day vulnerabilities. Once risks are detected, the system automatically activates such as traffic isolation and security policy adjustments, compressing threat response from hours of traditional manual handling to seconds, achieving closed-loop automation of security capabilities.

3. Device Health Management: Predictive Maintenance Prevents Business Interruption

Through time-series anomaly detection algorithms, the cloud-based intelligent management platform continuously analyzes device metrics such as CPU utilization, capacity consumption, and temperature to identify potential failure patterns, including disk failure and memory leakage trends. AI provides early warnings of hardware risks, shifting operations and maintenance from post-failure repair to pre-failure intervention, minimizing business downtime to the greatest extent possible.

4. Intelligent Operation and Maintenance System: AIOps Drives the Upgrade of O&M Paradigms

Through the 7x AIOps intelligent O&M hub and the 7x Service system, the platform integrates multi-source data including logs, traffic, and performance metrics to break down data silos prevalent in traditional frameworks. Leveraging machine learning, it delivers capabilities such as dynamic threshold alerting, root cause correlation analysis, and automated deployment. This frees O&M staff from repetitive tasks to focus on strategic optimization and value creation, truly transforming the O&M model from passive firefighting to proactive prevention.

Supported by the above measures, the overall network architecture has undergone a profound transformation from passive response to proactive prediction, and from human-dependent to algorithm-driven operations. The ultimate outcome is an intelligent network that is autonomous and controllable, secure and reliable, and equipped with continuous evolutionary capabilities.

(III) Implementation Plan

Figure 1 7x ONE IT application innovation Cloud Management Platform

Core Architecture of 7x ONE:

Intelligent Control Layer (the Brain): Deployed on a SaaS model, the 7x SD-WAN IT application innovation cloud management platform enables centralized, unified control over enterprise network devices, links, policies, and security. The platform has obtained National Level-3 Classified Protection Certification and IT application innovation certification, demonstrating its autonomy and controllability.

Intelligent Forwarding Layer (the Arteries): The 7x SD-WAN intelligent backbone network serves as the global data forwarding plane. It incorporates the self-developed 7x PATH intelligent path selection algorithm, which monitors the status of global POPs in real time and automatically selects the optimal, lowest-latency transmission path for each business flow.

Intelligent Access Layer (the Terminals): Deployed at each site, the new-generation intelligent IT application innovation CPE devices adopt KylinOS and are equipped with 4G/5G wireless backup links, self-developed firewalls, and SM series encryption chips, ensuring secure, reliable, and compliant edge access.

Figure 2 7x ONE Service — One-Stop Operation and Maintenance Service

Core Functions of 7x ONE Service:

Asset Management: Asset tree structure definition, asset scanning, integration with electronic approval workflows, asset information configuration, and more.

Monitoring & Alerting: Network and performance monitoring, traffic analysis, log analysis, end-to-end observability, diverse static and dynamic alerting mechanisms, etc.

Automated O&M : Automated inspections, automated job execution, automated QoS, fault self-recovering, and other services that unlock operations productivity (with a built-in ticketing system).

Intelligent Reporting: Intelligent O&M reports and an O&M visualization cost center dashboard.

A Full-Platform Solution built for Shanghai Pharmaceutical Holdings Co., Ltd., based on the 7x ONE.

Intelligent WAN: The 7x SD-WAN backbone network, consisting of PoP nodes and intelligent probes covering 53 major cities nationwide, delivers agile and stable WAN connectivity. It supports multiple WAN access types including Internet, ADSL, MSTP/MPLS, and 4G LTE, ensuring access quality of end-to-end business.

High-Availability Network: Branch offices adopt an HA deployment model with a dual-Internet and dual-device redundancy architecture to eliminate single points of failure and ensure continuous business application.

Internet and 4G Full-Network Compatibility: Built on the 7x SD-WAN intelligent backbone network, independent pharmacies adopts a dual-link high-availability design featuring Internet and 4G full-network compatibility, enabling intelligent redundancy switching to ensure uninterrupted store operations and deliver an optimal application access experience.

Agile Deployment and Second-level Provisioning: Leveraging the flexibility of software-defined networking, zero-touch deployment, and second-level SD-WAN service provisioning, it provides a foundation for rapid future business expansion of pharmacies.

Global Visualized Management: A unified cloud management platform enables centralized management and orchestration of all 7x CPE devices. Paired with visualized O&M components, it supports real-time monitoring of device status and link conditions.

Round-the-Clock Network Inspection: The 7x after-sales O&M team provides round-the-clock uninterrupted network inspection services. Through intelligent WAN probes, the system monitors the entire network in real time, promptly notifying on-site personnel of any anomalies to prevent reactive operations and maintenance.

III. Project Highlights

In terms of integration and innovation of technology, the platform is built around SDN technology, deeply integrating three cutting-edge technologies including SD-WAN, SASE, and NGFW to create a cloud-network collaborative security architecture. This architecture breaks down the traditional boundaries between networking and security, achieving deep synergy between intelligent traffic scheduling and security protection. It builds a all-round, multi-layered cloud-network security defense barrier for enterprises to address complex and dynamic business scenarios.

In terms of intelligent O&M management, the AI-powered integrated intelligent operations hub and 7x ONE Service possess core capabilities of global visualization, automated collaborative management, and data-driven intelligence. Together, they help enterprises transcend the limitations of traditional O&M models, enabling a leap from passive response to proactive value creation.

In terms of IT application innovation empowerment assurance, the company actively responds to the national IT application innovation strategy by achieving full-chain localization adaptation. Both the cloud management platform and edge intelligent IT application innovation CPE devices leverage domestically developed autonomous and controllable technologies, fully adapted to domestic operating systems, chips, and other core components. The platform has has obtained National Level-3 Classified Protection Certification and IT application innovation certification, ensuring security and controllability throughout the entire network lifecycle, eliminating reliance on foreign technologies, and consolidating a solid foundation for data security in critical industries.

With its three-in-one competitive advantage of technology integration, IT application innovation empowerment, and intelligent O&M, 7x Networks helps enterprises achieve secure and efficient overseas expansion and seamless multi-cloud transformation, serving as a trusted partner in enterprise digital transformation.

IV. Implementation Results

The deployment of this solution has delivered significant economic and social benefits for Shanghai Pharmaceutical Holdings Co., Ltd.:

1. Localization and IT application innovation Compliance: The SD-WAN IT application innovation networking solution customized by 7x Networks for Shanghai Pharmaceutical Holdings fully meets the requirements of localization and IT application innovation policies. By adopting autonomous and controllable software and hardware technologies, it reduces reliance on external technologies while enhancing network security and controllability.

2. 7x PATH AI-Driven Intelligent Path Detection: Through its self-developed intelligent algorithm, the system automatically detects network paths and selects the optimal PoP for connection, significantly improving network transmission efficiency and reducing latency and packet loss. Business network downtime rate has been reduced by over 75%.

3. Comprehensive Security Protection: The solution is equipped with a rich suite of security capabilities including intrusion detection, DDoS mitigation, WAF and more. It supports SM series encryption algorithms, delivering refined control and comprehensive protection to ensure secure and compliant data transmission, meeting the high standards of the pharmaceutical industry.

4. Intelligent O&M Platform: The transition from traditional O&M to AI-driven intelligent O&M is enabled by an integrated intelligent operations hub built on AIOps technology. This delivers global visualization, automated collaborative management, and data-driven intelligence, raising O&M efficiency while reducing operational costs.

5. O&M Service: Leveraging the diverse expertise of O&M specialists combined with the 7x AIOps intelligent operations platform, the solution rapidly empowers users' O&M scenarios through services such as automated

inspections, automated job execution, automated QoS, an AI-powered intelligent ticketing system, and automated ticket analysis. It automates repetitive O&M tasks, unlocking operations productivity.

The successful practice of this case demonstrates high replicability and broad promotion value.

Broad Industry Applicability:

1. Industries with high compliance requirements: Such as government affairs, energy, healthcare, and other sectors which are subject to strict regulatory oversight regarding data sovereignty and security.

2. Organizations with multi-branch structures: Such as retail, education, manufacturing, logistics, and other industries with extensive offline stores, which can rapidly realize unified network management and agile deployment.

3. Enterprises undergoing digital transformation: Those actively embracing multi-cloud and hybrid cloud strategies, or expanding businesses overseas, requiring global network interconnection and intelligent assurance.

High Technical Replicability:

1. High productization degree: The 7x ONE platform offers standardized, modular services, allowing flexible configuration based on enterprise scale with short implementation cycles.

2. Excellent ecosystem compatibility: The software-defined core architecture is highly open and adaptable to a wide range of domestic software and hardware environments, lowering system migration barriers.

3. Mature delivery model: The platform provides full-lifecycle services covering solution design, deployment, and around-the-clock intelligent O&M to ensure successful project implementation.

Strong Referential Value: With its three-in-one competitive advantage of technology integration, IT application innovation empowerment, and intelligent O&M, 7x Networks helps enterprises achieve secure and efficient overseas expansion and seamless multi-cloud transformation, serving as a trusted partner in enterprise digital transformation. It offers a proven, comprehensive paradigm and solution for SMEs facing similar challenges in localization replacement and digital transformation, delivering strong demonstration effects and significant industrial promotion value.

Case 31

Application of Multi-Sensor Fusion Perception Technology in Production Scenarios of Enterprises in Bulk Commodity Supply Chain

Shanghai Zhixinchuang Information Technology Co., Ltd.

I. Basic Enterprise Information

Established in May 2023, Shanghai Zhixinchuang Information Technology Co., Ltd., is certified as a Shanghai Technology-based SME. The company focuses on industrial AI applications for enterprises across the entire bulk commodity value chain including mining, steel, coal chemical, and ports, covering the full chain of production, manufacturing, processing, logistics transportation. The founding team brings together seasoned professionals from Chinese tech giant, well-known startups, and university research institutes with years of AI application experience, possessing deep expertise in practical deployment. The company’s flagship product, the Multi-Sensor Fusion Belt Surface Health Management AI Monitoring System, is one of the few solutions capable of simultaneously monitoring belt surface damage and belt longitudinal tear. The product was awarded the Third Prize in the 10th Qingpu Entrepreneurship Competition. To date, this AI solution has been deployed and operationalized at multiple renowned bulk cargo ports across China.

Leveraging multi-sensor fusion technology, integrated fusion algorithms, and robust engineering design, the system delivers real-time monitoring and early warning for various types of belt damage, including bulges, holes, wire exposure and repairs, and longitudinal belt tearing. It enables real-time, high-frequency, high-precision intelligent detection and identification of belt conditions, significantly improving detection accuracy. The deployment of this product has delivered substantial outcomes for users in areas such as equipment safety management, hazard-free operating environments for personnel, cost reduction and efficiency improvement, and digital transformation.

II. Case Introduction

(I) Background and Needs

1. Research Background

As a common equipment for mineral transportation, belt conveyors play a critical role in ensuring the continuity and stability of production operations. Manual inspection has been the primary means of detecting longitudinal belt tearing timely. However, manual inspection exposes workers to occupational diseases such as pneumoconiosis and even life-threatening safety risks. Moreover, with the gradual disappearance of the demographic dividend, labor shortages for inspection positions have already emerged in the Yangtze River Delta and the southeastern coastal regions. If no inspection personnel are present on site, various belt conveyor anomalies can lead to severe direct and indirect economic losses.

Given current and future trends, leveraging artificial intelligence to foster new quality productive forces has become a key priority in China’s economic development. AI offers a pathway to improve working conditions for inspection personnel, prevent significant direct and indirect economic losses, drive enterprise digital transformation, advance cost reduction and efficiency improvement, and accelerate the adoption of cutting-edge technologies.

In March 2024, the Central Economic Work Conference explicitly called to develop the digital economy and accelerate the development of artificial intelligence. AI has become a driving force for economic and social development, and accelerating the development of next-generation AI is a strategic initiative for China to gain a competitive edge in global technological race.

It is essential to seize the strategic opportunities presented by the new round of technological revolution and industrial transformation, fully leverage China’s unique advantages of abundant application scenarios supported by an ultra-large-scale market, and accelerate scientific and technological innovation in AI to cultivate new momentum for economic growth by pursuing high-quality AI development and high-level AI application.

China will deeply empower high-quality industrial development, leverage AI to vigorously advance new industrialization, and drive the intelligent upgrading of traditional industries while promoting the industrialization and commercialization of AI applications. Centered on the needs of economic and social development, China will fully harness its advantages of vast data resources and enormous market applications, organically integrating technology with industry.

It's necessary to adhere to a demand-oriented, market-driven path of technological development, strengthen the primary role of enterprises in AI innovation, technology pathway selection, and formulation of industrial product standards, and promote the commercialization of technological achievements.

With the rapid advancement of AI technologies, intelligent belt health monitoring systems are now capable of real-time, high-frequency, high-precision identification of belt conditions. These systems offer a new approach to addressing the pain points in current inspection practices. Leveraging next-generation AI, they enable round-the-clock, all-round, and high-precision real-time detection of various belt defects including bulges, holes, wire exposure, repairs, and longitudinal tear.

2. Project Necessity

(1) Equipment Safety: Belt tear is difficult to detect, and a full-length tear can cause enormous losses. It is essential to promptly identify and handle minor tears or edge tears to prevent damage escalation. After a shutdown, inspectors examine belt damage especially bulges and wire exposure, which are early warning signs of potential tear.

(2) Economic Benefits: Beyond the direct economic loss caused by the belt itself, the lockout and production stoppage caused by unscheduled belt replacement incur even greater losses. Post-shutdown inspections consume substantial manpower and material resources, and also pose challenges such as labor shortages.

(3) Digitalization: Quantitative management system enables storage and query of historical anomaly, real-time visual monitoring and warning, and data statistical analysis.

(II) Introduction to Implementation Plan

In order to address the risks affecting production safety, equipment safety, production efficiency, and potential production interruptions during belt conveyor operations, this project conducts research based on multi-sensor fusion and AI-based recognition technologies to achieve real-time and continuous detection of longitudinal belt tear risks. Detected alert information is transmitted to monitoring staff via monitoring terminals, helping them promptly identify risks and take effective measures to reduce the impact of losses. The main research contents are as follows:

1. Research on high-precision intelligent detection methods for longitudinal belt tear.

The probability of a longitudinal tear occurring once in the lifecycle of a conveyor belt is approximately 20%. A conveyor belt worth millions or even tens of millions of yuan will be completely damaged in a very short time once a longitudinal tear accident occurs, resulting in enormous economic losses. Even if repairable, it requires considerable manpower and time, severely impacting normal production. This project employs high-frequency laser imaging to acquire cross-sectional morphological data of the belt, and uses AI algorithms to perform high-frequency detection and analysis of these data, automatically identifying longitudinal tears and edge damage and missing sections, thereby meeting the requirements for high-precision longitudinal tear detection.

2. Development of a belt conveyor abnormal operation monitoring and management system.

For the detected longitudinal tears, a monitoring and management system based on a B/S architecture is developed. The system enables a one-to-one mapping of belt conveyor layouts on the monitoring interface, helping supervisors intuitively see the current operating status, live images, abnormal fault types, and statistical distribution of anomalies across the entire site. In addition, the system supports historical anomaly data query and parameter configuration.

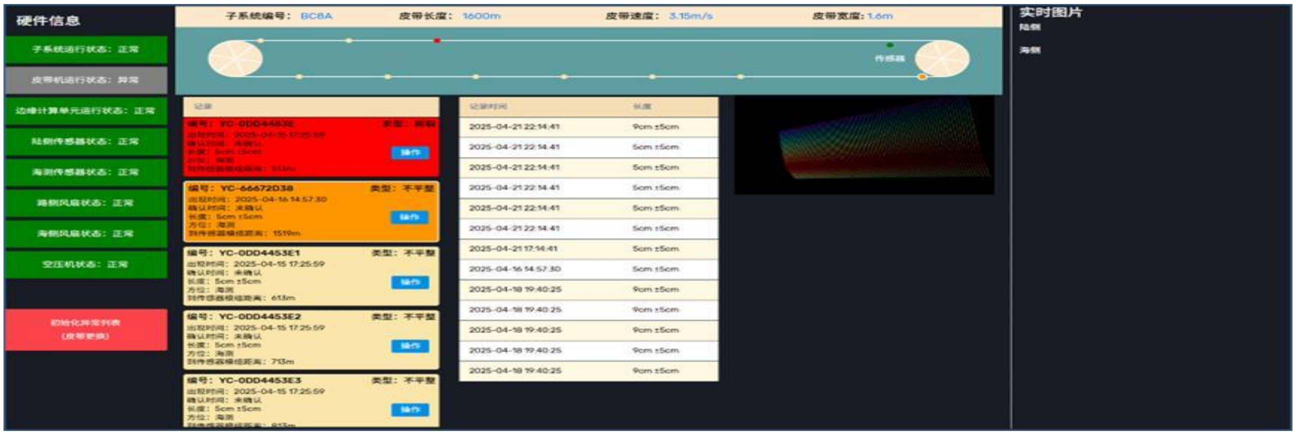


Figure 1 System Interface

The sensor assembly is installed between the upper and lower layers of the belts, and the installation position is shown in the figure below:

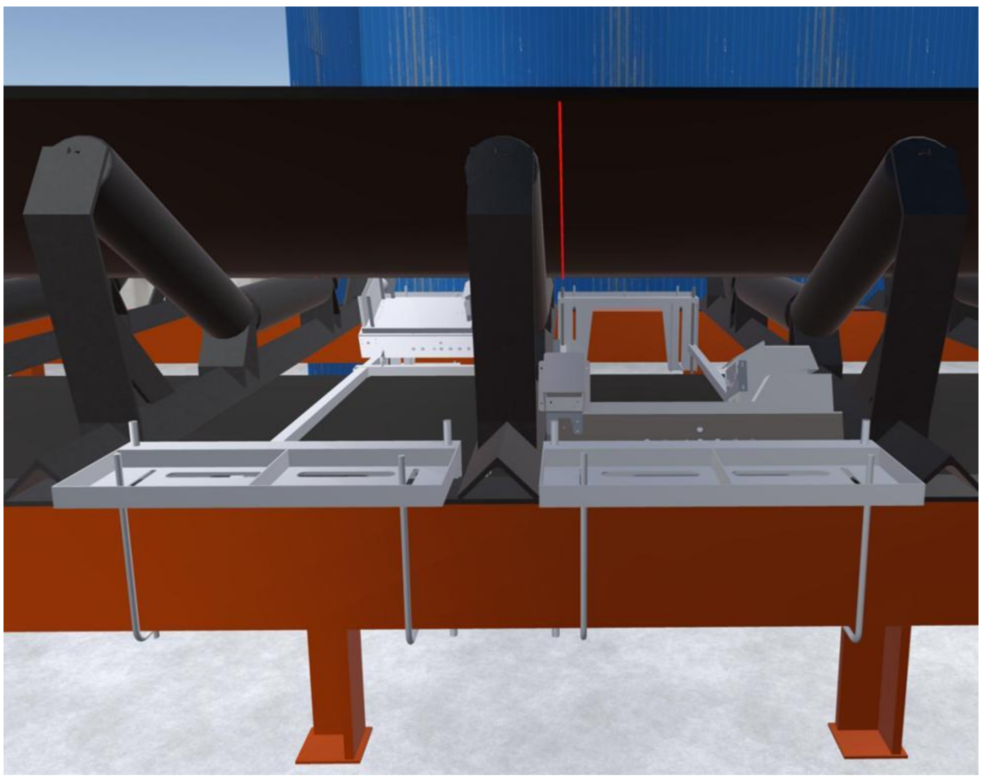


Figure 2 Data Acquisition

Once the sensing equipment is installed, on-site data collection can be carried out, including routine operational data, production anomaly data, and simulated anomaly data.

1. Routine operational data collection. Routine data refers to data recorded during normal production. This category features large volumes of collectible data, which helps the system learn the data characteristics under various normal operating conditions.

2. Production anomaly data collection. Production anomaly data is collected when various exceptions occur during routine operations. Data for longitudinal belt tearing in this category is characterized by high reliability and can truly reflect real abnormal conditions. The drawback, however, is that due to their low occurrence frequency, it is difficult to collect a sufficient variety and quantity of anomaly samples.

3. Sufficient anomaly data collection through simulation. Based on findings from research, along with specific images and video data, a collection scheme is designed to simulate risk scenarios to the greatest extent possible. Data

from this category closely resembles real conditions, offers high reliability and large volumes, making it well-suited for thorough model training to achieve optimal detection performance.

The system consists of four layers: hardware layer, algorithm layer, function layer, and application layer. Hardware layer includes sensing devices, computing devices, and network devices, providing the hardware basic for realizing all system functions. Algorithm layer adopts the company's proprietary algorithm platform, which incorporates a variety of basic perception algorithms for multi-sensor data fusion and serves as the technological core of the system. Function layer converts algorithm capabilities into detection capabilities. It sort out the judgment logic for all types of customer-concerned risks and conduct targeted optimization accordingly. Application layer serves as the interaction interface for monitoring staff and management staff, providing multiple operational interfaces, including monitoring, alerting, query, analysis, and configuration for users primarily through a B/S-architecture web client.

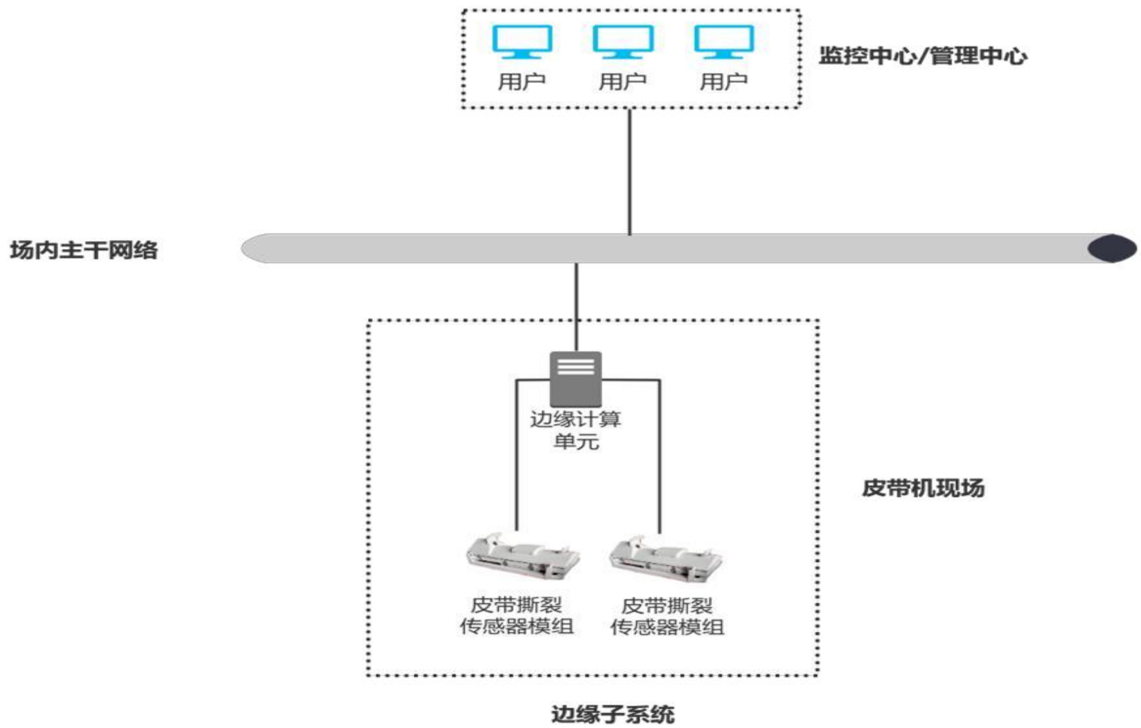


Figure 3 System Topology Diagram

Multi-Sensor Fusion Perception is a technology that integrates data from multiple sensors and combines them with algorithmic models to achieve more comprehensive, accurate, and robust perception of the environment or target objects. Its core advantage lies in leveraging the complementarity of different sensors to overcome the limitations of a single sensor, such as noise, blind area, and insufficient accuracy. This technology is widely applied in fields such as autonomous driving, robotics, industrial inspection, and smart cities.

1. Sensor Types and Complementarity

Vision sensors (high-speed industrial cameras): Support high-resolution images, but are heavily affected by lighting and weather conditions.

Laser sensors: Delivers high-precision 3D point cloud data, but comes with higher costs and performance degradation in rainy or snowy weather.

2. Fusion Levels

Data-level fusion: Directly fuses raw sensor data (such as alignment of point clouds and image pixels).

Early fusion: Input raw data directly into the network (such as concatenation of point clouds and images).

(III) Process Route

1. Algorithm & System R&D and Device Improvement

For multimodal fusion detection combining image and LiDAR Point Cloud Data, feature extraction is required not only for image data but also for processed laser data.

(1) View-based Network Model

A view-based network model segments the input point cloud data to generate multiple views. Each view may represent a small region within the point cloud or a subset of the entire point cloud, depending on the specific application scenario. The model then extracts features from each view and fuses the features across different views to obtain a holistic feature representation of the entire point cloud. Based on these feature representations, the model can detect objects within the point cloud and predict their 3D positions, sizes, and other parameters.

(2) Point-based Network Models

Raw point cloud data exists as unorganized point, while convolutional neural networks require organized input formats. Therefore, most previous approaches converted point cloud data into multi-view images. For example, the MV3D model transforms point cloud data into multi-view images before feeding them into 2D CNNs for deep learning. However, such data conversion introduces redundancy and inevitably leads to the loss of critical information during the conversion process. To overcome these issues, the PointNet network and its enhanced version PointNet++ have been developed.

Point-based network models address the challenge of sparsity and disorder inherent in point cloud data compared to regular image data. They can flexibly extract features at different scales and receptive fields, achieving significantly higher detection accuracy than view-based methods. However, they still face certain limitations, including excessive parameter, heavy computational load, high capacity consumption, and difficulty in meeting real-time performance demands.

(IV) Scenario Requirements

Belt conveyors are widely used in ports, mines, power plants, metallurgy, and other sectors, all of which share the same pain points: high risks associated with manual inspection, significant losses from downtime, and low-level digitalization. Taking ports as an example, belt conveyors at bulk cargo terminals must operate around the clock. Once a tear occurs, the entire loading and unloading process is disrupted. Mine belt conveyors are often deployed in underground or open-air dust environments, where manual inspection poses a high risk of pneumoconiosis. In power plants, belt conveyors transport coal and downtime directly affects power generation efficiency. These scenarios share a consistent demand for round-the-clock, high-precision, and low-cost intelligent monitoring, providing a broad market foundation for cross-industry replication of the system.

III. Project Highlights

(I) Key Technological Breakthroughs

1. Technological Innovation:

The system employs non-contact detection technology for belt operation anomaly monitoring. Traditional contact-based detection methods typically rely on mechanical components, which can detect faults such as belt misalignment and hopper blockages. However, mechanical components used in contact-based detection are highly susceptible to failure and damage under harsh conditions such as high dust, water flushing, and aggregate impact. The company adopts non-contact detection technology throughout, eliminating the use of mechanical components, resulting in longer service life of equipment and more stable detection outcomes.

Fusion perception technology is employed to detect belt operation anomalies. In the past, technologies relying solely on image recognition were often affected by various factors, failing to guarantee detection rates and resulting in high false-positive rates. Our company adopts fusion perception technology—including laser-image fusion, infrared-visible light fusion, and more—to achieve accurate detection of belt operation anomalies under harsh working conditions while reducing false-positive rates.

2. Key Features:

(1) The system provides real-time monitoring and identification of longitudinal tearing and damage on belt conveyors. It can extract multiple features of longitudinal tearing, including overlaps and cracks, as well as various abnormal conditions. Abnormalities on the belt rubber surface manifest as short-length damage and uneven surface, including bulges, scratches, surface breakages, and edge tears. The system also effectively detects subtle spatial depth variations of anomalies, capable of precisely identifying tears as small as 2-3 mm. It covers a maximum belt width of 2,400 mm and a maximum belt speed of 9 m/s.

(2) The product is equipped with a self-cleaning dust removal system, ensuring long-term operation without dust accumulation. It is also equipped with mud-drip shielding structure and anti-mud-dripping design to keep sensor windows clean. As a standardized product, it enables rapid deployment and adapts to confined installation spaces.

(II) Uniqueness of Application Scenarios

Enterprises in bulk commodity industrial chain face complex material transfer routes and harsh operating environments, making it difficult for manual inspection to comprehensively and promptly identify various risks. Equipment damage can significantly disrupt production. Large lumps and foreign objects mixed in bulk materials can cause longitudinal belt tear, resulting in substantial losses.

IV. Implementation Results

(I) Economic Benefits

Deployment Site: Qingdao Port

The core functions of this product include longitudinal tear detection, anomaly classification, anomaly localization, anomaly analysis, and early warning. The comparison of data before and after using this product in conveyor belt maintenance and early warning is shown in the figure below

Function Point	Before Deployment	After Deployment
Anomaly Classification	Manual inspection, 24 hours per week	No human labor required
Anomaly Localization	Positioning anomaly, 6 hours per location	No human labor required
Anomaly Analysis	Unavailable	Full statistics of existing anomalies, full-time tracking from occurrence to repair
Anomaly Early Warning	Unavailable	Real-time tracking of anomalies changes, full-time monitoring of high-risk anomalies

(II) Social Benefits

1. Replicability

(1) Core industry pain point: An industrial gap for stable and effective solutions

For a long time, 24/7, all-round, high-precision real-time detection of longitudinal belt tearing has been widely recognized as a critical pain point and technical challenge in the industry. To date, no mature, operationally stable, and commercially scalable industrial solution has emerged on the market. Most enterprises remain in the "trial-and-exploration" phase, with core technologies yet to achieve commercial deployment. This market gap means that an intelligent monitoring system that is the first to break through the technical bottleneck will gain a first-mover advantage in the competition, positioning itself for rapid replication and market capture.

(2) Comparative technology maturity of different generations: The disruptive advantage of fourth-generation technology

Belt Tear Detection	Physical contact sensing	Inspection robot	Fixed Camera image	Fixed Multi-Sensor Fusion Perception Modules
Environmental & Operational Impact	Vulnerable to mud splashing and small stones	Prone to missed detections	Adversely affected by mud splashing, dust, etc.,	Barely affected
Algorithm Effectiveness	None	Operational mode leads to failure	Low effectiveness	Highly effective
Functional Extensibility	None	None	Difficult	Highly effective for belt holes, damage, bulges, etc.

(3) Bottlenecks in the promotion of the first three generations of technology: Core obstacles to replicability

From the standpoint of industrialization practice, although the first three generations of technical solutions have been trialed by some enterprises, none of them have achieved large-scale deployment. The core obstacles are concentrated in the following three areas:

Poor environmental adaptability: On-site conditions such as dust, mud splashing, and low temperatures can cause sensor malfunction, shorten robot endurance, and blur camera images, making it impossible to meet the requirements of core application scenarios such as ports and mines.

Insufficient algorithmic accuracy: In cases of fewer samples, such as rare and unusual types of damage, traditional algorithms cannot achieve precise identification, and are prone to false and missed detections, undermining user confidence in practical application.

Limited functionality: Inability to achieve comprehensive detection or 3D Analysis, solving only isolated issues rather than providing enterprises with an integrated solution, thereby reducing the comprehensive value of replication.

The fourth-generation AI technology based on multi-sensor fusion precisely addresses the above bottlenecks, clearing the technical obstacles to large-scale replication.

The system can follow a commercialization path covering demonstration projects, regional expansion, and nationwide coverage, gradually achieving large-scale replication.

Case 32

MASS+AI Implementation Case for Empowering a New Education Ecosystem

Diguniu (Shanghai) Technology Co., Ltd.

I. Basic Enterprise Information

Established in 2024, Diguniu (Shanghai) Technology Co., Ltd., is China's first education enterprise to adopt a MaaS + AI Modular Middle Platform Architecture and has completed the filing of generative AI services. Its R&D team comprises 20% education specialists and 40% algorithm engineers, working alongside frontline teachers and industry experts to develop over 300 scenarios covering teaching, learning, research, evaluation, and management. The company provides one-stop solutions for schools across all grade level. Its Agent Factory enables large-scale and replicable deployment. Since its market launch, the product has been adopted by over 500 schools, serving more than 100,000 teachers and students, and has received wide recognition from numerous AI demonstration schools and regional education authorities.

II. Case Introduction

(I) Background and Needs

In recent years, the rapidly developing artificial intelligence technology has become a key driving force in educational transformation. To implement national strategic plan, the Ministry of Education of China has successively issued *the Notice on Strengthening AI Education in Primary and Secondary Schools (2024) and the Outline of the Plan for Building a Leading Country in Education Nation (2024–2035)*, explicitly calling for the deep integration of AI with education and the establishment of a K-12 AI education system, with the goal of achieving basic universal AI education in primary and secondary schools by 2030. In 2025, nine ministries and commissions including the Ministry of Education further proposed taking education digitalization as a breakthrough point to drive AI-empowered curriculum reform, teaching evaluation, and teacher development. They also issued *the Guidelines for General AI Education in Primary and Secondary Schools* and *the Guidelines for the Use of Generative AI*, to standardize technical application and address ethical compliance and information security.

Actively responding to national policies, the Shanghai Municipal Government has issued *the Action Plan for Empowering High-Quality Basic Education Development with Artificial Intelligence (2024–2026)*, focusing on curriculum development, teaching innovation, and the enhancement of digital literacy among teachers and students. Shanghai has taken the lead in offering local AI courses for fourth-grade primary and seventh-grade secondary students, while also deepening AI teaching at the high school level. Meanwhile, by building intelligent laboratories, developing educational large language models, and piloting dual-teacher classrooms model, the city aim at creating a demonstration hub for "AI + Education" providing replicable practical experience for the rest of the country.

Against this backdrop, the Diguniu AI Education Platform was launched, with the goal of leveraging AI to address issues such as imbalanced educational resources, low teaching efficiency, and the lack of personalized learning, thereby supporting the implementation of the national strategy of educational digitalization.

User Segmentation Needs Analysis Table

用户角色	核心需求	传统方案痛点
学校管理者	·教学质量可视化监控 ·师资效率优化 ·家校协同管理	依赖人工统计，数据滞后且误差率高

教师	·智能备课(教案/课件生成) ·作业自动批改与学情分析 ·课堂互动工具	重复劳动占比高， 缺乏个性化指导能力
学生	·自适应学习路径 ·实时答疑与错题巩固 ·创造性思维培养	统一教学进度，无法因材施教

(II) Implementation Plan

1. System Development Content

The Diguniu AI Teaching LLM System is designed to build an intelligent education ecosystem covering the entire chain of platform, resources, teaching, and governance. Centered on the Diguniu AI Education Platform, it integrates an adaptive learning engine and intelligent teaching tools to establish a dynamic resource repository that supports AI-powered precise content push. Through a precision teaching model, the system reshapes the entire teaching process, enabling AI-enhanced smart lesson preparation, hybrid instruction, and personalized assignment feedback. At the same time, it builds a three-tier collaborative governance system involving government, school, and family, leveraging data-driven approaches to achieve scientific decision-making, precise management, and home-school collaboration. Ultimately, the system aims to break the educational Impossible Trinity and propel education into a new era of high-quality, personalized, and balanced development empowered by intelligence.

2. System Framework

The Diguniu AI Education Platform adopts a three-layer integrated architecture:

Infrastructure Layer: Built on a self-developed education-specific LLM with hundreds of billions of parameters, integrated with subject knowledge graphs (covering 3,000+ knowledge points from K-12 to higher education).

Capability Layer: Provides six major categories of API interfaces (natural language processing, multimodal generation, data analytics, etc.).

Application Layer: Develops 20+ teaching agents (e.g., AI teaching assistant, intelligent grading officer, curriculum designer, etc.).

3. Core Functional Modules

模块	功能描述	解决的问题
AI 备课系统	·5 分钟生成跨学科融合教案 ·自动匹配课程标准与德育目标	教师备课效率提升 3 倍
智能作业系统	·支持数学公式/作文/编程题批改 ·错题归因分析并推荐相似题	批改准确率 98%,节省 70%时间
学情追踪系统	·动态生成学生能力雷达图 ·预警潜在学习困难点	帮助教师早期干预落后生
AI 创意工坊	·文生图工具辅助艺术创作 ·代码生成器支持 STEAM 教学	激发学生创新能力

4. AI-Powered Closed-Loop Teaching Process

Pre-class: The teacher inputs the keyword “Grade 8 Physics: Buoyancy (PEP edition)”. The AI automatically generates an interactive lesson plan incorporating experiment videos and real-life examples.

In-class: Students answer questions via tablets. The AI analyzes accuracy rates in real time and pushes supplementary materials (e.g., if 65% of students select the wrong answer to option B, a targeted micro-lesson on that knowledge point is triggered).

Post-class: The system automatically generates personalized learning reports. The parent portal synchronous receives updates on the student's weak points along with suggested at-home practice activities.

5. Summary of Key Problem Resolution

传统教育痛点	AI 解决方案	量化效果
教学资源不足	云端共享百万级试题、教案库	资源获取效率提升 90%
个性化缺失	基于大模型的自适应学习	学生专注度提高 40%
管理低效	数据驾驶舱实时监控教学管理	管理决策速度提升 60%

6. Page Layouts by User Terminal Role

School Administrator Portal: Five core modules: AI Dialogue; Dashboard (visual analytics of teaching and learning data for teachers and students); AI Assistant; Data Management (User Management, Function Management, Home-School Connect); and Personal Center.

Teacher Portal: Eight core modules: AI Dialogue; AI Creative Workshop; AI Assistant; Course Center (Assignment Center and Resource Center sub-modules); Member Management; Function Management; Home-School Connect; and Personal Center.

Student Portal: Six core modules: AI Dialogue; AI Creative Workshop; AI Assistant; Course Center; Learning Center; and Personal Center.

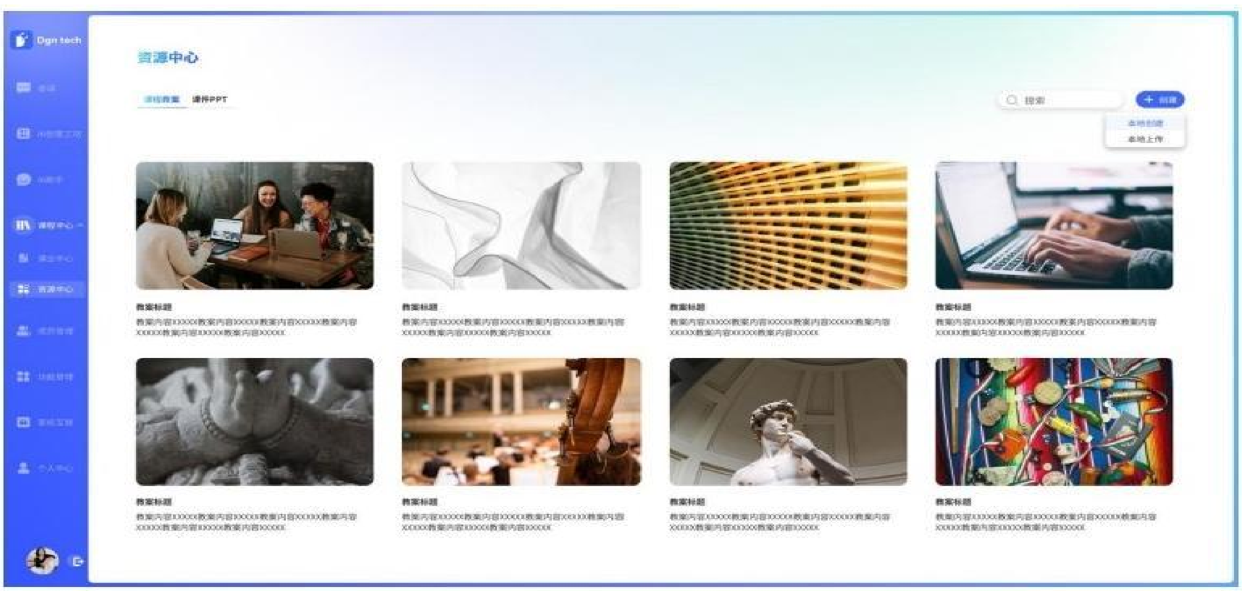


Figure 1 Teacher Terminal Interface

III. Project Highlights

(I) Key Technological Breakthroughs

1. Innovation in Education-Specific Vertical LLMs

Billion-parameter professional training: Built on the Transformer architecture, the model undergoes continuous pre-training on 1.2 TB of education-specific corpus, such as textbooks, test questions, classroom recordings, achieving 92% accuracy in mathematical problem-solving and 89% consistency with expert evaluators in essay scoring.

Multimodal fusion capabilities: Breaking through the limitations of text-only single-modality, supporting formula recognition, motion capture from experiment videos (error <3%), intelligent mind map generation, and other complex scenario requirements.

Dynamic knowledge distillation: Compressing the LLM to 7B parameters, delivering millisecond-level response within 10ms on ordinary GPU servers to meet real-time classroom interaction demands.

2. Adaptive Learning Engine

Cognitive diagnostic model: Employs a hybrid architecture combining Bayesian Knowledge Tracing (BKT) and Deep Knowledge Tracing (DKT), capable of pinpointing student knowledge gaps with an accuracy of ±5% within five tests .

Intelligent resource matching algorithm: Based on the Felder-Silverman learning style model, it delivers video, audio, and interactive experiment content to visual, auditory, and kinesthetic learners respectively, improving resource utilization by 60%.

(II) Scenario Uniqueness

The scenario uniqueness of the Diguniu AI Education Platform is primarily reflected in the following aspects:

1. Full-Process Intelligent Closed-Loop Teaching

The platform establishes a complete AI-empowered closed loop covering lesson preparation, classroom instruction, and evaluation. During lesson preparation, AI automatically generates instructional designs incorporating multimedia resources based on curriculum standards, offering three differentiated versions for teachers to choose from. In class, the system analyzes student feedback in real time and automatically adjusts the teaching pace, ensuring over 80% of students master the core knowledge points. In the post-class evaluation stage, the platform not only automatically grades assignments but also generates personalized learning analytics reports, precisely identifying each student’s weak knowledge areas.

2. Subject-Deeply-Adapted Intelligent Applications

Targeting the distinct characteristics of different subjects, the platform has developed specialized intelligent tools. In Chinese language teaching, AI can automatically annotate key classical Chinese characters and phrases, providing their allusions and modern interpretations. For mathematics, it supports dynamic reasoning demonstrations for geometry problems, presenting multiple solution approaches. For science courses, it offers a virtual experiment environment where students can conduct high-risk chemistry experiments through drag-and-drop operations. These deeply subject-adapted intelligent applications significantly enhance teaching efficiency and safety.

3. Multimodal Teaching Support Capabilities

The platform breaks through the single-interaction mode of traditional educational software, integrating multiple interaction modalities including text, voice, and image. Teachers can quickly access teaching resources via voice commands, and students can input math formulas by handwriting for automatic recognition and grading. Particularly in arts courses, students need only describe their creative concepts, and the AI generates corresponding visual sketches or musical snippets, greatly expanding students’ creative horizons.

4. Personalized Learning Support System

Based on students’ learning behavior data, the platform generates precise student capability profiles. The system intelligently recommends appropriate learning resources and practice questions based on each studen’s cognitive characteristics and learning pace. For students with learning difficulties, the AI automatically breaks down complex knowledge points and provides step-by-step tutoring; for students with advanced learning abilities, it pushes advanced enrichment content, meeting differentiated learning needs.

5. Regional Teaching-Research Collaboration Platform

The platform provides intelligent support tools for regional teaching and research. By aggregating teaching data and resource usage from teachers within the region, AI can automatically identify excellent teaching practice cases and generate reusable teaching templates. Instructors and researchers can quickly understand the teaching status of schools across the region, identify common issues, and organize targeted teaching-research activities.

These scenario innovations make the platform not merely a collection of teaching tools, but a complete intelligent teaching ecosystem. Its application significantly improves teaching efficiency, reduces teacher’s workload, and provides students with a more personalized learning experience, truly realizing the educational philosophy of teaching students in accordance to their aptitude.

6. Scenario Technology Support Examples

Teaching quality assessment: Filters relevant teacher teaching data and comprehensively evaluates teaching quality based on classroom performance, assignment grading, student feedback, and exam results.

Teaching delivery analysis: Filters relevant teacher teaching data and comprehensively analyzes teaching delivery based on instruction duration and course completion rates.

Learning analytics: Filters student data and comprehensively analyzes learning status based on course completion, assignment and exam submission, and online learning time.

Assignment analysis: Conducts data analysis based on student submission rates and assignment accuracy rates.

Exam analysis: Monitor performance trends across terms. Use score change patterns to inform a comprehensive review, enabling targeted improvements in content delivery and instructional focus to address identified weaknesses.

Health data analytics: Leverages data from health wristbands to monitor student health and physiological indicators.

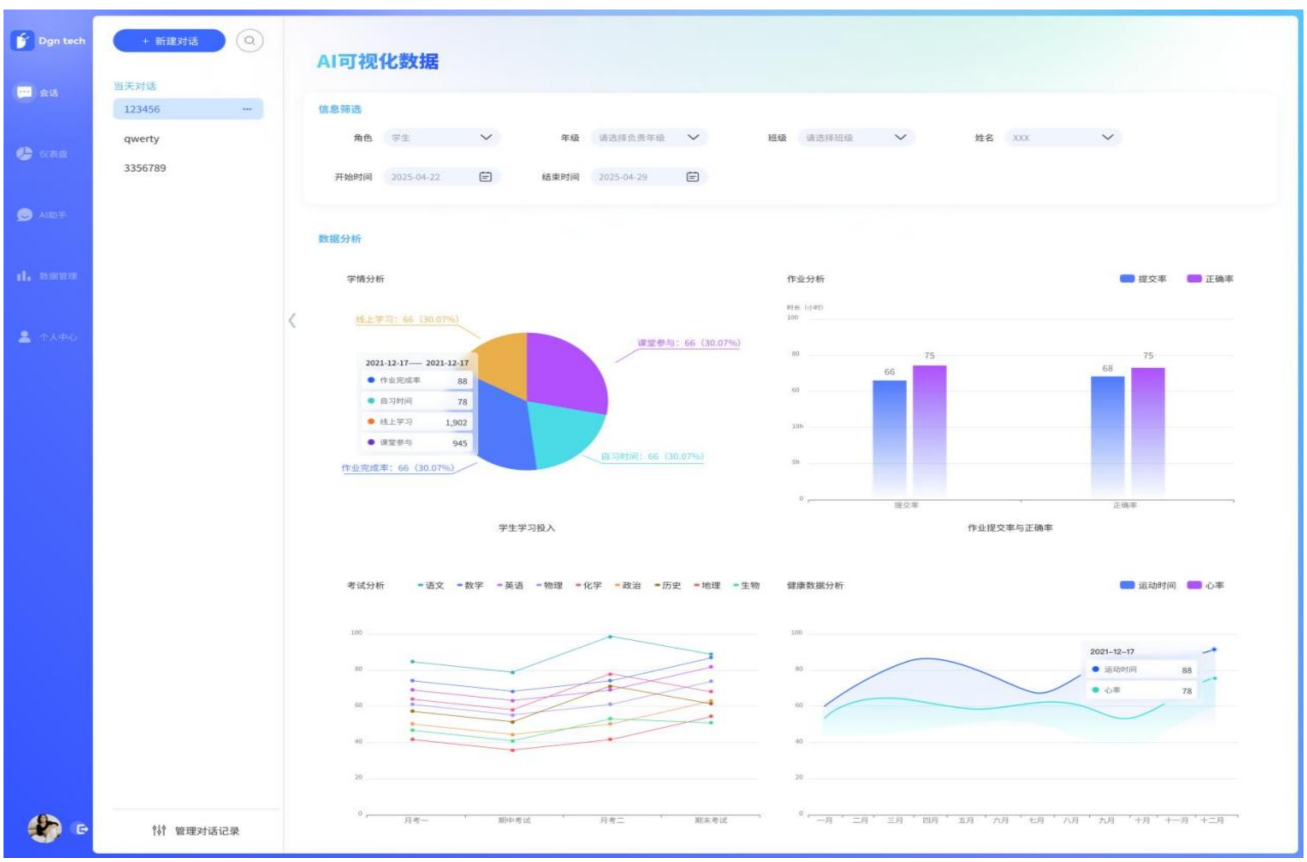


Figure 2 Learning Analytics Page

(III) Educational Model Restructuring

1. Teacher Role Upgrade

From lecturer to learning designer: Approximately 80% of teachers' repetitive tasks are automated, allowing them to shift toward curriculum customization and innovation capability cultivation.

New human-AI collaboration paradigm: AI handles standardized tasks such as dictation grading, while teachers focus on higher-order thinking development such as guiding classroom debates.

2. Teaching Organization Innovation

AI-powered dynamic level-based class rotation: Real-time adjustment of differentiated instruction groupings based on AI diagnostic results.

Borderless learning communities: Connects multiple experimental schools nationwide to co-build AI research and learning projects, enabling cross-school teacher-student teams to complete innovative topics.

3. Evaluation System Breakthrough

Process-oriented data analytics: Analyzes multiple dimensions including classroom participation quality and collaborative learning contributions, generating a digital portrait reflecting the simultaneous development of moral, intellectual, physical, aesthetic, and labor education).

Integrated five-domain education evaluation: Real-time analysis of data from moral education practice, physical training, aesthetic creation, labor education, and more, automatically generating a comprehensive competency development radar chart.

Dynamic competency mapping: Leverages knowledge graph technology to correlate and analyze discrete data including exam scores, assignment performance, and enrichment activities and visually present the evolutionary path of students' core competencies.

Developmental feedback mechanism: Automatically pushes a monthly Student Learning Growth Recommendation Report , including:

Strengths reinforcement strategies; (e.g., "Your spatial reasoning ability exceeds 90% of your peers. Robotics programming is recommended.") Customized resource packs for areas requiring improvement. (e.g., "Your ability to apply functions needs strengthening; 3 interactive micro-lessons and 5 targeted practice questions have been matched.")

IV. Implementation Results

Since its launch in 2024, the Diguniu AI Education Platform has achieved rapid promotion and large-scale application nationwide, with its advanced technical architecture, comprehensive functional modules, and deep adaptation to educational scenarios. The project has been deployed in over 500 primary and secondary schools, serving more than 100,000 teachers and students, and has become one of the outstanding platform cases in AI-empowered basic education in Shanghai and across China. Significant results have been achieved in economic returns, technological innovation, and social benefits.

(I) Economic Benefits: Achieving A Win-Win Situation and Driving Digital Upgrading of the Education Industry

In terms of economic benefits, the platform has not only delivered substantial commercial returns for the company but also created tangible value for schools, teachers, and families. The platform adopts a composite business model of "SaaS + Hardware + Content Services", combining subscription-based services with customized deployment to provide flexible solutions for educational institutions at all levels. The following are quantitative results from two typical partner organizations:

According to statistics, one educational technology company in Shanghai that adopted our platform achieved approximately 200% year-over-year revenue growth within six months of going live, with a user retention rate exceeding 85%, demonstrating strong market competitiveness and customer loyalty.

More importantly, the platform has significantly reduced educational operational costs for schools. At a secondary school in Baoshan District, teachers previously spent substantial time on tasks covering lesson preparation, assignment grading, and learning analytics. On average, each teacher needed to spend more than 15 hours per week dealing with non-teaching matters. Leveraging AI-powered functions such as intelligent grading, automated test paper generation, and personalized lesson plan creation, teachers have been able to save over 70% of their repetitive workload, redirecting their efforts toward curriculum design and personalized student guidance. According to feedback from this pilot school, the overall work efficiency of teachers has increased by approximately 60% after platform adoption. Regional calculations indicate that each school can save hundreds of thousands of yuan annually in personnel costs.

(II) Social Benefits: Promoting Education Equity and Supporting the Development of a Leading Country in Education

The social benefits brought by the Diguniu AI Education Platform is reflected across multiple dimensions, particularly in promoting educational equity, improving teaching quality, and cultivating future talent.

First, the platform effectively alleviates the imbalance in educational resource distribution. Through cloud deployment and standardized interfaces, schools in remote areas can also access the same high-quality teaching resources and AI-assisted tools as those in first-tier cities.

Second, the platform facilitates the implementation of the concept of "teaching students in accordance of their aptitude". Through precise learning analytics and personalized learning path recommendations, the system can tailor customized study plans for each student. Data shows that intensity of focus of student has generally increased by 40%, and mastery of weak knowledge points has risen to over 85%.

Third, the platform strengthens home-school collaboration and regional teaching-research coordination. Parents can view their children's learning progress, error records, and learning growth recommendations in real time via the mobile app. Education authorities, meanwhile, leverage the data dashboard for teaching quality monitoring and policy evaluation, forming an intelligent integrated ecosystem combining government agencies, schools.

Case 33

Agora Conservation AI Helping SMEs Rapidly Build High-quality, Low-latency Intelligent Scenarios

Shanghai Agora Technology Co., Ltd.

I. Basic Enterprise Information

Founded in 2014, Agora is a pioneer in global real-time audio and video cloud services, delivering the best possible experience for audio and video interactions between people and between people and intelligent agents. Developers can simply call the Agora API to build various real-time interactive scenarios within their apps, such as conversational AI, audio and video calls, and live streaming. The Agora API has empowered more than 20 industries, including AI, social live streaming, education, gaming, IoT, finance, healthcare, and enterprise collaboration, across more than 200 application scenarios.

On June 26, 2020, Agora Inc., the parent company of Agora, was successfully listed on the NASDAQ under the stock ticker symbol “API”. As of December 31, 2024, Agora had over 931,000 registered applications worldwide. Its monthly service volume exceeded 70 billion minutes (calculated based on channel participants).

Agora has launched the world’s first conversational AI engine, empowering developers to build real-time voice dialogue experiences based on any large language model, and has created the real-time audio and video network known as the Software-Defined Real-time Network (SD-RTN™). Agora’s technology services cover more than 200 countries and regions worldwide, with clients including industry giants, unicorns, and startups such as Xiaomi, Momo, Douyu, Bilibili, red note, and Yalla; Agora’s technology is also adopted by well-known global companies such as HTC VIVE, The Meet Group, and Bunch.

(2)Introduction to Implementation Plan

1. System Construction Content: Build End-to-End Real-Time Dialogue Capabilities

Agora Conversational AI Engine delivers out-of-the-box intelligent voice interaction solutions for SMEs. Its core development consists of three major modules. First, the real-time voice processing layer integrates high-precision ASR and TTS, along with acoustic echo cancellation and noise reduction algorithms, supporting stable audio input and output in multiple languages and complex environments. Second, the large model scheduling and adaptation platform is compatible with mainstream large language models such as Tongyi Qianwen, GPT-4 and Zhipu GLM through the unified interface abstract layer, to realize “develop once, deploy across multiple models”. Third, the middleware for business system integration offers lightweight SDK and standardized API, which can seamlessly connect with enterprise backgrounds including CRM, order systems and knowledge bases, upgrading AI from a “Q&A assistant” to an intelligent gateway capable of executing operational business operations.

2. System Architecture Design: A Layered and Decoupled Architecture Based on the TEN Framework

The system is built on TEN Framework, a leading global open-source framework, and features an overall three-tier microservices architecture: the access layer supports multiple terminals such as Web, APP, IoT devices and multiple modal inputs such as voice, text, and images; the intelligent processing layer includes core modules such as LLM agent, context memory management, and emotional intonation control, ensuring semantic coherence and natural expression; the execution layer leverages Agora’s global RTE network to enable low-latency, high-concurrency command issuance and result feedback. This architecture offers high scalability and low-code deployment capabilities, breaking the traditional limitation of “one assistant tied to one model” and supporting flexible combinations of functional modules to adapt to different industry scenarios.

II. Case Introduction

(I) Background and Needs

The Outline of the 14th Five-Year Plan for National Economic and Social Development of the People’s Republic of China and the Long-Range Objectives Through the Year 2035 clearly proposes strengthening basic research, tackling key core technologies, and conducting application demonstrations in the field of artificial intelligence, as well as promoting the extensive application of artificial intelligence technologies across various industries and sectors. AIGC represents a cross-domain development model that integrates artificial intelligence with big data, cloud computing, 5G and other technologies. Its development depends on three fundamental infrastructures: computing power, algorithms and data. This case can be integrated with various large model platforms to deliver real-time voice dialogue featuring lower latency and a more natural immersive experience, thereby creating new opportunities for scenarios including remote collaboration, online education, social interaction, live streaming and games.

Figure 1 Product Architecture of Agora's Conversational AI Engine

3. Core Business Logic: Realizing the closed-loop covering perception, Understanding , Decision, Execution , and Feedback

The system operates following this five-step closed-loop process: a user initiates interaction through voice or images; the system invokes the ASR or CV module to parse input data and conducts intent comprehension combined with user identity metadata and historical context information; the large language model generates natural language responses and judges whether it is necessary to call business systems; if operational actions are required, backend interfaces will be automatically triggered for execution; finally, results are returned in the form of emotionalized TTS or graphics and texts. Supported by Agora's RTE network, the average end-to-end response latency of the entire link is less than 300 milliseconds, ensuring smooth interaction, efficient operation and a lifelike user experience.

III. Project Highlights

(I) SD-RTN™ Technology for Real-Time Virtual Transport Networks with QoS Guarantees

Agora' Software-Defined Real-Time Network (SD-RTN™) is built on the underlying Internet infrastructure. It enables the creation of virtual real-time networks with guaranteed Quality of Service (QoS) through software-defined algorithms, specifically designed to deliver quality comparable to dedicated lines for real-time audio and video services and conversational AI engines.

Network Architecture: SD-RTN™ adopts an architectural framework with Edge Node access and Core Node control. All nodes are deployed across different data centers on the global Internet. Different nodes collaborate with each other to jointly complete path measurement, data forwarding, transmission mode control, access selection, permission authentication and other tasks. SD-RTN™ adopts cloud control algorithms that comprehensively take into account indicators including access selection, high scalability, high availability and load balancing, so as to provide the best QoS quality guarantee for real-time audio and video services.

Global Coverage: SD-RTN™ deploys access nodes worldwide and builds a global overlay network through software technologies. At present, SD-RTN™ has been deployed across 255 backbone data centers spanning five continents around the globe. Every node within SD-RTN™ serves as an access point, offering nearby high-speed access for all Agora users across the globe. Its backend does not rely on a fixed domain name system list, and always finds the best connection route in real time based on actual network conditions, providing dynamic network logic. No matter when and where the user is, the business is within reach. Meanwhile, SD-RTN™ intelligently transmits data packets over multiple independent optimized paths to secure a high transmission success rate within the shortest time window. The first to arrive packet is used, and any lost or delayed packets will be ignored.

Intelligent Routing: SD-RTN™ adopts an intelligent routing mechanism. All edge nodes periodically send test data packets to one another on a second-level cycle and report relevant data to core nodes in real time. With deep learning algorithms, the central controller synthesizes network metrics including packet loss, latency and jitter to navigate an optimal path for real-time audio and video services. Compared to the public Internet, SD-RTN™ significantly improves routing refresh rates, effectively bypassing congested sections of the network to provide users with the most "real-time" audio and video interactive experience.

(II) Full-Link Intelligent Audio Processing Technology Based on Acoustic Echo Cancellation and Noise Suppression

In conversational AI application scenarios, audio quality directly determines the fluency and credibility of human-machine interaction. Agora has built an end-to-end full-link intelligent audio engine relying on its self-developed 3A audio processing algorithms, namely AEC (Acoustic Echo Cancellation), ANS (Automatic Noise Suppression) and AGC (Automatic Gain Control).

Acoustic Echo Cancellation: By analysing the correlation between the audio signal collected by the microphone and the audio signal played by the speaker, it estimates the impulse response of the acoustic echo path. Then, based on the estimated echo path impulse response and the speaker's playback signal, it estimates the echo signal captured by the microphone, and finally subtracts the estimated echo from the microphone audio through inverse filtering, achieving echo cancellation.

Automatic Noise Suppression: Various technical means are used to reduce unwanted noise in the audio signal. During the audio capture phase, the noise reduction system first acquires the raw audio signal and conducts preliminary processing with dedicated algorithms, including mean filtering, median filtering, adaptive filtering and more. Such algorithms remove noise components by analyzing statistical characteristics of audio signals. Further noise reduction processing is performed after audio capture, where post-processing filters cover frequency-domain filtering, time-domain filtering, hybrid filtering and other solutions. Agora's AI-powered noise reduction model is trained on massive audio datasets exceeding one trillion minutes in total duration. It effectively suppresses more than 100 types of stationary and non-stationary noises while preserving human voice integrity. It delivers crisp and clear audio experience even under low signal-to-noise ratio conditions or scenarios with dense overlapping human voices.

Automatic Gain Control: AGC dynamically adjusts input and output volume using an adaptive gain algorithm to resolve volume fluctuations caused by varying distances between users and microphones or differences among devices, keeping the intensity of voice signals stably within the optimal range (-20 dB to 40 dB). It automatically amplifies human voice volume in noisy environments and reduces gain in quiet scenarios to prevent audio clipping and distortion, enabling flexible dynamic response. Combined with voice spectrum preservation technology, it avoids speech distortion triggered by excessive amplification. In essence, AGC dynamically tunes audio gain according to the energy level of incoming audio signals: it lowers the gain when signal energy is high to prevent distortion from over-loud audio, and raises the gain when signal energy is low to ensure voices remain intelligible. Agora's self-developed AGC algorithm delivers clear sound pickup up to a distance of 10 meters without the need for professional audio equipment.

(III) Selective Attention Lock-on

Human-machine dialogue in open environments is subject not only to noise interference but also disturbance from ambient background speech. Typical scenarios include high-noise settings such as exhibitions, shopping malls and railway stations, as well as home environments with conversations from other family members or voices coming from televisions. Such interference often prevents AI from locking onto the intended speaker, resulting in disrupted conversations. Selective Attention Locking is a core concept in cognitive psychology and human-computer interaction. It refers to the capability of a system like human or machine to actively focus on specific targets while disregarding irrelevant information amid information overload.

When applied to conversational AI, Selective Attention Locking centers on locking onto the target speaker and minimizing the impacts of noise and interfering signals. Its main implementation approaches are outlined as follows:

Voiceprint Recognition: Using characteristics including timbre, intonation and speaking rate, the AI system can reliably distinguish and identify different speakers. Powered by deep learning-based voiceprint recognition, Agora's conversational AI engine boasts high-precision recognition performance. It can accurately locate the target speaker even in noisy surroundings and block 95% of interfering background human voices.

Spatial Perception and Beamforming: With microphone arrays and beamforming algorithms, directional sound pickup is performed according to the spatial position of the sound source, attenuating audio signals from non-target users.

Attention Mechanism Modeling: The Transformer attention mechanism is integrated into speech recognition and semantic understanding models, enabling the model to focus on target signals and filter out interference. For instance, in meeting scenarios, an AI assistant only responds to voice commands issued by the host or differentiates speakers via voiceprints to generate highly accurate meeting minutes. In multi-member households, personalized voice control and tailored services can be delivered to each family member based on their unique voiceprints. In venues like exhibitions and shopping malls, the system locks onto the user's voice to prevent false triggers caused by background noise.

(IV)Intelligent Interruption Technology

In human-machine voice interaction, intelligent interruption serves as a core capability for smooth and natural conversations, allowing users to cut off the AI's speech stream at any moment while it is speaking. This technology is built upon an in-depth restoration and reproduction of real human dialogue patterns. To pursue the ultimate goal of delivering interactions as natural as face-to-face communication between humans, equipping AI with the capability to gracefully handle user interruptions has evolved from a desirable add-on feature into an indispensable core technical requirement.

Voice Activity Detection Interruption: It judges whether the user intends to speak by analyzing the signal intensity and spectral features within the audio stream in real time, and is implemented through two mechanisms: speak-to-interrupt and duration-triggered interruption. For speak-to-interrupt mode, the AI speech is halted immediately once the instantaneous human voice energy exceeds a preset threshold (e.g., SNR > 0.8). This mode delivers the fastest response speed yet is prone to false triggers. Duration-triggered interruption activates only when the user's voice persists longer than a predefined time threshold, ranging from 300 ms to 1000 ms, effectively preventing false interruptions caused by filler utterances such as "um" and "ah".

Intent-Based Voice Interruption: Combined with Natural Language Understanding technology, this approach identifies keywords or urgent intentions contained in user input. Keyword matching triggers interruption based on preset keywords (such as "stop", "customer service"). It supports two triggering logics: one triggers interruption once keywords appear anywhere in speech (e.g., "Xiaodu Xiaodu, turn on the light"); the other only activates when keywords are located at the beginning of the utterance (e.g., "stop playing"), which reduces false judgments. Intent recognition calculates intent probability through the model and cuts off AI output when the intent confidence level exceeds the threshold. It enables precise control over interruption conditions and applies to strongly intent-oriented scenarios such as smart home voice commands.

Agora's conversational AI engine has been completely rebuilt using its proprietary RTC technology stack, with end-to-end optimization spanning audio capture, ASR recognition, LLM inference, and TTS synthesis, achieving an average end-to-end response latency of less than 300 milliseconds. This performance enables the AI to achieve near-human conversational fluency in highly real-time scenarios such as spoken language practice, call center customer service, and live streaming interactions. It also supports the precise recording of speech interruption timestamps, ensuring contextual continuity after interruptions and significantly enhancing the user experience.

IV. Implementation Results

With continuous technological advances, conversational AI is evolving at an accelerated pace toward miniaturized large models, cross-lingual comprehension, emotional intelligence, commonsense reasoning and other directions. Meanwhile, its engineering capabilities keep improving in terms of development efficiency, deployment convenience, operation and maintenance automation, and cost optimization, which strongly fuels the coordinated development of the entire industrial ecosystem. During this evolution process, audio and video communication serves as critical infrastructure for real-time human-machine interaction and plays an irreplaceable role. According to the report China Video Cloud Market Tracker, H1 2024 released by International Data Corporation, Agora ranks first in the audio and video communication cloud service (RTC/RTE) market with a market share of 33.1%. Statistics from data.ai also show that Agora leads the market share ranking for conversational AI engines. Leveraging this competitive edge, Agora has established in-depth partnerships with numerous enterprises including SenseTime, MiniMax and Mass Spectrometry Technology. These collaborations have not only generated substantial economic value, but also boosted employment and innovation across the upstream and downstream industrial chains.

(I) Changing Human-Machine Interaction Modes

The advent of conversational AI has shifted interaction from GUI to natural language. Users can communicate with machines in everyday language just as they talk with other people. For example, when using an intelligent voice assistant, if a user says, “Help me look up flights from Beijing to Shanghai tomorrow.”, the assistant can comprehend the intent and quickly return relevant flight information. In smart home scenarios, once the user issues the command “Turn on the living room light and adjust its brightness to 50 percent”, smart devices will execute the instruction accordingly. This natural language-based interaction aligns better with human communication habits, removes operational constraints imposed by GUI, and enables users to interact with devices more effortlessly and efficiently. Visually impaired users are able to operate devices via voice commands to search for information and run software applications. Elderly users do not need to learn complicated operations; they can access digital services simply through natural language conversations, such as chatting with smart speakers, retrieving daily information, and controlling smart home appliances. In regions lacking adequate educational resources, students can leverage conversational AI learning tools to ask questions in natural language and receive curriculum explanations, homework tutoring and other academic support, making up for shortages of teachers. In this way, advanced technologies benefit broader groups of people. Regardless of physical condition, age or geographical location, people can gain equal access to information and technological services, narrowing the digital divide and advancing social equity.

(II) Improving Efficiency and Quality of Life

In the office sector, intelligent office assistants can interpret natural language commands. For instance, given the instruction “Schedule a meeting with the sales team at 3 p.m. next Friday and prepare the sales report for the previous quarter”, the assistant automatically completes tedious tasks including meeting scheduling and report generation, cutting down manual operating hours. In the customer service industry, intelligent customer service powered by conversational AI responds to customer inquiries in real time and handles frequently asked questions. On e-commerce platforms, for example, intelligent agents swiftly answer users’ questions about product details and logistics tracking status, easing the workload of human customer service staff. It enables round-the-clock service, boosting customer satisfaction and corporate operational efficiency. One municipal government service platform adopted large models and carried out colloquial optimization on the Q&A knowledge base covering 1,974 government service items. As a result, its intelligent customer service system can accurately understand and reply to users’ colloquial inquiries about administrative procedures, with an accuracy rate exceeding 90%.

(III) Promoting Industrial Innovation and Economic Growth

Conversational AI Agents gives birth to new business models and changes the industrial ecosystem. In intelligent marketing, companies use conversational AI to analyse customer dialogues, precisely identify needs, and carry out personalised marketing. In intelligent customer service scenarios, it can help enterprises reduce labour costs by 30%. As AI-powered toys gain popularity, toy companies have seen explosive order growth—a Dongguan toy company once stated that after adding AI functionality to their plush toys, orders exceeded 2 million units. In the medical industry, intelligent triage systems interact with patients to understand symptoms, make preliminary diagnoses, and recommend departments. Medical robots can collaborate with doctors, record medical histories, query medical literature, and assist in diagnosis. Patients can also access health management advice and disease prevention knowledge through AI chatbots, improving the accessibility and efficiency of medical services. In the financial sector, intelligent customer service answers client questions about account management and wealth management products. Meanwhile, by analysing customer dialogue data, it assesses credit risks, enhancing the quality and security of financial services and promoting innovation in the financial industry.

Parts of this article are referenced from *202 White Paper on the Development of Conversational AI released by Agora*.

Case 34

A New Energy Power Generation Forecasting Agent Based on a Meteorological Foundation Model

Dalang (Shanghai) Intelligent Technology Co., Ltd.

I. Basic Enterprise Information

Founded by senior professionals from leading energy groups, the Chinese Academy of Sciences, the University of California, and other institutions, together with AI scientists and architects from top research organizations, Dalang Intelligent is a technology innovation enterprise specializing in “Energy + Artificial Intelligence”.

The company boasts strong technical credentials: R&D personnel account for 90% of the workforce, and 80% of the team hold master’s or doctoral degrees. It has published over 40 SCI journal papers in the field of artificial intelligence. Its intellectual property portfolio includes 2 granted invention patents (with 5 more under application), 2 utility model patents, and 7 software copyrights. The company has passed ISO 9001 and ISO 27001 certifications, and is recognized as a Technology-based SME and a “Double-Soft” Certified Enterprise (software enterprise and software product certifications). It has also won the Silver Award in the National Machinery Industry Design Innovation Competition, been selected for Shanghai’s cultivated professional service provider list for “AI + Manufacturing”, and serves as a sci-tech innovation partner for Xuhui District’s “Artificial Intelligence +”initiative.

Leveraging the deep integration of general large models and industry-specific small models, Dalang Intelligent has built an MaaS platform for the energy industry, delivering “AI+” solutions that help enterprises optimize the entire chain of energy production, trading, and operation & maintenance.

The company has delivered several benchmark projects both domestically and internationally: it developed an AI power forecasting system for Ivanpah, the world-renowned solar thermal power station; designed AI trading strategies for electricity spot markets for power generation and retail companies across multiple regions; built an AI intelligent O&M system for China’s most efficient coal-fired power plant; and created a gas dispatching intelligent agent for a gas company with a 150-year history.

II. Case Introduction

(I) Background and Needs

Situated in a subtropical monsoon climate zone and highly sensitive to climate change, Shanghai has experienced a warming rate of 0.18℃ per decade over the past 150 years. Extreme weather events, such as heatwaves, super typhoons (e.g., Typhoon Bebinca and Typhoon Pulasan in 2024), and torrential rainstorms occur frequently, severely impacting energy infrastructure. In 2024, extreme heat pushed Shanghai’s peak grid load beyond 40 million kilowatts for the first time. Meanwhile, the output of new energy sources like wind and solar fluctuated dramatically, and traditional dispatching systems lagged in response, further heightening the risk of power supply-demand imbalances.

From a policy standpoint, the Outline for High-Quality Meteorological Development issued by the State Council and the China Meteorological Science and Technology Development Plan (2021 - 2035) explicitly call for the deep integration of AI with meteorology.

From a technical perspective, traditional meteorological forecasting models suffer from accuracy bottlenecks, relying on low-resolution numerical simulations and sparse observation stations makes it difficult to capture micro-meteorological processes such as typhoons and turbulence. In contrast, meteorological large models like Fengwu, Pangu-Weather, and Fuxi can identify complex weather patterns, offering the technical foundation needed for accurate new energy power generation forecasting.

(II) Functional Requirements

Shenergy (Group) Co., Ltd. and Shenergy Co., Ltd. plan to accelerate the deployment of digital technology innovation, aiming to build a high-quality energy industry ecosystem. Focusing on the core scenario of renewable energy power forecasting, they will develop a forecasting agent that covers both wind and photovoltaic power generation across ultra-short-term, short-term, and medium-term time horizons, built on a domestically developed meteorological foundation model. Through this multi-timescale precision forecasting capability, the initiative seeks to enhance the refined management and decision-making of power generation enterprises in production operations, electricity trading, and equipment maintenance, ultimately achieving the core goals of cost reduction, efficiency improvement, and revenue assurance.

The deployment targets two wind farms and one photovoltaic power station under Shenergy Group and Shenergy Co., Ltd.

(III) Implementation Plan

Implemented in cooperation with Shenergy (Group) Co., Ltd. and Shenergy Co., Ltd., the new energy power generation forecasting intelligent agent system includes the following components:

1. Ultra-short-term Forecasting (next 4 hours, rolling execution at 15-minute intervals): Enhances real-time dispatching responsiveness and the economic performance of grid assessments. Accurate predictions allow power enterprises to respond swiftly to dispatch instructions, effectively manage the stochastic fluctuations of wind and solar output, and reduce penalties caused by forecasting deviations. It also enables intelligent scheduling of equipment maintenance during periods of low wind or irradiance, minimizing generation losses and optimizing operational costs.

2. Short-term Forecasting (next 1 day, at 15-minute intervals): Optimizes generation schedules and trading strategies. It provides the crucial basis for formulating intraday generation plans that better align with load demand. High-accuracy forecasting is the cornerstone of power trading, supporting spot market bidding and enabling more profitable trading strategies by anticipating electricity price fluctuations.

3. Medium-term Forecasting (next 10 days, at 15-minute intervals): Supports medium- and long-term trading and O&M planning. Accurate medium-term weather and generation forecasts serve as an essential reference when signing medium- and long-term power purchase agreements and declaring generation schedules. They also supply the data needed for advance planning of reserve capacity and maintenance outages, thereby reducing operational risks at the planning level.

The system adopts a layered architecture that spans the entire process from data input to forecast output. The data input layer ingests real-time data for ultra-short-term forecasting, historical and forecast data for short-term forecasting, and long-term and seasonal data for medium-term forecasting. The data processing layer performs real-time anomaly detection, multi-source data fusion, and long-term trend and cycle analysis correspondingly. The feature engineering layer extracts spatiotemporal variation features for ultra-short-term, meteorological features for short-term, and seasonal and climate pattern features for medium-term. The prediction model layer employs a real-time adaptive model for ultra-short-term, ensemble learning for short-term wind and solar power forecasting, and time-series and deep learning models for medium-term. Finally, the prediction output layer delivers power forecasting results at 15-minute intervals for ultra-short-term, hourly for short-term, and daily for medium-term up to 10 days ahead.The forecasting output layer ultimately delivers power forecasts at 15-minute granularity for ultra-short-term, 24-hour granularity for short-term, and 10-day granularity for medium-term horizons.

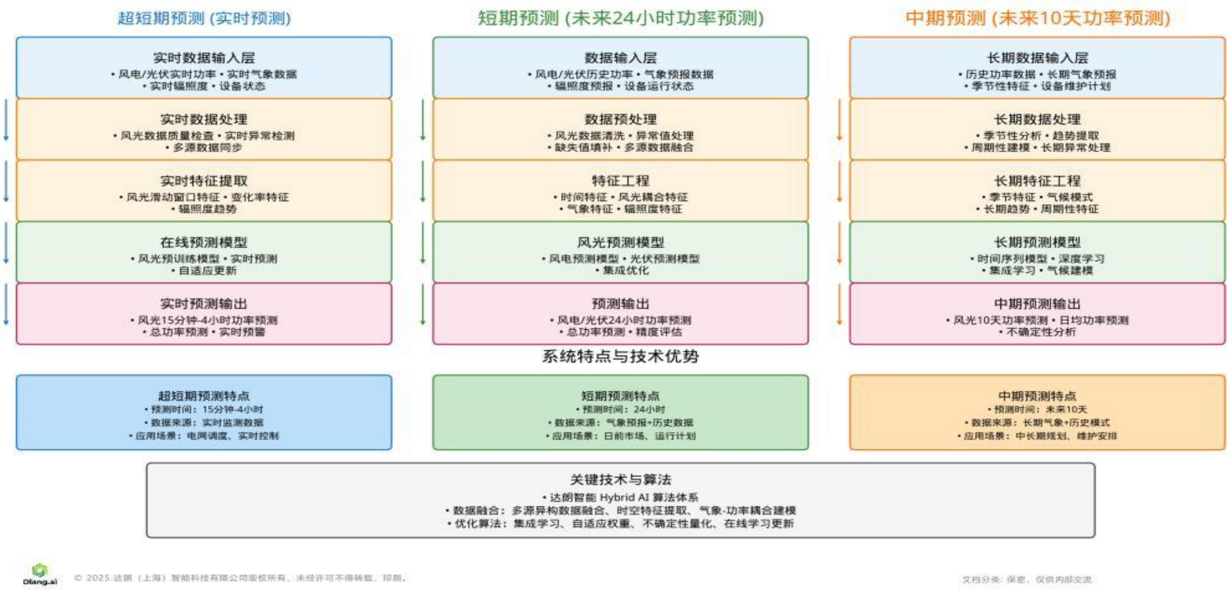


Figure 1 Architecture of the New Energy Power Generation Forecasting System

The system employs a hybrid AI algorithm framework that integrates data-driven approaches with physical constraints such as wind turbine power curves and photovoltaic conversion equations. Technological breakthroughs at the data layer are achieved through the fusion of multi-source heterogeneous data (meteorological large models, SCADA data, etc.), spatio-temporal feature extraction, and weather-to-power coupling modeling. Algorithm optimization is accomplished via ensemble learning, adaptive weighting, uncertainty quantification, and online learning updates. The final architecture employs an LSTM/Transformer primary model to capture temporal features and an XGBoost auxiliary model to handle nonlinear meteorological factors, forming a multi-task learning framework that supports both power forecasting and meteorological variable inversion.

In terms of business logic and system strengths, the solution is fully adaptable to new energy power stations across East China, covering power forecasting needs ranging from ultra-short-term grid dispatching and short-term day-ahead markets to medium- and long-term planning. It adopts an intelligent agent architecture based on MCP (Model Context Protocol) to realize a closed loop of “data request - command issuance – result feedback”, and it supports dynamic expansion. Forecasting accuracy is ensured through transfer learning, customized loss functions, extreme-scenario training, and continuous multi-indicator optimization. Additionally, the system integrates seamlessly with Shenergy Group’s AI middle platform, on-site SCADA systems, and large-screen visualization tools. Comprehensive documentation and training programs are provided to guarantee smooth business implementation.

III. Project Highlights

(I) Core Technological Breakthroughs

1. Breakthrough in accuracy bottlenecks through three-source data fusion: Innovatively integrated by the project team, the fusion system combines domestic meteorological large models (Fengwu, Pangu, Fuxi), refined grid data from local meteorological stations, and observation data from new energy power stations of Shenergy Group. The project adopts a grid alignment algorithm to eliminate spatial deviations, solving the long-standing accuracy deficiencies of traditional single-data forecasting methods. In extreme weather scenarios, the forecasting error of wind and solar power output is reduced by 15%-20%, and the forecasting accuracy across all time scales steadily meets industry standards, effectively supporting the rigid implementation of power generation plans.

2. Hybrid-driven model balancing accuracy and robustness: Adopted by the system, the innovative architecture integrates data-driven methods and physical constraints. The data-driven layer leverages LSTM/Transformer models to capture temporal features and XGBoost algorithms to process non-linear meteorological factors, while the physical constraint layer incorporates wind turbine Betz limit and photovoltaic photoelectric conversion equations. Combined

with multi-task learning that takes power forecasting as the main task and meteorological inversion as the auxiliary task, the model is well adapted to scenarios with sparse data under extreme weather conditions.

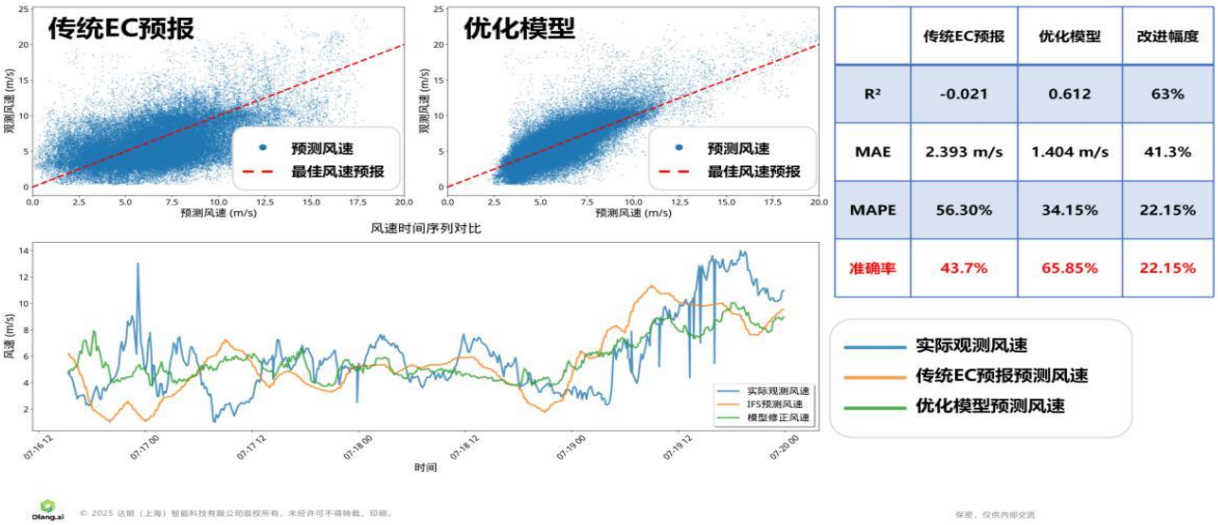


Figure 2 The optimized model achieves a 22.15% increase in accuracy rate compared with traditional EC forecasting

3. MCP-Agent architecture breaking collaborative barriers: Built based on the Model Context Protocol (MCP) by the project team, the new energy power generation forecasting intelligent agent encapsulates meteorological and energy models into standardized services. It realizes a closed-loop operation covering environment perception, decision planning, and execution feedback, and breaks the limitations of cross-model communication barriers and static rule mechanisms in traditional systems.

4. Full-stack domestic adaptation for safety and controllability: Supported by the computing power of Shenergy Group's AI middle platform, the system adopts domestic general large models including Qwen3-32B and domestic professional meteorological large models (Fengwu, Pangu, Fuxi). All data is stored in independent servers of meteorological departments and the group's middle platform, fully complying with domestic IT application innovation requirements and realizing independent and controllable core technologies and data security.

(II) Replicable and Promotable Experience

1. Resolution of common industry problems with wide scenario adaptability: The system effectively solves prevalent pain points in the new energy industry, including forecasting deviations in extreme weather, multi-source data silos, and insufficient data of newly built power stations. Through transfer learning and fine-tuning with a small amount of station data, it can quickly adapt to new scenarios. It is applicable to wind power and photovoltaic power stations in all climate zones across China, with superior adaptability especially in areas frequently affected by extreme weather such as typhoons, high temperature and cold waves.

2. Reusable technologies and tools:

Data governance tools: A full set of self-developed tools including sliding window spatio-temporal feature construction, correlation coefficient-based feature screening, and interpolation-based abnormal value completion can be directly reused for different projects, greatly reducing data preprocessing costs and cycle time.

Model framework templates: The hybrid-driven architecture, which combines LSTM and Transformer networks with physical constraints and adopts the MCP-Agent framework, does not require extensive customization. Only fine-tuning of sub-model parameters based on the site type, whether wind or photovoltaic, is needed.

Domestic selection scheme: For the localized technology selection, priority is given to the open-source domestic Qwen3-32B large model and domestic meteorological foundation models such as Fengwu and Pangu. This approach reduces external procurement costs and technology dependence while aligning with enterprise Xinchuang requirements.

3. Standardized implementation process: Adopted by the project, the three-phase promotion method is highly replicable for enterprises in the energy industry:

Month 1: Demand research and data preprocessing, including docking and standardized governance of meteorological and power station data;

Month 2: Scheme design and model development, including building hybrid models and Agent architecture, and completing model training and optimization;

Month 3-4: System integration, online deployment and optimization, including interface development, operational testing and final project delivery.

4. Promotable cooperation model: The project replicates a tripartite cooperation model consisting of implementer (Dalang Intelligent), initiator (energy enterprise) and data partner (local meteorological bureau/meteorological data company). In this model, the implementer provides technical landing capabilities, the initiator puts forward business demands and provides computing power resources, and the data partner supplies localized high-precision meteorological data. This forms a collaborative system with clear rights and obligations and complementary resources.

IV. Implementation Results

(I) Economic Benefits

1. Significant reduction in assessment fines and direct profit improvement: Applied to 2 wind farms and 1 photovoltaic power station under Shenergy Group and Shenergy Co., Ltd. in Jiangsu Province, the system raises the power generation forecasting accuracy to over 90%. It reduces annual assessment fines under the “Two Detailed Rules” by 1.5-3.5 million RMB, accounting for 40%-60% of the total traditional assessment expenditure. As the most direct and stable benefit, it eliminates rigid expenditure caused by forecasting deviations and is directly converted into operating profits of power stations.

2. Reduction in wind and solar curtailment and optimization of asset returns: High-precision short-term forecasting provides reliable data support for power grid dispatching, helping the power grid optimize new energy consumption. The wind and solar curtailment rate of the power station cluster is reduced by 2%-5%, generating an additional annual power generation benefit of 800,000-2 million RMB. By improving forecasting accuracy to optimize quotations in the electricity spot market, the trading deviation assessment loss is reduced by 20%-40%, cutting annual losses by 500,000-1 million RMB and effectively guaranteeing power generation revenue.

3. Dual optimization of O&M efficiency and trading strategies for cost reduction and efficiency improvement: Precision power forecasting supports predictive maintenance arrangement, enabling accurate deployment of fault maintenance windows during low power generation periods and reducing unplanned outage losses, saving 300,000-600,000 RMB in annual operation and maintenance costs. Medium-term forecasting provides a key decision-making basis for the three power stations to declare medium and long-term contractual power, controlling the contract deviation rate within 3% and avoiding additional power purchase costs or fines caused by data deviations.

(II). Technological Innovation Achievements

1. Breakthrough in forecasting accuracy: In extreme weather scenarios, the forecasting error of wind and solar power output is narrowed by 15%-20%, and the forecasting accuracy across all time scales steadily meets industry standards, strongly supporting the rigid and accurate implementation of power generation plans.

Forecasting Time Scale	**Wind Power Forecasting Accuracy**	**Photovoltaic Power Forecasting Accuracy**
Ultra-short-term (4H)	≥92%	≥97%
Short-term (1 Day)	≥90%	≥93%
Medium-term (10 Days)	≥85%	≥88%

2. Improvement of data collaboration efficiency: The integration cycle of multi-source heterogeneous data is shortened from the original 72 hours to 4 hours, and the response delay of model invocation is controlled within 500ms, realizing efficient and real-time data analysis and model output.

(III) Social Benefits

1. Safeguarding energy security and empowering the scientific implementation of the dual-carbon goals: Meteorological large models provide high-precision and long-time-series wind and solar resource forecasting, converting meteorological uncertainty into manageable operational risks. Professional industry models further convert meteorological data into accurate power output data, making green power a predictable and programmable reliable energy source for the power grid and gradually replacing fossil energy. The project helps a 500MW power station cluster reduce carbon emissions by 36,000 tons annually, directly serving the national carbon neutrality and carbon peaking goals.

2. Reducing social energy consumption costs and realizing inclusive economic benefits: By improving forecasting accuracy, the intelligent agent significantly reduces the standby capacity and regulation costs required by the power grid to balance the volatility of new energy power generation. The saved system costs ultimately benefit all power users and promote the decline of green power tariffs. By reducing wind and solar curtailment and delivering more low-cost green power to the grid, the system stabilizes electricity prices and fundamentally lowers the overall social power consumption cost.

3. Enhancing power grid resilience and ensuring stable power supply for people's livelihood: The agent's ultra-short-term forecasting capability gives grid dispatchers a forward-looking view, much like a telescope that peers into the future, enabling early warning of sharp rises and drops in wind and solar power output 10-30 minutes in advance. It allows dispatching departments to deploy standby resources in advance, greatly enhancing the power grid's resilience in response to extreme weather and emergencies and ensuring power supply reliability. During the extreme high temperature in Shanghai in 2024, the pilot power stations adjusted power generation plans in advance relying on the model, helping the power grid steadily withstand the peak load of 40 million kilowatts and reducing power outage duration by more than 30%.

Case 35

Integrated Implementation Practice of Foundation Model and Algorithm Security Assessment

Shanghai Zhidaxin Technology Co., Ltd.

I.Basic Enterprise Information

Established in July 2025, Shanghai Zhidaxin Technology Co., Ltd. has a team originating from the institute of cybersecurity at Shanghai Jiao Tong University, a national first-class cybersecurity demonstration college. The institute ranks first in Asia in the field of information security. The team has long been committed to the innovation and practical application of artificial intelligence technology related to public security, actively cooperated with Shanghai Public Security Bureau, Cyberspace Administration of China and other units to carry out scientific research, and achieved demonstration value in the direction of police revitalization through science and technology and cyberspace security.

The team performed outstandingly at the 2024 Global AI Attack and Defense Challenge and won the "Outstanding Solution Award", and was invited to exhibit at the cybersecurity exhibition of the 2024 World Internet Conference in Wuzhen, comprehensively demonstrating the latest research progress of multi-mode large model security. The team works closely with the Network Security Corps of the Municipal Public Security Bureau to jointly build an intelligent policing platform and promote the construction and application of the intelligent research and judgment system of police big data. This has supported the applications for the Shanghai Public Security Bureau's "Revitalizing Policing through Science and Technology" team and for the Ministry of Public Security Science and Technology Award. At present, the company has about 20 employees, accounting for more than 60% . It has a complete technical system and successful industry cases in AI algorithm security assessment, model fingerprint and notarization, watermark traceability, data desensitization and synthesis, automatic security assessment, AI compliance integration service and other aspects.

II. Case Introduction

(I) Background and Needs

With the rapid popularization of generative artificial intelligence in government affairs, finance, medical care, education, industrial manufacturing and other fields, the role of AI model in content generation, decision support and process automation is increasingly prominent. But at the SME level, this process is constrained by multiple realities. Based on the in-depth analysis of the needs of these real landing scenarios, Zhidaxin Technology proposed and built a systematic solution.

Figure 1 Necessity of large model filing

1. Regulatory pressure and compliance demand coexist

At the policy level, *the Interim Measures for the Management of Generative Artificial Intelligence Services, Regulations on the Management of Algorithm Recommendation, Network Security Law* and other regulations have been issued intensively. For AI application enterprises, models must pass multi-dimensional evaluation such as algorithm filing, security assessment, data traceability and content compliance review before going online. In particular, the uncertainty of the content output of the generative model makes it urgent for enterprises to have an independent and executable security compliance tool chain to meet the requirements of policy audit.

2. The disconnect between the speed of business innovation and security capabilities Small and medium-sized enterprises pursue “quick launch and quick verification” in AI innovation, but they often lack the support of security and risk control team. The traditional manual review and compliance outsourcing method has a long cycle and high cost, which cannot adapt to the speed of AI product iteration. Enterprises hope to implement low-threshold, self-service model security assessment through an automated and reusable AI security assessment platform to ensure the compatibility of innovation pace and security requirements.

3. Differentiated risks in multi-industry and multi-modal scenarios

Different industries face different compliance priorities:

(1) In the field of government affairs and education, focus on content authenticity and political security;

(2) In medical scenarios, privacy protection and professional advice risks are emphasized;

(3) In the financial field, the focus is on algorithm fairness and fraud prevention;

(4) In the industrial and manufacturing links, it is necessary to ensure the traceability and stability of model decisions.

These differentiated demands make enterprises urgently need a set of configurable and extensible security evaluation system, which can quickly generate compliance documents and evaluation reports for different vertical scenarios.

4. Responsibility for content generation and lack of forensic ability

In the large-scale application of AI-generated content, responsibility division and evidence storage have become the core problems of legal compliance. Enterprises hope to have the ability of full-link storage from model generation to content output to ensure that they can “prove their innocence” by technical means in case of violation or infringement disputes.

5. Systematic requirements for data asset governance

Data is both the “fuel” of the model and the biggest source of risk. Enterprises generally lack the ability to assess data provenance compliance, classify sensitive data, and control access rights. In view of this pain point, Zhidaxin Technology proposed an integrated scheme of “data map + desensitization + synthetic data” to help enterprises reconstruct data usage boundaries under the premise of security. Zhidaxin Technology’s AI service security compliance assessment platform is designed for these typical scenarios. Zhidaxin Technology has long been deeply engaged in six services, including “large model security assessment, algorithm security assessment, model fingerprint and notarization, data desensitization and synthesis, and AI all-in-one machine security assessment”, forming a “one-stop” compliance and technical solution covering the whole life cycle of large models. It transforms complex policy requirements, algorithm evaluation, data governance and judicial storage into modular, automatic and landing product capabilities, so that SMEs can quickly have security defense and compliance governance capabilities with limited resources, and achieve a fundamental shift from passively responding to regulatory requirements to proactively building trust.

(II) Introduction of implementation plan

1. System construction content

The core module of security assessment is to build a two-layer detection system that combines security test question bank and AI recognition model. The test question bank covers 5 categories and 31 small categories of illegal scenes, such as political sensitivity, sexual violence and discrimination. The AI recognition model supports multi-modal input, such as text, image and voice, and performs compliance detection for generated content through AI recognition mode. Optimize model security to achieve red line clearing before going online.

The built-in automatic attack process simulates more than 10 hacker attack methods such as prompt word attack, role playing and jailbreak. The model is reinforced by command alignment technology such as security fine-tuning. Meanwhile, the privacy disclosure detection engine is integrated to prevent personal privacy disclosure by using de-identification strategy.

Algorithm transparency evaluation module integrates interpretable algorithm, quantifies model feature contribution and decision deviation, outputs algorithm fairness analysis report, generates compliance supporting materials that meet regulatory requirements, and solves the regulatory trust problem caused by algorithm black box.

Compliance governance module and data governance module provide semi-automated data mapping tools to perform dynamic desensitization of sensitive fields, including shielding, replacement, encryption and other policies, and integrate synthetic data generation engine to supplement training data for scenarios with insufficient samples and ensure data compliance.

Automatic generation module of filing materials: built-in regulatory mapping templates such as *Interim Measures for Generative Artificial Intelligence Service Management* and *Provisions on Algorithm Recommendation Management* can automatically generate core documents such as *Algorithm Filing Form, Safety Assessment Report and Risk Rectification List*, covering the filing requirements of algorithm recommendation, generative AI and other types of services.

The fingerprint identification system generates a unique fingerprint (resistant to fine-tuning disturbances) for each model and supports bidirectional traceability between mode and output content.

Traceability and storage module

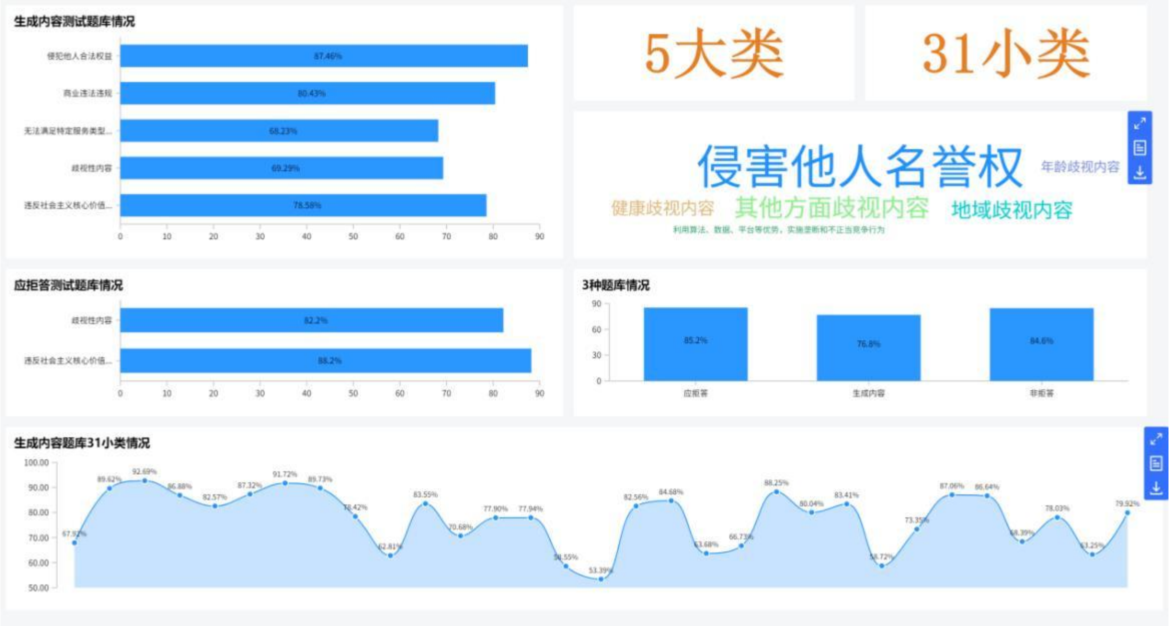


Figure 2 Overall test overview

(III) System architecture

Basic layer: The system is deployed in a private server environment and uses a hybrid architecture of relational database and distributed data storage to store structured data such as evaluation configuration and compliance indicators to ensure the security of core data.

Core layer: it contains four core engines: multi-mode recognition algorithm, interpretable algorithm, adversarial attack algorithm and synthetic data generation algorithm. The algorithm model is based on the transformation of scientific research achievements of School of Cyberspace Security, Shanghai Jiao Tong University. The integrated model includes fingerprint generation algorithm, digital watermark embedding algorithm, data desensitization algorithm and permission control engine to construct a full-link security protection system.

Functional application layer: provides core evaluation functions such as multi-mode content detection, adversarial testing, algorithm transparency evaluation, and supports custom test scenarios and index thresholds. Including record material generation, compliance review, rectification tracking and other functions, to achieve full closed-loop management of compliance process; It provides data governance planning such as data mapping, desensitization of sensitive data, and generation of synthetic data.

Interaction layer: Provides Web management background and API interface for interaction. The Web background supports visual configuration and report generation, and the API interface supports seamless connection with the existing AI system of the enterprise.

Technical architecture features: It provides three interactive modes: Web management background, API interface and mobile terminal alarm entrance. The Web background supports visual configuration and report generation, the API interface supports seamless docking with the existing AI system of the enterprise, and the mobile terminal supports real-time security alarm push. The mechanism of master-slave backup, load balancing and automatic fault transfer is adopted to ensure the stable operation of the system for 7×24 hours, and the test task processing delay is less than 500ms.

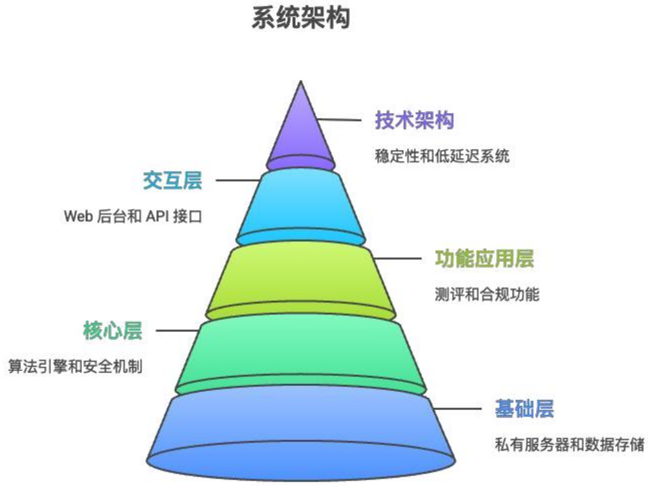


Figure 3 System architecture

(IV) Business logic

Security compliance lifecycle process

1. Diagnosis stage: data tools were used to sort out the source and flow path of model training data to identify data compliance risks. Automated scanning tools are used to detect security risks such as model content risk control shortcomings and adversarial vulnerabilities.

2. Evaluation stage: Based on the industry compliance template, perform four core evaluation tasks: multi-mode content compliance evaluation, resistance attack test, algorithm transparency evaluation and data security evaluation, and generate evaluation reports containing 87 subdivision indicators.

3. Reinforcement stage: According to the problems found in the evaluation, strengthen the model’s resistance through security fine-tuning and instruction alignment; Adopt data desensitization and de-identification strategies to fix data security vulnerabilities. Decision bias is reduced by algorithm optimization.

4. Monitoring stage: Monitor the output content of the model in real time after it goes online, identify new violation risks based on the dynamically updated rule library, automatically alarm and record logs when exceptions are triggered.

5. Storage stage: Key data such as model fingerprint, generated content watermark, evaluation report and monitoring log are stored on blockchain to form a full-link traceable evidence chain.

Figure 4 Compliance assessment results

Through the above multi-dimensional AI technology application and systematic governance, Zhidaxin Technology helps customers fundamentally build security compliance capabilities covering the whole life cycle of the model, and realizes the transformation from “passive response to supervision” to “active defense and trusted operation”.

III. Project Highlights

(I) Basic support platform construction

Linkage of three open source platforms: Integrate the resources of three platforms, “large model security range”, “large model security evaluation list” and “AI security vulnerability Library”, and access more than 2 million large model related vulnerability data in total, providing real-time vulnerability analysis and security protection support for evaluation.

The security test question bank comprehensively covers 5 categories and 31 small categories of violation scenarios, including political sensitivity, sexual violence and discrimination, and has built more than 180,000 security test question banks.

Figure 5 Large model safety evaluation list

(II) Flexible application scenarios

Differentiated adaptation of the industry: dynamically adjust the weight of evaluation indicators, compliance rules and report templates for different scenarios such as government affairs (focusing on political security and content authenticity), medical care (focusing on privacy protection and professional risk), finance (focusing on algorithm fairness and fraud prevention), and industrial manufacturing (focusing on decision traceability) through the industry parameter configuration module.

Multi-mode model adaptation: built-in model adaptation interface supports automatic recognition of different types of models such as large language model, visual language model, Wensheng graph model and Wensheng video model, and calls corresponding evaluation tools and detection algorithms to ensure the accuracy of evaluation.

(III) Core Technological Breakthroughs

The technological innovation achieved by Zhidaxin Technology in this project is not a simple function stack, but comes from the deep understanding and systematic practice of the nature of AI security and industrial implementation requirements.

1. The introduction of the concept of active defense

Traditional AI security management is mostly passive remediation, but the “AI security Red and Blue confrontation system” proposed by Zhidaxin Technology takes extreme simulation and dynamic attack and defense as the core, realizes the “immune growth” of AI model through automatic generation of confrontation samples and security reinforcement training, and truly puts security defense in the model training stage.

2. Algorithm explainability and transparency mechanism

In view of the supervision and trust problems caused by the “black box” of AI model, the project introduced interpretable algorithms such as LIME and SHAP, and built an algorithm transparency analysis engine. The system can automatically generate feature contribution report and deviation quantitative index, so that the risk and decision logic of the model can be audited and tracked, providing data support for the fairness and compliance of the algorithm.

3. Integration innovation of model fingerprint and digital watermark

For the first time in the industry, the “model fingerprint” and “digital watermark” technology are deeply combined, and the dual uniqueness identification of AI model and output content is realized through encryption signature and blockchain storage. The mechanism not only supports judicial-level content traceability and evidence confirmation, but also establishes a digital trust infrastructure with credibility for the AI industry.

4. Compliance data generation and governance capability

The project pioneered the integrated engine of “data map + automatic desensitization + synthetic data”, which generates high-quality samples in line with business distribution laws on the premise of ensuring data security and meets the needs of model iteration and differentiated training. This establishes a new balance path between privacy protection and model optimization for small and medium-sized enterprises.5.Reusable security assessment and rule base system.

5. Reusable Security Assessment and Rule Base System

By modular encapsulation of evaluation scripts, rule engines and violation sample libraries, a “reusable security baseline” is formed. Enterprises can quickly call in subsequent projects to complete security compliance re-testing and continuous optimization under different products and industry scenarios, greatly improving the evaluation efficiency and versatility.

6. Judicially credible chain of digital evidence

By cooperating with third-party notary institutions and blockchain technology service providers, Zidaxin Technology has established an end-to-end digital evidence chain to realize the full-link association of model fingerprint, content watermark and notarization certificate. This mechanism enables AI products to have the legal guarantee ability of forensics, traceability and tamper-proof, becoming the most innovative technical support in AI compliance system.

(IV) Scenario Uniqueness

1. Focus on compliance pain points of small and medium-sized enterprises

Unlike traditional security assessment solutions for large organizations, Zhidaxin Technology designs AI security compliance solutions that can be deployed with low barriers based on the actual capabilities and cost structure of smes. The system transforms tedious regulations interpretation, material preparation and algorithm evaluation into modular tools, enabling innovative enterprises to meet regulatory requirements with limited resources.

2. Cross-industry migrable architecture

The security assessment and compliance template of the project can be adapted to the specific requirements of different fields such as government affairs, medical care, finance and Internet. The platform realizes differentiated compliance mapping of the industry through parametric configuration, enabling the same technology system to be efficiently reused in different vertical industries, and significantly improving universality and promotion.

3. In-depth combination of actual combat and scientific research achievements

The team is from the School of Cyberspace Security, Shanghai Jiao Tong University, and has been involved in public security system science and technology projects and national scientific research tasks for a long time. The evaluation method, red-blue confrontation strategy and synthetic data algorithm adopted in the project are all derived from the engineering transformation of scientific research results, which give consideration to both technological advancement and practical maturity.

4. Dual Characteristics of High Trustworthiness and High Efficiency

The project adheres to the principle of “zero tolerance” in security, while achieving “high efficiency” through tool chain automation. In practice, it only takes 8-10 weeks for customers to go online from start-up, and the testing, rectification and evidence collection processes can be completed in a closed loop, greatly shortening the traditional compliance review cycle.

IV. Implementation Results

(I) Economic benefits

During the filing of the large model of Shanghai Wuyin Transition Technology Co., LTD., the cycle was significantly shortened, directly reducing the time cost. Through the automatic tool chain of evaluation, the traditional manual compliance review cycle of 12-16 weeks was compressed to 8-10 weeks, and the cycle shortening rate was more than 30%. At the same time, the number of roundtrip communication between the enterprise and the regulatory authorities is reduced, the average rework rate of a single project is reduced by less than 3% from 25% of the traditional mode, and the human input is reduced by about 40%(according to the calculation of the 20-person R&D team, a single project can save 80 people/day of working hours), which greatly reduces the cost of the case.

Operational risk costs are greatly reduced to avoid violation losses, and the probability of model output of politically sensitive, pornographic and violent high-risk violation content is minimized through multi-modal content risk control and adversarial testing. Avoid a single administrative fine of up to 500,000 yuan and loss of service removal (refer to the punishment standard of *the Interim Measures on the Management of Generative Artificial Intelligence Services*).

Improve reusability and reduce long-term investment to form templating evaluation assets (including 12 sets of industry compliance templates and 31 types of test scripts). More than 85% of evaluation processes and materials can be directly reused during iteration of similar AI products, and the labor input of new projects can be reduced by 55%).

(II) Social Benefits

Practice has proved that small and medium-sized enterprises are fully capable and have a path to control cutting-edge AI technology and build safe, reliable and competitive innovative applications on the basis of regulations. We believe that this case has an important “pioneer” demonstration significance in promoting the universal and standardized development of AI technology. So that many science and technology enterprises rushed to a vast frontier to explore in the “AI+” space.

The AI compliance threshold of small and medium-sized enterprises is significantly reduced: the platform supports self-service evaluation of small and medium-sized enterprises, and the compliance start-up cost of a single enterprise is reduced from 200,000 yuan of traditional outsourcing to 100,000 yuan, and the threshold is reduced by more than 55%. It has helped many small and medium-sized enterprises in gaming and education in Minhang District of Shanghai to complete AI model filing, and the filing pass rate reaches 100%.

Popularization of compliance ability: Through the mode of “AI security Public service station”, more than 2 million vulnerability data and free evaluation tools have been opened to enterprises in the park, serving more than 50 enterprises in total, and the coverage rate of AI security evaluation of small and medium-sized enterprises has increased from 18% before the service to 65%.

Accelerated construction of AI industry compliance ecology: the “5-step closed-loop evaluation process” and “87 compliance indicators” precipitated by the project provide standardized reference for 200+AI enterprises in the region;

Risk prevention and control effectiveness: The “Large Model Security evaluation list” has covered 218 multi-mode models and 11 dialogue models, promoting the average resistance rate of models in the industry to increase from 62% to 89%, effectively reducing the risk of AI technology abuse.

Promotion of universal benefit of trusted AI technology: it helps enterprises realize the transformation of “from passive compliance to active defense”. 80% of small and medium-sized enterprises in the service object have the security management ability of the whole life cycle of AI model for the first time; At the same time, the low-cost application of synthetic data, model fingerprint and other technologies will be promoted, increasing the compliance rate of AI model training data in small and medium-sized enterprises from 45% to 91%, and accelerating the safe implementation of AI technology in government affairs, medical care and other livelihood fields.

Case 36

Lingang AIDC Computing-power and Electricity Coordination Platform

Shanghai Damao Technology Company Limited

I. Basic Enterprise Information

Founded in March 2021, Damao Technology is the first high-tech enterprise in China that applies the energy large models to the collaborative optimization of computing power and electricity. Relying on its full-stack self-developed energy large model, the company has built an adaptive energy operating system and developed innovative products including an integrated computing-power and electricity coordination platform. It provides comprehensive energy optimization solutions for high-energy-consumption computing infrastructure such as intelligent computing centers, general computing centers and supercomputing centers, facilitating the sustainable development of the digital economy. Compared with traditional policy algorithms, Damao's energy large model boasts powerful generalization and reasoning capabilities in vertical industries, which greatly improves the real-time performance, accuracy and interpretability of energy system balance strategy generation, with a prediction accuracy of over 90% and a decision-making accuracy of more than 95%.

Centered on its self-developed energy large model, Damao Technology's computing-power and electricity coordination platform integrates precise forecasting, dynamic decision-making and real-time dispatching. It realizes accurate coordination between computing power and electric power, fully leverages China's advantages in energy structure, cuts operating costs of the AI industry, boosts the consumption of clean energy, and enhances the country's core technological competitiveness.

At present, Damao Technology has established in-depth cooperation with leading enterprises including SenseTime, Contemporary Amperex Technology Company Limited (CATL), Sugon and Fujian Big Data Group, and has also secured investment from Puquan Capital (CATL Capital), the exclusive industrial investment platform under Cambricon, SenseTime and CATL. The company's energy large model has completed the filing from the Cyberspace Administration of China. Its implemented projects have respectively won the Big Data Industry Development Demonstration Project of the Ministry of Industry and Information Technology, the Top Ten Vertical Scenarios of Artificial Intelligence in Digital Shanghai 2024, and the 2025 Special Demonstration Project for Promoting High-Quality Industrial Development of Shanghai, leading the vertical large model track.

II. Case Introduction

(I) Background and Needs

Countries across the globe are ramping up investment in artificial intelligence technologies, which has not only fueled the rapid advancement of AI but also triggered an explosive growth of the computing power industry. As the core infrastructure underpinning AI development, intelligent computing center are springing up worldwide. Forecasts indicate that by 2030, the annual power consumption of intelligent computing centers will account for 5% to 10% of the total annual electricity consumption of the whole society, and their incremental power use will make up 27% to 77% of the national total newly added electricity consumption in that year.

To cope with such changes, the state has successively issued a series of policies mandating a higher proportion of new energy power consumption in data centers and lower PUE (Power Usage Effectiveness, the ratio of total energy consumed by a data center to energy consumed by its IT loads). In December 2023, the National Development and Reform Commission (NDRC) issued the *Implementation plan on Further carrying out the "East Data, West Computing" Project and speeding up the construction of a National Integrated Computing Power Network*. The document proposes the integrated and coordinated development of computing power and green power, and advances the construction of a nationwide integrated computing power network featuring interconnected dispatching, universal accessibility, environmental friendliness and security. In 2024, the NDRC and two other central authorities jointly released policies to support the coordinated construction of data centers and new energy bases. In March 2025, the NDRC, the National Energy Administration (NEA) and other relevant authorities jointly issued the *Opinions on*

Promoting the High-Quality Development of the Renewable Energy Green Power Certificate Market. At the special press conference on Green Electricity Certificates, the National Data Administration unveiled China's green power consumption target for data centers: China will strive to achieve an 80% proportion of green power in the newly-built data centers of national hub nodes by 2025, and further raise the green power usage ratio of data centers on this basis. On May 21, 2025, the NDRC and the NEA jointly issued the *Notice on Matters Concerning the Orderly Promotion of the Development of Green Electricity Direct Connection*, which provides new top-level guidance on raising the proportion of green power consumed by intelligent computing centers and advancing green electricity direct connection.

As China's largest single intelligent computing center, SenseTime Lingang intelligence computing center incurs energy costs of over 100 million yuan each year. Nevertheless, the workload of intelligent computing centers is highly volatile, prone to sudden surges and hard to predict, resulting in disorderly electricity consumption. Power generation from green energy sources such as wind power and photovoltaic power is also random. Therefore, the key for intelligent computing centers to save energy and boost green power consumption lies in establishing a linkage between computing power and energy, and realizing bidirectional dispatching of computing power and energy by combining the forecasts of both sides.

Based on its full-stack self-developed energy large model, Damao Technology has built an adaptive energy operating system and developed a variety of innovative applications including an integrated computing-power and electricity coordination platform. The company delivers complete energy optimization solutions for high-power-consuming computing infrastructure such as intelligent computing centers, general computing centers and supercomputing centers, advancing the sustainable development of the digital economy.

Figure 1 Integrated Computing-Power and Electricity Coordination Solution of Damao Technology

(II) Implementation Plans

Faced with the randomness of workloads in intelligent computing centers and the volatility of new energy power generation, bidirectional dispatching of computing power and energy is the key to cutting operating costs and boosting the consumption of clean energy.

Built on the practical idea of two-way empowerment between computing power and electricity, the Lingang AIDC Computing-Power and Electricity Coordination Platform integrates Damao Technology's energy large model, SenseTime's self-developed multimodal foundational large model and CATL's energy storage technology. Leveraging the integrated capabilities of the heterogeneous computing elastic scheduling platform and the dynamic flexible energy dispatching platform, combined with full-volume data from SenseTime's intelligent operation and maintenance platform, the system forms a full-chain intelligent dispatching framework covering "data-algorithm-strategy-execution-verification". Relying on a two-way application closed loop of "computing load follows electricity supply" and "electricity supply adapts to computing demand", it achieves precise matching between computing power and energy.

Figure 2 36MWh Lithium Battery Energy Storage System of Lingang AIDC

Damao's Computing-Power and Electricity Coordination Platform delivers all-round support for the operation and management of group-level intelligent computing centers via a series of functional modules. Its core capabilities cover management, dispatching, response, power trading, operation & maintenance, and value-added services.

Figure 3 Core Functional Modules of Damao Computing-Power and Electricity Coordination Platform

III. Project Highlights

Built on foundational large models, the Damao energy large model undergoes pre-training using energy industry datasets and automated labeling tools. It is then fine-tuned to adapt to industrial traits, regional features and diverse application scenarios. Paired with mounted knowledge bases, it constitutes a full-fledged LM+Agent energy large model architecture.

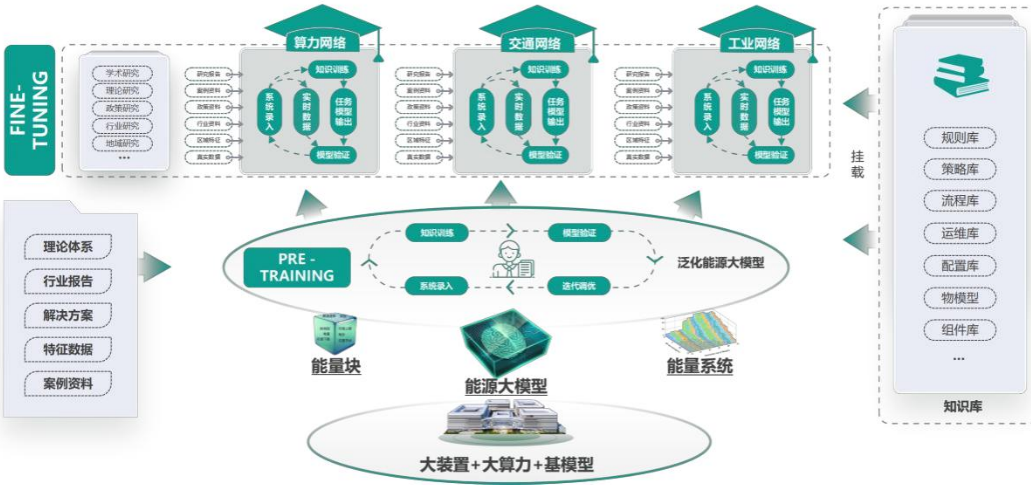


Figure 4 Technical Architecture of Damao Energy Large Model

Leveraging advanced algorithms including deep learning and big data analytics, Damao Technology's energy large model enables precise modeling of power systems supported by TB-scale datasets. With a parameter scale reaching the hundred-billion level, the model is capable of capturing complex correlations among various system variables, such as new energy generation characteristics, load fluctuation patterns, meteorological impacts and grid state evolution. This delivers dynamic, accurate global forecasting and real-time intelligent optimization of the entire power system. Compared with traditional policy algorithms, Damao's energy large model boasts powerful generalization and reasoning capabilities in vertical industries. It greatly improves the real-time performance, accuracy and interpretability of energy system balance strategy generation, with prediction accuracy exceeding 90% and decision accuracy above 95%.

(I) Core Technological Breakthroughs

1. Cross-Domain and Multi-Dimensional Data Fusion

One of the core technological innovations of Damao Energy Large Model lies in cross-domain and multi-dimensional data fusion. By interconnecting multi-dimensional data including power operation monitoring, segmented industrial operation monitoring, meteorological data, and equipment asset lifecycle management, the model enables seamless information interconnection across disparate systems. Such cross-domain data fusion not only comprehensively enhances the power system's responsiveness to complex operational issues, but also delivers more holistic and precise decision support for real-time grid dispatching, energy management and power market transactions.

2. Application of Long-Cycle and High-Quality Data

Long-cycle high-quality data constitutes one of the core strengths in the R&D of energy large models. Trained on 18 months of high-quality data supplied by scenario operators, the energy large model can deeply learn energy characteristics across various segmented scenarios to deliver accurate predictions. Long-cycle data provides comprehensive historical information, allowing the model to identify long-term trends and short-term fluctuations of energy systems and further elevate forecasting accuracy for power loads, equipment operating statuses and new energy generation output. This innovation greatly reinforces the model's decision-support capability under uncertain operating environments, empowering power systems with scientific dispatching and emergency response capacity amid dynamically evolving market landscapes.

3. Industry-Original Unique Data Structure

Drawing on expertise from electricity and energy system specialists, the project team designed a proprietary "energy block" data structure dedicated to energy balance models based on the fundamental "energy block" theory of power trading systems. This data structure describes and characterizes energy via discretization, which renders energy data processing more logical and interpretable, while substantially cutting the computational cost of model training and inference. Thanks to the innovation of this data structure, the energy large model can efficiently process massive datasets while maintaining high forecasting accuracy and computational efficiency.

4. Cloud-Edge Collaboration and Edge Agent Integrating Training and Inference

Another key innovation of the energy large model is the cloud-edge collaborative edge agent integrating training and inference. Integrating perception, forecasting, decision-making, dispatching and learning into one unit, this agent features self-optimization capability with unified training and inference workflows. Even under off-grid conditions, the edge agent can still complete local balance solution calculations. By integrating technologies including the cloud-edge collaboration architecture, energy block model extraction, and model computation, the demand for data transmission and computational workload during the global solving process is drastically reduced, thus improving the operational efficiency of the entire system. This innovation enables the model to still achieve efficient computation when facing large-scale complex data from the global balancing tasks of the power system, and enhances the real-time response and autonomous regulation capabilities of the energy large model in complex scenarios.

5. Real-Time Dynamic Dispatching Strategies for Energy Operation Scenarios

This project adopts real-time dynamic dispatching strategies to optimize the operational efficiency of power systems. Conventional power dispatching strategies are generally static, failing to flexibly cope with fluctuations in power demand and the intermittency of new energy sources. The Computing-Power and Electricity Coordination Platform built on the energy large model constructs a correlation model between computing power and energy by analyzing historical data of intelligent computing centers, including intrinsic energy data, user energy consumption characteristics and energy balance rules. Supported by this innovative technology, the platform delivers accurate forecasting of computing power and IT equipment loads every 15 minutes, and can dynamically adjust and optimize dispatching strategies within a 5-minute interval. It automatically generates optimal dispatching strategies in real time and executes them via cross-system linkage, realizing precise computing power forecasting, load forecasting, as well as strategy generation and revision. This capability comprehensively improves prediction accuracy and generalization performance under complex scenarios, fully demonstrating the great potential of computing-power-energy coordination technology.

(II) Scenarios Uniqueness

1.Uniqueness of Application Innovation

Damao Technology is the first high-tech enterprise in China to apply energy large models to the field of Computing-Power and Electricity Coordination optimization. Relying on its full-stack self-developed energy large model, the company has built an adaptive energy operating system and developed innovative products such as the integrated computing-power and electricity coordination platform. It provides comprehensive energy optimization solutions for high-energy-consumption computing infrastructure including intelligent computing centers, general computing centers and supercomputing centers, facilitating the sustainable development of the digital economy.

While overseas players are betting heavily on building nuclear power plants on the energy supply side, Shanghai Damao Technology Co., Ltd. has actively responded to national strategies through the innovative "AI + Energy" scheduling mode. Leveraging large model algorithms of artificial intelligence, the company has pioneered a new paradigm of computing load regulation driven by power peak-shaving signals. Computing-power load dispatching can establish a digital twin correlation between computing tasks and energy supply. With technologies including artificial intelligence, Internet of Things and blockchain, it models and analyzes the power consumption patterns of distributed computing nodes, and formulates optimal dispatching strategies accordingly. From the energy perspective, time-space shifting of computing load curves is adopted to match the fluctuation of new energy power generation, which markedly boosts the absorption rate of intermittent power sources such as wind and solar power. From the economic perspective, electricity price signals are utilized to guide computing centers to shift power consumption to off-peak hours, cutting energy costs for computing facilities including data centers.

2.Uniqueness of Model Innovation

The closed-loop model of "technical solutions + commercial verification + capital empowerment" pioneered by Damao Technology has witnessed the industrial transformation driven by the coordinated development of computing power and energy, and will also help the company grow into a key player in the global intelligent transformation of energy.

At present, the technical solutions of Damao Technology have been implemented and verified in benchmark projects such as SenseTime Lingang AIDC in Shanghai. Through precise energy dispatching, the solutions not only reduce electricity costs, but also greatly improve the self-balancing capacity of power systems, effectively coping with energy consumption shocks caused by fluctuations in computing load. In addition, the company has aggregated 100 MW of virtual power plant resources and operates computing centers with a total capacity exceeding 30,000 Petaflops, evolving into an “intelligent hub” that connects power grids and energy assets.

In the summer of 2025, the deployed projects of Damao Technology achieved a regulation response capacity of 2.14 MW solely relying on cold storage tanks. Such regulating capacity is equivalent to the scheduling capability of one conventional transformer (the maximum standard capacity of a 0.4 kV civil transformer is 2,500 kW). This project has become one of the typical cases promoted by State Grid at the virtual power plant conference held in Shanghai.

By focusing on the application and development of the Energy Large Model platform and computing-power and electricity coordination technology, Damao Technology has won the favor of numerous industry giants: it obtained strategic equity investment from SenseTime and Cambricon in 2022, reached a strategic cooperation with Schneider Electric in 2024, and in 2025, it even received direct capital support from industrial ecosystem enterprises such as CATL.

The value of this solution lies not only in regulating the supply and demand of energy matched with computing power, but also in enabling computing centers to participate in the electricity market as new market entrants relying on their load regulation capacity, so as to address the randomness and volatility accompanying new energy substitution. The load of data centers can be converted into "adjustable resources", which helps raise the consumption ratio of renewable energy and further advance the national Dual Carbon Goals.

For investors, the “model + simulation” system delivers matching value between investors and end users. It resolves the contradiction plaguing traditional energy storage investment — conservative forecasting paired with insufficient user returns. It provides critical references for commercial investment decisions and holds potential to spawn brand-new business models.

IV. Implementation Results

Against the backdrop of an era where energy transition intersects with the digital revolution, the launch of Damao’s computing-power and electricity coordination platform represents not merely a technological breakthrough, but also a profound upgrade to models of social governance and value creation. Viewed from the macro perspective of overall social well-being, its value extends far beyond simple technical services, mainly embodied in three core dimensions: consolidating energy security bottom lines, optimizing the energy mix and boosting economic benefits. Together, they inject strong momentum into advancing the green, secure and efficient development of society.

Damao Technology’s AIDC Computing-Power and Electricity Coordination Platform has been deployed at SenseTime Lingang Intelligent Computing Center, China’s first Class 5A intelligent computing center. By constructing intelligent infrastructure featuring deep integration of artificial intelligence and energy, this platform not only provides a practical energy-saving and cost-reducing solution for SenseTime Lingang intelligent computing center to advance into a new stage of “low cost, high efficiency and sustainability”, but also delivers implementable, replicable and promotable solutions for the green upgrading of AI infrastructure such as other intelligent computing centers, supercomputing centers and general-purpose computing centers.

(I) Economic Benefits

Based on its full-stack self-developed energy large model, Damao Technology has built an adaptive energy operating system and an integrated computing-power and electricity coordination platform to cut the operating costs of computing centers and boost the consumption of clean energy. By combining the energy large model with on-site intelligent operation and maintenance to carry out various energy-saving measures, the solution empowers intelligent computing centers with traditional air-cooled architectures to optimize their annual PUE from above 1.7 to below 1.28. In 2025, the project further constructed a 36 MWh lithium battery energy storage system, expanding its adjustable capacity to cover 100% of the entire intelligent computing center. Deployed with Damao's computing-power and electricity coordination platform and CATL's energy storage power station, the project cut its per-kilowatt-hour electricity cost by over 6.5%, achieving an annual electricity cost saving of more than ten million yuan.

The balancing capability of the energy large model can be further amplified under the source-grid-load-storage and direct green power connection model. First, minute-level forecasting and correction strategies guarantee the safe operation of the energy system. Second, source-load-storage matching enables the utilization rate of clean energy to exceed 80%. Third, it realizes guaranteed capacity control and docking with the spot market, lifting the project's economic efficiency by more than 10%.

(II) Social Benefits

1. Firmly Secure the Safety Bottom Line

Traditional energy systems rely on stable power generation from fossil fuels. As a high proportion of intermittent and volatile clean energy sources such as wind and solar power are integrated into the power grid, operational risks and uncertainties of the system have risen sharply. Leveraging minute-level forecasting and correction strategies, Damao's computing-power and electricity coordination platform safeguards the safe operation of the energy system.

2. Optimize the energy structure

China is committed to building a clean, low-carbon, safe and efficient energy system. However, the weather-dependent nature of clean energy poses huge challenges to its large-scale grid integration. The core capability of Damao's computing-power and electricity coordination platform lies in realizing real-time dynamic matching and coordinated interaction among the four major links of source, grid, load and storage. It supports high penetration of clean energy, optimizes the energy consumption structure of intelligent computing centers, and raises the proportion of new energy utilization.

3. Improve economic benefits

The value of the platform is reflected not only in macro structural optimization, but also directly translates into microeconomic benefits. On the one hand, through accurate load forecasting and optimized dispatching, the platform guides users in capacity control to cut their demand charges while easing the grid's investment pressure during peak periods, thereby achieving cost reduction. On the other hand, the platform opens up channels for market-based power transactions. It aggregates adjustable loads and distributed energy resources to operate as a virtual power plant that participates in the electricity spot market and ancillary service markets. Additional revenues can be generated by providing peak shaving, frequency regulation, and grid demand response services.

In summary, the value of Damao's computing-power and electricity coordination platform resides in its successful integration of cutting-edge AI technologies and urgent energy demands, delivering a comprehensive solution that is technologically advanced, economically viable and environmentally friendly. It is not merely a tool to reduce costs and boost efficiency for intelligent computing centers, but also a critical infrastructure driving the coordinated and symbiotic development of digitalization and green transformation across society at large, with broad prospects for large-scale popularization and application.

Case 37

An Efficient Shipping Operations Hub Driven by Multimodal Foundation Models

Shanghai WallTech Information Technology Co., Ltd.

I. Basic Enterprise Information

Established in 2013 and registered in Hongkou District, Shanghai, Shanghai WallTech Information Technology Co., Ltd. (WallTech) is a limited liability company solely invested by a legal person from Hong Kong, Macao, or Taiwan interests. WallTech is a leading developer and operator of SaaS cloud service platforms for cross-border logistics. The company focuses on cloud-based software services in the field of international freight forwarding and cross-border e-commerce logistics, and is committed to driving the practical application of SaaS technology within the industry, aspiring to become the most trusted platform for international logistics customers. Through more than a decade of steady growth, WallTech has built a product portfolio centered on two major platforms: CargoWare, a cloud service platform for international freight forwarding logistics, and eTower, a collaborative cloud service platform for cross-border e-commerce logistics. By integrating advanced technology with deep industry expertise, WallTech empowers international logistics enterprises in their digital transformation and upgrading, supporting their global expansion while effectively improving operational efficiency and minimizing business costs.

II. Case Introduction

(I) Background and Needs

The international shipping industry is currently at a critical juncture where policy-driven intelligent upgrading is accelerating and the limitations of traditional operating models are becoming increasingly evident. On one hand, national policies such as *the Next Generation Artificial Intelligence Development Plan* are promoting the deployment of AI in the real economy and empowering the transportation and shipping sectors, with clear targets for intelligent transformation. As a world-class international shipping hub, Shanghai is expected to take the lead in building smart ports and advancing digital shipping. On the other hand, the constraints of traditional shipping operations have become increasingly apparent, demanding technological breakthroughs. There are four specific pain points:

1. Low efficiency: Manual processing of tasks such as booking requires 4–5 hours and is highly dependent on operators' performance. Data flow across different process stages and different enterprises is hindered, relying on manual data handoffs and further extending cycle times.

2. High costs: Operations require substantial manpower, leading to high labor and error-correction costs. The reliance on paper documentation incurs costs for procurement, storage, and other overheads, while also posing risks of document loss. Additionally, the process also causes high carbon emissions which are not aligned with the trend toward green development.

3. High risks: Traditional OCR-based document recognition achieves accuracy below 90%, increasing the risk of customs declaration errors. Manual review is difficult to cover multilingual orders and regulatory requirements such as dangerous goods supervision, making it prone to oversights and misinterpretation.

4. Low digitalization: The lack of unified data standards has created data silos between different systems, requiring manual synchronization and verification. Without an integrated intelligent solution, achieving full-process digital management remains a challenge.

As a leading developer and operator of SaaS cloud service platforms for cross-border logistics, Shanghai WallTech Information Technology Co., Ltd. has leveraged its technological expertise and industry experience to launch the Shanghai WallTech AI-Integrated Application Platform project, aiming to address these industry pain points and support in building it into an international shipping hub.

(II) Implementation Plan

This project deeply integrates two core technologies: AI-powered intelligent document recognition and AI agents, to build a closed-loop system encompassing data collection, intelligent processing, and efficient execution. The specific technical solutions are as follows:

(1). AI-Powered Intelligent Document Recognition

The project adopts an architecture that combines a vision encoder with a large language model. The vision encoder enhances image quality through preprocessing, and then uses a convolutional neural network to extract visual features, such as text and tables, from document images. The LLM, integrated with a logistics knowledge base, receives the feature vectors and performs semantic understanding and key information extraction. The two are deeply integrated through a modality interface, overcoming the limitations of traditional OCR that relies solely on visual features, and significantly improving the capability to recognize complex documents.

(2) Prompt-Driven Recognition Mechanism

Instead of building fixed templates for each document format, the system relies on prompts that incorporate logistics terminology, business logic, and definitions of key fields to guide the recognition process. For example, a bill of lading prompt includes core fields such as shipper, consignee, port of loading, and bill of lading number, along with associated rules. When the model receives both the document image and the prompt, it can accurately extract the target fields. Even with non-standard bills of lading from different shipping lines, the system can still automatically identify critical information, greatly enhancing flexibility and adaptability.

(3) Logical Supplementation and Reasoning Capabilities

The model can supplement or correct missing or erroneous document information based on the recognized data and industry knowledge. For example, when information about quantities or weights is missing from a packing list, the system can calculate them based on the packaging specifications and the total number of cartons. If it detects a discrepancy between the subtotal amount and the unit price multiplied by quantity, it will flag the error and issue a prompt. In addition, the model can verify the consistency of business data across documents. For example, if discrepancies are found between the cargo information on the bill of lading and that on the commercial invoice, a compliance alert will be triggered to ensure data consistency and accuracy.

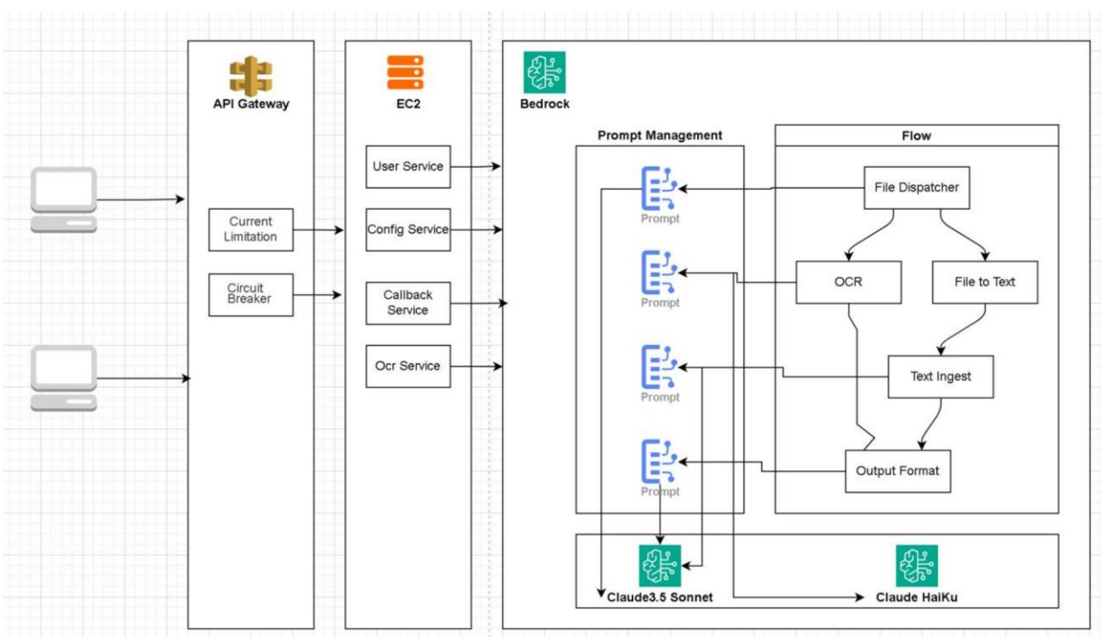


Figure 1 AI-Powered Document Recognition Workflow

(III) AI Agents

(1) Interaction Model Innovation

AI agents break away from the traditional menu-driven interaction model of conventional software, adopting a dual-mode approach that combines natural language conversation with visual operations. Users can issue business instructions directly in natural language. For example, issue the instruction that “Change the consignee of order SHSExxxxx to xxxxx International Trading Co., Ltd.” , then the agent uses natural language processing to interpret the intent, extract key information, and generate an action plan. At the same time, a visual interface displays the operational process and results, such as a comparison of the order before and after modification, enabling users to verify the changes and avoid misunderstandings. In addition, the agent supports multi-turn conversations and can respond accurately based on historical context. For instance, when a user later asks about the logistics status of that order, there is no need to restate the order number.

(2) Intelligent Reasoning and System Integration

Automated API parameter generation: The agent leverages the reasoning capabilities of Claude 3.5 to convert natural language instructions into API call parameters. The system comes pre-integrated with API documentation and parameter definitions for external systems, including the CargoWare platform, shipping line systems, and customs databases. The agent first parses the instruction to determine which interface to call, then extracts and converts the parameter values according to the interface requirements.

Cross-system data collaboration: The platform enables real-time data synchronization between CargoWare and external systems, breaking down data silos. Taking booking as an example, after a user issues a booking instruction, the agent first creates an order in CargoWare and extracts key data. It then calls the shipping line’s space availability API to identify the optimal space option, and completes the process through the booking confirmation API. Once the booking is confirmed, information such as the booking confirmation number from the shipping line is automatically synchronized back to CargoWare, updating the order status. Simultaneously, the agent calls the customs pre-declaration API to submit order information in preparation for customs clearance. The entire process requires no manual switching between systems.

(3) Lightweight Deployment and Elastic Scalability

Cloud-native architecture design: The project is built on cloud services of Amazon Web Services (AWS), utilizing components such as Lambda, RDS, and S3. Lambda supports on-demand code execution without the need for server management. When business traffic increases, the system automatically scales computing resources to handle more requests; when traffic decreases, resources are released automatically, avoiding waste and reducing IT operational costs. RDS provides a highly available database service with features such as data backup and multi-Availability Zone deployment, ensuring the security and reliability of business data. S3 is used to store large volumes of document images, business data, and system logs, offering high scalability and low cost to meet long-term data storage requirements.

No-code integration capability: The system provides low-code and no-code plug-in interfaces along with visual configuration tools. Customers can configure business processes and system integration rules through simple drag-and-drop and selection operations, without the need for professional programming skills.

(4) Additional Functional Features

Multilingual services: The agent supports real-time multilingual interaction and document translation. Users can issue instructions in languages such as Chinese, English, and Japanese, and the agent automatically identifies the language and responds accordingly. For multilingual documents, the agent can translate them into the user’s preferred language and extract key information for standardized processing.

Sentiment analysis and user experience optimization: The system integrates sentiment analysis technology to identify users’ emotional states through their language input. When dissatisfaction or anxiety is detected, for example when a user repeatedly asks why an order has not yet been confirmed, the agent will prioritize that user’s request, provide more detailed explanations and solutions, and synchronize the feedback to the enterprise customer service team for further manual follow-up, thereby enhancing user experience and satisfaction.

Intelligent BI for decision-making: The agent works in coordination with an intelligent BI system to perform in-depth mining and analysis of business data, including order volume, processing efficiency, and customer satisfaction. It generates visual reports and trend forecasts using charts such as bar charts, line charts, and pie charts.

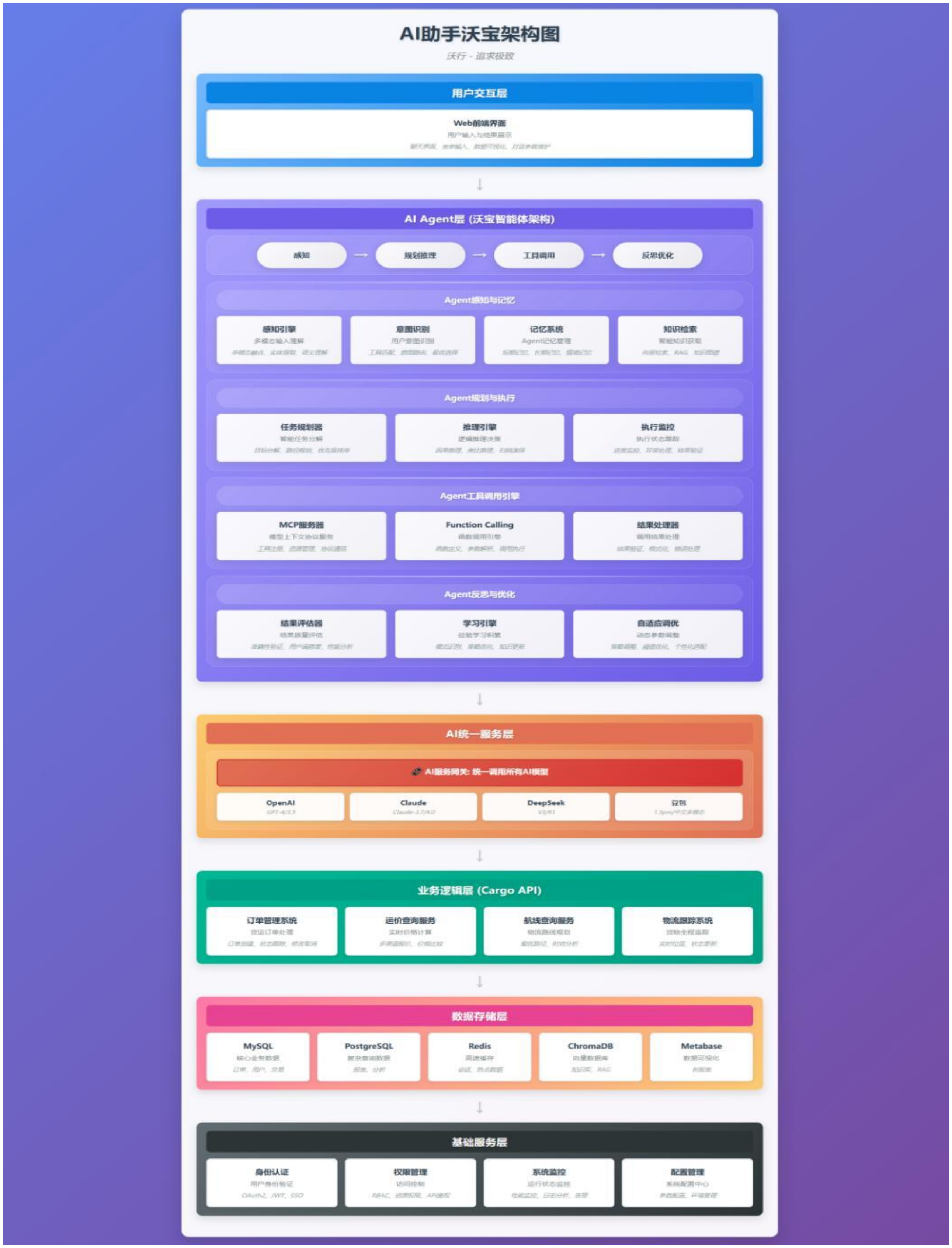


Figure 2 AI Agent Architecture Diagram

(IV) Application Scenarios

(1) International Freight Forwarding

When handling various documents such as bills of lading, customs declarations, and insurance policies, freight forwarders can use AI-powered intelligent document recognition to automatically audit up to 20 types of documents,without the need for manual verification on a case-by-case basis, reducing manual review workload by over 50%. AI agents can also complete operations such as booking and order modifications rapidly. For example, a booking process that traditionally took 4 to 5 hours can now be completed in just 15 minutes. In addition, the system provides real-time logistics tracking and keeps customers promptly informed of shipment status, significantly improving service responsiveness and customer satisfaction.

(2) Cross-Border E-Commerce Logistics

Challenges in cross-border e-commerce logistics include handling multilingual orders and integrating with logistics systems across multiple countries. This project enables precise extraction of key data, including cargo details and recipient information from orders in different languages through AI-powered document recognition. After the AI agent automatically standardizes the data, it interfaces with overseas warehouse systems and local courier systems to enable rapid order fulfillment. In addition, the AI agent can automatically adjust operational workflows based on the shipping regulations and business practices of different countries to ensure compliance. For example, when processing e-commerce orders for the European and U.S. markets, the system automatically aligns with local customs declaration requirements.

(3) Port and Multimodal Transport

Port operators can use AI-powered document recognition to quickly extract information from ship documents, cargo manifests, and other documentation. After integration, this data helps optimize vessel scheduling, berth allocation, and yard management, improving overall port operational efficiency. In multimodal transport scenarios, the system can automatically identify cross-modal documents such as railway waybills and road transport contracts, and intelligently match bulk vessel shipping information with overland transport schedules to reduce cargo transit times. For example, an international logistics company that deployed this solution saw a marked improvement in cargo transshipment efficiency on the Shanghai–Nanjing inland waterway route.

III. Project Highlights

(I) Technological Breakthroughs

WallTech AI-Integrated Application Platform project focuses on addressing the key technical pain points of the international freight forwarding industry, achieving critical breakthroughs across three dimensions, including AI application, technology integration, and architectural design. The project has established a technological innovation system tailored to industry needs, as demonstrated in the following four areas:

1. Innovation in AI-Powered Document Recognition

It breaks through the limitations of traditional OCR by integrating a multimodal large model with a logistics knowledge base. It supports the processing of multiple document types and languages, and offers logical supplementation and risk alert capabilities. The system can adapt to new scenarios through prompts, without requiring large volumes of training samples.

2. Innovation in AI Agents

It uses natural language to drive business operations, enabling automated API parameter generation and cross-system data collaboration. It supports full lifecycle management of orders and can adjust workflows according to the shipping regulations and business practices of different countries. The cloud-native architecture ensures elastic scaling, while low-code interfaces lower the barrier to customer adoption.

3. Innovation in Technology Integration

It builds a closed-loop system combining AI-powered document recognition and AI agents, achieving full-process automation from data collection and processing to execution. It fills the gap in cross-modal data processing and business automation within the shipping industry, driving deep integration between technology and business operations.

4. High-Concurrency and High-Performance Architecture Design

Through distributed computing architecture optimization, the system supports high-concurrency scenarios where over 100 documents should be handled per second. Both recognition accuracy and response time can meet real-world business requirements, making the system well-suited for peak-period operations of large logistics enterprises and platforms.

(II) Scenario Uniqueness

1. Replicable Practices for Promotion

Practical technology selection: In order to solve core industry pain points, it selects AI technologies such as multimodal recognition and natural language processing that are tailored to business needs, rather than pursuing cutting-edge technology for its own sake. This ensures practical deployment and makes the solution replicable across document-intensive industries such as logistics and trade.

Lightweight architecture design: By adopting the cloud-native architecture to ensure elastic scaling, core functions can be broken down into modular components that can be flexibly combined according to the needs of different industries and enterprises. This reduces adaptation costs during promotion and makes the solution suitable for rapid adoption by small and medium-sized enterprises.

Pilot-based iterative optimization model: The project starts with small-scale pilots to collect user feedback, then optimizes system functions and adaptability accordingly before gradually expanding deployment. This approach minimizes the risks of large-scale implementation and can be applied to a wide range of AI innovation projects.

2. Reusable Tools and Solutions

Core tools: The multimodal document recognition system (built on Claude 3.5, supporting multilingual processing and logical verification) and the AI agent interaction system (driven by natural language, supporting automated API generation and cross-system collaboration) can be directly applied to scenarios requiring extensive document processing.

General-purpose solution: For document-intensive and process-driven business scenarios, the project offers a replicable architecture combining AI document recognition for data collection and AI agents for process automation. By integrating industry knowledge bases to refine prompts and business rules, the solution enables rapid deployment of intelligent systems, helping enterprises achieve digital transformation at low cost.

IV. Implementation Results

The Shanghai WallTech AI-Integrated Application Platform project has been successfully deployed in collaboration with well-known logistics enterprises including China Post, Shanghai Yuesheng, Shanghai Bashi Yuexin, Shanghai Chunhong, Shanghai Jiebei, and UBI. Focusing on the core pain points of the international freight forwarding industry, the project has achieved significant breakthroughs across four key dimensions: efficiency improvement, cost optimization, market and customer benefits, and industry and social benefits. It not only addresses traditional operational challenges through technological innovation, but also provides strong support for the digital transformation of the industry and the pursuit of green and sustainable development, delivering dual value for both business and society. The specific outcomes are as follows:

(I) Efficiency Improvement: AI technology has been used to optimize document processing, business workflows, and industry collaboration, delivering comprehensive efficiency gains across all operations.

1. Document recognition efficiency: Accuracy reaches 96% or higher, surpassing traditional OCR accuracy of 85%–90%, while reducing labor costs. The system can process over 100 documents per second, meeting high-concurrency demands.

2. Business processing efficiency: The efficiency of processes such as booking has increased by over 80%, with the time required reduced from 4–5 hours to just 15 minutes. AI agent response times have improved from hours to seconds, with around-the-clock service availability.

3. Industry collaboration efficiency: The project promotes standardization of shipping data, strengthens Shanghai's shipping hub infrastructure, improves efficiency of multimodal transport transshipment, and shortens cross-regional logistics cycles.

(II) Cost Optimization: Cost optimization has been achieved across both operational and environmental dimensions, balancing enterprise benefits with green development.

1. Operational costs: By reducing the workload of manual review and data entry, the enterprise can reduce costs by over 30%. The cloud-native architecture supports elastic scaling, reducing the investment in IT infrastructure and hardware maintenance costs. The total project investment amounts to approximately 5.3 million yuan, including 491,600 yuan for cloud infrastructure rental and 4.8 million yuan for labor cost, with overall investment kept within manageable levels.

2. Environmental costs: The project promotes paperless operations, reducing the consumption of paper documents and lowering energy consumption and carbon emissions associated with paper production, transportation, and processing, thereby helping the enterprises achieve green and low-carbon development for enterprises.

(III) Market and Customer Benefits: The project leverages existing platform resources to expand customer coverage while enhancing customer loyalty and brand competitiveness through high-quality services.

1. The customer coverage is based on the CargoWare platform, which connects with 2,000 international freight forwarding companies and serves over 80,000 users, The eTower platform serves over 10,000 cross-border e-commerce sellers, with a daily order processing capacity exceeding 2 million, and an annual package volume of over 606 million pieces, demonstrating broad market reach..

2. Customer satisfaction: The AI intelligent body provides precise and efficient services, enhancing customer satisfaction, reducing churn rate, and strengthening customer loyalty. It also boosts brand competitiveness by providing customized solutions to meet customers’ special needs.

(IV) Industry and Social Benefits

The project has received strong industry recognition while contributing to the shipping development, green transformation, and public safety in Shanghai, creating multi-facetesd social benefits:

Industry recognition: Shanghai WallTech has been honored with awards including the 2024 Shanghai Innovative Potential Project for Promoting Modern Shipping Services and the 2023 Top 10 Shipping Pioneer Enterprises of the Year. Its CEO, Guo Shunri, was named to the 2023 China Shipping Celebrities List, reflecting broad recognition of the company’s technology and achievements.

Social benefits: The project supports Shanghai’s development as an international shipping hub and enhances the global standing of the Port of Shanghai. It promotes green and low-carbon development, driving sustainable transformation across the industry. It also strengthens dangerous goods identification and compliance verification capabilities, enhancing port clearance security. In addition, the project has indirectly stimulated job creation and tax revenue growth, supporting broader socioeconomic development.

Case 38

Dynamic Collaborative Control of HVAC Equipment Clusters in Electronics Manufacturing Plants Using AI Algorithms and Its Application Practice

Shanghai Electronic Engineering Design & Research Institute Co., Ltd.

I. Basic Enterprise Information

Founded in 1964, the company's core businesses cover engineering consulting, design, technological R&D, and EPC general contracting, delivering full-lifecycle integrated engineering services. It has obtained numerous honors including High-tech Enterprise of Shanghai and Specialized, Refined, Unique & Innovative SME of Shanghai, and owns an enterprise technical center as well as a master designer studio.

Over the past decade, rooted in the electronics manufacturing sector, the company has undertaken design work for a full spectrum of modern industrial manufacturing projects, including integrated circuits, lithium/sodium batteries, photovoltaic cells, high-end electronic components, and smart factories. Meanwhile, with forward-looking strategic layout and innovative capabilities, it has yielded remarkable R&D achievements in artificial intelligence and digital technologies, boasting distinctive technical strengths and featured business lines. It has evolved into a renowned domestic electronic engineering design enterprise and technical consulting institution.

As of the end of 2024, the company employed 622 staff, including top-tier industry specialists, forming a complete echelon of technical professionals. As its core functional department, the Engineering Technology Research Institute assembles technical experts and young engineers with profound professional expertise. Centered on its Shanghai headquarters and supported by six domestic branches and multiple overseas offices, the company has greatly expanded its business channels dominated by electronics manufacturing engineering design. Each design division synchronously advances design work and technical R&D to boost overall competitiveness. The company has also established industry-university-research collaborative partnerships with universities and research institutes to build joint innovation capacity.

The company holds 52 independent intellectual property rights, has won 41 national, provincial and ministerial-level awards, and has participated in compiling 54 national and industrial technical standards.

II. Case Introduction

(I) Background and Needs

1. Project Background

Production workshops of electronics manufacturing plants impose extremely stringent requirements on ambient temperature and humidity to guarantee stable and accurate environmental parameters. Their HVAC systems are required to operate continuously 24/7 to ensure uninterrupted production. Nevertheless, HVAC systems suffer severe high energy consumption, accounting for over 35% of the plant's total annual power consumption, which drives up operating costs and significantly impairs corporate economic benefits.

Against this backdrop, the company organized R&D teams to carry out targeted technical breakthroughs. Following thorough investigation and technical demonstration, the team adopted an energy-saving IoT control box technology centered on AI algorithms. Taking the legacy HVAC system of an electronics manufacturing enterprise as the testbed, comprehensive and in-depth intelligent retrofitting was conducted integrating theoretical analysis and on-site practice. The project successfully realized energy consumption optimization and efficiency improvement, facilitating the enterprise's green and low-carbon transformation and upgrading.

2. Scenario-Specific Requirements

Production scenarios of electronics manufacturing factories generally feature three core characteristics that impose stringent control requirements on HVAC systems:

(1) Variable production shifts: Day-shift and night-shift output differ by up to 40%, triggering frequent fluctuations in workshop cooling load. Conventional fixed control strategies fail to adapt to load variations in real time.

(2) Diverse environmental disturbances: Outdoor temperature and humidity fluctuate drastically across seasons (maximum summer temperature: 38°C; minimum winter temperature: -5°C), with a daily diurnal temperature swing exceeding 10°C, which directly deteriorates the operational efficiency of HVAC equipment.

(3) Large-scale equipment with strong coupling effects: The HVAC system comprises 13 chillers, 30 cooling towers and 30 water pumps. Strong coupling exists among all equipment units (e.g., chiller efficiency is subject to cooling tower water temperature; variations in pump flow rate will collaboratively affect cold energy transmission). The traditional "independent single-device control" paradigm easily results in the dilemma of "local optimum ≠ global optimum", leading to excessive overall energy consumption.

(II) Core Pain Points Addressed via AI Technology

Targeting the above scenario demands, this project leverages AI technology to resolve three pivotal technical challenges:

1. Dynamic strategy generation: Conventional fixed control strategies cannot flexibly cope with shift adjustments and environmental fluctuations. In contrast, AI algorithms deliver robust real-time analytical capabilities to mine nearly one year of historical load data of the enterprise, covering detailed records of production schedules, environmental parameters and equipment operating status. Combined with real-time collected workshop temperature, humidity and cooling load data, AI algorithms automatically generate optimized equipment start-stop schemes and parameter adjustment combinations via comprehensive analysis. This not only maintains workshop ambient conditions compliant with production specifications, but also minimizes energy consumption to meet energy conservation and emission reduction targets.

2. Multi-equipment collaborative optimization: Complex industrial production environments feature tight coupling between interconnected equipment, making single-device control unable to achieve system-wide optimal performance. To address this limitation, AI algorithms innovatively construct a multi-agent reinforcement learning model that breaks the constraints of traditional single-unit control. Powered by this model, AI realizes cluster collaborative management of core equipment including chillers, cooling towers and water pumps, effectively eliminating energy waste caused by overoperation of partial equipment. Such collaborative optimization elevates the operational efficiency of individual equipment and substantially improves the comprehensive performance of the entire system, providing solid technical support for corporate energy conservation and emission reduction.

3. Operating condition self-adaptation：Equipment performance degrades over service life — for instance, chillers with 5 years of operation may suffer a 15% efficiency drop. Coupled with seasonal environmental variations, the energy consumption characteristics of equipment keep evolving. AI algorithms exhibit outstanding self-learning capacity: they automatically update model parameters based on daily operating data and adjust control strategies in real time to accommodate dynamic operating conditions. This intervention-free self-adaptive mechanism guarantees long-term stable energy-saving performance, drastically reduces the workload of operation & maintenance (O&M) personnel, and enhances corporate operational efficiency and management standardization.

(III) Implementation Plan

Driven by industrial energy conservation initiatives and the Dual Carbon Policy, the energy-saving solution featuring AI algorithm dynamic collaborative control for HVAC equipment clusters in electronics manufacturing plants reconstructs the energy efficiency management logic of HVAC systems through the integration of AI algorithms and IoT technologies.

System Overview: Integration of Equipment Collaboration and Data-Driven Operation

The system covers full-link core HVAC equipment including multiple chillers, cooling towers and water pumps, and realizes collaborative scheduling of equipment clusters via artificial intelligence. Supported by a traversal optimization simulation database for optimal energy-saving operation that provides massive training samples for AI algorithms, the system achieves a 5%–15% improvement in overall system energy efficiency, forming a closed-loop energy-saving architecture spanning the equipment layer to the system layer.

System Architecture: Dual-Layer Collaboration of Edge Computing and Cloud Optimization

The project adopts an "edge computing + cloud optimization" dual architecture: An AI algorithm execution box is deployed on the edge side. Based on the Modbus TCP communication protocol, it collects high-frequency input parameters such as low-temperature cooling load, medium-temperature cooling load and outdoor wet-bulb temperature, and outputs optimized control instructions in real time (including the number of online chillers and operating frequency of cooling towers) to realize millisecond-level strategy issuance. On the cloud side, massive data training and model iteration continuously optimize edge-side algorithms to adapt to equipment aging and dynamic operating conditions, sustaining long-term stable energy-saving effects.

Business Logic: Workflow from Data Input to Global Optimal Operation

Built on multi-dimensional operating condition data, AI algorithms break the bottlenecks of traditional control frameworks: The input layer collects dynamic data including production load and environmental parameters. After analysis via multi-agent reinforcement learning and self-learning models within the algorithm layer, the global optimal operation strategy for equipment clusters is generated to resolve the collaborative optimization dilemma of "local optimum ≠ global optimum". For example, targeting the coupling relationship between chillers and cooling towers, the algorithm dynamically adjusts the number of running chillers and cooling tower operating frequency to minimize overall energy consumption while satisfying production temperature control requirements, ultimately realizing intelligent and energy-efficient operation of the entire equipment cluster.

Figure 1 System Panoramic Architecture and Business Logic of the Energy-Saving Box

The full implementation cycle of this pilot project lasts 12 weeks, divided into four phases centered on AI model development and system deployment:

Pre-implementation Preparation (1 week): Collect historical data, including energy consumption, environment, and production schedules, to build training datasets; define core AI algorithm indicators in accordance with enterprise standards (response time ≤ 1 second, energy-saving rate ≥ 5%, instruction delivery success rate ≥ 99.99%).

AI Model Development & Deployment (4 weeks): Adopt the dual edge-cloud architecture. Deploy lightweight models on the edge for real-time data processing and strategy issuance, and deep reinforcement learning models on the cloud for parameter optimization. Train load prediction and equipment collaborative models, complete 3-day single-unit verification after passing simulation tests, and embed the models into high-performance control boxes.

Installation & Commissioning (3 weeks): Upload real-time data to fine-tune models and limit load prediction deviation within 8%; gradually expand control coverage to the full system and optimize equipment linkage response time to 0.8 seconds.

Trial Operation & Iterative Optimization (4 weeks): Activate self-learning mode to update model weights daily, and verify performance weekly via the Similar Day Comparison Method. The energy-saving rate stabilized above 5% in the 4th week, fulfilling all expected targets.

In addition, Shanghai Aurora Digital Technology Co., Ltd. participated in the project's technical collaboration, undertaking the R&D, training and deployment of AI algorithms. The algorithms are developed based on Spring Boot, Python and Spring AI frameworks, supporting dynamic model updates to guarantee the advancement and stability of the technical solution.

III. Project Highlights

(I) Core Technological Breakthroughs

1. Innovative Edge-Cloud Collaborative AI Architecture: Breaking the limitations of traditional standalone control architectures, lightweight edge-side models deliver millisecond-level strategy issuance with a measured delay of 0.6 seconds, meeting the real-time control demands of industrial scenarios. Cloud-side deep reinforcement learning models iteratively optimize edge-side parameters on a daily basis to accommodate long-term operating fluctuations such as equipment aging and seasonal shifts, ensuring the energy-saving rate remains stably above 5% — fully complying with enterprise specifications mandating algorithm response time ≤ 1 second and energy-saving performance ≥ 5%.

2. Precise Parameter System Construction: Tailored to the operating characteristics of HVAC equipment clusters, a refined set of input and target parameters is established: Input parameters cover low/medium-temperature cooling load and outdoor wet-bulb temperature (measurement error ±0.1°C); target parameters include total power consumption, chiller power, cooling tower power and cooling pump power. The Modbus TCP protocol enables high-frequency data acquisition at 500 times per second with 99.99% data accuracy, providing high-quality data foundations for AI algorithm optimization, and aligning with the parameter configuration requirements of the low/medium-temperature strategy control module specified in software manuals.

3. Optimized Performance Verification Methodology: Based on the Similar Day Comparison Method, an innovative Similar Operating Condition Comparison Method is developed. By locking core variables such as cooling load and outdoor temperature & humidity, the number of comparison samples is reduced (from 3 traditional similar days to 1), shortening the energy-saving effect verification cycle from 7 days to 3 days with energy-saving rate calculation deviation ≤ 1%. This method complies with the national standard *GB/T 31349-2014 Technical Requirements for Measurement and Verification of Energy Savings—Central Air Conditioning Systems*, significantly accelerating project acceptance efficiency and cutting waiting costs for enterprises.

(II) Scenario Uniqueness

1. Targeting the Hidden Energy Consumption Pain Point of Manufacturing Industries: Unlike mainstream AI applications focused on intuitive and explicit industrial scenarios such as production workflow and quality inspection, this project innovatively targets HVAC systems in manufacturing plants — a high-energy-consumption field that is frequently overlooked as "hidden energy loss". Deep integration of AI technology unlocks massive untapped energy-saving potential in this domain, effectively filling the gap of AI applications for energy conservation in auxiliary industrial production systems, and creating a new pathway for manufacturing enterprises to reduce operating costs and boost comprehensive economic benefits.

2. Adaptability to Complex HVAC Cluster Control Scenarios: A multi-agent collaborative model is custom-built for HVAC clusters consisting of 13 chillers, 30 cooling towers and 30 water pumps. The model effectively addresses complex inter-equipment coupling issues and achieves system-wide optimal control via intelligent collaborative optimization. This innovative solution boasts strong cross-industry adaptability: beyond electronics manufacturing, it can be replicated directly for machinery processing, chemical production and other industries with HVAC equipment to improve equipment operational efficiency and overall system performance.

3. Low-Threshold Deployment Design: The AI model features a robust self-learning module, eliminating reliance on professional AI engineers for complex on-site operations. Specifically, plant operators only need to upload new operating condition data (e.g., load data generated after new production line expansion) via the control box software interface, and the AI model will automatically complete self-adaptive learning. The entire adjustment process is efficient and convenient, with adaptive parameter tuning finished within 7 days to rapidly accommodate new production environments and demands.

4. High Compatibility of Hardware Equipment: The energy-saving IoT control box adopted in the system delivers outstanding compatibility and flexibility. It supports a wide voltage range (9–19 V) to maintain stable operation under diverse voltage environments. Meanwhile, it supports multiple communication protocols including Modbus and MQTT,

enabling seamless docking with enterprises' legacy HVAC control systems without large-scale hardware retrofitting. This design lowers the technical threshold for project deployment and reduces overall implementation costs.

IV.Implementation Results

(I) Economic Benefits

Taking Yangzhou Canadian Solar Solar Cell Co., Ltd. as a demonstration case, the project generates remarkable economic gains in two dimensions after deployment:

1. Reduced energy consumption costs: Optimized control via AI algorithms stabilizes the overall energy-saving rate of the HVAC equipment cluster above 5% (verified via the Similar Day Comparison Method specified in GB/T 31349-2014). Based on the enterprise's annual HVAC power consumption of 21 million kWh and industrial electricity price of 0.5 RMB/kWh, the annual electricity cost saving reaches 1.05 million RMB.

2. Cut labor costs: Conventional HVAC systems require 2 O&M staff working shifts to monitor equipment status and adjust parameters. The AI system realizes fully automatic control and fault early warning, enabling the reduction of 1 dedicated O&M personnel. Calculated based on the local manufacturing average monthly salary of 8,000 RMB, annual labor cost savings approximate 100,000 RMB.

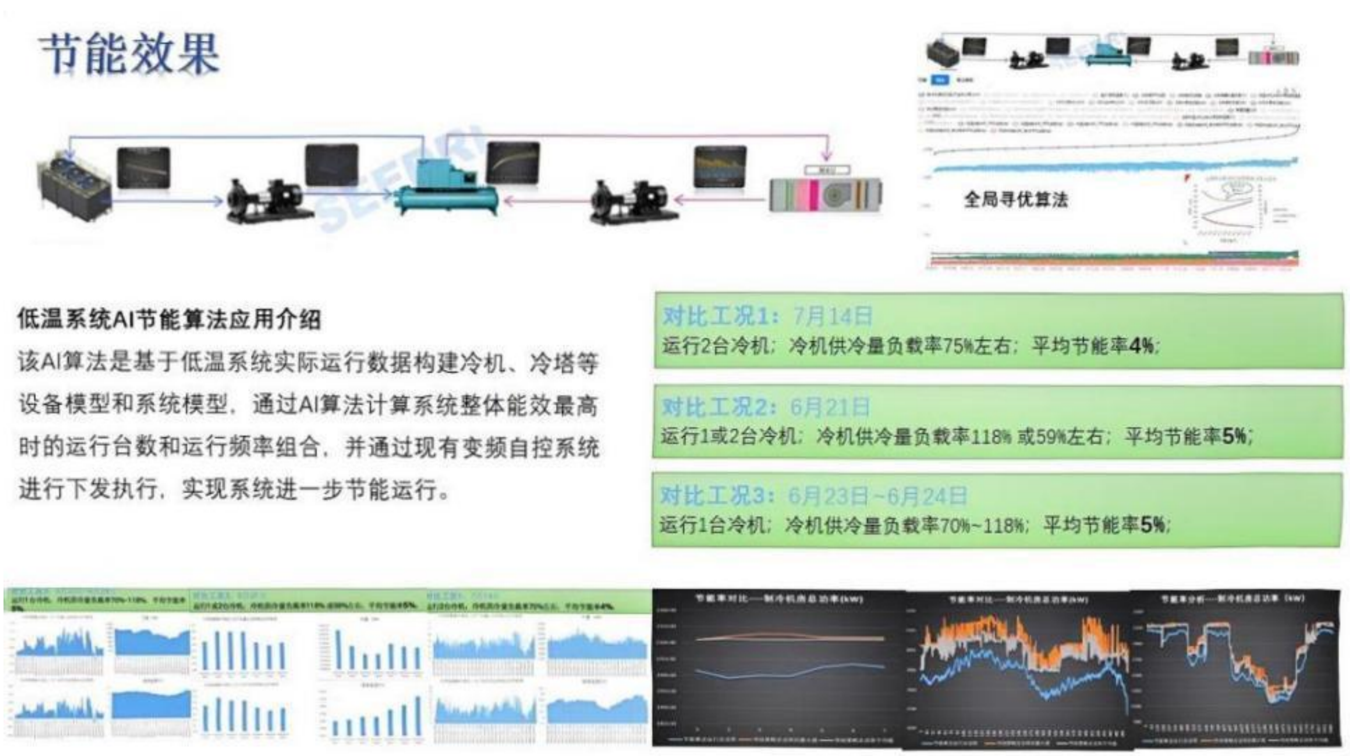


Figure 2 Energy-saving control and its effects

The total investment of this pilot project amounts to 500,000 RMB (including 400,000 RMB for AI algorithm R&D, 50,000 RMB for hardware adaptation, and 50,000 RMB for on-site investigation). The annual total cost savings for enterprises reach 1.15 million RMB, demonstrating prominent economic returns.

(II) Core Technical Innovations

1. Innovative Edge-Cloud Collaborative Control Architecture: The edge side delivers millisecond-level response with a measured strategy issuance delay of 0.6 seconds, satisfying the real-time performance requirements of industrial scenarios. The cloud side optimizes model parameters monthly, maintaining an energy-saving rate ≥ 5% even under equipment aging conditions (e.g., 15% efficiency degradation of chillers). This breaks the rigid limitations of traditional control systems and realizes long-term stable energy-saving performance.

2. Self-Adaptive Scenario Algorithm: An AI model supporting customized operating condition adaptation is developed. Enterprises upload new operating condition data (e.g., new production lines, adjusted production shifts) via the software's model training module, and the model completes parameter updates within 7 days without dedicated professional teams, lowering the technical application threshold.

3. High-Reliability Safety Control: The AI algorithm embeds a safety constraint module that restricts equipment operating parameters within safe thresholds (e.g., chiller operating frequency ≥ 30 Hz), preventing equipment overload or faults induced by algorithm optimization, with a 100% strategy execution success rate. Meanwhile, the AI model supports a maximum data processing throughput ≥ 50,000 records per second and operational stability ≥ 99.99%, guaranteeing reliable system operation.

(III) Social Benefits

1. Facilitating the Delivery of Dual Carbon Targets: Based on the annual power saving of 1.05 million kWh from this project, it corresponds to a reduction of 350 tons of standard coal consumption (calculated at 0.33 kg standard coal per kWh) and 875 tons of carbon dioxide emissions, providing a deployable AI solution for the green and low-carbon transformation of manufacturing industries.

2. Replicable Industry Reference Framework: This pilot project forms a closed-loop solution covering "data acquisition (500 times/second) – AI load prediction (load accuracy ≥ 90%) – collaborative control (equipment linkage response ≤ 1 second) – performance verification (similar operating condition method)". The solution can be directly replicated for HVAC retrofitting of factories in machinery, chemical and other industries beyond electronics manufacturing, possessing high promotion value.

3. Driving Industrial Transformation from Experience-Based Control to Data-Driven Control: Conventional HVAC cluster control relies on O&M personnel's subjective experience, leading to delayed and low-precision parameter adjustment. This project realizes data-driven decision-making via AI algorithms, lifting control precision from ±2°C to ±0.5°C. This not only optimizes energy consumption but also improves workshop ambient stability to indirectly guarantee product quality (e.g., every 1°C rise in temperature & humidity deviation in photovoltaic cell production workshops reduces product yield by 0.3%), supporting the upgrading of manufacturing industries from experience-driven to technology-driven operation.

(IV) Intellectual Property Achievements

During the implementation of this pilot project, the company collaborated with Shanghai Aurora Digital Technology Co., Ltd. to develop core technologies including AI energy-saving control algorithms and edge-cloud collaborative architectures. Two software copyrights have been obtained: *Smart Chip Energy-Saving Optimization Control Box Software V1.0 and Factory Multi-Dimensional Energy Statistics Analysis & Energy-Saving Optimization System V1.0*, consolidating the company's technical advantages in the field of industrial AI energy conservation.

Case 39

A New "Data-driven" Inclusive Finance Model Based on Federated Learning and an Intelligent Decision Engine.

Shanghai Weicheng Enterprise Credit Investigation Co., Ltd.

I. Basic Enterprise Information

Shanghai Weicheng Enterprise Credit Investigation Co., Ltd., was established in 2015. It originated from a research and development team composed of Ph.D. graduates from prestigious universities including Shanghai University of Finance and Economics, East China Normal University, The Hong Kong University of Science and Technology, and Johns Hopkins University. The company is a big data credit investigation agency specializing in financial product innovation, financial risk control, industrial resource exploration, and credit data services. It has established the Yangtze River Delta Data Model Research Laboratory and an Expert Workstation, and possesses massive data resources. Through data collection, processing, storage, analysis, and visualization, it provides data products and services to banks, government agencies, and industrial parks.

The qualifications obtained by the company include: Enterprise Credit Agencies Filing with the PBOC (Filing No. 01028), Level 3 Classified Protection of Cybersecurity Level 3 (the highest level) Certification from the Ministry of Public Security, National High-Tech Enterprise designation, ISO/IEC 27001 Information Security Management System certification, ISO/IEC 20000 Information Technology Service Management System certification, ISO 9001 Quality Management System certification, as well as Shanghai-level designations including Double-Soft Enterprise, Specialized annd Sophisticated SMEs, "Little Giant Cultivation" Enterprise, "High-Tech Enterprise", and Domestic Technology Vendor Adaptation Enterprise.

II. Case Introduction

(I) Background and Needs

In April 2024, the General Office of the State Council issued *the Implementation Plan for Coordinating the Development of Financing Credit Service Platforms and Improving Financing Convenience for Micro, Small, and Medium-Sized Enterprises*. The plan aims to enhance convenience of financing for MSMEs by strengthening channels for credit information aggregation and sharing, optimizing information aggregation and sharing mechanisms, and deepening the application of credit data. It emphasizes safeguarding information security and the legitimate rights and interests of business entities, and encourages financial institutions to transform their operational philosophies, optimize financial services, and prevent and control financial risks, so as to deliver high-quality financial services to MSMEs.

In early 2024, Secretary Chen Jining issued an important directive regarding SME lending: "We must pay attention to the overlooked minority and solve the financing and loan issues faced by enterprises." As a solution to this challenge, Overall Credit Facility for Industrial Parks are gaining increasing attention from SMEs.

In response to the State Council's Implementation Plan and relevant Shanghai municipal initiatives, efforts are being made to address the persistent challenges of difficult and expensive financing for SMEs. Currently, the Shanghai Enterprise Service Cloud has covered 1.8 million enterprises across the city, while Hongkou District alone has 38 industrial parks hosting over 50,000 enterprises. These enterprises face three common challenges: First, lengthy procedures: enterprise account opening takes 14 days, and loan approval requires 3–5 working days. Second, information silos: financial institutions struggle to verify real-time operational data, resulting in a rejection rate exceeding 60% for asset-light enterprises. Third, lagging services: the traditional model where enterprises took the initiative to seek financing has led to a first-time loan rate of less than 30%.

(II) Implementation Plan

1. System Architecture

Adopting a layered, decoupled architectural design, with data-driven functionality at its core, the platform enables a full-chain closed loop from data aggregation to financial service delivery. The overall business logic follows the pathway of data integration, intelligent decision-making, process automation, and service delivery.

(1) Data Layer (Privacy Computing Hub)：It serves as the digital foundation. By employing federated learning, it tries to build a secure, cross-departmental, cross-regional data fusion space, enabling usable but invisible access to multi-source heterogeneous data related to government affairs, finance and industry, thereby providing compliant, high-quality data asset for upper-layer applications.

(2) Model Layer (Intelligent Decision Engine): It's core function is risk pricing. A model suite is deployed, comprising pre-credit dynamic scoring models, anti-fraud models, risk pricing models, and others. Built on machine learning algorithms, these models perform real-time analysis on over 400 dimensional tags provided by the data layer, delivering precise credit assessments and risk pricing to turn complex financial rules into executable algorithms.

(3) Application Layer (Intelligent Interaction Platform): It provides a unified portal for enterprises, parks, and banks. It features application functions including one-click pre-qualification check, intelligent product matching, and RPA (Robotic Process Automation) workflow automation, transforming the decisions made by model-layer into concrete financial services, thereby achieving streamlined user experience and standardized processes.

2. Business Logic

We stick to progress step by step from basic to advanced , prioritize user experience via technological empowerment to build a complete digital financing solution from user engagement touchpoint to closed-loop data system. The implementation plan not only resolves specific technical docking challenges but also achieves paradigmatic innovation in process reconstruction, data fusion, and ecosystem collaboration, systematically advancing through four phases:

Phase One: Digitalization of Frontend Touchpoints—Build a Seamless QR Code Application Access

The core objective of this phase is to lower the application threshold for enterprises by fully digitizing the traditional offline application process. Using a product QR code as the key touchpoint to connect the platform and banking services, enterprise users can check their pre-approved credit limit and access the bank’s application page by scanning the QR code with one click, eliminating the cumbersome offline document submission process.

Phase Two: Automation of Core Interactions—Build a Unified Identity Authentication Systemwith Password-free Login

This phase is dedicated to breaking through the critical bottleneck of data flow, enabling secure and seamless identity transition for enterprise users between the platform and banking systems. Enabled password-free login capability eliminates repeated entry of the five core enterprise credentials, including enterprise name, Unified Social Credit Code, name, ID number, and mobile numbe, delivering a truly one-stop service experience for users.

Phase Three: Intelligent Risk Control Decision-Making—Enable Online Pre-Credit Approval and Precise Credit Limit Return

This phase upgrades pre-credit decision-making from an offline, manual, experience-driven model to an online, automated, data-driven intelligent model. The platform periodically pushes enterprise lists to partner banks via API interfaces. After the banks invoke their risk control models, the pre-approved credit limit data is returned to the platform in real time. Rather than attempting to unify or replace banks’ internal risk control models, the company have innovatively developed a model alignment middleware. This middleware serves as a translator, accurately mapping the platform’s standardized enterprise profiles (risk indicator dataset) to the risk control logic of different banks, preserving banks’ autonomy in risk control while ensuring high accuracy and comparability of pre-credit results on the platform.

Phase Four: Integrated Business Closed Loop—Enable Automated Loan Data Return

This phase aims to complete the full data closed loop of financing services, automating post-loan management. After credit approval and loan disbursement, banks return key loan data, such as approved credit limit, disbursement amount, loan status, to the Xinyidai platform via standardized interfaces in real time. This phase achieves a fully automatic data population, liberating account managers from trivial administrative workloads and providing the platform with authentic, real-time big data for financing effectiveness analysis and precise policy targeting.

III. Project Highlights

The AI Financing Assistant for SMEs, exemplified by Hongkou Instant Loan (an express financing service rolled out by Hongkou District exclusively for enterprises registered within the district) and the Shanghai Enterprise Service Cloud Financing Service Zone, focuses on addressing the core challenges faced by MSMEs, such as cumbersome financing procedures, information asymmetry, and low service efficiency. The application comprehensively covers the entire MSME loan application, risk control optimization, and the entire approval process, delivering inclusive financial services through AI technology. The platform features a one-stop operation model covering enterprise, park, and financial institution portals, enabling full-chain AI empowerment across data, matching, process, and risk control.

(I) Technological Innovation 1: Paradigm Innovation

This project is not a simple application of federated learning technology. It represents the first large-scale deployment of a federated learning paradigm with localized data and mobile models within a provincial-level inclusive finance platform in China. It fundamentally overturns the traditional data aggregation mindset, constructing a virtual data space spanning 37 government departments and covering over 400 dimensional tags. In this space, what is transmitted is not sensitive raw data, but encrypted model parameters and gradients, achieving a trusted closed loop of usable but invisible data. This provides a replicable, standardized solution to the fundamental tension between security and efficiency in the application of government data to financial services. In terms of technical implementation, we have built the core of the platform: the Privacy Computing Hub. Serving as the platform’s cornerstone of trust, this hub incorporates complementary technologies such as secure multi-party computation (PSI) and homomorphic encryption, working in synergy with federated learning. This not only safeguards data privacy during the training phase but also ensures the absolute security of business secrets, such as patents and core tax data, for example, those of a energy enterprise, throughout the entire process, including authorized queries and model inference. This establishes a technical and engineering benchmark for the large-scale use of government data in highly sensitive financial risk control scenarios.

Figure 1 Data Foundation Architecture

(II) Technological Innovation 2: Model Innovation

This project moves beyond the traditional flat credit assessment model based on individual enterprises, and introduces the pioneering concept of a park-level enterprise risk profile. By constructing the five-dimension data framework, covering enterprise chain, facility chain, industry chain, investment-financing chain, and talent chain. We have not merely superimposed these data sources. Instead, within the Yangtze River Delta Data Model Laboratory, we have leveraged AI technologies such as graph computing and network analysis to build a digital twin that dynamically reflects the collaborative, competitive, and symbiotic relationships among enterprises within an industrial park. This model achieves a multi-dimensional leap from individual assessment to ecosystem assessment. It systematically evaluates the robustness of industrial chains, the synergy of industrial clusters, and potential regional risk transmission pathways. This transforms park-level collective credit from a simple aggregation of credit limits into a strategic decision-making tool grounded in deep insights into the soundness of the entire micro-economic ecosystem. It provides an unprecedented, dynamic, and quantifiable decision-making foundation for innovative financial services such as supply chain finance and industrial cluster financing. This shifts financial services from merely supporting businesses

that are already thriving, to proactively identifying and empowering those with real potential.

Figure 2 Park-Level Enterprise Risk Profile – Enterprise Chain

(III) Technological Innovation 3: Architectural Innovation

The core architectural innovation of the project lies in our decision not to take the difficult path of overthrowing or replacing banks' existing risk control systems. Instead, we have uniquely designed and implemented an intelligent risk control middleware. This middleware functions as a powerful "risk control protocol converter". Its core algorithms are capable of understanding and learning the decision-making logic of different banks' risk models, and transforming the unified credit view formed by the platform based on massive government data into standard signals that can be directly understood and accepted by the internal systems of 11 major banks. This innovation neatly solves the last-mile challenge in inclusive finance. It leverages the immense advantage of government data to expand the breadth of risk identification to cover credit invisible enterprises, while also deepening risk assessment to uncover hidden value, all while fully respecting each bank's autonomy over their risk control systems. Through this middleware, we have upgraded the platform-bank connection from complex, customized system coupling to an efficient, standardized capability plug-in. This has reduced bank integration costs by over 70% and increased pre-credit approval accuracy to 90%, achieving a win-win for the ecosystem rather than a zero-sum competition.

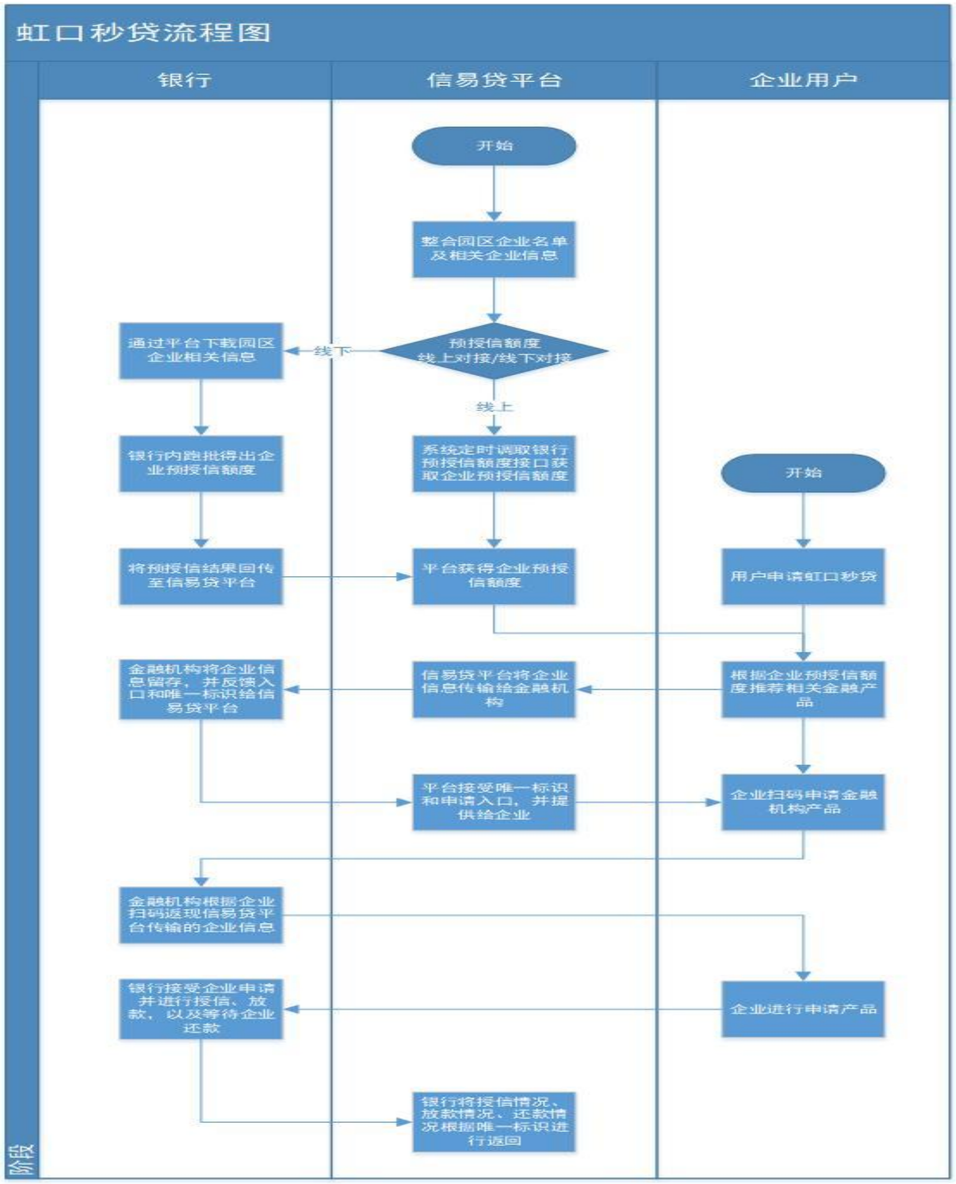


Figure 3 Business Process Flow

IV. Implementation Results

First, the coverage of inclusive finance has expanded, serving over 10,000 enterprises, with the first-time borrower rate rising to 65% , compared to the traditional rate of under 30%. Second, process efficiency has been optimized, with an 80% reduction in the account-opening cycle, and the entire loan process can be completed in as few as 8 hours. Third, the project has set an industry benchmark, having been selected as an outstanding case by the National Development and Reform Commission (NDRC) and replicated across 11 cities in the Yangtze River Delta region.

Taking the Hongkou Instant Loan project as an example, significant economic benefits have already been achieved. By the first half of 2025, the program served 4,600 enterprises across 77 park buildings, further bringing financing services closer to enterprises. 123 enterprises successfully obtained loans totaling 430 million yuan. For enterprise, account-opening costs have been reduced by 50%, and bank operating costs have decreased by 30% including a reduction in manual approval costs.

In terms of efficiency gains, the pre-credit approval rate has jumped from 30% to 90%. The account-opening cycle has been compressed from two weeks to just 2-3 days, and loan approvals are processed instantly. The platform processes applications in an average of under 1 minute, realizing an 80% efficiency improvement. In the future, policy interest subsidies, such as the Hongqi Loan program, are expected to further reduce enterprise financing costs by 15%, helping to increase enterprise income and reduce burdens.

In terms of social benefits, the case has been selected as an outstanding example in China (Shanghai) of innovative practices for optimizing the business environment, enhancing regional financial inclusivity. The project offers high replicability and scalability: the technical solution is modular (e.g., the pre-credit engine), the platform is highly compatible with rapid API integration, initial investment is shared between government and enterprises, and operational costs remain manageable. It is well-suited for expansion into scenarios such as supply chain finance and green finance, providing a replicable pathway for digital inclusive finance nationwide.

Case 40

Zhihe AI: Building a Legal Technology AI Pioneer Through Triple Empowerment of "Data - Algorithms - Scenarios"

Shanghai Zhihe Network Technology Co., Ltd.

I. Company Profile

Founded in 2014, Shanghai Zhihe Network Technology Co., Ltd. (hereinafter referred to as Zhihe) is one of China's leading providers of legal technology, legal education, and legal information services. Centered on artificial intelligence technologies, the company has established a technology-driven legal ecosystem through a portfolio of brands, including Zhihe AI, Zhihe Standards Center, Zhishi, and Bangshang.

Driven by the dual engines of intelligent technology and professional excellence, Zhihe delivers efficient and reliable legal technology products and services to law firms, legal and compliance departments of large and medium-sized enterprises, SMEs, and legal professionals. Its core offerings cover AI-assisted legal research, contract review, workflow optimization, compliance and risk management, as well as value-added services including think tank research, brand enhancement, continuing legal education, and AI-powered legal marketing.

In addition, the Zhihe Standards Center specializes in standards development and consulting services in areas such as corporate compliance, artificial intelligence, data elements, overseas business expansion, ESG, and sustainable development, helping enterprises address increasingly complex compliance challenges.

II. Case Introduction: Deep Integration of Law and Technology

Against the dual backdrop of digital transformation and judicial reform, the legal services industry is undergoing unprecedented change. Traditional legal service models have long faced challenges such as low operational efficiency, high service costs, and uneven distribution of professional resources.

The year 2023 marked the beginning of the generative artificial intelligence era. AI applications experienced explosive growth worldwide, rapidly permeating every aspect of daily life. At the same time, breakthroughs in foundation models—such as GPT-4 and ChatGLM—opened new pathways for the intelligent transformation of legal services.

In mid-2025, Shanghai Zhihe Network Technology Co., Ltd. (hereinafter referred to as Zhihe) launched the Zhihe AI Platform, an intelligent legal application platform powered by foundation model technologies. The platform provides AI-driven legal research, AI contract review, AI legal document drafting, and other professional services, enabling enterprises to obtain high-quality legal support more efficiently and at lower cost while promoting the intelligent upgrading and broader accessibility of legal services.

(I) Background

From a policy perspective, the National Public Legal Service System Development Plan (2021–2025) issued by China's Ministry of Justice has progressively advanced the goal of increasing the number of practicing lawyers to 750,000 by 2025, further intensifying competition within the legal services market. Large law firms continue to strengthen their competitiveness through economies of scale and specialized practice divisions, while small and medium-sized firms seek sustainable development by focusing on niche markets.

Meanwhile, Shanghai has introduced a series of supportive policies, including the Several Measures of Shanghai Municipality for Further Expanding AI Applications, to accelerate the development of new quality productive forces through artificial intelligence. These initiatives encourage the application of domain-specific foundation models and provide strong policy support for intelligent innovation in the legal technology sector.

From an application perspective, the legal services industry's demand for digital transformation stems from three major challenges.

First, China's legal services market has maintained rapid expansion. By the end of 2024, the market size had reached approximately RMB 300 billion, with a compound annual growth rate of 14.3% between 2019 and 2024—significantly outpacing GDP growth during the same period. This growth has been fueled by the continuous introduction of new regulations, technology-enabled productivity improvements, and rapidly increasing corporate compliance requirements.

Second, the industry faces a structural imbalance in resource allocation. China's leading "Red Circle" law firms account for approximately 17.1% of total industry revenue, while many small and medium-sized firms continue to rely heavily on low-margin consumer legal services, such as private lending disputes and divorce litigation. Supported by stronger financial resources, leading firms have already digitalized nearly 80% of their business processes, whereas digital adoption among smaller firms remains significantly lower. This has created a vicious cycle in which insufficient digitalization leads to lower productivity, weaker business performance, and consequently even less capacity to invest in digital transformation.

Third, traditional client acquisition in the legal profession still depends largely on passive channels, including personal and professional networks (such as family referrals, alumni associations, chambers of commerce, and industry organizations), paid traffic acquisition through search engines and online platforms, and referrals from intermediary legal service platforms. These conventional approaches often limit business expansion while increasing customer acquisition costs.

(II) Solution Overview

The Zhihe AI Platform is designed to address the industry's long-standing divide between digitally advanced leading law firms and digitally underdeveloped small and medium-sized legal service providers. According to the Shenzhen Lawyers Industry Development Report (2024), only 30.54% of law firms in Shenzhen have implemented basic office automation (OA) systems, while more than 80% rely primarily on standardized SaaS products. Most remain at the preliminary stage of process digitalization, with limited capability to unlock the value of legal data or deploy intelligent applications.

To overcome these limitations, Zhihe abandons the conventional linear evolution from informatization to digitalization and ultimately intelligent transformation. Instead, it adopts a "Data + Algorithms + Scenarios" architecture, integrating legal expertise with AI capabilities into a unified intelligent service platform. This architecture establishes a direct value chain from data resources to legal practice, providing advanced AI solutions for leading institutions while significantly lowering the adoption threshold for smaller firms. In doing so, the platform enables users to bypass intermediate stages and directly realize the value of intelligent legal services.

1. Core Architecture: A Three-layer Framework Supporting Intelligent Legal Services

The Zhihe AI Platform is built upon a three-layer architecture comprising the Data Layer, Technology Layer, and Application Layer, which together form a comprehensive foundation for intelligent legal services.

Data Layer: The platform is powered by a high-quality legal knowledge base containing more than 6 million laws and regulations, 150 million judicial cases, 300 million records of enterprises and organizations, and 500 million practical legal insights collected from across the Internet. Through version management, structured annotation, and compliant data acquisition mechanisms, these resources form a comprehensive legal data infrastructure covering virtually all legal scenarios and providing reliable knowledge support for AI-powered legal applications.

Technology Layer: The technology layer integrates a series of advanced AI technologies, including Retrieval-Augmented Generation (RAG), multi-agent workflows, and domain-specific legal embedding models. It combines hybrid retrieval techniques—including BM25 lexical retrieval, vector-based semantic retrieval, and cross-encoder reranking—with intelligent information extraction and consistency verification algorithms, enabling highly efficient transformation of legal data into intelligent legal services.

Application Layer: Focusing on core legal practice scenarios, the platform integrates four flagship AI applications: AI Contract Review, AI Legal Research, AI Legal Document Drafting, AI Document Question Answering. These capabilities are complemented by online professional training and community-based knowledge-sharing services, creating an integrated service ecosystem that combines tools, knowledge, and professional resources within a single platform.

Figure 1 Product Technical Architecture

2. Business Logic: Empowering the Entire Legal Workflow and Delivering Practical Value

Guided by the mission of making legal services more accessible through AI, the Zhihe AI Platform establishes an end-to-end business workflow extending from data input to value creation. Every stage is designed around real-world legal practice, addressing not only surface-level issues such as low efficiency and high costs, but also deeper structural challenges arising from the uneven distribution of legal resources.

(1) Data Processing: Secure Integration of Public and Private Data for Instant Access

The data processing stage serves as the starting point of the platform's workflow. Through standardized technical procedures, Zhihe securely integrates publicly available legal resources with users' proprietary documents, enabling both categories of data to be rapidly and securely utilized throughout subsequent AI applications.

Public legal resources—including legislation and judicial precedents—are first subjected to comprehensive version management covering enactment, amendment, and repeal histories. After data cleansing, they are stored in structured databases while vector indexes are generated to support millisecond-level semantic retrieval.

Meanwhile, users' private documents undergo encrypted processing, optical character recognition (OCR), document chunking, and vector embedding before being stored in tenant-specific vector databases. This architecture enables AI applications such as contract review and legal research to instantly retrieve both authoritative legal sources and enterprise-specific knowledge. For example, when reviewing a contract, the platform can simultaneously access relevant statutes, comparable judicial precedents, and the user's internal contract templates within seconds.

(2) Intelligent Interaction: Delivering Professionally Verifiable Results

The intelligent interaction module constitutes the core of the platform's operational workflow. It is designed around the actual working habits of legal professionals, enabling users to obtain professional, verifiable results through a streamlined interaction process.

Enterprise users may access the platform through APIs/SDKs, which support seamless integration with existing business systems, while individual users can operate through an intuitive web-based console. After submitting a request—for example, uploading a contract and selecting Contract Review—the platform automatically initiates a multi-agent collaborative workflow.

The Query Understanding Agent first interprets the user's request and analyzes both the user's role (such as lawyer, in-house counsel, or government official) and the applicable business scenario (including contract review, legal research, or document drafting). It then decomposes the request into a series of executable subtasks. For instance, a contract review request is divided into tasks such as clause legality assessment, risk identification, and revision recommendation generation.

Next, the Hybrid Retrieval Agent retrieves relevant information from both the public legal knowledge base and the user's private knowledge repository. By combining BM25 lexical retrieval, vector-based semantic retrieval, and cross-encoder reranking, the system identifies the most relevant legal authorities. When reviewing a construction contract, for example, the platform simultaneously retrieves applicable provisions of the Civil Code of the People's Republic of China, comparable judicial precedents, and the user's internal contract templates.

The retrieved materials are then processed by the Evidence Aggregation Agent, which validates their reliability by confirming, for example, whether cited judgments are legally effective and whether referenced statutes remain in force. The validated materials are organized into a structured Legal Authority List.

Finally, the Answer Generation Agent produces responses that conform to established legal practice standards. Every cited authority is linked directly to its original source, allowing users to review the underlying legal basis with a single click and ensuring both professional reliability and full traceability of the generated results.

III. Core Technical Highlights: A Dedicated Legal AI Engine for Accurate and Controllable Intelligent Outputs

The Zhihe AI Platform is built upon a dedicated legal AI engine centered on RAG and multi-agent workflows, enabling highly accurate, controllable, and fully traceable outputs tailored to legal practice.

Within the legal research workflow, the Query Understanding Agent first decomposes user queries into multiple specialized tasks, including statutory retrieval, precedent retrieval, and fact extraction. The retrieval stage employs a three-level hybrid retrieval mechanism comprising BM25 lexical matching, legal-domain semantic vector retrieval, and cross-encoder reranking to identify the most relevant legal evidence with high precision.

During the evidence aggregation and reflection stage, candidate statutes and judicial precedents undergo alignment and conflict detection to eliminate contradictory or unreliable authorities before answer generation.

Based on the validated legal evidence, the foundation model generates responses following standard legal reasoning structures. Every conclusion is accompanied by traceable citations through embedded footnotes or direct hyperlinks to the original legal sources. Practical articles and online reference materials are treated only as supplementary evidence and are incorporated only after evaluation by a credibility scoring mechanism and source whitelist, thereby ensuring the authority and reliability of the final output.

Following the technological upgrades introduced in the second quarter of 2025, the platform's multimodal foundation model capabilities were further enhanced in several key aspects.

First, improvements in document understanding and multimodal Document Visual Question Answering (DocVQA) enable the platform to directly interpret tables, tables of contents, and footnotes contained in lengthy PDF documents, significantly reducing citation errors caused by OCR noise. Second, an expanded context window, together with an intelligent paragraph selection mechanism, substantially reduces document fragmentation during legal research, minimizing information loss while improving reasoning continuity across long documents. Third, enhanced function-calling stability has significantly reduced missing or incorrect citations during both retrieval and reference insertion.

Internal evaluations based on a benchmark dataset containing more than 2,000 manually annotated samples demonstrate that evidence traceability has improved by 10–15%, while the F1 score for long-document question answering has increased by 5–8 percentage points. At the same time, average response latency has been reduced by 15–25% through the combined optimization of intelligent caching and batch processing.

Rather than conducting full pre-training of a general-purpose foundation model, Zhihe adopts a domain adaptation strategy based on compliant foundation models specifically optimized for legal applications. Through Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO), the platform adapts foundation models to legal question answering, legal reasoning structures, and professional citation styles. In parallel, proprietary legal-domain embedding models and cross-encoder reranking models—developed through knowledge distillation and contrastive learning techniques—further enhance retrieval performance.

To improve practical legal task performance, the platform employs weakly supervised learning combined with limited manually annotated data to optimize models for contract clause classification, legal risk identification, and fact and timeline extraction. In addition, terminology alignment and LoRA-based style adaptation significantly improve the quality and consistency of legal translation.

The platform also supports multi-model routing, enabling compliant access to both domestic cloud-based foundation models and selected international models. Based on cost, latency, and regulatory requirements, the routing engine dynamically selects the optimal model for each task.

To ensure continuous quality improvement, Zhihe has established a biweekly iteration cycle supported by a comprehensive Quality Gate mechanism. Each iteration proceeds through a complete development pipeline—from product requirement definition, joint technical and legal review, prompt engineering and data preparation, offline integration testing, and limited-scale gray release, to full production deployment—with every stage driven by the practical needs of legal professionals.

Operational teams continuously contribute real legal cases and judicial documents, while engineering teams review unsuccessful cases on a weekly basis and produce reproducible prompt optimization plans and data correction proposals. Both teams jointly take responsibility for key performance indicators, including evidence traceability, answer accuracy, average response latency, and user retention, ensuring that every technological iteration remains closely aligned with real-world legal practice.

IV. Implementation Outcomes: Dual Gains in Efficiency Improvement and Social Value

(I) Economic Benefits and Social Value

The Zhihe AI Platform received immediate and widespread recognition upon its launch. At the 2025 World Artificial Intelligence Conference (WAIC), it was awarded the "Global SME AI Recommended Solution Award". In addition, under the guidance of the Municipal Bureau of Justice and hosted by the Shanghai Bar Association, the platform also received the honor of "Most Favored LegalTech Product among Shanghai Lawyers" at the 2024 Legal Technology Product Competition.

1. Core Value for SMEs

Cost reduction and efficiency improvement: Significantly reduces the cost of legal risk prevention while improving contract processing speed and quality.

Capability democratization: Enables SMEs without dedicated legal teams to obtain enterprise-grade contract risk management capabilities.

Transaction acceleration: Speeds up contract circulation and supports business expansion.

2. Transformational Value for the Industry

Productivity liberation for SMEs: Frees enterprises from time-consuming contract-related work, allowing resources to shift toward high-value activities such as product development and marketing.

Compliance enhancement: Helps enterprises identify contractual risks and achieve compliant operations.

3. Broader Industrial and Social Value

Optimized business environment: Reduces legal friction costs in SME operations and enhances overall market vitality.

LegalTech advancement: Establishes a benchmark application for deep integration of AI and legal services, driving industry innovation.

Support for the real economy: Empowers SMEs with technological tools to strengthen their operational foundation and support high-quality development.

(II) Success Case

Zhihe Company, in collaboration with the Minhang District Justice Bureau, developed an AI-powered contract review platform for SMEs. At the 2025 WAIC "AI for SME – Empowering High-Quality SME Development" Forum, the Zhihe AI contract review system was recognized for its technological innovation, practical effectiveness, and scalability, and received official commendation as an outstanding solution.

The system adopts a dual-engine architecture combining large language models with legal expert rule-based engines, effectively addressing three core pain points in SME contract management: shortage of professional legal resources, low review efficiency, weak risk-control capability. With this system, the cost of reviewing a single contract has been reduced to less than10% of traditional methods, while processing time has been shortened from days to minutes. Through SaaS deployment and API integration, SMEs can obtain enterprise-level contract risk control capabilities without investing in high-end infrastructure.

In practical application, the platform integrates large language models with accumulated expertise from legal professionals and corporate legal departments, providing diversified intelligent contract review services to SMEs in Minhang District. This fully demonstrates the value of inclusive, technology-driven empowerment.

(III) Future Outlook

The competitiveness of the legal services industry is shifting from "manual clause refinement capability" to "AI tool mastery capability."

For the legal profession, AI's most immediate benefit lies in eliminating time-intensive tasks. Previously, summarizing judicial opinions or lengthy testimony records required significant manual effort; now such tasks can be completed within seconds using AI tools. More importantly, AI liberates lawyers from routine administrative work, allowing them to focus on core professional judgment and deep case analysis. AI can precisely identify legal boundaries with unprecedented accuracy, yet it cannot predict judicial interpretation. It provides clear constraints, but final judgment still requires human discretion.

Therefore, human–machine collaboration will become the inevitable pathway for achieving both effective risk control and commercial value in the future legal services industry. The emergence of the Zhihe AI Platform marks a new phase in the company's development. The company's vision is to become "an accelerator of the rule of law in China." Its mission is "to empower legal accessibility through AI and advance legal progress through technology."

The Zhihe AI Platform is not only a unified customer-facing entry point, but also an integrated service ecosystem that consolidates: legal and regulatory case databases, enterprise information search services, the "AI four core modules" (legal research, contract review, document drafting, and Q&A), ZhiShi online legal training platform. It represents a major step forward for the company in the LegalTech sector.

In the future, the platform will continue to evolve, integrating more advanced technologies to further enhance intelligence and efficiency, providing legal professionals with more accurate and effective services. Guided by its mission of "AI empowering legal inclusion and technology advancing the rule of law," the company will continue to move forward and contribute to shaping the future of the legal industry.

Case 41

AI Classroom Teaching Analytics Supporting Teachers' Professional Development

Oriental Cable Network Co., Ltd.

I. Company Profile

Founded in December 1998, Oriental Cable Network Co., Ltd. is a nationally recognized high-tech enterprise, a Shanghai Municipal Enterprise Technology Center, and one of China's leading integrated operators offering comprehensive telecommunications services. Its business portfolio spans interactive high-definition television, broadband services, 5G communications, enterprise data services, smart city solutions, and the Internet of Things (IoT).

The company has strategically focused on four core development priorities: strengthening political communication capabilities, promoting mainstream cultural values, enhancing public cultural services, and safeguarding strategic information resources. Leveraging the distinctive advantages of 5G broadcasting, Oriental Cable Network has established an integrated ecosystem encompassing specialized applications, operational services, and intelligent infrastructure, positioning itself as an important component of China's next-generation cultural communication network and digital infrastructure.

As Shanghai's comprehensive information service provider, the company is responsible for building and operating key municipal digital infrastructure, including the city's "One Network for Government Services" and "One Network for Unified Urban Management" initiatives. By providing robust digital infrastructure and information technology services, Oriental Cable Network plays a significant role in supporting Shanghai's smart city development and accelerating the city's digital transformation.

In the education sector, the company integrates 5G and Artificial Intelligence (AI) technologies to develop a comprehensive portfolio of AI-powered education solutions under the vision of "AI Empowering Every Educational Scenario and Building a New Intelligent Education Paradigm." Its product portfolio includes an AI-powered smart campus platform for intelligent management, resource integration, and educational services; a large AI model that assists teachers in delivering precision instruction; an AI-based precision teaching platform capable of dynamically analyzing classroom activities and learning performance to support personalized learning; and immersive XR metaverse classrooms designed to enhance science education. Through technological innovation, the company is driving the digital transformation of teaching while establishing a new data-driven model for educational research and high-quality education development.

II. Case Introduction

(I) Background

Against the backdrop of deepening educational reform, conventional classroom evaluation methods continue to rely heavily on subjective experience and expert judgment. Such approaches often lack precision, provide limited support for data-driven teaching research, and fail to meet the growing demand for teachers' professional development and balanced regional education.

Meanwhile, the rapid advancement of artificial intelligence and big data technologies has created new opportunities for comprehensive analysis of multimodal classroom data, including video, audio, and behavioral information. These technological advances make it possible to establish intelligent teacher-support tools capable of transforming classroom observations into objective, measurable, and actionable insights.

Based on this vision, the project adopts the authoritative teaching research framework developed by the Chinese Academy of Educational Sciences (CAES) and employs digital technologies to analyze classroom teaching data systematically. Through intelligent assessment of classroom interaction quality, instructional strategies, and teaching effectiveness, the platform provides teachers with targeted recommendations for continuous improvement while supplying regional education authorities with reliable data to support evidence-based decision-making. Ultimately, the project establishes a continuous improvement cycle of "Diagnosis – Improvement – Validation," driving innovation in teaching evaluation and promoting scientific enhancement of educational quality.

(II) Solution Overview

To support intelligent classroom improvement initiatives conducted by the Shanghai Pudong Institute for Education Development (Lingang Branch), Oriental Cable Network developed an AI-powered classroom teaching analytics platform centered on data-driven evidence, intelligent diagnosis, and collaborative teaching research.

Built upon the authoritative teaching evaluation framework of the Chinese Academy of Educational Sciences and supported by multimodal AI technologies, the platform transforms classroom teaching quality assessment into a comprehensive, data-driven analytical process. By integrating classroom observation, intelligent analysis, and collaborative teaching research, it establishes an innovative regional teaching research model that promotes continuous improvement in instructional quality.

The solution is driven by the integration of 5G connectivity, Artificial Intelligence, and big data analytics, enabling the collection and analysis of multimodal classroom information in real time. It supports a wide range of educational applications—including personalized learning, teacher professional development, scientific education, and smart campus management—thereby creating a new paradigm for AI-enabled educational innovation.

1. Technical Architecture and Core Capabilities

The project integrates twelve core AI technologies, including speech recognition, behavioral analytics, and multimodal data processing, to establish a comprehensive AI-powered classroom analytics framework.

By collecting and integrating multimodal classroom data—including teaching practices, instructional design, classroom interactions, video, audio, and behavioral trajectories—the platform leverages big data analytics to construct a comprehensive digital teaching profile for every classroom session. It performs second-level intelligent diagnostics across twelve key dimensions, including classroom interaction quality and instructional strategy effectiveness, and automatically generates both Teacher Competency Radar Charts and Classroom Optimization Reports, providing reliable data support for precision teaching research and continuous instructional improvement.

2. Implementation Path and Practical Deployment

Data Collection and Model Development: The project was implemented across 11 primary and secondary schools (including elementary, junior high, and senior high schools) affiliated with the Lingang Branch of the Shanghai Pudong Institute for Education Development. A total of 48 classroom sessions were observed, with multimodal teaching data collected simultaneously to establish a foundational database for evaluating teacher competencies and classroom instructional quality.

Intelligent Diagnosis and Feedback: Using AI-powered analytics, the platform evaluates multiple aspects of classroom performance, including classroom interaction intensity, knowledge explanation duration, and student participation. Based on these analyses, the system automatically generates personalized recommendations for instructional improvement, enabling teachers to complete a continuous “Diagnosis – Improvement – Validation” cycle supported by objective data.

Innovation in Teaching Research: The project also introduced innovative collaborative teaching research practices. A total of 21 cross-school lesson comparison activities were organized, during which AI-generated comparative analyses of instructional strategies enabled teachers to refine lesson design and optimize teaching methodologies.

In addition, five centralized professional development workshops were conducted, involving more than 100 teachers, to promote practical application of the platform and strengthen teachers’ capabilities in educational data interpretation and AI-assisted instructional improvement.

3. Project Outcomes and Value

Since its implementation, the project has produced significant improvements in both teaching efficiency and educational research. Teachers’ lesson preparation efficiency has increased by approximately 45%, while the productivity of teaching research has more than doubled. Through in-depth analysis of 48 classroom sessions, the platform accurately identified three common instructional challenges, including insufficient classroom interaction diversity and inadequate differentiated instruction. Targeted optimization strategies subsequently led to measurable improvements in classroom effectiveness.

More importantly, the project has established a scalable and replicable model integrating technology empowerment, data-driven diagnosis, and collaborative teaching research. This model provides scientific support for improving educational quality across regions while promoting innovation in AI-enabled teacher professional development.

III. Project Highlights

Building an Intelligent Engine for a New Teaching Evaluation Paradigm

Powered by artificial intelligence and big data technologies, and built upon the professional classroom evaluation framework developed by the Chinese Academy of Educational Sciences, the project digitally reconstructs classroom teaching data—including instructional design, teaching practices, and classroom interactions—into structured, measurable teaching intelligence.

Through AI-driven analysis, the platform not only uses data to validate teaching practices but also uses data to enhance learning outcomes, enabling more precise improvement of teachers' instructional capabilities while providing robust evidence for scientific teaching research.

Key Product Advantages

1. Comprehensive AI Technology Framework: A full-stack AI architecture provides robust technical support for intelligent teaching research and classroom analytics.

2. Comprehensive Multisource Data Collection: Multimodal data from diverse classroom scenarios are integrated to construct comprehensive digital teaching profiles.

3. Multidimensional Intelligent Diagnosis: The platform analyzes classroom performance from multiple perspectives, enabling precise identification of instructional challenges and evidence-based recommendations.

4. User-friendly Design: An intuitive operational workflow allows the platform to adapt easily to diverse classroom teaching environments while minimizing deployment complexity.

Built upon the authoritative teaching research framework of the Chinese Academy of Educational Sciences, and integrating multimodal AI technologies with big data analytics, the platform converts classroom teaching activities and lesson designs into measurable digital teaching profiles.

By combining speech recognition, behavioral analytics, and twelve core AI technologies, it delivers second-level intelligent diagnostics covering classroom interaction quality, instructional strategy effectiveness, and other critical indicators. The platform automatically generates Teacher Competency Radar Charts and Classroom Optimization Reports, supporting innovative teaching research scenarios such as cross-school lesson comparison and collaborative instructional improvement.

Practical implementation has demonstrated significant improvements in teaching research efficiency, with lesson preparation efficiency increasing by approximately 45% and teaching research productivity more than doubling. Through a continuous data-driven improvement cycle, the platform establishes a new paradigm for evidence-based educational research and supports balanced regional educational development.

IV. Implementation Outcomes

This project takes technical empowerment in education as its core, leveraging multimodal AI and big data analysis to achieve innovative transformation in teaching evaluation methods, generating significant economic and social value.

From an economic perspective, it helps teachers improve lesson preparation efficiency by 45%, significantly increases the output efficiency of teaching research outcomes, and reduces the labor and time costs of traditional teaching preparation.

From a social value perspective, the system covers more than 100 teachers, completes 48 in-depth lesson evaluations and 21 rounds of cross-school collaborative teaching research, forming precise teaching diagnosis. Through AI-generated "teacher competency radar charts" and optimization suggestions, the project promotes targeted improvement across 48 lessons, enabling optimization of teaching strategies. By building multimodal data-driven teaching profiles and collecting teacher teaching data across 11 districts for primary and secondary schools, it supports a "diagnosis–improvement–validation" closed loop, accelerating the equalization of educational quality and forming a replicable smart teaching model.


校级AI综合报告
教师数据
课堂数据
思维数据
数据排名
教师教学设计能力
教师教学设计能力详情
教师参与详情
教师综合能力分析
教学设计与课堂实施相关度
AI课堂分析总数
35
33
12
7


校级AI综合报告
高阶思维问题
初阶思维问题
课堂问-答-评数据
学生个体应答
教师反馈
23.4%
76.6%
77.78%
22.22%
42.55%
教师高阶思维引导数据排名
年级-教师高阶思维引导数据排名
学科-教师高阶思维引导数据排名

Part Three
AI Applications

This section brings together benchmark cases that vividly depict how small and medium-sized enterprises, even without abundant resources, can examine their own businesses with a strategic eye and precisely identify pain points in core areas such as production processes, quality control, and customer service as entry scenarios for AI. They clearly demonstrate the complete closed from strategic planning, scenario anchoring, technology selection and implementation, to final value assessment. These practices fully prove that successful AI applications do not lie in the coolness of the technology, but in its deep alignment with business needs.

- Market Insight & Demand Analysis (Case 42)
- Supply Chain Planning & Procurement (Case 43)
- Manufacturing & Quality Inspection (Cases 44-45)
- Marketing, Sales & Channel Management (Case 46)
- Marketing, Sales & Channel Management (Case 46)
- After-sales Service & Customer Operations (Cases 47-48)
- Financial Management & Risk Control (Case 49)
- Others (Cases 50-53)

Case 42

"Nezha Goes Global": An AI-Powered Cross-border Legal Services Platform

Lantai Partners

I. Company Profile

Lantai Partners is a comprehensive Chinese law firm committed to the principles of professional excellence, internationalization, and appropriately scaled development. Dedicated to becoming a leading law firm serving Chinese enterprises worldwide, the firm provides full-spectrum legal services covering virtually every area of commercial law.

Since its establishment, Lantai has actively supported China's Belt and Road Initiative, providing legal services to numerous Fortune Global 500 companies and leading Chinese enterprises—including Tencent, ZTE, China Construction Third Engineering Bureau, Xiaomi, and many others—as they expand into international markets. To strengthen its cross-border legal capabilities, the firm established the Lantai Global Legal Academy, following its strategic philosophy of "Three Markets" while developing expertise across 22 practice areas and more than 200 specialized legal service teams.

Lantai maintains an extensive domestic and international service network. In China, the firm operates 34 offices across major cities, while internationally it has established more than 30 overseas offices spanning 26 countries, including Singapore and India. As one of China's most internationally connected law firms along the Belt and Road, Lantai continues to expand its global legal service footprint.

Beyond legal practice, the firm actively promotes international legal collaboration through professional exchange programs, including the "Middle East New Opportunities" forum and the "Rainbow Lawyer" training series. It has also pioneered an equity partner system and received the Jinfeng Award for Innovation in Law Firm Management, reflecting its commitment to continuous professional excellence and high-quality legal services.

Leveraging its extensive international legal service network and practical experience, Lantai has accumulated substantial expertise in cross-border legal affairs, particularly in supporting Chinese enterprises' global expansion. Through close collaboration with lawyers in multiple jurisdictions, the firm delivers customized legal compliance solutions tailored to local regulatory environments and clients' specific business needs.

II. Case Introduction

(I) Background

With the rapid growth of the global digital economy, Chinese enterprises are accelerating their international expansion across sectors including e-commerce, intelligent manufacturing, and financial services. However, entering overseas markets exposes companies to increasingly complex legal environments characterized by constantly evolving compliance requirements, high cross-border legal service costs, and demanding response times.

These challenges are particularly significant for small and medium-sized enterprises (SMEs), which often lack sufficient legal resources and international expertise to manage regulatory requirements across multiple jurisdictions or respond promptly to overseas legal risks.

For many years, China's cross-border legal service market has faced several structural challenges. First, enterprises often encounter significant information asymmetry when attempting to understand foreign legal systems and regulatory requirements. Second, businesses frequently struggle to identify legal professionals with the appropriate jurisdictional expertise and industry experience. Third, the absence of effective mechanisms for international legal risk data sharing and early warning makes it difficult to establish proactive compliance management. Finally, the high cost of traditional legal services places substantial financial pressure on SMEs seeking overseas expansion.

Against this backdrop, artificial intelligence has emerged as a key driver for transforming legal services. AI-enabled legal technologies not only improve operational efficiency and reduce service costs but also provide critical technological support for building a more intelligent, accessible, and globally connected legal service ecosystem.

(II) Solution Overview

The project aims to build an intelligent, globally oriented legal service platform powered by large language models (LLMs), creating a sustainable digital ecosystem for cross-border legal services that effectively supports Chinese enterprises expanding overseas while accelerating the digital transformation of the international legal service industry.

The platform incorporates advanced technologies spanning natural language processing, legal knowledge graph construction, workflow orchestration, and intelligent document processing, enabling comprehensive digital management of cross-border legal service processes. By integrating these technologies, the platform transforms conventional legal services into a more accurate, efficient, and intelligent support system.

Upon deployment, the platform reshapes the traditional legal service model from an experience-driven, reactive approach into a data-driven, AI-powered, proactive service model. It serves as a bridge connecting Chinese enterprises with an international network of legal professionals while providing 24/7 intelligent online legal consultation, automated document generation, precise lawyer matching, and professional referral services.

Supported by Lantai's global legal network, the platform has expanded its service coverage from 35 jurisdictions specializing in intellectual property protection to more than 50 key overseas investment destinations, representing a substantial step toward the deep integration of artificial intelligence and legal services.

(III) Introduction to Implementation Plan

1. Project Overview

The project is led by Lantai Partners in collaboration with Shanghai SenseTime Technology Development Co., Ltd. and iFlytek Co., Ltd. to develop China's first large language model (LLM)-based legal service platform dedicated to supporting Chinese enterprises in overseas expansion. Its flagship product, the "Nezha Goes Global" Cross-border Legal Services Platform, pioneers a hybrid service model that combines AI intelligence with professional legal expertise, providing comprehensive, accurate, and around-the-clock legal support for businesses operating internationally.

The platform currently covers legal regulations in 35 countries and regions across nine major practice areas, including corporate compliance, intellectual property protection, international trade, and cross-border dispute resolution. Through an efficient service workflow, enterprise users can submit legal questions arising from overseas business operations, after which the AI system performs real-time retrieval and semantic matching across specialized legal databases to generate accurate responses based on applicable legal provisions.

Where AI-generated answers are insufficient to resolve complex legal issues, the platform automatically connects users with experienced cross-border legal professionals, enabling customized legal consultation and end-to-end service delivery. Designed for enterprises, lawyers, investors, and other users engaged in international business, the platform represents a comprehensive intelligent legal ecosystem supporting every stage of overseas expansion.

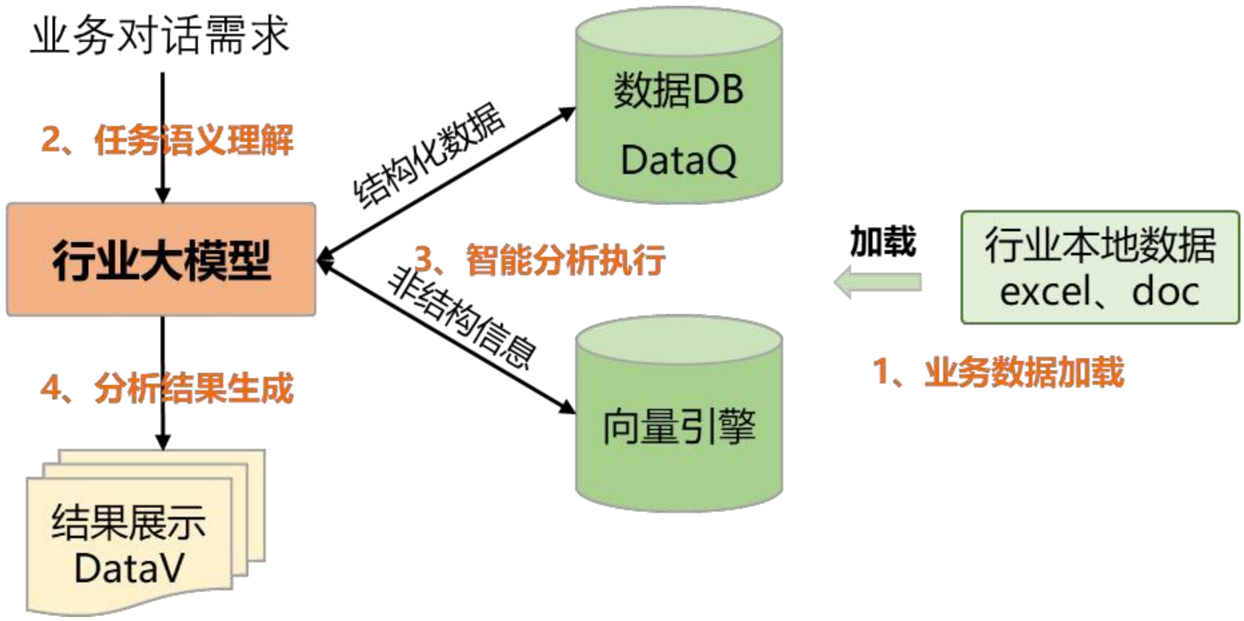
The project adopts the Qwen-plus large language model developed by Alibaba Cloud as its foundational AI engine. Built upon an integrated architecture combining large language models, vertical legal knowledge bases, and multimodal interaction, it supports legal research, intent recognition, multilingual legal consultation, expert collaboration, and numerous other intelligent legal services.

The project has significant demonstration value in advancing AI-powered international legal cooperation. As China's first 24-hour AI legal platform specifically designed for overseas business scenarios, it fills an important gap in intelligent cross-border legal services. Compared with conventional digital legal tools, the platform establishes a collaborative ecosystem linking enterprises, lawyers, and global legal resources. Supported by 130 professional legal service tags, a 10-category qualification verification system, and an international legal expert network, the platform has become a leading example of AI-enabled legal innovation in China.

2. Technical Architecture

The platform adopts an integrated architecture based on three core components: Large Language Models (LLMs), Vertical Legal Knowledge Bases, Multimodal Human–AI Interaction. Powered primarily by the Qwen-plus model and enhanced through industry-specific knowledge engineering, the system incorporates multidimensional information fusion technologies to construct comprehensive legal knowledge repositories.

These repositories are integrated with vector retrieval technologies to support efficient interaction between AI models and structured legal knowledge, enabling intelligent legal research, user intent recognition, legal consultation, document generation, and expert collaboration. The overall processing workflow consists of four major stages.



3. Application Architecture

The platform adopts a layered architecture consisting of a front-end WeChat Mini Program and mobile application, connected to an AI Agent service layer supported by large language models and vector databases. Users interact with the system through natural language, submitting legal questions directly through the client interface. The AI Agent then invokes large language models and specialized legal knowledge bases to perform legal question answering, intelligent legal retrieval, compliance review, legal recommendation, and document generation.

The overall platform architecture is organized into five layers: Interaction Layer, Application Layer, AI Agent Layer, Model Capability Layer, Infrastructure Layer. This modular architecture provides excellent scalability, deployment flexibility, and support for future expansion of intelligent legal services.

4. Product Functions

子系统	功能模块	功能描述
微信小程序端	注册	提供微信小程序端多类用户的注册相关功能
	问答	提供用户问答交互的相关功能
	转人工	提供用户由智能体转人工服务的相关功能
	律师集市	提供已登记律师信息的查询功能
	设置	提供微信小程序的相关设置功能
后台管理端	数据分析	提供应用使用的分析功能
	用户管理	提供用户的增删改查相关功能
	会话明细	提供用户会话明细管理的相关功能
	账号管理	提供账号管理的相关功能
	角色管理	提供角色权限管理的相关功能
	预设管理	提供标签管理与系统首页的相关预设功能

III. Project Highlights

(I) Business Innovation

The project introduces China's first 24-hour AI legal large language model platform specifically designed for enterprises expanding into international markets. By pioneering an integrated "AI + Legal Rules" service model, the platform constructs a dedicated cross-border legal knowledge system using private-domain legal knowledge engineering technologies.

Drawing upon the practical experience of hundreds of lawyers at Lantai Partners, the platform has established specialized legal knowledge bases covering 36 countries and regions and nine professional legal practice areas, including corporate compliance, intellectual property, and cross-border dispute resolution.

Its legal knowledge repository currently contains more than 52,000 evaluated data records and over 140 authoritative legal knowledge documents, enabling intelligent retrieval and precise matching of legal regulations across multiple jurisdictions while effectively addressing the challenges of fragmented cross-border legal information.

(II) Technological Innovation

The platform pioneers a deeply integrated architecture combining LLMs, Vertical Legal Knowledge Bases, and Multimodal Interaction, creating a new paradigm for AI-powered cross-border legal services.

Through multidimensional information fusion—including enterprise profiles, geographic location, historical interactions, business scenarios, and legal risk indicators—the platform significantly improves user intent recognition and contextual understanding, effectively addressing the long-standing challenges of inaccurate legal consultation and inefficient service matching.

Powered by the Qwen-plus large language model and jointly developed with Alibaba Cloud, the platform adopts a RAG architecture to substantially improve both the accuracy and timeliness of AI-generated legal responses while minimizing the risk of hallucinations commonly associated with large language models.

The combination of LLM reasoning, domain-specific legal knowledge, and vector retrieval enables highly accurate legal consultation, multilingual legal analysis, intelligent document generation, and cross-jurisdiction legal knowledge retrieval.

(III) Service Innovation

The project moves beyond the limitations of conventional legal information retrieval systems by establishing a collaborative Enterprise–Lawyer–Global Legal Network service ecosystem.

Supported by a rigorous ten-category lawyer qualification verification mechanism and a 130-dimensional professional tagging system, the platform connects enterprises with carefully screened legal professionals across multiple jurisdictions.

Unlike traditional legal service platforms that merely provide lawyer directories, the system enables a closed-loop service model integrating AI consultation, professional lawyer matching, and expert legal services. This approach ensures both service quality and professional reliability while significantly reducing information asymmetry between enterprises and legal practitioners.

(IV) Business Model Innovation

The platform introduces an innovative dual-engine service model combining intelligent AI tools with professional legal experts. On one hand, AI provides free or low-cost legal consultation services, substantially lowering the entry barriers for enterprises—particularly SMEs seeking international expansion. On the other hand, strict lawyer qualification reviews and service quality evaluation mechanisms ensure that complex legal matters are handled by experienced professionals, preserving the high standards expected of premium legal services.

This hybrid model achieves an effective balance between accessibility, service efficiency, and professional quality, while creating a sustainable commercial ecosystem for AI-enabled legal services.

(V) Advanced System Architecture

Built on the Spring Boot framework, the platform integrates Bailian Knowledge Base, the Qwen large language model, and PostgreSQL/Redis databases into a robust five-layer architecture consisting of: Interaction Layer, Application Layer, AI Agent Layer, Model Capability Layer, Infrastructure Layer.

The architecture supports multimodal interaction, high-concurrency processing, flexible deployment, and secure data management. Its modular design provides a scalable foundation for future functional expansion and continuous technological evolution.

IV. Implementation Results

(I) Business Impact

The successful deployment of the platform has fundamentally transformed the traditional legal service model, replacing an experience-driven, offline, manually coordinated workflow with an intelligent, online, and precision-matching service ecosystem. Its business achievements are reflected in several key dimensions.

1. Significant Improvement in Service Efficiency and Accessibility

The platform’s core 24/7 AI legal consultation service eliminates geographical and time constraints, enabling users to obtain immediate legal support regardless of location. Traditionally, enterprises needed to identify appropriate lawyers, arrange consultations, negotiate fees, and schedule appointments before receiving legal advice—a process that could take hours or even days. With the new platform, users simply submit legal questions through the WeChat Mini Program, and the AI system delivers preliminary legal analysis within seconds based on regulations from more than 35 jurisdictions and multiple professional legal domains.

According to operational statistics, the AI assistant now handles thousands of legal consultations each day, with an average response time of less than three seconds. Initial consultation cycles that previously required several hours—or even several days—have been shortened to just a few minutes, improving overall service efficiency by more than 90%.

2. Reduced Operating Costs and Lower Compliance Expenses

For Lantai Partners, the AI platform now undertakes more than 70% of standardized and repetitive legal consultation tasks, allowing lawyers to focus on more sophisticated, high-value legal services such as cross-border mergers and acquisitions, major commercial litigation, and complex compliance matters.

This optimized allocation of professional resources significantly improves operational efficiency while enhancing overall service quality. For enterprise clients, the platform offers a tiered service model that combines free AI consultation with premium professional legal services. Routine legal inquiries—including company registration, labor law, and standard contract templates—can be resolved through AI at little or no cost, dramatically reducing legal expenditures during the early stages of overseas market expansion.

Internal evaluations indicate that enterprises can reduce the cost of obtaining basic legal information by more than 60% compared with traditional legal service models.

3. Intelligent Matching of Global Legal Resources

One of the greatest challenges facing enterprises expanding overseas is identifying the most appropriate legal professionals. The platform's Lawyer Marketplace addresses this challenge through its rigorous qualification review mechanism and refined professional tagging system. Enterprise users can search and compare lawyers according to destination country, practice area, language capability, industry expertise, and historical service performance, much like selecting products through an e-commerce platform.

This transparent and standardized matching process substantially reduces information asymmetry and shortens the time required to identify suitable legal counsel from several days to only a few minutes. To date, the platform has successfully connected Chinese enterprises with lawyers from more than 30 countries, facilitating hundreds of successful online legal collaborations while effectively solving the "last mile" challenge of international legal services.

4. Advancing the Standardization of Legal Services

The platform is more than a conventional legal information tool; it is a continuously evolving intelligent legal knowledge hub. Leveraging RAG and other AI technologies, the platform continuously transforms the practical experience of Lantai Partners and its affiliated lawyers—including case precedents, legal interpretations, regulatory updates, and best practices—into structured, reusable digital knowledge assets that are incorporated into its legal knowledge base.

This ongoing knowledge accumulation not only enhances the accuracy, consistency, and timeliness of AI-generated legal advice, but also establishes a standardized workflow and service interface for cross-border legal services. By reducing dependence on individual expertise and converting professional knowledge into scalable digital capabilities, the platform lays a solid foundation for the modernization and industrialization of legal services.

(II) Economic Benefits

The project's economic value extends beyond the platform's direct revenue generation. More importantly, it establishes a sustainable business ecosystem capable of creating long-term value for enterprises, legal professionals, and industry partners.

1. Legal Service Revenue

The platform generates recurring revenue through multiple channels, including: Paid online legal consultation services for enterprise clients; Membership subscription services that provide access to premium AI consultation, document download privileges, and additional value-added features; Platform service fees and annual membership fees paid by participating lawyers.

These diversified revenue streams provide a stable commercial foundation for the platform's long-term development.

2. Data-driven Value-added Services

The platform accumulates extensive high-quality legal consultation records, user behavior data, and case databases during daily operations.

These valuable data assets can be further developed into customized industry compliance reports, international market risk analysis services, and specialized legal intelligence products for research institutions, government agencies, and large enterprises, creating new business opportunities beyond conventional legal consulting.

3. Legal Resource Commercialization

As the platform's global lawyer network continues to expand—with a long-term objective of serving more than 2,000 registered lawyers—additional business models such as legal service commissions, premium lawyer recommendations, and professional marketing services are expected to become important sources of revenue growth.

Based on current business projections, the project targets annual operating revenue exceeding RMB 7 million, with an estimated return on investment (ROI) of no less than 4. Through effective cost control and efficient operations, the platform is expected to achieve break-even within two years, establishing a healthy and self-sustaining business model.

(III) Social Benefits

1. Supporting Chinese Enterprises in Global Expansion

The platform provides SMEs with affordable, efficient, and accessible legal services, significantly lowering the legal barriers associated with overseas expansion.

Through real-time legal risk monitoring and intelligent compliance support, enterprises are better equipped to identify and mitigate potential legal risks before they develop into costly disputes.

2. Expanding Access to Legal Services

By overcoming geographical and language barriers, the platform enables enterprises and individual users worldwide—particularly those participating in the Belt and Road Initiative—to obtain high-quality legal information and professional services more conveniently than ever before.

3. Accelerating the Digital Transformation of the Legal Industry

Through intelligent legal knowledge management and AI-powered service tools, the project promotes the transformation of legal practice from an experience-driven profession to a data-driven intelligent service industry.

The platform not only improves lawyers' productivity but also raises the overall quality, efficiency, and consistency of legal services across the industry.

4. Supporting National Development Strategies

The project actively supports the implementation of the Belt and Road Initiative by providing Chinese enterprises with reliable legal assistance for overseas investment, international trade, and cross-border business operations.

By strengthening international legal cooperation, the platform contributes to the development of a more secure and efficient global business environment for Chinese enterprises.

5. Demonstrating an Innovative AI-enabled Legal Service Model

As one of China's pioneering AI-powered legal service platforms, the project provides a highly replicable technological and operational model for the legal industry.

Its successful integration of large language models, legal knowledge engineering, intelligent legal consultation, and professional legal services offers valuable experience for establishing future industry standards while accelerating the digital transformation of law firms, particularly small and medium-sized legal practices.

Case 43

From Traditional Operations to Intelligent Transformation: Gangyin E-Commerce Harnesses AI to Drive the Digital and Intelligent Transformation of the Steel Supply Chain

Shanghai Gangyin E-Commerce Co., Ltd.

I. Company Profile

Shanghai Gangyin E-Commerce Co., Ltd. (stock code: 835092), commonly known as Gangyin E-Commerce, was founded in 2008. It is one of China's leading B2B e-commerce platforms for the steel industry and among the country's earliest trillion-yuan industrial e-commerce platforms.

Leveraging e-commerce as its foundation and industrial big data as its core capability, the company integrates upstream and downstream supply chain resources to build a comprehensive industrial ecosystem encompassing market information, online trading, supply chain finance, warehousing and processing, logistics and distribution, and Software-as-a-Service (SaaS) solutions.

Committed to the integration of Internet technologies, big data, cloud computing, the Internet of Things (IoT), artificial intelligence, and the steel industry, Gangyin E-Commerce continuously promotes digital transformation across the industrial value chain. Through technological innovation, the company strives to reduce transaction costs, improve circulation efficiency, and create a safer, more transparent, and more efficient steel trading ecosystem.

To date, the platform has established strategic partnerships with more than 350 steel mills, operates over 500 cooperative warehouses nationwide, and connects with more than 50 logistics service providers, forming a nationwide supply chain service network.

The platform serves over 190,000 enterprise customers, with annual online transaction volume exceeding RMB 600 billion in 2024 and settlement volume reaching RMB 220 billion. Gangyin E-Commerce has ranked among China's Top 100 Internet Enterprises for nine consecutive years, making it one of the country's most influential industrial Internet platforms.

II. Case Introduction

(I) Background

The steel industry is a pillar of China's real economy, and the modernization of its supply chain plays a vital role in improving resource allocation efficiency and supporting high-quality economic development. However, conventional steel circulation systems continue to face numerous structural challenges, including cumbersome transaction procedures, severe information asymmetry, underdeveloped credit mechanisms, elevated financial risks, inefficient logistics coordination, and fragmented supply chain management.

Traditional steel trading relies heavily on manual order processing, paper contracts, and offline verification procedures. These practices not only reduce operational efficiency but also increase transaction costs while limiting the scalability and responsiveness of supply chain services.

In response to these challenges, Gangyin E-Commerce has built an integrated digital supply chain platform centered on industrial Internet technologies. By integrating market information, online transactions, supply chain finance, warehousing, logistics, and SaaS applications, the company has established a comprehensive digital ecosystem for the steel industry.

In recent years, China has introduced a series of national policies encouraging the deep integration of artificial intelligence with the real economy. In August 2025, the State Council issued the Action Plan for AI Plus, calling for the accelerated development of AI-enabled industrial applications and new digital service models.

Against this policy backdrop, Gangyin E-Commerce has actively incorporated artificial intelligence into every stage of its business operations. By developing an AI-enabled intelligent supply chain platform covering trading, finance, logistics, warehousing, and customer service, the company seeks to improve operational efficiency, reduce overall supply chain costs, and accelerate the digital and intelligent transformation of the steel circulation industry.

(II) Key Operational Challenges

Despite continuous digitalization efforts, the platform continued to face six major operational challenges:

1. Low Transaction Efficiency

Traditional manual order processing required significant human intervention. Paper-based contract generation, review, archiving, and approval procedures were time-consuming, limiting platform scalability and reducing customer experience.

2. Insufficient Risk Control

The absence of real-time credit evaluation and dynamic risk warning mechanisms increased the likelihood of fraudulent transactions, financing risks, and bad debt exposure. Conventional risk management lacked the capability to identify emerging risks through continuous data analysis.

3. Complex Financial Workflows

Processes such as payment collection, invoice management, and voucher processing relied heavily on manual operations, resulting in low processing efficiency, higher error rates, and increased financial management costs.

4. Inefficient Logistics Coordination

Logistics information could not be tracked in real time. Delayed responses to transportation exceptions frequently led to cargo disputes and contractual disagreements, increasing overall logistics costs.

5. Low Warehouse Management Efficiency

Warehouse receiving, dispatching, and inventory management depended largely on manual operations. Data entry errors and delayed inventory updates reduced warehouse efficiency while increasing management costs.

6. Limited Customer Service Capacity

Customer inquiries, complaints, and after-sales support required substantial manual effort. Service response times and customer satisfaction were therefore constrained by the availability of customer service personnel.

III. Project Highlights

Driven by artificial intelligence technologies, Steel Silver E-commerce has built an intelligent service system covering the entire steel industry value chain. The project adopts a modular architecture design supported by a data middle platform, integrating advanced technologies such as machine learning, natural language processing (NLP), Internet of Things (IoT), and big data analytics. It enables an end-to-end intelligent upgrade from transaction execution, risk control, to supply chain coordination.

By deeply integrating the MCP protocol, the system ensures seamless collaboration and data interoperability among various intelligent agents, forming a comprehensive AI-enabled service solution.

(I) AI Trading Assistant: Full Automation of Inquiry and Order Processing

In the trading process, an intelligent solution covering inquiry, order placement, and contract management is built based on rule engines and machine learning algorithms.

Leveraging natural language processing technology, the system can interpret user inquiry requests in real time and automatically connect internal resource pools with external market data. It provides business users with accurate and timely steel pricing and available resource information, significantly improving inquiry response efficiency and customer satisfaction.

Meanwhile, the platform uses self-developed robotic automation to enable automatic order placement in steel mill systems, optimizing order processing workflows, reducing manual data entry, and lowering error rates.

By introducing smart contract technology, the system automatically generates standardized contract documents based on transaction terms, product specifications, and pricing conditions, while also performing automated contract review. This greatly improves transaction efficiency, shortens the trading cycle, and eliminates bottlenecks in supply chain collaboration.

(II) AI Risk Control Assistant: Building a Full-Lifecycle Dynamic Risk Management System

In risk management, the system integrates multi-source internal and external data to build customer risk profiles and credit line evaluation models.

The platform fuses multi-dimensional real-time data streams, including user identity and qualification information, historical transaction patterns, trading behaviors, real-time transaction dynamics, and corporate credit records, to dynamically assess customer risk levels.

The system covers the entire transaction lifecycle:

Pre-transaction stage: qualification review, credit rating, and dynamic credit limit adjustment

During transaction stage: real-time fraud detection, behavior monitoring, and fund flow tracking

Post-transaction stage: performance evaluation and abnormal settlement alerts

For overdue delivery, overdue payments, and invoice delays, the system performs real-time risk monitoring and dynamic credit adjustment.

It also integrates the “Fengbao” enterprise intelligence platform to strengthen risk warning and coordination mechanisms. In cooperation with credit reporting institutions such as “Tianyancha,” the system ensures the authenticity and reliability of enterprise data.

In addition, the platform integrates the E-signature system (E-Sign), using OCR-based intelligent contract comparison to detect abnormal clauses in real time. Combined with digital signatures, timestamps, and blockchain-based notarization, the system ensures secure and compliant contract execution, significantly enhancing the intelligence and trustworthiness of the risk control framework.

(III) AI Financial Assistant: Enabling Automated Financial Process Transformation

In the financial domain, the system integrates APIs with automated workflow engines to achieve deep connectivity with the “Leping Platform.”

It enables intelligent transformation of financial operations, including: automatic parsing of payment instructions; electronic invoice recognition and management; automatic voucher generation and reconciliation.

These capabilities reduce manual intervention, improve financial processing accuracy and timeliness, and significantly enhance enterprise capital turnover efficiency.

(IV) AI Logistics Assistant: Real-Time Visibility and Intelligent Logistics Management

In logistics operations, IoT devices and tracking data are integrated to build an intelligent logistics monitoring platform.

The system aggregates: vehicle GPS location and status; transportation plan details; cargo condition data; historical routes and delivery time performance; warehouse geographic information; third-party traffic and weather data.

By introducing route optimization algorithms and anomaly detection models, the system continuously monitors transportation trajectories, issues early warnings for potential risks, and enables intelligent dispatching of vehicles. It supports real-time risk alerts and efficient emergency response during transit, significantly reducing cargo loss rates and dispute occurrences, while ensuring safe and controllable logistics operations throughout the entire process.

(V) AI Warehouse Assistant: Enhancing Precision in Warehouse Management

In warehouse operations, the system connects with the “Cloud Warehouse Digital Link” platform to integrate real-time inventory data from nationwide partner warehouses.

It consolidates: SKU-level inventory information; inbound and outbound records; electronic warehouse receipt status. By introducing automated robotics systems, the platform enables intelligent warehouse operations, reduces operating costs, and significantly improves inventory turnover accuracy and operational efficiency.

(VI) AI Customer Service Assistant: 24/7 Intelligent Customer Support

In customer service operations, a 24/7 intelligent support system is established.

Using natural language processing technology, the system understands customer intent, automatically responds to inquiries, and generates standardized replies.

This effectively reduces the workload of human customer service staff while improving response speed and customer satisfaction, delivering a continuous, efficient, and intelligent service experience.

IV. Implementation Results

Through deep integration of AI technologies across its entire business chain, Steel Silver E-commerce has achieved significant improvements in operational efficiency and cost optimization across trading, risk control, finance, logistics, warehousing, and customer service.

This transformation not only strengthens the platform’s operational capabilities and service quality, but also provides a practical benchmark for the digital and intelligent transformation of the steel supply chain industry.

(I) Comprehensive Improvement in Operational Efficiency and Industry Benchmarking

Through deep AI integration, the platform’s operational efficiency has reached an industry-leading level.

The AI Trading Assistant processes more than 30,000 steel mill orders annually, automatically generates over 750,000 contracts, and issues more than 820,000 delivery orders, significantly improving end-to-end transaction automation.

In 2024, Steel Silver E-commerce achieved a trading volume exceeding 60 million tons and a total transaction value surpassing RMB 220 billion. The platform’s per capita service efficiency reached 5.94 million tons, representing an 11.89-fold improvement compared with traditional steel trading enterprises.

The AI Financial Assistant processed more than 1.56 million payment transactions, generated over 540,000 accounting voucher, and handled more than 510,000 invoicing operations within a year. Average financial processing time was reduced by 30%, significantly improving both financial efficiency and capital turnover.

(II) Intelligent Upgrade of Risk Control System Ensuring Transaction Security

The AI Risk Control Assistant generated more than 30,000 risk alerts annually, automatically reduced prepayment credit limits over 30,000 time, and reviewed more than 1 million tons of cargo automatically.

The platform has significantly reduced bad debt risk and substantially improved transaction security, creating a safer and more reliable trading environment.

(III) Logistics and Warehousing Optimization Enabling End-to-End Controllability

The AI Logistics Assistant conducted over 800,000 trajectory monitoring operations annually and issued more than 100,000 parking risk alerts.

Logistics anomaly response time has been reduced to within 5 minutes, while cargo ownership dispute rates have significantly declined, enabling real-time visibility, controllability, and efficient coordination across the logistics process.

The AI Warehouse Assistant completed over 150,000 robotic inbound and outbound operations annually, significantly improving warehouse operational efficiency and accuracy.

(IV) Customer Service Optimization Delivering High-Quality Service Experience

The AI Customer Service Assistant handled a large volume of standardized inquiries, reducing overall reliance on human agents by 10%, while increasing customer satisfaction to over 95%.

With uninterrupted 24/7 intelligent service and automated responses, the system delivers a highly efficient, standardized, and consistent customer service experience.

Through the construction of an AI-driven intelligent assistant system, Steel Silver E-commerce has achieved a full-chain upgrade of intelligent steel circulation. The platform not only significantly enhances operational efficiency and risk control capabilities, but also provides comprehensive digital empowerment for upstream and downstream enterprises in the industrial supply chain.

The company's AI application practice fully demonstrates the enormous potential of deep integration between artificial intelligence and traditional industries, offering a valuable reference for the digital transformation of conventional sectors.

Looking ahead, Steel Silver E-commerce will continue to deepen the development and application of AI technologies in the steel vertical domain, expand industrial internet scenarios, and promote the transformation of the steel industry from "digital empowerment" to "intelligent-driven development." This will contribute significantly to the construction of a modern industrial system.

Case 44

AI-Based Intelligent Defect Detection through Multi-Source Information Fusion for Additive Manufacturing

Shanghai Yunjin Additive Technology Co., Ltd.

I. Company Profile

Shanghai Yunjin Additive Technology Co., Ltd. specializes in additive manufacturing services and customized equipment development. The company is recognized as a National High-Tech Enterprise, a Specialized, Refined, Distinctive and Innovative (SRDI) SME, and a qualified supplier for military equipment manufacturing.

Leveraging its strong capabilities in aerospace additive manufacturing, the company has established long-term research collaborations with leading institutions including the Chinese Academy of Engineering Physics, Shanghai Jiao Tong University, and Tongji University.

The company operates an intelligent additive manufacturing production facility covering approximately 1,100 square meters, employs around 120 staff members, of whom more than 30% are dedicated to research and development, and possesses comprehensive capabilities spanning product design, simulation verification, process development, quality inspection, and full-process manufacturing technologies.

Maintaining a strong innovation-driven strategy, the company invests heavily in research and development, with annual R&D expenditure accounting for approximately 12% of revenue, while maintaining annual revenue growth exceeding 20%.

II. Case Introduction

(I) Background

Laser Powder Bed Fusion (L-PBF) has become one of the most important additive manufacturing technologies because of its high design flexibility, short production cycles, excellent material utilization, and ability to manufacture complex geometries.

The technology has been increasingly adopted in high-end manufacturing sectors such as aerospace, satellites, missiles, aviation engines, medical devices, and precision industrial equipment. For example, the global market for aerospace additive-manufactured components has already reached RMB 26 billion, while China accounts for approximately 17% of the global aerospace additive manufacturing market.

China’s additive manufacturing industry has maintained rapid growth in recent years. It is projected that the domestic market will reach approximately RMB 63 billion by 2025, with an average annual growth rate of around 24%. The industry’s expanding application scenarios and accelerating product iteration cycles continue to create enormous market opportunities for additive manufacturing technologies.

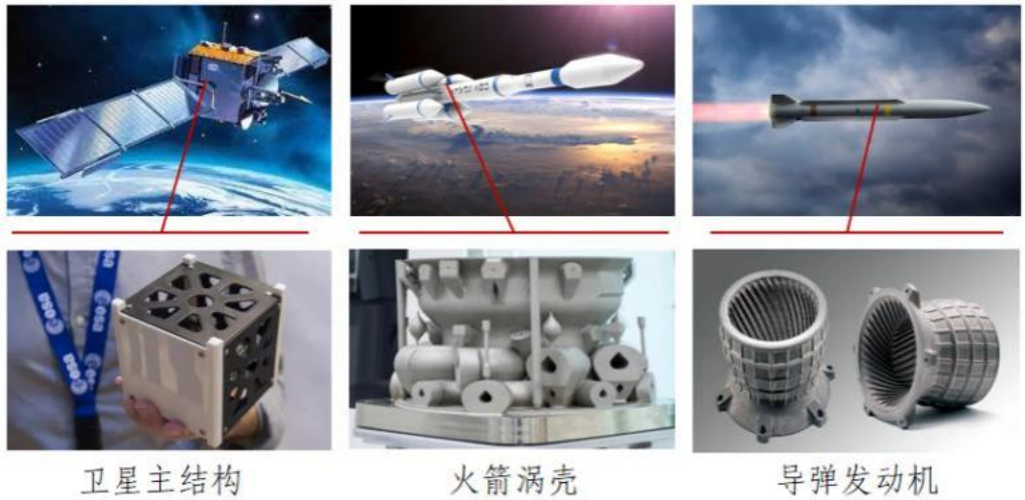


Figure 1 Typical Additively Manufactured Parts in the Aerospace Field

In powder-bed additive manufacturing, auxiliary support structures are critical to the forming quality of cavity and overhanging features. However, the fabrication and subsequent removal of these support structures significantly extend both the manufacturing and post-processing time of components, thereby increasing overall production costs. During the post-processing of aerospace load-bearing pressure components, poor accessibility of cantilevered cavity features often makes it impossible to remove support structures completely, which directly affects key service performance indicators such as flow resistance.

In addition, for aerospace bracket components, the forming time and material weight of support structures can account for up to 80% of the entire product. Taking a satellite bracket as an example, Figure 2 compares the processing time between traditional printing methods and support-free additive manufacturing processes. Under conventional printing methods, the total processing time is approximately 29 hours, whereas with support-free additive manufacturing, the post-processing time is reduced to 14 hours, effectively cutting the overall production time by half.

With support-free additive manufacturing, the printing stage eliminates the time required for support generation, and post-processing only requires the removal of a small amount of residual support along with a reduction in polishing workload. Therefore, support-free additive manufacturing holds significant importance for cost reduction and efficiency improvement in aerospace component manufacturing.

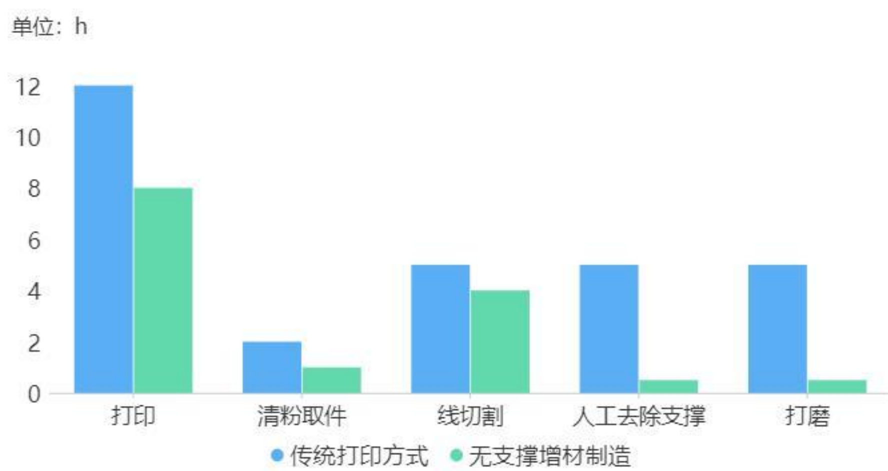


Figure 2 Comparison of Production Process Time for Satellite Brackets

Despite these advances, quality assurance remains one of the major bottlenecks restricting large-scale industrial adoption.Current manufacturing processes still rely heavily on offline inspection methods. During fabrication, defects such as pores, cracks, warping, spheroidization, slag inclusions, and incomplete fusion may occur. These defects significantly affect the mechanical properties, reliability, and service life of metal components.

For critical industries such as aerospace, where structural integrity and reliability are paramount, the inability to detect manufacturing defects in real time greatly limits the industrialization and commercialization of additive manufacturing technologies.

(II) Implementation Plan

The research content of this project is divided into three main parts.

First, the development of a real-time morphology and temperature in-situ monitoring device. A binocular structured-light measurement system and an infrared thermal imaging system are employed to collect real-time data on powder bed morphology, geometric dimensions, and temperature field distribution.

Second, the development of a data fusion algorithm. A deep separable convolution-based image fusion method with an adaptive pixel-weighting strategy is proposed, combined with sparse tensor modeling to enable fast and efficient integration of morphology–temperature monitoring data.

Third, the construction of an intelligent defect recognition and localization algorithm. Based on a lightweight deep learning model, feature mapping of morphology–temperature coupled information is used to establish defect correlation analysis, enabling intelligent defect detection and precise localization.

The detailed research framework is shown in Figure 3.

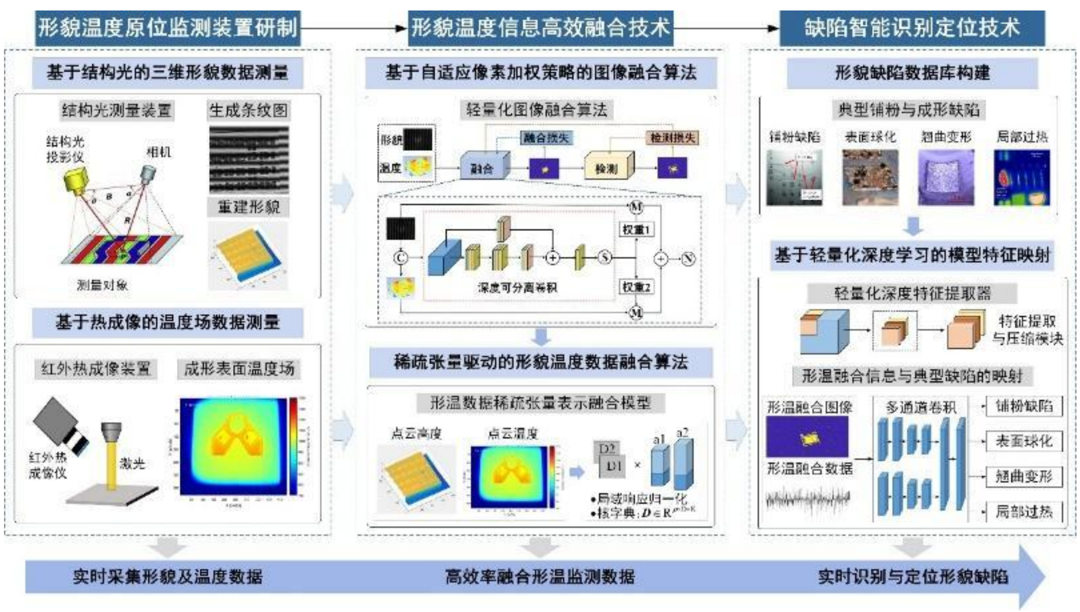


Figure 3 Project Research Content

1. Development of an In-Situ Morphology and Temperature Monitoring System

Real-time online monitoring of morphology and temperature field information is the foundation for intelligent defect identification.

The project established a binocular structured-light measurement platform and developed a method for rapid 3D profile reconstruction, enabling real-time monitoring of the powder bed and component surface morphology in three dimensions.

In addition, an infrared thermal imaging platform was constructed. A pixel-width recognition and calibration algorithm was developed, along with an adaptive temperature image stitching and enhancement mechanism, enabling continuous tracking of temperature field evolution characteristics, as shown in Figure 4.

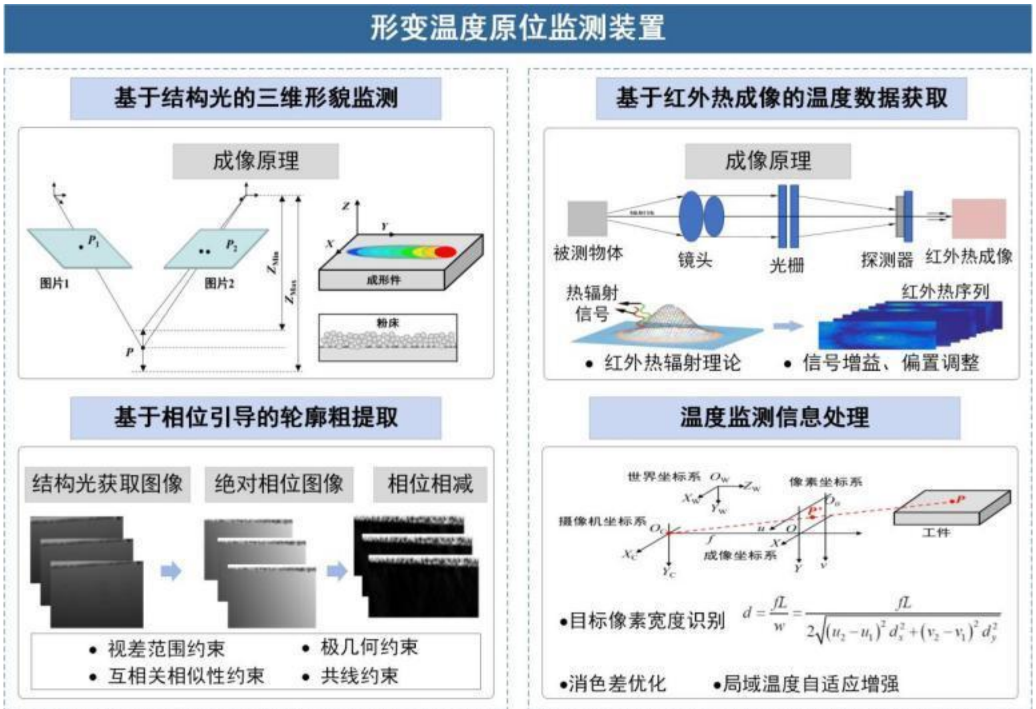


Figure 4 Design Approach for the In-Situ Morphology and Temperature Monitoring System

The in-situ morphology and temperature monitoring system consists of one structured-light camera and one infrared camera. Within a working area of approximately 200 × 200 mm, the structured-light camera is used to acquire morphology point cloud data for monitoring geometric defects such as warping and collapse.

The infrared camera is responsible for capturing thermal images to monitor temperature-related defects, such as overheating and abnormal heat accumulation.

2. Morphology–Temperature Image Data Fusion

Morphology and temperature information alone cannot sufficiently characterize defect features.

To address this limitation, the project develops an adaptive weighted fusion algorithm and designs a lightweight separable convolution-based feature extraction module, enabling efficient fusion of morphology–temperature monitoring images.

To tackle the challenge of multi-source heterogeneous data fusion, a sparse tensor-driven multi-source fusion algorithm is further developed, enabling fast and accurate integration of morphology and temperature information, as shown in Figure 5.

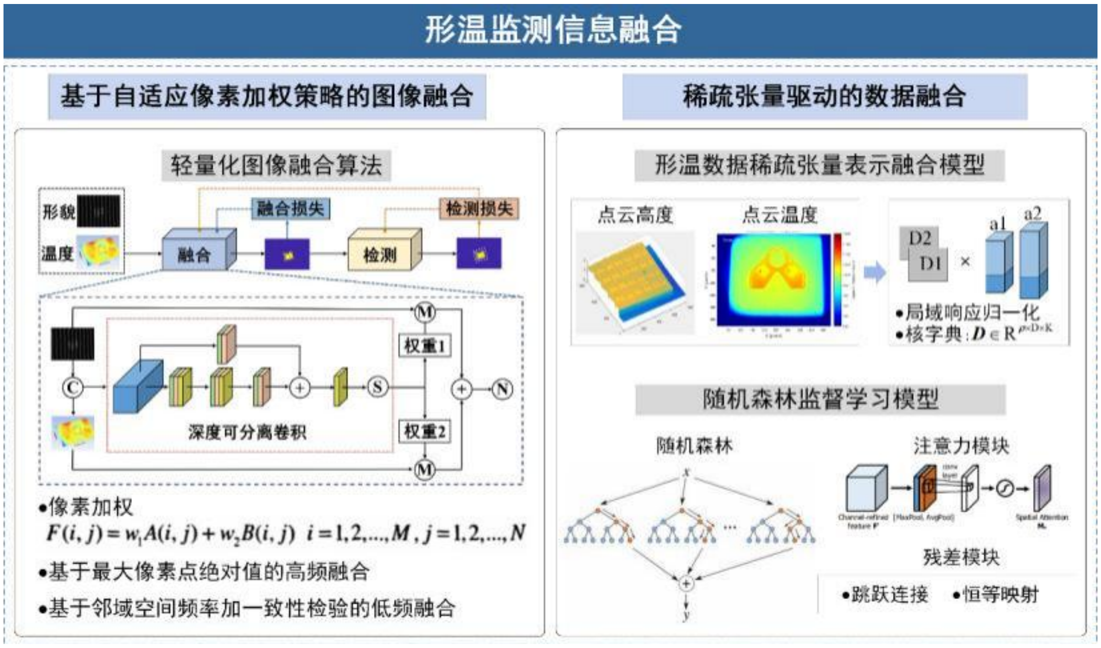


Figure 5 Morphology–Temperature Information Fusion Framework

3. Intelligent Recognition and Localization of Forming Defects

This project adopts YOLO (You Only Look Once) for rapid detection and classification of multiple targets within images, enabling efficient localization and identification of defects.

During actual imaging, due to the imaging principles of infrared thermal cameras, reflection phenomena may occur. The infrared radiation intensity received by the camera increases at reflective surfaces, resulting in false overheating signals. Therefore, such erroneous thermal artifacts must be filtered out. Based on this preprocessing step, temperature-related defect locations are annotated accordingly.

Online characterization of forming defects suffers from issues such as low accuracy and limited dimensional representation. To address these challenges, this task develops a multi-channel convolutional mapping algorithm and a lightweight defect feature compression and extraction algorithm, establishing a mapping relationship between morphology–temperature information and forming defects.

A defect database is further constructed to enable efficient acquisition of morphology–temperature data and rapid defect recognition and localization, thereby providing a solid foundation for forming quality control, as shown in Figure 6.

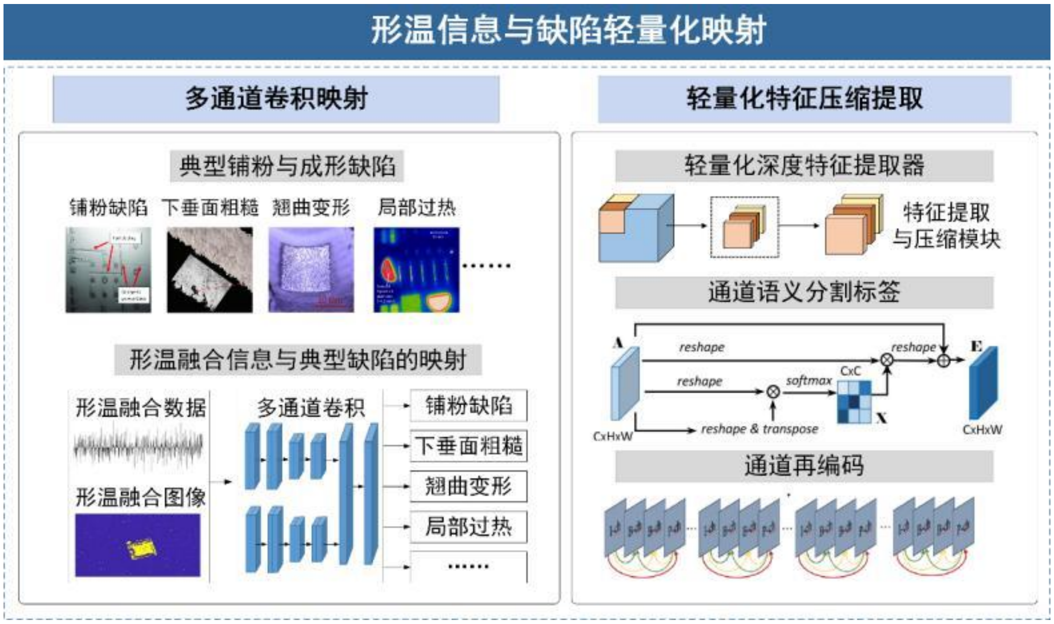


Figure 6 Mapping Between Morphology–Temperature Monitoring Information and Typical Defects

Point cloud data is acquired using a 3D scanner and can accurately represent the geometric morphology of an object's surface. In order to identify defects such as warping, depressions, or protrusions, a normal vector-based analysis method is typically adopted.

Compared with traditional curvature-based methods, the computation of normal vector angles is more intuitive and computationally efficient, making it particularly suitable for real-time processing of large-scale point cloud data. At the same time, this method exhibits a certain degree of robustness to noise, enabling it to preserve fine geometric features while avoiding misjudgments caused by isolated outlier points.

By setting an appropriate threshold for normal vector angle deviation, potential defect regions can be quickly identified, providing an important basis for subsequent defect classification, area estimation, and visualization.

After completing the recognition of temperature and morphology defects, the system further designs an active intervention module, aimed at helping operators fully understand the distribution and severity of defects, thereby improving the interpretability and practical value of detection results.

The system maps defect information from the image coordinate system to the physical space and converts it based on pixel resolution to calculate the actual area of temperature anomaly regions. For morphology defects, a triangulation-based surface reconstruction method is used to estimate the surface area of each detected defect region within the point cloud.

Finally, all defect information is aggregated and analyzed to generate a comprehensive statistical report, including defect types and total affected areas, providing both data support and visual evidence for subsequent quality evaluation and maintenance decision-making.

4. Integration of Defect Intelligent Monitoring and Recognition System

The project developed a support-free laser powder bed fusion additive manufacturing system, and integrated the defect intelligent monitoring and recognition system into the equipment.

The image acquisition system collects data twice during each powder spreading and printing cycle, specifically after powder spreading and after printing. The collected data are cached in the main memory and processed via an FPGA platform, before being transmitted as feedback signals to the control-side PC, enabling evaluation of powder bed quality and printing quality.

Uniform backlighting is used inside the chamber to reduce interference caused by metal powder reflectivity. A high-resolution camera captures the dynamic powder spreading process to avoid motion blur. A structured-light camera projects stripe patterns onto the powder layer, and surface morphology is reconstructed by analyzing stripe deformation to obtain height information.

The system also develops a defect intelligent monitoring and recognition module. When a new defect is detected, a pop-up alert is triggered, displaying key information such as defect type, location coordinates, and area size. This active alert mechanism significantly enhances system intelligence, enabling operators to immediately understand equipment status and take timely corrective or maintenance actions.

(III) Challenges Before Implementation

Intelligent defect recognition is a prerequisite for forming quality control and improvement. In support-free additive manufacturing processes, defects such as powder bed defects and surface spheroidization significantly affect forming quality. At the same time, due to the absence of support constraints, formed components are more prone to deformation.

Process monitoring and quality control in support-free printing is therefore a critical research direction and a common challenge in both academia and industry.

However, current defect recognition still faces several difficulties:

1. Monitoring equipment must simultaneously capture both temperature and morphology information;

2. Efficient fusion methods for morphology–temperature data are lacking;

3. Single-modal features are insufficient for accurate defect identification, requiring multi-source fusion-based intelligent recognition methods.

These challenges lead to low detection efficiency and poor accuracy, making it difficult to meet the requirements of future equipment development and industrial manufacturing.

(IV) AI-Driven Problem Requirements

1. Develop a rapid 3D reconstruction method for powder beds and formed components, and establish an online parameter self-optimization mechanism for vision measurement systems to address reflectivity interference and high computational cost in in-situ morphology measurement.

2. Develop a pixel-level, lightweight, and efficient fusion algorithm for morphology–temperature monitoring images, enabling registration and integration of equipment-side measurement data, and addressing the challenge of high-confidence fast fusion of multi-source data.

3. Develop a rapid intelligent defect recognition and localization algorithm based on hybrid temperature–morphology features to improve accuracy and efficiency, while addressing the mechanism of multi-scale feature evolution affecting defect formation and development.

This case focuses on the urgent challenge of in-situ defect monitoring and intelligent recognition in metal support-free additive manufacturing. It develops an in-situ multi-sensor temperature–morphology detection system, proposes an efficient multi-source data fusion algorithm based on deep separable convolution and sparse tensor modeling, and implements intelligent defect recognition based on deep learning and multi-modal feature fusion. The results provide important technical support for aerospace manufacturing of complex components and demonstrate strong application value in additive manufacturing for aerospace engineering. This project represents a typical application of artificial intelligence in advanced manufacturing, addressing long-standing challenges in process quality monitoring, particularly in support-free printing scenarios, and holds significant industrial and scientific importance.

III. Project Highlights

(I) Key Technological Breakthroughs

1. Key Technology for Position-Free Disturbance Powder Spreading

To address the disturbance of component forming posture caused by the contact force of scraper-based powder spreading, this project proposes a novel position-free disturbance-free powder spreading approach based on a composite process of “ultrasonic powder feeding + scraper compaction + vacuum powder suction.”

Specifically: Ultrasonic powder feeding is used to avoid scratching or mechanical disturbance caused by traditional scraper pushing; Scraper compaction is applied to improve powder bed density; Vacuum powder suction is used to ensure precise and controllable powder layer thickness.

2. Key Technology for Intelligent Defect Localization and Recognition

To meet the demand for rapid defect identification and localization in additive manufacturing processes, this project adopts a deep separable convolution-based image fusion algorithm and a sparse tensor-driven data fusion algorithm to achieve fast alignment and fusion of morphology and temperature information.

Based on a lightweight deep learning feature-mapping model, a morphology defect database is constructed, enabling intelligent defect recognition and localization.

(II) Uniqueness Description

The technical achievements of this project are highly transferable and can be replicated across multiple fields. The analysis is carried out from three dimensions: application scenarios, tools and software, and overall solution framework.

1. Application Scenarios

This project implements a defect recognition and monitoring method based on multi-source information fusion, particularly achieving cross-dimensional fusion between temperature and morphology data.

This requirement is not limited to additive manufacturing; it also exists in machining, welding, and forming processes. Therefore, the results of this project can be extended to similar manufacturing domains to address multi-source data fusion challenges in defect detection.

2. Tools and Software

The project independently developed an in-situ defect monitoring and recognition system for additive manufacturing. This software can: run independently on equipment terminals, be integrated into machine control systems, and be deployed on cloud platforms as a service-based system.

The developed models and algorithms for data fusion and defect recognition can be rapidly migrated to similar scenarios. With only a small amount of retraining data, the system can achieve strong defect recognition performance.

3. Solution Framework

This project represents a typical application of defect detection and quality control in manufacturing processes.

For enterprises undergoing digital transformation, this solution provides a valuable reference model. Its application promotes the deployment of support-free printing technology in high-end equipment manufacturing and further enhances the manufacturing capability of advanced industrial systems. It also provides strong support for the development of new quality productive forces (new productive capabilities) and technological innovation.

IV. Implementation Results

Aerospace is a strategic industry closely related to national security. At present, powder bed additive manufacturing has already been applied in aerospace mass production. However, aerospace manufacturing urgently requires support-free additive manufacturing technology to meet batch production demands of key models.

The rapid defect recognition and localization technology developed in this project addresses critical technical bottlenecks in support-free additive manufacturing and breaks technological restrictions imposed by advanced manufacturing countries, providing a solid foundation for the independent development of controllable support-free additive manufacturing technologies.

The project outcomes are expected to play a leading role in the manufacturing of complex aerospace components and enable large-scale and batch application of domestically developed high-end additive manufacturing equipment in the aerospace sector.

(I) Rocket Engine Volute (Turbopump Casing)

The turbopump of a rocket engine is a key component of reusable launch vehicle propulsion systems. The volute is a typical internal cavity and overhanging structure.

Traditional manufacturing relies on a combination of additive manufacturing and vacuum diffusion bonding. To improve forming efficiency, an integrated support-free printing process is adopted; however, defects such as cracks may occur during printing, leading to failure in meeting performance requirements.

The AI-based multi-source information fusion in-situ defect detection technology enables online monitoring of forming quality during the printing process, effectively preventing quality issues caused by powder bed defects. After adopting this technology, crack detection accuracy can reach over 99.5%.

This project supports the development of independently designed support-free powder bed additive manufacturing equipment and integrated process systems. The overall manufacturing efficiency of rocket engine volute components is improved by more than 50%, manufacturing cost is reduced by more than 50%, and delivery cycles are shortened by more than 30%, providing strong support for reusable launch vehicle development in China.

(II) Aviation Engine Bearing Casing

The bearing casing of an aero-engine is a critical component and typically features a large-scale thin-walled structure. To avoid deformation, traditional additive manufacturing processes require extensive support structures.

To improve forming efficiency and reduce post-processing workload, a support-free printing process is adopted. However, this makes the component more prone to stress-induced deformation during printing.

The AI-based multi-source information fusion in-situ defect detection technology enables real-time monitoring of forming quality during the printing process, preventing quality issues caused by powder bed defects. After applying this technology, detection accuracy for defects such as porosity and deformation reaches over 99.2%.

The project supports the development of independently designed support-free powder bed additive manufacturing equipment and integrated process systems. The overall manufacturing efficiency of aero-engine bearing casings is improved by more than 48%, manufacturing cost is reduced by more than 45%, and delivery cycles are shortened by more than 25%, providing strong support for the development of China’s aviation industry.

Case 45

Enhancing Quality Prediction and Precision Control in High-End Manufacturing through AI and Multimodal Data Fusion

Shanghai Micro Electronics Technology Co., Ltd.

I. Company Profile

Shanghai Micro Electronics Technology Co., Ltd., founded in September 2000, is a high-tech enterprise specializing in the R&D and manufacturing of fiber-optic gyroscopes and related optical components. Recognized as a National High-Tech Enterprise and a Shanghai “Specialized, Refined, Distinctive, and Innovative (SRDI)” Enterprise, the company has established a strategic presence across the entire fiber-optic gyroscope industry chain. It provides complete solutions covering fiber-optic gyroscopes, optical devices, optical modules, and integrated fiber-optic navigation systems.

Manufacturing is carried out in strict accordance with national military standards, and the company’s products are widely applied in aerospace, aviation, marine engineering, resource exploration, intelligent equipment, and other high-end industrial sectors, where they have earned an excellent reputation for reliability and performance.

Over more than two decades of technological innovation, the company has mastered core technologies in fiber-optic gyroscopes and key optical components while establishing a comprehensive management system encompassing R&D, manufacturing, quality assurance, and after-sales service. These capabilities have laid a solid foundation for the continued advancement of China’s high-end optical manufacturing industry.

II. Case Introduction

(I) Background and Business Requirements

In the context of accelerating global technological revolution and industrial transformation, China is actively implementing the strategy of becoming a manufacturing powerhouse, promoting the manufacturing sector toward intelligent and green development.

As a core embodiment of national industrial competitiveness, the high-end equipment manufacturing industry plays a critical role in ensuring the security of industrial and supply chains. With the rapid development of aerospace, low-altitude economy, and related sectors, market demands for core components such as fiber-optic inertial navigation systems have become increasingly stringent in terms of customization, reliability, and delivery efficiency.

As a leading enterprise in China’s fiber-optic inertial navigation field, the company’s core products—such as fiber-optic gyroscopes and IMUs—are characterized by high technological intensity and complex manufacturing processes. Traditional production and quality management models rely heavily on manual experience and are unable to meet the high-dimensional and massive data analysis requirements under the trend of multi-variety and small-batch customization. This significantly restricts the company’s ability to improve quality and efficiency.

To address these challenges, the company initiated this project with the goal of deeply integrating artificial intelligence technologies, breaking through key bottlenecks such as quality data analysis and risk prediction, and building intelligent operational capabilities. During the advancement of digital and intelligent factory construction, two critical application gaps have been identified:

First, insufficient depth in quality data analysis. Traditional methods rely on manual experience and single-parameter threshold judgment, making it difficult to systematically extract hidden patterns from high-dimensional and nonlinear data. This limits knowledge accumulation and inheritance in quality management.

Second, insufficient forward-looking capability in quality risk prediction. Existing quality control systems are mainly based on final inspection, lacking trend prediction and early warning mechanisms. This makes it difficult to identify and intervene in quality issues at an early stage, thereby constraining cost control and delivery assurance for high-value products.

(II) Implementation Plan

To address these challenges and support high-quality development, the company launched an AI-enabled quality analytics improvement project. The core objective is to build a data-driven and AI-empowered intelligent operational system, promoting the digitalization and intelligence transformation of the entire manufacturing inspection process.

AI technologies have been deeply integrated into quality management and fault knowledge construction, playing a key role in the following areas:

1. Intelligent Quality Diagnosis and Analysis

AI significantly enhances the ability to interpret high-dimensional and nonlinear quality data. Through machine learning algorithms, the system can automatically identify complex relationships between process parameters and product quality. When quality fluctuations occur, it can rapidly locate key influencing factors, reducing root cause analysis time from several days to minutes-level response.

This greatly improves analysis efficiency and standardization, reduces dependence on individual experience, and lays a solid foundation for knowledge accumulation and reuse.

2. Predictive Quality Control

AI enables the transformation of quality management from post-event inspection to pre-event prediction and early warning. Based on time-series analysis and deep learning algorithms, the system can monitor process parameter trends in real time and identify subtle anomaly signals, enabling early warning of quality risks.

This capability allows engineers to proactively intervene before problems occur, improving process control accuracy and providing an effective pathway for reducing defect rates and optimizing quality costs.

III. Project Highlights

(I) Key Technological Breakthroughs

To address core challenges such as difficulty in tracing product quality fluctuations and low efficiency in root cause analysis, this project builds an intelligent quality management system driven by data intelligence. The system is built upon deep integration of multiple enterprise systems, including MES (Manufacturing Execution System), DAS (Data Acquisition System), and TPM (Total Productive Maintenance), forming a comprehensive data foundation that covers the entire manufacturing process.

On this basis, artificial intelligence technologies are deeply introduced to establish a closed-loop integration between data and AI-driven analytics, enabling end-to-end intelligent empowerment. Through the organic integration of system-level data fusion and AI-based analysis, the project successfully transforms traditional quality management from a post-event traceability model into a new intelligent paradigm featuring: real-time in-process warning, and predictive pre-process control.

The specific innovative applications in core business scenarios are reflected in the following aspects:

1. Intelligent Testing and Data Fusion

Figure 1 Flexible Testing System

Figure 2 Data Acquisition System for Inertial Measurement Products

Through the flexible testing system, a digital twin-based test plan is generated, which automatically controls the inertial measurement product data acquisition system (DAS), as well as environmental equipment such as temperature chambers and turntables, to execute tests and collect data. The system then performs deep data fusion between this high-precision test data and MES production process data as well as TPM equipment status data, constructing a complete “Man–Machine–Material–Method–Environment” (5M1E) data chain.

When quality fluctuations occur, the system is able to accurately trace them down to specific production batches, production lines, workstations, operators, and supplier batches, enabling fine-grained traceability.

2. Fault Mechanism Analysis Based on Association Rule Mining

An innovative association rule mining algorithm is introduced to systematically analyze historical fault data and maintenance records. The algorithm can automatically identify key rules, for example: When fault pattern A occurs, the effectiveness confidence of maintenance action B reaches 95%." In this way, implicit expert experience is transformed into quantifiable and reusable decision knowledge.

This approach not only improves the efficiency of diagnosing complex quality issues, but also helps uncover weakly correlated feature combinations that are often overlooked in traditional analysis yet hold predictive value, thereby providing algorithmic support for continuously enriching the fault knowledge base.

3. Quality Early Warning and Diagnosis Based on Multimodal Data Fusion and Deep Learning

By constructing deep learning predictive models for multi-source continuous parameters—such as temperature curves, vibration spectra, and test waveforms—and integrating multimodal data including production, equipment, and environmental information, the system achieves forward-looking quality risk early warning and systematic diagnosis.

This technological approach overcomes the limitations of traditional single-source analysis by leveraging deep neural networks for cross-domain feature extraction and correlation modeling. It not only detects subtle anomalies in the production process that are difficult for human observation—enabling early warning—but also reveals deep interdependencies across heterogeneous data sources, providing a comprehensive perspective for root cause analysis.

The resulting model possesses continuous learning capabilities, allowing it to improve progressively with accumulated data and gradually establish an adaptive, predictive quality management system with proactive intervention capability.

3. Building an Intelligent Quality Closed-Loop Management System Driven by an "AI Middle Platform"

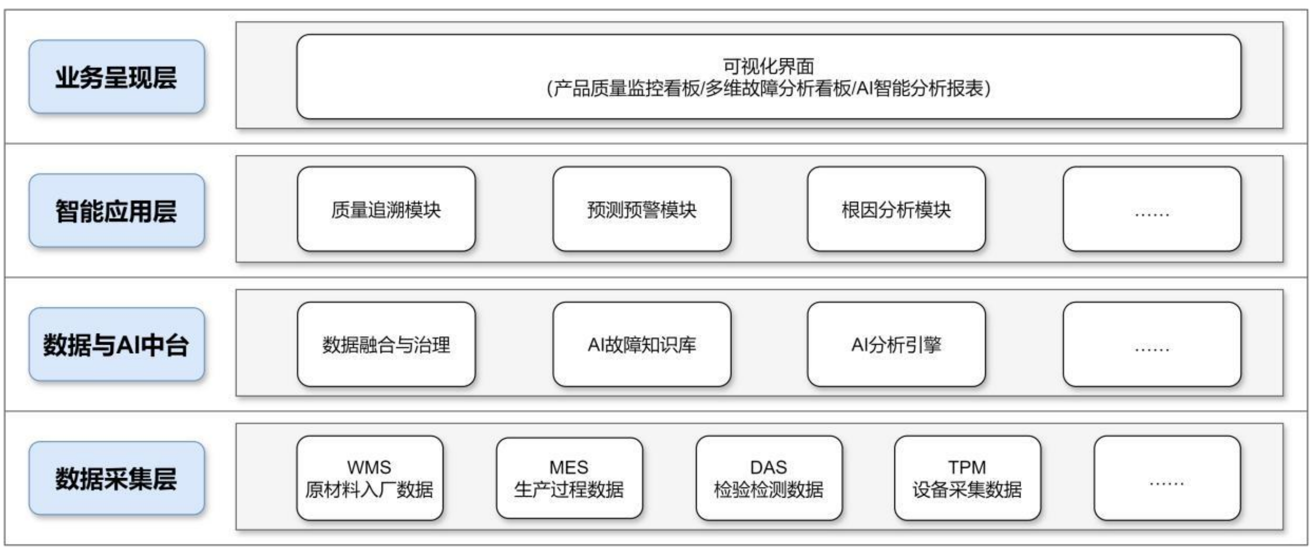


Figure 3 Intelligent Quality Management Platform

Based on your provided text, here is the publication-style English translation (consistent with prior AI middle-platform / industrial AI documentation tone): Centered on the construction of an enterprise-level "AI Middle Platform", this system aims to systematically enhance the intelligence level of quality management.

The AI middle platform provides standardized capabilities such as machine learning, model training, and algorithm management, and encapsulates complex AI technologies into reusable service components (e.g., anomaly detection, root cause analysis, etc.). This significantly lowers the barrier to AI adoption and provides a robust platform foundation for the large-scale and inclusive deployment of AI capabilities. On this basis, the system establishes an end-to-end quality closed-loop management mechanism, spanning:

"Data Sensing → AI Analytics → Decision Support → Knowledge Accumulation."

If you want, I can also help you unify all "AI 中台" related paragraphs into a single standardized architecture description section for the full report.

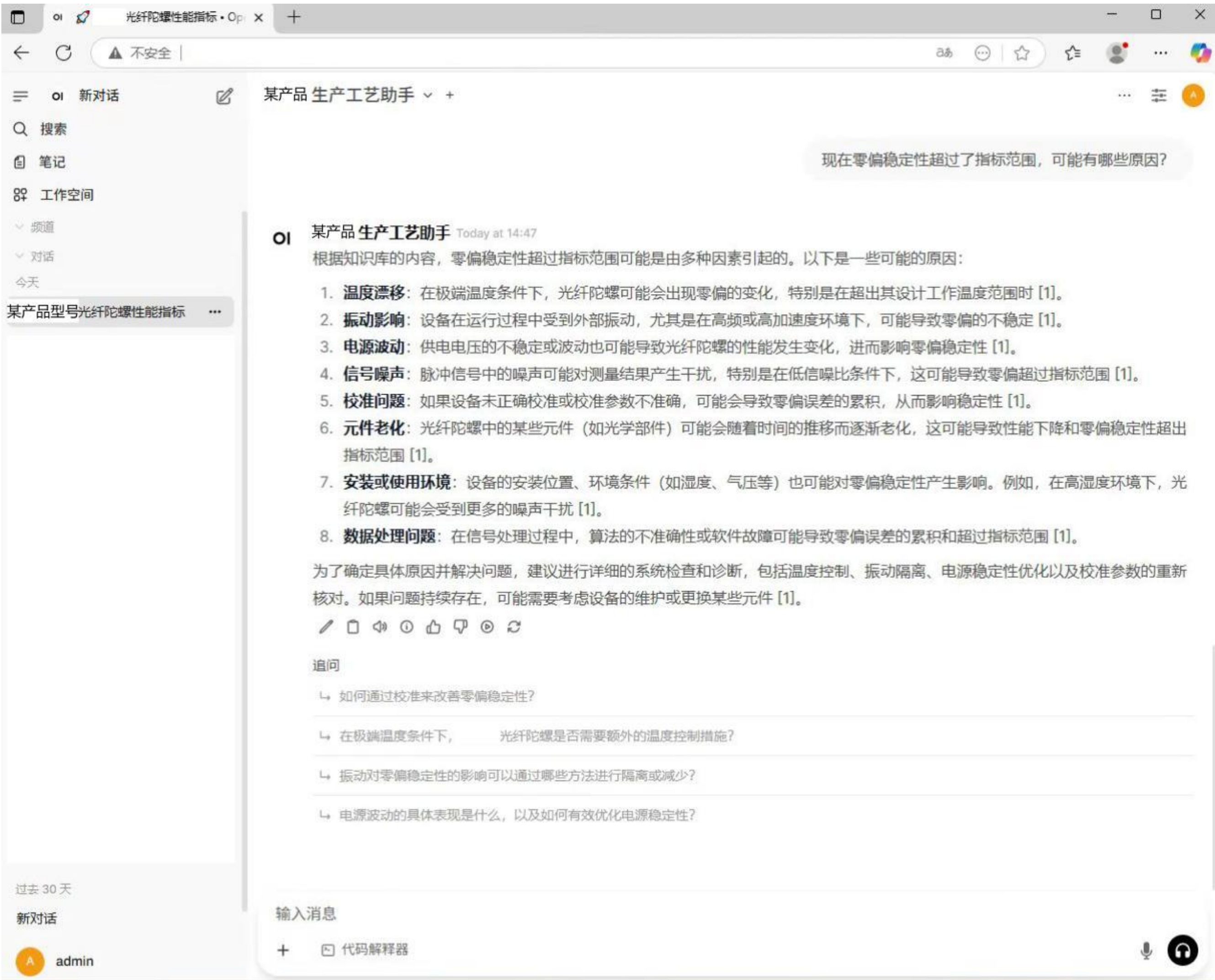


Figure 4 Intranet AI-Based Fault Knowledge Base Q&A System

The AI analysis results are not only used for real-time decision support, but are also continuously accumulated into the enterprise fault knowledge base, where they are transformed into core knowledge assets, including diagnostic rules, analytical cases, and predictive models. This closed-loop architecture ensures that the system continuously self-optimizes with the accumulation of production practice data, ultimately forming a data-driven and continuously evolving learning-oriented quality management system.

(II) Scenario-Specific Uniqueness

This project has achieved a fundamental transformation in quality management within the high-end manufacturing sector, demonstrating significant scenario-specific uniqueness and fully reflecting its distinct industrial value. It provides a replicable implementation pathway for industrial transformation and upgrading. The uniqueness is reflected in the following five aspects:

1. Paradigm Innovation in Quality Management

The project breaks through the limitations of traditional quality management by shifting the core paradigm from: "standard-based inspection" and "experience-based judgment" to data-driven prediction and decision-making. By leveraging AI technologies, it establishes an objective data-driven analytical system, eliminating reliance on individual experience and intuition, significantly improving scientific rigor and reliability, and achieving a qualitative leap in quality management methodology.

2. Benchmark AI Empowerment for High-End Manufacturing

Targeting the high-precision and high-complexity characteristics of fiber-optic gyroscope manufacturing, the project successfully validates the effectiveness of AI in addressing extreme manufacturing quality challenges. In complex scenarios involving intricate physicochemical processes and precision assembly, AI algorithms demonstrate exceptional accuracy and reliability, providing a replicable implementation pathway for the intelligent upgrade of China's high-end equipment manufacturing industry.

3. Enterprise-Level AI Operating System

Through the construction of an AI Middle Platform, AI capabilities are elevated from isolated pilot applications to a continuously operable enterprise core capability. This platform-based architecture differs from traditional project-based implementations, laying a solid foundation for continuous iteration and large-scale deployment of AI technologies, and forming a core competitive advantage in digital transformation.

4. Innovative Knowledge Management Mechanism

The project has built an AI-driven quality knowledge base containing more than 3,000 rules and cases, effectively addressing long-standing issues in manufacturing such as "knowledge silos" and "experience loss." It transforms expert knowledge into reusable and transferable digital assets, pioneering a new paradigm of knowledge management in manufacturing industries.

5. New Human–Machine Collaborative Decision-Making Model

The project implements a collaborative mechanism of "AI-driven analysis and human-led decision-making." The AI system is responsible for tasks such as data mining and pattern recognition, while human experts focus on strategic judgment and innovation optimization. This collaborative paradigm effectively integrates computational intelligence with human expertise, ensuring both scientific rigor and decision efficiency, and providing an important methodological foundation for enterprise intelligent transformation.

IV. Implementation Results

Through the deep empowerment of quality management enabled by artificial intelligence technologies, this project has achieved significant outcomes in both economic benefits and social value.

(I) Economic Benefits

From an economic perspective, AI-driven precise traceability and intelligent root cause analysis have improved quality traceability efficiency by 90%, while reducing complex problem analysis cycles by 65%, significantly lowering fault diagnosis and troubleshooting costs. Based on AI-powered predictive early warning models, quality issue detection is shifted further upstream in the production process, resulting in a 28% reduction in quality-related costs. In addition, the comprehensive energy consumption per unit output value has decreased by 12.5%, achieving a dual improvement in cost efficiency and green manufacturing performance.

(II) Social Value

From a social perspective, the project provides a replicable implementation pathway for the intelligent transformation of the high-end equipment manufacturing industry, promoting the large-scale application of AI technologies in industrial scenarios. By enhancing the high-quality, large-scale, and stable supply capability of inertial navigation products, the project effectively strengthens the autonomy and controllability of China's industrial chain in strategic sectors such as national defense security and the low-altitude economy.

The carbon dioxide emissions per unit output value have decreased by 15.2%, representing a concrete practice of supporting China's "dual carbon" goals in the manufacturing sector. During implementation, the project also cultivated a group of interdisciplinary and highly skilled professionals. Through an industry–academia–research–application collaboration mechanism, it contributed valuable experience to ecosystem development and technical standardization.

The project has built an AI-driven quality knowledge base containing more than 3,000 rules and cases, transforming tacit expert experience into transferable digital assets and enabling a 40% reduction in quality improvement cycle time.

These achievements fully demonstrate the critical role of technological innovation in enhancing the core competitiveness of the manufacturing industry and supporting national strategic development objectives.

Case 46

An AI-Powered Intelligent Assistant for Multi-Scenario Applications in the Industrial Internet

Shanghai Ouyeel Co., Ltd.

Shanghai Ouyeel Co., Ltd. (Stock Code: 6676.HK) was established in 2012 and is headquartered in Shanghai. As one of China's leading industrial Internet platform providers, the company leverages cloud computing, big data, and artificial intelligence to build an integrated industrial service ecosystem covering online trading platforms, logistics, warehousing and processing, SaaS applications, and supply chain services.

Committed to digital transformation in the steel industry through technological innovation, Ouyeel has developed into one of China's most influential industrial Internet platforms. In 2024, the company's annual transaction value exceeded RMB 500 billion.

The company has accumulated extensive expertise in AI and big data, holding more than 40 patents (including 12 invention patents) and over 180 software copyrights. Its AI technologies have been widely applied in intelligent manufacturing, industrial Internet platforms, smart customer service, and digital supply chain management. Multiple projects have been recognized as benchmark applications for industrial digital transformation at both municipal and national levels.

By continuously integrating industrial Internet technologies with artificial intelligence, Ouyeel is building a new-generation intelligent industrial service platform that promotes the digital, intelligent, and high-quality development of China's manufacturing and supply chain industries.

I. Background and Business Requirements

(I) Growing Challenges in Traditional B2B Models

The B2B industrial internet aims to achieve deep interconnection across multiple links in the industrial chain through platform-based integration, enabling efficient and dynamic balancing between supply and demand. However, in the field of bulk commodities such as steel, the industry is characterized by complex transaction chains, heterogeneous data sources, frequent price fluctuations, and high requirements for transaction timeliness. As a result, traditional manual processing methods are increasingly unable to meet business demands.

The industry currently faces several common pain points: Resource pricing acquisition still relies heavily on manual operations (phone calls, messaging apps, and platform listings), resulting in low efficiency and poor real-time responsiveness; Data formats are diverse and heterogeneous, making high-quality integration and rapid feedback difficult to achieve; Bulk commodity prices are highly volatile, requiring strong predictive analytics capabilities; As business scale expands, processes such as quotation inquiries, contract generation, and reconciliation become increasingly complex, and manual handling is prone to errors and inefficiency, negatively impacting customer experience and retention.

(II) Growing Value of Platform Data Assets

As an industrial internet platform, Zhaogang.com has accumulated a vast amount of structured, semi-structured, and unstructured data, totaling over 300TB, covering the entire business lifecycle, including quotation inquiries, transaction fulfillment, logistics, warehousing, contracts, and documentation. Leveraging its strong capabilities in large language models, machine learning, and computing infrastructure, the company has established a solid foundation for building a vertical industry large model and possesses multiple real-world, scenario-driven applications for implementation.

II. Implementation Plan

(I) Execution Strategy

This application is built upon the long-term data accumulation and AI R&D capabilities of Zhaogang.com in the steel industrial internet sector. Based on a general-purpose large model foundation, the system conducts NLP training and instruction fine-tuning using domain-specific data advantages in steel trading and B2B industrial materials. This enables the development of a vertical B2B industrial large model with strong capabilities in understanding and analyzing industrial material domain knowledge.

In addition, by integrating instant messaging tools such as QQ and WeChat, the system provides query-based and conversational interaction services for both customers and internal employees, including: real-time pricing information, intelligent quotation and inquiry handling, data services, and industry and enterprise knowledge services. The solution fully covers core business scenarios such as procurement, trading, fulfillment, and knowledge services, enabling end-to-end intelligent transformation and value creation across the entire workflow.

Figure 1 System Architecture Diagram

“AI Procurement Assistant” and “AI Trading Assistant” To address the inefficiency caused by sales and customer service personnel frequently switching between multiple platforms, the system leverages the natural language understanding capabilities of large models and knowledge graph technology to enable: intelligent parsing of customer requirements, automatic opportunity mining, and 24/7 automated quotation services. The system can automatically match inventory and generate orders, improving transaction efficiency by more than 10 times. In addition, through AI-based tagging during customer interactions, it builds structured user profiles, significantly enhancing customer repurchase rates and transaction activity.

At the same time, an “AI Order Fulfillment Assistant” is developed to support complex workflows such as contract generation, payment confirmation, and warehouse inbound processing. The system integrates OCR recognition, RPA automation, and a risk-control large model, enabling automated processing of business documents and operational workflows, thereby significantly improving end-to-end collaboration efficiency. In the knowledge service layer, the company has developed and widely deployed “Fatmao Xiaomi (胖猫小蜜)”, an internal intelligent assistant. By combining the DeepSeek large model with enterprise knowledge base construction, the system provides an intelligent Q&A and learning assistant for employees across the organization.

(II) Implementation Steps

In the early requirement analysis stage, the company conducted a comprehensive investigation to identify key application scenarios, including: lead acquisition, supply-demand matching, intelligent quotation, channel performance review, and customer lifecycle management. Based on these scenarios, evaluation benchmarks for system development were defined.

During the model development stage, after completing data cleaning and standardization, the R&D team initiated large model training and fine-tuning based on industry data, forming the underlying technical foundation for subsequent applications.

In the pilot (grey-scale) phase, high-frequency and high-demand scenarios such as quotation handling and contract recognition were selected for limited deployment testing. Through real-world validation, model performance was evaluated and algorithm parameters were optimized to ensure business applicability.

As the effects became increasingly evident, the system entered the full-scale deployment stage, covering all business links and forming an end-to-end business closed loop through coordinated operation of multiple AI agents. In the continuous iteration stage, the system is continuously optimized based on channel performance dashboards and operational metric feedback, ensuring stable and efficient performance under business expansion and market changes.

III. Application Highlights

Zhaogang.com has achieved multidimensional technological breakthroughs in the industrial internet sector, mainly in three areas: vertical large model customization, multimodal technology integration, and business scenario implementation adaptation.

1. Deep Customization of Vertical Large Models

To overcome the adaptation limitations of general-purpose large models in industrial scenarios, the system builds a domain-specific model based on a large model foundation. It integrates a steel industry-specific corpus, including: historical quotation data from over 12,000 suppliers, more than 20 million product attribute records, and over 5 years of transaction and fulfillment data. The training process adopts a two-stage approach:

(1) Domain-adaptive pretraining: A mixed corpus of general and industrial data is used to embed domain knowledge such as steel terminology, specifications (material, model, tolerance, etc.), and trading rules (credit terms, logistics clauses, etc.) into model parameters.

(2) Instruction fine-tuning: A high-quality dataset of over 5,000 annotated samples is constructed for key tasks such as quotation parsing, demand matching, and contract clause recognition. Using LoRA (Low-Rank Adaptation), the model improves its understanding accuracy of industrial requirements to over 90%, without modifying the core architecture of the large model.

2. Multimodal Technology Integration

To address the coexistence of multiple data formats in the steel trading industry—such as IM chat records, Excel quotations, PDF contracts, and logistics documents—the system builds an integrated technology stack combining: “AI large model + intelligent image segmentation + OCR + knowledge graph” This breaks the technical barriers between unstructured text, structured data, and business documents. The intelligent segmentation module identifies image structure features and splits complex document images into lightweight sub-images, reducing processing complexity.

The OCR module integrates text detection and character recognition, optimized for low-quality documents (blurred, tilted, etc.), achieving over 85% accuracy in key field extraction (e.g., shipping numbers, consignee information). Through RPA integration, extracted data is automatically synchronized into ERP systems, eliminating manual entry.

Meanwhile, the large model and industry knowledge graph work collaboratively. Leveraging relationships such as “product–specification–customer,” the system extracts key requirements (material, quantity, delivery location, etc.) from customer inquiries with over 90% accuracy.

As a result, intelligent quotation response time is reduced from 15 minutes to seconds, significantly improving multimodal data processing efficiency and business responsiveness.

3.Business Scenario Deployment via AI Agents

Through systematic construction of an AI Agent architecture, the system tightly integrates technology with business workflows. A total of 12 specialized AI modules are developed, each equipped with a perception–decision–execution closed loop. For example: The AI Procurement Agent monitors supplier messages in real time via QQ/WeChat, automatically triggering quotation parsing, inventory matching, and price updates. Other agents interconnect through data sharing mechanisms; for instance, after the Quotation Agent generates an order, it automatically synchronizes with the Warehouse Agent to trigger inbound preparation. This end-to-end automation resolves inefficiencies and data fragmentation caused by traditional cross-system manual operations.

IV. Implementation Results

Focusing on core functions of market sales and channel management, and leveraging long-term data accumulation and AI R&D capabilities in the steel industrial internet sector, Zhaogang.com has launched a series of intelligent applications, including: “AI Procurement Assistant”; “AI Trading Assistant” ; “Fatmao Cloud AI Manager”; “Fatmao Xiaomi”. These applications are built on a system of 12 AI Agents, fully covering procurement, trading, fulfillment, and knowledge service scenarios, enabling end-to-end intelligent transformation and value creation.

Figure 2 Systematic Architecture of AI Agents

In the procurement process, the company fully deployed the “AI Procurement Assistant” in 2024 to address the inefficiencies of traditional manual processing of supplier quotations, which is typically time-consuming and inefficient. Through intelligent parsing and automated consolidation, the system transforms fragmented supplier quotations into available sellable inventory. Combined with an automated price comparison algorithm and an AI real-time tracking mechanism, it enables rapid updates and precise management of quotation information.

The system is currently connected to more than 12,000 suppliers, capable of automatically processing and updating approximately 20 million product records per day, with over 10 million messages processed daily. It saves an average of about 1 hour per person per day in price consolidation time, while maintaining an overall quotation accuracy rate of over 95%.

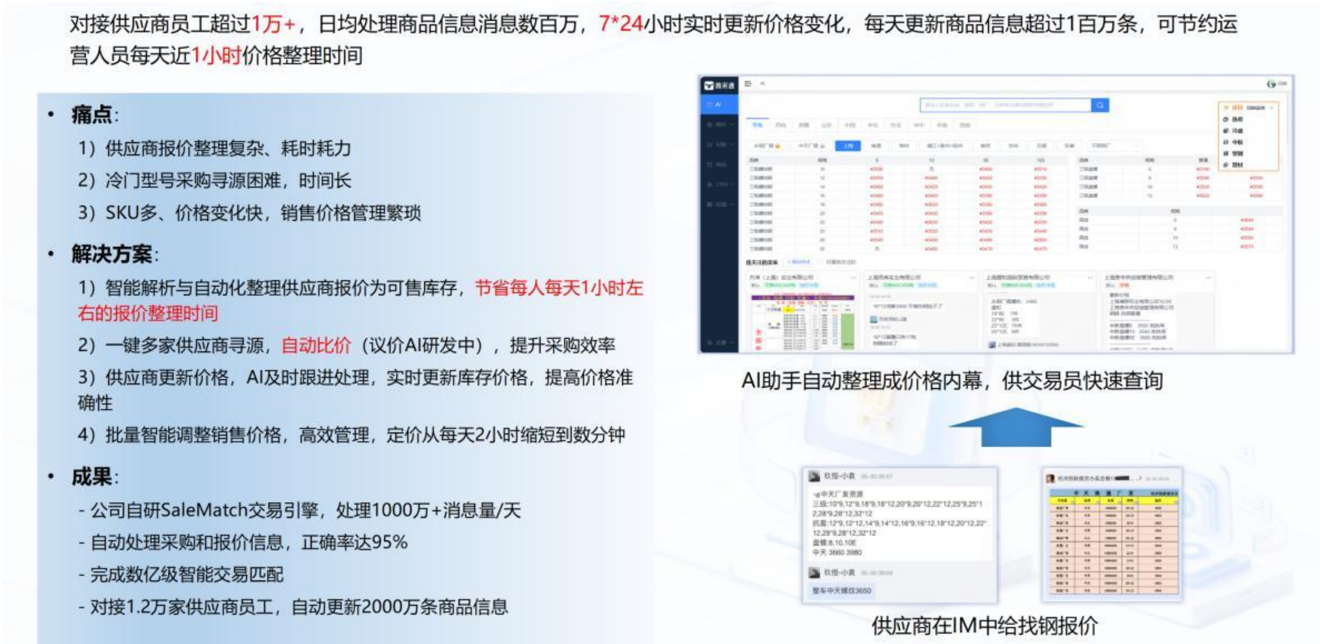


Figure 3 AI Procurement Assistant (Available Resource Management)

In the transaction process, the company promotes the “AI Trading Assistant” to platform users to address inefficiencies caused by sales and customer service staff frequently switching between multiple platforms, which results in low operational efficiency. By leveraging the natural language understanding capabilities of large language models and knowledge graph technology, the system enables: intelligent parsing of customer requirements, automatic opportunity mining, and 7×24-hour automated quotation services.

The system can automatically match inventory and generate orders, improving transaction efficiency by more than 10 times. In addition, through AI-based tagging during customer interactions, it builds structured user profiles, significantly enhancing customer repurchase rates and overall transaction activity.

Figure 4 AI Trading Assistant (Intelligent Trading Customer Service)

In the fulfillment process, the company has launched and scaled the “Fatmao Cloud AI Manager”, covering complex workflows such as contract generation, payment confirmation, logistics tracking, and warehouse inbound management.

The system leverages OCR recognition, RPA automation, and a risk-control large model to enable automated processing of business documents and operational tasks.

Implementation results show that the average processing time per order has been reduced by 7.03 minutes. The system has annually processed over 140,000 contracts and business documents, handled more than 4.365 million payment transactions, monitored over 400,000 logistics tracking events, covered more than 100 warehouse centers, and completed over 10,000 inbound confirmations, significantly improving end-to-end collaboration efficiency across the entire fulfillment chain.

Figure 5 AI Order Fulfillment Assistant

In the knowledge service domain, the company has launched “Fatmao Xiaomi (胖猫小蜜)”, which has been fully deployed across the organization for enterprise-wide use. By integrating the DeepSeek large model with the enterprise knowledge base system, the platform provides an intelligent Q&A and learning assistant. It can automatically convert documents and training materials into a continuously updated knowledge base, ensuring real-time knowledge accumulation and evolution.

The system offers employees on-demand access to an “AI master craftsman” service, enabling instant knowledge retrieval and guidance. With this tool, the competency level of new employees can be rapidly elevated within a few months to that of professionals with 3–5 years of industry experience, significantly shortening talent development cycles while establishing a sustainable internal knowledge management and reuse mechanism.

Overall, the implementation of the application not only covers the full lifecycle scenarios of enterprise market sales and channel management, but also delivers measurable economic and managerial benefits. The overall efficiency of procurement and sales processes has improved by more than 50% on average, while manual error rates have significantly decreased. Transaction accuracy remains consistently above 95%. The annual volume of processed information related to contracts, orders, payments, and logistics has exceeded ten million-level scale, substantially reducing labor costs and operational risks for the enterprise.

Meanwhile, the improvement of the knowledge management system and the AI-empowered employee development model have provided a new pathway for sustainable enterprise growth and industry enablement. At the same time, the implementation of this application fully demonstrates the strong empowering effect of the digital economy on the real economy. Leveraging industrial large models, multimodal technologies, and an AI Agent framework, Zhaogang.com has successfully achieved comprehensive digital and intelligent transformation across key industrial processes, including procurement, sales, logistics, warehousing, and customer service.

By transforming massive accumulated industrial data into intelligent decision-making factors, the system enables more efficient resource allocation, greater transparency in transactions, and smoother fulfillment coordination. This further reduces supply chain costs and significantly improves the operational efficiency of the industrial system.

Case 47

AI-Enabled Deep Integration for Intelligent Full-Life-Cycle Management of Hydrogen Fuel Cells

Shanghai Lianji Technology Co., Ltd.

I. Company Profile

Shanghai Lianji Technology Co., Ltd. is a high-tech enterprise incubated by Shanghai Automotive Industry Corporation (SAIC Group). Specializing in hydrogen fuel cell technologies, the company has established a comprehensive industrial layout covering the R&D, manufacturing, sales, and engineering services of fuel cell stacks, systems, and core components.

Committed to providing efficient, reliable, and clean new-energy solutions, Lianji's products are widely used in passenger vehicles, commercial vehicles, logistics and heavy-duty transportation, distributed power generation, energy storage, and hydrogen-powered engineering equipment.

As a national high-tech enterprise, Shanghai Lianji has become an important force driving the industrialization of hydrogen fuel cells and supporting China's transition toward clean energy.

II. Case Description

(I) Project Introduction

Against the backdrop of global energy transition and China's strategic goal of achieving "dual carbon" (carbon peaking and carbon neutrality), hydrogen energy and fuel cell technologies are regarded as critical components for building a clean, low-carbon, safe, and efficient modern energy system. However, the large-scale commercialization of fuel cell technologies faces multiple challenges, including long R&D cycles, high manufacturing costs, insufficient system durability, and complex operation and maintenance services.

At the core of these challenges lies the inherent complexity of fuel cell systems, which involve multidisciplinary coupling across electrochemistry, materials science, thermophysics, and fluid dynamics. The R&D, testing, and operational processes generate massive volumes of multi-source heterogeneous data. Traditional manual experience and fragmented data analysis methods are no longer sufficient to support rapid iteration and precise optimization.

As a leading enterprise in the industry, Shanghai JIE Hydrogen Technology Co., Ltd. has experienced the following systemic bottlenecks during its rapid business growth, which significantly constrain technological breakthroughs and market expansion:

1. Insufficient Intelligence in R&D, Limiting Iteration Speed

(1) Stack-level challenges (domestic material limitations)

The catalyst layer, known as the "heart" of fuel cells, has a micro-porous structure that directly determines reaction efficiency and lifespan. Traditional methods rely on manual interpretation of scanning electron microscope (SEM) images, which are not only extremely inefficient (over 30 minutes per sample), but also highly subjective with limited accuracy (below 80%), significantly slowing down the development of high-performance domestic catalyst materials.

(2) System-level challenges (short component lifespan)

Inside fuel cells, complex multi-physics coupling occurs among water, gas, heat, and electricity. Taking liquid water at the cathode outlet as an example, its erosion effect on high-speed rotating components (such as air compressors and expanders) is a key factor affecting system durability. However, traditional methods cannot accurately quantify droplet size, distribution, and velocity, resulting in insufficient data support for durability design and frequent erosion-related failures.

(3) Test bench-level challenges (large deviation from real-world conditions)

Energy management strategies for vehicles rely heavily on test bench conditions. However, traditional standard test cycles significantly deviate from real-world driving scenarios, leading to hydrogen consumption estimation errors exceeding 20%. As a result, optimized strategies often fail to deliver expected performance in real applications.

(4) Vehicle-level challenges (low diagnostic efficiency)

When vehicle failures occur, diagnosis heavily depends on expert experience. For typical faults such as “low single-cell voltage,” diagnostic analysis often takes more than 40 minutes, resulting in prolonged vehicle downtime and reduced operational efficiency and customer satisfaction.

2. Severe Data Silos and Low End-to-End Collaboration Efficiency

Lack of unified multi-source heterogeneous data governance. R&D data, bench test data, vehicle operation data, and after-sales service ticket data are scattered across different systems, departments, and even local files, with mixed structured and unstructured formats. The absence of a unified data governance and collaboration platform prevents data from flowing effectively across R&D and after-sales processes. As a result: After-sales intelligent support systems lack coordination, Customer feedback cannot be rapidly fed back into R&D optimization, A closed-loop system of “problem discovery → analysis → improvement → product upgrade” cannot be established.

(II) Scenario Requirements: Fuel Cell Full Lifecycle Intelligent Platform

To systematically address the above challenges, JIE Hydrogen independently developed and implemented an “AI + Hydrogen Deep Integration — Fuel Cell Full Lifecycle Intelligent Platform.” Rather than a single-point technological application, this project is a system-level engineering initiative driven by data and powered by AI, aiming to connect the entire value chain from microscopic material research to macroscopic fleet operation.

The platform integrates proprietary AI models and general-purpose large language models, and builds intelligent agent workflows based on a privately deployed Dify platform, forming two core sub-systems:

1. Domestic Fuel Cell Intelligent R&D Subsystem

This subsystem focuses on core technological breakthroughs, enabling intelligence and automation in:

material analysis; system simulation; operating condition construction; fault diagnosis.

2. Vehicle R&D and After-Sales Intelligent Collaboration Subsystem

This subsystem focuses on data value realization and cross-department collaboration by: building an enterprise knowledge base; deploying intelligent Q&A systems; improving after-sales response efficiency; and enabling feedback-driven R&D optimization

III. Project Highlights

The technological advancement of this project lies in the deep integration of multiple cutting-edge AI technologies with domain-specific fuel cell expertise, achieving several first-of-its-kind breakthroughs across multiple application scenarios.

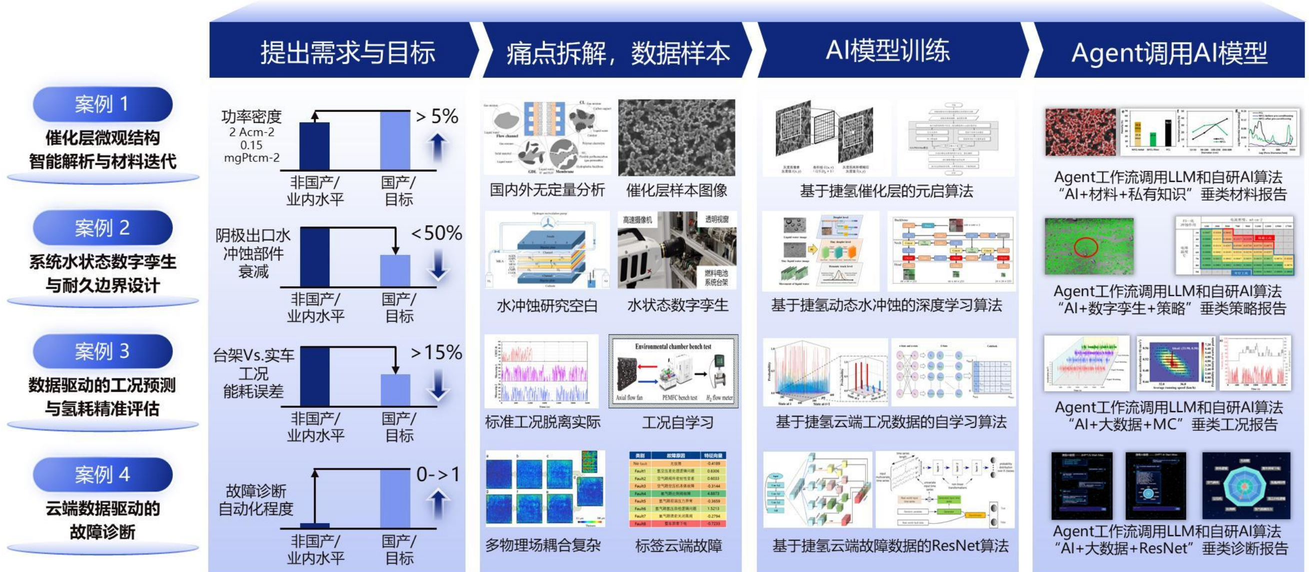


Figure 1 Domestic Fuel Cell Intelligent R&D Subsystem Technical Architecture (Cases 1–4)

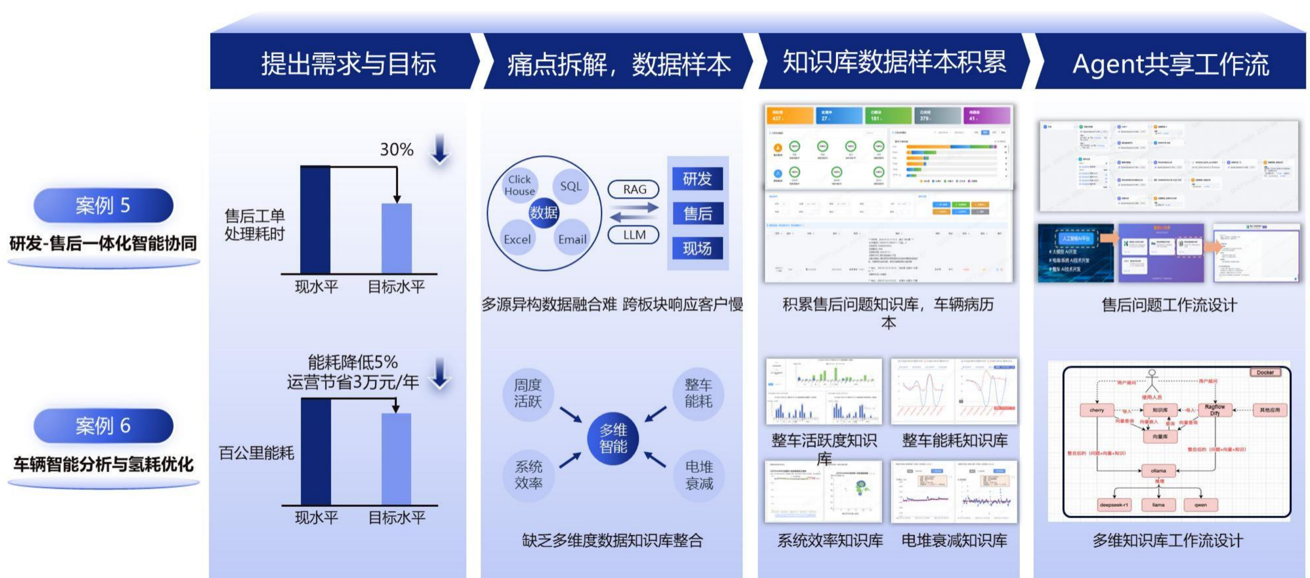


Figure 2 Vehicle R&D and After-Sales Intelligent Collaboration Subsystem Technical Architecture (Cases 5–6)

(I) Industry-First Innovations in Core Technologies

1. Industry-First AI Algorithm for Porosity Recognition in Fuel Cell Catalyst Layers

To address the global challenge of low-contrast and poorly distinguishable mesoporous structures in scanning electron microscopy (SEM) images of domestic catalyst layers, this work innovatively integrates Genetic Algorithms (GA), Particle Swarm Optimization (PSO), and Otsu thresholding methods to achieve high-precision adaptive quantification of porous microstructures.

The model achieves a porosity estimation error of less than 6%, reducing analysis time from 30 minutes to 30 seconds, with excellent global convergence and generalization capability. It can adaptively process SEM images under different material compositions and imaging conditions.

By integrating RAG (Retrieval-Augmented Generation) and LLM technologies, the system establishes correlations between material formulations and performance data, automatically generating optimization suggestions for key process parameters such as slurry composition.

This enables a true closed-loop AI-driven material development pipeline. The system has been successfully applied in the development of automotive-grade fuel cell catalyst layers, supporting the development of next-generation domestic electrospun catalyst materials. It improves membrane electrode assembly (MEA) output power by 9.4% while reducing cost by 24.6%. This is the first industrialized AI-based catalyst layer analysis system in China's fuel cell sector.

2. Industry-First AI Digital Twin Model for Liquid Water at Cathode Outlet in Fuel Cell Systems

This work constructs a novel digital twin framework integrating a YOLOv5 deep learning model enhanced with Convolutional Block Attention Modules (CBAM) and a Kalman filtering mechanism, enabling dynamic perception, tracking, and prediction of micro-droplets under high temperature, high humidity, and strong background noise conditions.

The system achieves an F1-score of 0.87–0.90, improves recall for droplets smaller than 300 μm by 36.5%, and reduces erosion effects by 72.8%. A durability boundary model covering 44 representative vehicle operating conditions has been established, providing a robust foundation for high-durability control strategy development. To date, zero erosion-related failures have been observed. This is the first AI system in China capable of dynamic liquid water perception and erosion quantification in cathode recirculation architectures.

3. Industry-First Data-Driven AI Driving Cycle Prediction and Hydrogen Consumption Evaluation Model

An improved two-dimensional Markov Chain Monte Carlo (MCMC) algorithm based on Metropolis–Hastings sampling is proposed to jointly encode vehicle speed and acceleration from real-world driving data, enabling probabilistic inference of realistic driving cycles.

The relative error of hydrogen consumption prediction is controlled within 1.20%–3.01%, while computational efficiency is improved by 12.6% compared with traditional Markov methods.

The system has been deployed to support economic evaluation and control strategy optimization for fuel cell buses in Shanghai's Minhang District.

4. Cloud-Based Data-Driven Fault Diagnosis System

To address the inefficiency of traditional expert-dependent and rule-based diagnostic systems, this work introduces a Residual Neural Network (ResNet)-based deep learning model for end-to-end feature extraction and intelligent pattern recognition from multi-dimensional time-series vehicle data.

After deployment, diagnostic time for typical faults such as "low single-cell voltage" is reduced from 40 minutes to under 5 minutes, improving efficiency by nearly 7×. This represents a breakthrough from manual diagnosis to AI-driven intelligent fault detection. The achievement received the Best Paper Presentation Award **at the Energy and AI International Conference in France.**

5. LLM-Based Integrated R&D and After-Sales Intelligent Collaboration System

A data middle platform is established to systematically consolidate after-sales issues, resolution workflows, and root cause analyses, forming a proprietary enterprise knowledge base.

By integrating general-purpose LLMs with RAG technology, an intelligent after-sales Q&A and decision support system is developed, featuring deep semantic understanding and reasoning capabilities for: rapid fault root cause analysis; automated solution generation; continuous knowledge evolution.

The average after-sales issue resolution time is reduced from 18 hours to 6 hours (−66.7%). From January to May 2025, after-sales costs decreased by 21.6% year-on-year, demonstrating significant cost optimization effects.

6. Intelligent Vehicle Analytics and Hydrogen Consumption Optimization

An AI agent workflow is designed to automatically analyze over 2,000 vehicles (including logistics vehicles, buses, tractors, heavy-duty trucks, freight trucks, and minibuses) within 3 hours, covering: energy consumption efficiency; fuel cell system performance; stack degradation analysis. A dynamically updated 4D vehicle knowledge base is constructed, enabling both longitudinal (week-over-week) and cross-vehicle performance comparisons.

By integrating domain expertise and historical strategy knowledge, the system automatically generates energy management optimization strategies.

Since deployment in 2024, vehicle energy consumption has decreased by 8.44%. Each vehicle saves approximately 27,000 CNY per year, resulting in total annual savings of approximately 54 million CNY for a fleet of 2,000 vehicles.

The system has been successfully deployed in major national demonstration projects, including China's Fuel Cell Vehicle Demonstration Program, providing a scalable and replicable intelligent solution for the industry.

IV. Implementation Results

1. Direct Cost Reduction

Material cost reduction: Replacement of imported materials with domestically developed catalyst layers reduces MEA cost per system by 24.6%.

After-sales cost reduction: From January to May 2025, after-sales costs decreased by21.6% year-on-year through predictive maintenance and intelligent diagnostics.

2. Operational Efficiency Improvement and Labor Cost Savings

R&D efficiency: Approximately 450 working hours per month saved across material analysis, component testing, bench testing, and fault diagnosis, equivalent to ~810,000 CNY annual labor savings.

After-sales efficiency: Average work order processing time reduced from 18 hours to 6 hours (3× improvement), saving ~576,000 CNY annually.

3. Customer Value Creation

Hydrogen consumption optimization: Fleet-wide hydrogen consumption reduced by 8.44%.

For a typical heavy-duty truck operating 100,000 km/year at 30 CNY/kg hydrogen, each vehicle saves ~27,000 CNY/year. For a fleet of 2,000 vehicles, total annual savings reach approximately 54 million CNY, significantly improving product competitiveness and customer satisfaction.

(II) Technological Innovation Achievements

The project has produced substantial scientific outcomes: 6 SCI/ESCI/EI indexed papers published; 2 international patent applications filed; Received the Best Paper Presentation Award at the Energy and AI International Conference (France); Awarded Shanghai Automotive Engineering Society Annual Paper Award; Included in the 60th Anniversary Special Issue of China SAE; Supported by the Shanghai Pujiang Talent Program. These achievements demonstrate significant influence across both academic and industrial communities.

子平台	案例	核心痛点	关键业务指标（应用前）	关键业务指标（应用后）	应用效果 业内领先度
国产燃料电池智能研发子平台	案例1 催化层微观结构智能解析与材料迭代	材料分析效率低、精度差、迭代慢	效率>30分钟/样本 精度<80%	耗时30秒/样本（效率提升60倍） 精度＞94%（精度提升18%）	支撑新一代国产催化材料开发，膜电极性能提升9.4%，国产材料降本24%。
	案例2 系统水状态数字孪生与耐久边界设计	系统内部状态不透明，关键部件耐久性差	冲蚀风险无法量化	冲蚀效应降低72.8%（0→1突破，衰减降低72.8%）	精准界定44类高耐久工况，为长寿命策略设计提供核心依据，目前冲蚀返修0例。
	案例3 数据驱动的工况预测与氢耗精准评估	台架工况与实际路况脱节，氢耗评估误差大	氢耗评估误差＞20%	氢耗预测误差≤±3%（重复实验减少75%）	使台架测试更贴合实际，每次耗时降低12小时，策略更精准，源头降低能耗。
	案例4 云端数据驱动的故障诊断	故障诊断高度依赖人工，效率低下	分析耗时＞40分钟/例	分析时间＜5分钟/例（0→1突破，效率提升8倍）	售后响应速度从“小时级”进入“分钟级”，保障客户运营，月节省256小时。
整车研发售后智能协同子平台	案例5 研发-售后一体化智能协同	数据孤岛，协同困难，工单处理效率低	18小时/工单	6小时/工单（耗时降低66.7%）	研发+售后闭环优化，2025年1月-5月售后成本同比降低21%。
	案例6 车辆智能分析与氢耗优化	整车性能提升缺乏数据指导，氢耗高	无自动化分析手段	年均氢耗降低8.44%（单车成本降低2.7万元/年/车）	提供并落实2000+辆车分析和优化措施，直接降低客户2.7万元/年/车运营成本。

- 6篇SCI/ESCI/EI论文
- 2篇国际专利
- 1项法国Energy and AI会议最佳报告奖
- 1次2024年世界人工智能大会特邀报告
- 1篇中国汽车工程学会60周年官微推文
- 1项上海市汽车工程学会年度论文奖项
- 1次2025年国际氢能展会捷氢代表报告
- 1次2024中国结构完整性联盟国际会议电池分会场主席及特邀报告

Figure 3 Comparison of Key Indicators Before and After Project Implementation

2. Driving Industry Technological Progress

This project has successfully overcome a series of “bottleneck” (chokepoint) technologies, including AI-based material microstructure analysis and system-level digital twin modeling, forming a proprietary AI R&D toolchain with independent intellectual property rights.

This achievement not only establishes JIE Hydrogen’s leading position in fuel cell technology, but also provides a replicable intelligent transformation paradigm for the broader domestic fuel cell industry, accelerating both domestic substitution and cost reduction & efficiency improvement across the sector.

(III) Social Value

1. Alignment with the 14th and 15th Five-Year Plans and Hydrogen Energy Strategy

By improving fuel cell performance, durability, and cost efficiency, this project strongly promotes the large-scale and commercial deployment of hydrogen fuel cell vehicles, providing critical technological support for deep decarbonization in the transportation sector.

2. Creation of Reusable Digital Assets for “Hydrogen + AI” Cross-Sector Applications

The AI algorithms, models, workflows, and enterprise knowledge base developed in this project constitute a valuable digital core asset system for JIE Hydrogen.

This integrated “AI + Industrial Systems” solution demonstrates strong cross-domain scalability and has already been successfully extended to multiple emerging application scenarios, including: hydrogen-powered UAVs (unmanned aerial vehicles); embodied intelligent robotics; zero-carbon industrial park logistics platforms; grid-side stationary hydrogen power generation systems. These applications are opening new growth pathways for the company and establishing a scalable hydrogen-energy-enabled intelligent industrial ecosystem.

Case 48

AI Agents and Domain Models Driving the Next-Generation Intelligent Cloud Operations System

Shanghai AnyHelp Network Technology Co., Ltd.

I. Company Profile

Shanghai AnyHelp Network Technology Co., Ltd. is a leading Chinese provider of cloud computing and artificial intelligence management solutions. The company employs over 700 professionals and helps enterprises and government organizations build intelligent cloud infrastructure, AI applications, and data governance capabilities.

Headquartered in Shanghai, the company also operates offices in Beijing, Shenzhen, Xi'an, Wuhan, Hong Kong, and other locations. It serves more than 3,000 enterprise customers across industries including telecommunications, manufacturing, healthcare, finance, and energy, maintaining long-term strategic partnerships with leading technology companies such as Huawei, Alibaba Cloud, Tencent Cloud, Microsoft, and AWS.

The company has continuously invested in cloud-native technologies and AI innovation, holding more than 200 software copyrights and over 100 invention patents. It has led or participated in the development of numerous national and industry standards related to cloud computing and intelligent operations. It has also received recognition from organizations including the China Academy of Information and Communications Technology (CAICT), Gartner, and IDC, and has been honored as a Shanghai High-Tech Enterprise, a "Specialized, Refined, Distinctive and Innovative" Enterprise, one of Shanghai's Top 100 Emerging Enterprises, one of Shanghai's Top 100 Service Enterprises, and one of Shanghai's Top 50 High-Growth Technology Enterprises.

II. Case Introduction

(I) Background Requirements

With the rapid advancement of artificial intelligence technologies, cloud computing has become a critical infrastructure supporting the digital economy. According to the Cloud Computing White Paper (2025)released by the China Academy of Information and Communications Technology (CAICT), the global cloud computing market reached USD 692.9 billion in 2024, representing a year-on-year growth of 20.2%. Multi-cloud deployment has now become the global mainstream architecture.

Against this backdrop, cloud operations and maintenance (Cloud O&M) has become the core capability ensuring stable enterprise IT system operation, directly affecting business continuity and service quality.

For example: A financial client's core trading system experienced a 3-minute outage due to delayed O&M response, resulting in direct economic losses exceeding RMB 8 million. A manufacturing enterprise suffered a 1-hour production line shutdown due to cloud resource scheduling failure, severely impacting order delivery. As a Cloud Managed Service Provider (Cloud MSP) with 18 years of industry experience, Anchain currently supports the secure and stable operations of more than 3,000 enterprise customers.

However, with continuous customer growth and expanding service scope, operational scenarios have become increasingly diversified. Customers differ significantly in organizational structure and technology stacks. Traditional O&M models—limited by manpower efficiency, experience dependency, and lack of standardization—can no longer meet large-scale operational demands.

In addition, general-purpose AI models demonstrate limited adaptability in O&M scenarios and cannot be directly applied. Therefore, technological innovation is required to balance efficiency, cost, and service quality in cloud operations.

(II) Case Implementation Overview

1. Project Construction Content

The project, titled "AI-Driven Intelligent Cloud Operations Engineering Practice", aims to build an intelligent O&M service management platform composed of domain-specific operational models and scenario-based intelligent agents.

Phase 1 focuses on addressing high-frequency, high-value pain points such as ticket processing, fault localization, and resource inspection: Fine-tune general large language models for the O&M domain; Develop a vertical model achieving >95% accuracy in ticket classification and knowledge retrieval.

Build a suite of intelligent agents including: Ticket assistant; Intelligent inspection system; Cloud resource query agent

Target outcomes: Automated ticket dispatch accuracy improved to >90%; Fault root cause analysis time reduced from 2 hours to within 30 minutes.

Meanwhile, a preliminary version of an SRE platform intelligence engine is constructed, integrating intelligent alert analysis capabilities.

Phase 2 focuses on capability enhancement and system scaling: Optimize vertical models using knowledge distillation techniques, improving performance while reducing cost.

Extend the platform intelligence engine to more complex scenarios such as: Cloud cost optimization; Security compliance management.

Target outcomes: Complex fault reasoning accuracy exceeding 90%; Engineer productivity improvement of over 50% for senior O&M personnel.

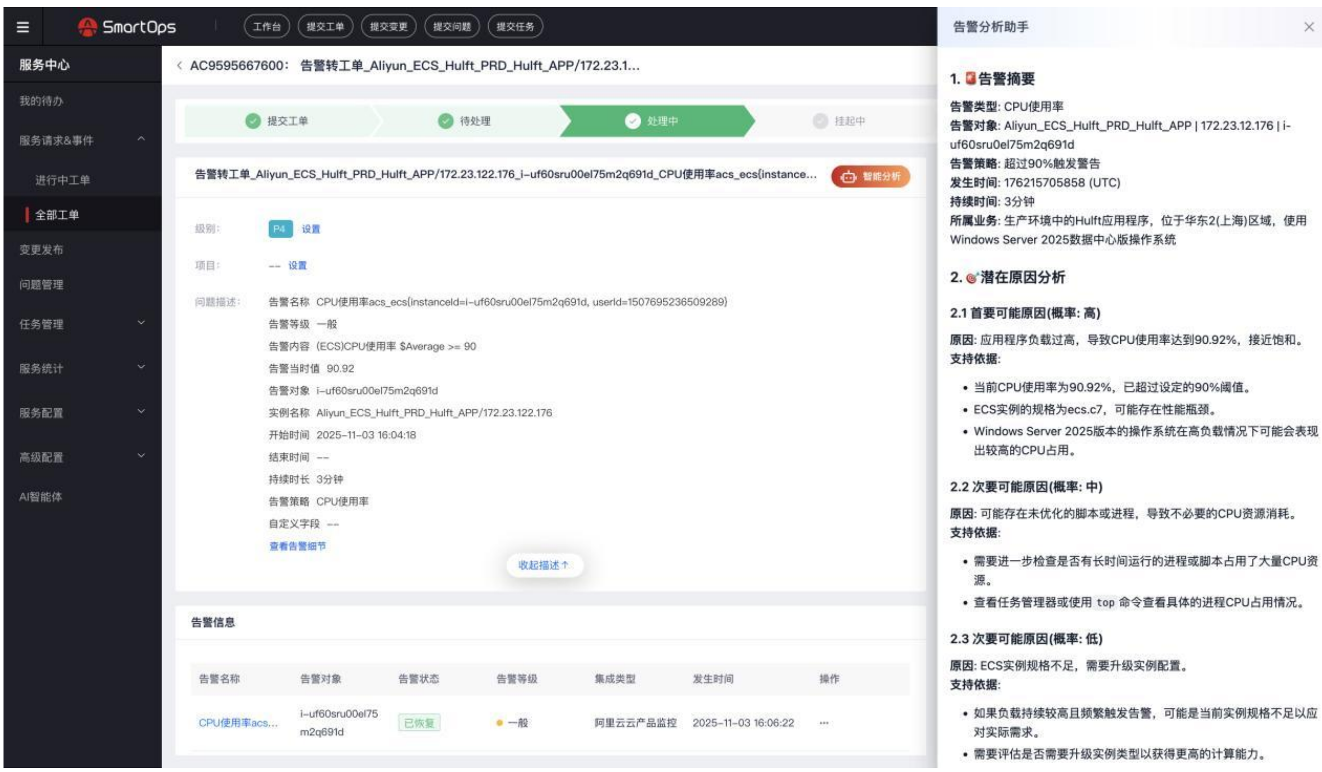


Figure 1 Alert Intelligent Agent: Helps Engineers Rapidly Identify Faults

2. System Architecture Overview

In terms of system architecture design, this project innovatively adopts a tri-layer integrated architecture of "O&M Intelligent Agents + Domain-Specific O&M Model + SRE Platform Brain", ensuring both system versatility and cross-component collaboration.

The bottom layer is the infrastructure and data layer, responsible for integrating monitoring data, logs, resource metrics from multi-cloud platforms, as well as internal knowledge bases. After rigorous data cleaning and governance, it provides high-quality "fuel" for upper-layer intelligence systems.

The core layer is the intelligent foundation layer, which consists of a domain-specific O&M large language model built through supervised fine-tuning and prompt optimization. Serving as the system's "professional knowledge engine", it deeply understands O&M terminology and contextual semantics, and plays a key role in improving decision accuracy across all intelligent applications while significantly reducing hallucination risks.

The execution layer is built on multi-agent frameworks such as AutoGen, and includes multiple specialized agents such as ticket-handling agents, inspection agents, and analysis agents. These agents operate both independently and collaboratively, acting as the system's "hands and feet" to interact with real-world environments and execute concrete O&M tasks.

The top layer is the central decision-making unit—the SRE Platform Brain, which functions as the system's "command center." It features end-to-end intelligence capabilities including perception, understanding, reasoning, decision-making, and self-learning. It continuously monitors system states, performs deep correlation analysis using domain models and knowledge graphs, and orchestrates downstream agents to execute tasks, thereby forming a complete closed-loop from perception to action. This decoupled layered architecture ensures high system flexibility, scalability, and continuous evolution capability.

3. Business Logic Overview

The system's business logic is tightly aligned with core O&M workflows and high-frequency operational scenarios, forming a data-driven, AI-empowered intelligent operation paradigm that enables closed-loop optimization.

Taking three core scenarios—incident handling, intelligent inspection, and fault prediction—as examples:

(1) Ticket Processing Workflow

After a user submits a ticket, the tagging agent immediately invokes the domain-specific model to perform semantic analysis and automatically generate precise labels. The workflow engine is then triggered to route the ticket intelligently or execute automated scripts for resolution. Engineers are supported by solution recommendations generated by the AI brain, enabling rapid problem resolution. The resolution process is fed back into the system as new knowledge, forming a continuous learning loop.

(2) Intelligent Inspection

The scheduling agent triggers inspection tasks according to predefined plans and coordinates multiple execution agents to collect state data from cloud resources and Kubernetes clusters in parallel. The analysis agent then performs anomaly detection and insight extraction, while the reporting agent generates standardized reports for proactive delivery. For critical anomalies, the system automatically creates high-priority tickets and routes them into the processing pipeline.

(3) Fault Prediction and Self-Healing Loop

At the core of the system is the SRE Platform Brain, which continuously perceives global system data. When multiple anomalies are detected, the system performs correlation analysis rather than issuing simple alerts, enabling rapid root cause identification. Based on predefined playbooks, it can automatically execute fault isolation or traffic switching, while simultaneously generating comprehensive reports including root cause, impact scope, and remediation recommendations to guide engineers in precise recovery.

This redesigned workflow significantly enhances operational efficiency and transforms O&M from reactive response to proactive prediction and autonomous healing.

Overall, the platform's core operational logic is driven by the SRE Platform Brain, which orchestrates more than ten specialized O&M intelligent agents through well-designed workflows, enabling rapid capability iteration and system expansion. Leveraging this core capability, the platform achieves efficient collaboration among user intent sensing, automated system execution, and human engineering operations.

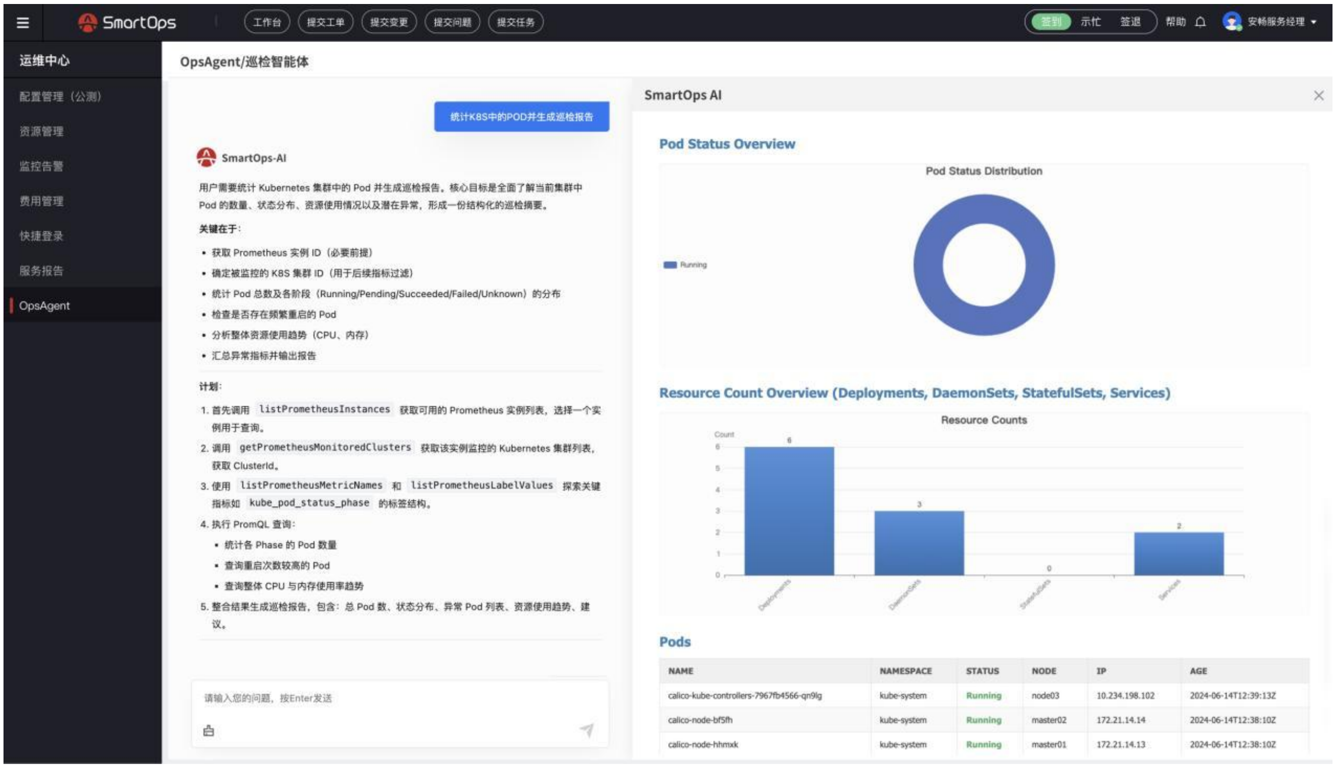


Figure 2 Inspection Intelligent Agent: Improves Inspection Frequency and Reduces Operational Risks

III. Project Highlights

(I) Key Technological Breakthroughs

1. Fine-tuning of a Domain-Specific O&M Model

The project selected Llama 3 (70B) and Qwen (13B) as base foundation models, and adopted a combined optimization strategy consisting of: SFT; Prompt Tuning; LoRA (Low-Rank Adaptation) lightweight fine-tuning.

The training dataset was constructed from internal operational data accumulated over the past three years, including: ~50,000 high-quality incident tickets; ~20,000 event records; ~5,000 knowledge base entries. The dataset was split into an 80:20 training-to-test ratio and trained on GPU servers. Training loss was stabilized below 0.05, effectively preventing overfitting.

After fine-tuning: Ticket and incident classification accuracy improved from 65% → 95% (≈ +30% improvement over mainstream general-purpose models such as DeepSeek). Knowledge retrieval matching accuracy improved from 60% → 93%

2. Multi-Agent Framework Construction

A multi-agent O&M system based on the AutoGen framework was developed, forming a structured “Scheduling – Execution – Analysis – Reporting” architecture.

Each agent plays a specialized role:

Scheduling Agent: allocates inspection and operational tasks

Execution Agents: collect multi-source data (cloud resources, Kubernetes, monitoring systems)

Analysis Agents: perform anomaly detection and root cause analysis

Reporting Agents: automatically generate standardized HTML/PDF reports and deliver results

In addition, an MCP client integration was implemented to enable seamless connectivity with cloud platforms, supporting continuous monitoring of: EC2 instances; S3 storage buckets; RDS databases; Kubernetes nodes, pods, and controllers.

3. SRE Platform Brain Development

- The system builds a full lifecycle intelligence capability: Perception → Understanding → Reasoning → Decision-making → Self-learning
- Key enhancements include: Integration of multiple AI models and rule-based business logic
- Knowledge update cycle shortened from weekly → daily
- Introduction of multimodal analysis and knowledge graphs
- Complex fault reasoning accuracy exceeding 90%
- Model deployment cost reduced by >30% via lightweight architecture and dynamic scheduling
- Hallucination rate controlled within 5%

(II) Scenario Uniqueness

1. Industry Adaptability

This solution integrates SRE engineering with AI, representing the first deep "AI + SRE" operational practice in China's cloud MSP industry. It supports multi-cloud environments and serves clients across industries such as finance, manufacturing, and retail.

2. Full Role Coverage

The system provides role-specific intelligent agents for different user types, including: Service desk operators; DevOps engineers; FinOps engineers. Key applications include: Ticket assistants; Inspection assistants; FinOps reporting agents. The system supports both natural language interaction and rule-based automation.

3. Practical Deployment Value

The platform focuses on high-frequency operational pain points such as: Ticket handling; Incident response

; System inspection; Resource query.

All functions have been validated in both internal testing and production environments, ensuring strong adaptability across different MSP clients and operational scales. It is directly reusable for similar Cloud MSP organizations.

IV. Implementation Outcomes

After deployment, the platform significantly improves operational efficiency and personnel productivity at Anchain. It reduces model inference costs, increases proactive service coverage, and enhances customer experience, generating an estimated direct economic benefit of approximately RMB 1.5 million.

(I) Productivity Improvement

- Ticket Handling: AI-based tagging agents and automated workflows significantly improve ticket classification and dispatch accuracy, reducing average response time by nearly two-thirds.
- Fault Diagnosis: By integrating monitoring, logs, and resource data, root cause analysis time is reduced to 45 minutes, with significantly improved knowledge reuse.
- Intelligent Inspection: Daily inspections increased to 3 cycles per day, with: 85% improvement in report generation efficiency; 95% coverage of optimization recommendations
- Cloud Query Efficiency: Natural language-based query system significantly reduces retrieval time.
- FinOps Support: Domain models assist senior engineers in generating cost optimization reports, reducing workload by approximately 30%.

(II) Economic Benefits

- O&M staffing requirements reduced by 60%
- Onboarding time reduced from 3 months → 1.5 months
- Improved reuse efficiency of senior engineers
- Estimated annual labor cost savings of ~RMB 1.5 million
- 70% reduction in large-scale model invocation costs through vertical fine-tuning

- Reduced customer complaints and improved service response time
- Increased customer retention rate by 5%
- Reduced system downtime and operational losses
- Potential future revenue via external commercialization of the domain model

(III) Social Value

- Provides a replicable transformation benchmark for AI-driven MSP operations
- Accelerates shift from manual operations → AI + automation
- Shortens training cycles and improves workforce capability in AI-enabled O&M
- Enhances resilience of critical industries (finance, manufacturing, energy)
- Supports stable digital infrastructure for the broader digital economy

Case 49

End-to-End Intelligent Financial Digital Management Based on Large Language Models, RPA, and OCR

Tiannong Digital Life Technology Co., Ltd.

I. Company Profile

Tiannong Digital Life Technology Co., Ltd. is a specialized technology enterprise established by China Telecom Tianyi Digital Life Technology Group to develop digital lifestyle services. Leveraging China Telecom's extensive ecosystem, the company focuses on smart communities, digital home services, cloud-based business platforms, and AI-enabled digital transformation.

Recognized as a National High-Tech Enterprise and a Shanghai Enterprise Technology Center, the company has applied for more than 600 patents, participated in the formulation of several national standards, and obtained multiple international certifications, including CNAS accreditation and CMMI Level 5 certification.

Headquartered in Shanghai with branches in Guangdong, Jiangsu, and Beijing, the company employs nearly one thousand staff members, with an average employee age of 36. Technical personnel account for approximately 53% of the workforce, while employees holding bachelor's degrees or above represent more than 90%.

Guided by the vision of "Serving Every Household", the company has built an integrated digital ecosystem covering cloud services, smart communities, and AI-powered digital lifestyles. Through artificial intelligence, big data, cloud computing, and Internet technologies, it provides comprehensive digital services for government agencies, enterprises, and consumers, contributing to the digital transformation of both urban and rural communities.

II. Case Overview

(I) Business Requirements

With rapid business expansion and increasingly stringent regulatory supervision, the company's existing financial processes rely heavily on manual invoice review, Excel-based ledgers, and multiple specialized MSS systems within China Telecom Group. This has led to issues such as inconsistent data standards, delayed financial closing, and insufficient real-time risk visibility.

To implement China Telecom Group's strategic initiative of "Cloudification, Digital Transformation, and Intelligent Empowerment (云改数转智惠)", and to address the core requirement of integrating finance and business data as well as achieving "daily clearing and monthly closing" with comprehensive risk prevention, the company built an end-to-end intelligent financial digital management system based on Tianyi Cloud. In this system:

- AI capabilities serve as the "brain"
- RPA (Robotic Process Automation) serves as the "limbs"
- OCR (Optical Character Recognition) serves as the "sensing layer"

Together, they form an integrated intelligent financial system that addresses high-frequency and critical financial pain points:

1. Data Silos

Multiple business and financial systems (CRM, asset management systems, payment systems, billing systems, and financial management platforms) were built independently, resulting in severe data silos.

- Data extraction requires direct access to source tables, with complex cleaning processes and low development efficiency.
- Financial and business data (orders, receipts, payments, assets, costs, etc.) are scattered across multiple systems, making analysis repetitive and non-reusable.
- Unified data modeling across systems is not possible, limiting rapid reporting and ad-hoc analysis.

2. Low Efficiency in Manual Review

Financial operations such as reimbursement review, bank statement processing, travel expense verification, and revenue recognition require reviewing nearly 10,000 documents per month.

- Over 200 mapping relationships and 20 audit checkpoints must be verified.
- Processes require switching across multiple systems without full API integration.
- Manual review is prone to omissions, delays, and inconsistencies, consuming significant labor resources.

3. Contract Processing Efficiency and Risk Issues

With business expansion, contract volume has grown exponentially.

- Manual extraction and entry of key contract information (amount, counterpart entity, payment terms, etc.) takes about 30 minutes per contract.
- High repetitive workload consumes large amounts of manpower.
- Human errors (e.g., missing penalty clauses, incorrect contract amounts, inconsistent invoice data) introduce significant compliance and financial risks.

4. Difficulty in Policy and Process Retrieval

Traditional financial support relies heavily on manual consultation of policy documents and operation manuals.

Employees often face difficulties such as:

- "Cannot find information"
- "Cannot get clear answers"
- "Cannot understand procedures"

This significantly reduces operational efficiency.

5. Lag in Financial Data Analytics

Financial analysis traditionally depends on SQL queries or complex scripting written by specialized analysts.

- Response time is slow
- Decision-making efficiency is low
- Unable to support real-time business insights

(II) Project Objectives

This project aims to build a smart financial system with:

- A unified business-finance data foundation (data lakehouse)
- AI-powered intelligent applications as the core engine

The overall objective is: Integrate all core business systems into a unified data foundation, build standardized data services, and develop reusable data assets; on this basis, deploy four major AI scenarios at scale to achieve significant improvements in financial efficiency, risk control, and data intelligence capabilities.

Four Key AI Application Goals

1. AI + RPA Automation

- Automate at least four financial processes (e.g., initial invoice review)
- Improve efficiency by 70%+
- Achieve 95%+ accuracy in document processing

2. Intelligent Contract Parsing

- Reduce key contract extraction time from hours → minutes
- Achieve ≥95% accuracy for key fields (amount, entities, bank accounts, etc.)

3. Intelligent Q&A for Policies and Processes

- Reduce user dissatisfaction ("dislike" rate) below 5%

- Provide real-time and accurate procedural guidance

4. Self-Service Financial Analytics (Ask Data)

- Cover five domains: contracts, cooperation, revenue, accounting, and reimbursement
- Achieve ≥95% query accuracy
- Enable business users to perform data analysis without SQL

(III) Implementation Plan

1. Overall Architecture

Based on China Telecom Group's M-domain systems and the company's AI and IT capabilities, a smart financial-business integration system is built.

(1) Infrastructure Layer (Foundation Layer)

- Built on the company's operations and security infrastructure
- Establishes an IaaS–PaaS integrated monitoring system
- Leverages AI platform + data platform capabilities to:

  Unify financial data elements

  Enable secondary development of AI applications

(2) Core Capability Layer

- Develops two major capability systems:

  Operational digital capabilities

  Management digital capabilities

  Integrates with China Telecom M-domain systems to supplement financial data and capabilities

(3) Capability Exposure Layer

- Provides standardized interfaces for cross-platform and cross-module capability access
- Ensures unified service exposure across systems

(4) Application Layer

- Builds applications across three major categories:

  Production scenarios

  Operational scenarios

  Intelligent financial scenarios

- Focuses on AI-enabled capabilities such as:

  Intelligent business processing

  Policy and regulation Q&A

  Intelligent data querying

  Rule-based intelligent parsing

Figure 1 Overall Technical Architecture Diagram

2. AI + RPA Digital Employee Technical Solution

This solution builds a low-code automation platform based on a four-layer architecture, enabling a flexible and composable “digital employee” system.

(1) Presentation Layer

- Provides an Electron-based visual client
- Supports intuitive user interaction and workflow configuration
- Enables business users to design automation processes without coding

(2) Core Service Layer

- Integrates a workflow designer based on SVG process canvas
- Includes DAG (Directed Acyclic Graph) validation mechanism to ensure process correctness
- Incorporates intelligent scheduling algorithms to optimize execution efficiency and resource allocation

(3) Backend Support Layer

Contains:

- Process parsing engine
- Code generation engine
- Supports natural language → workflow model conversion
- Enables automatic generation of executable automation logic from business descriptions

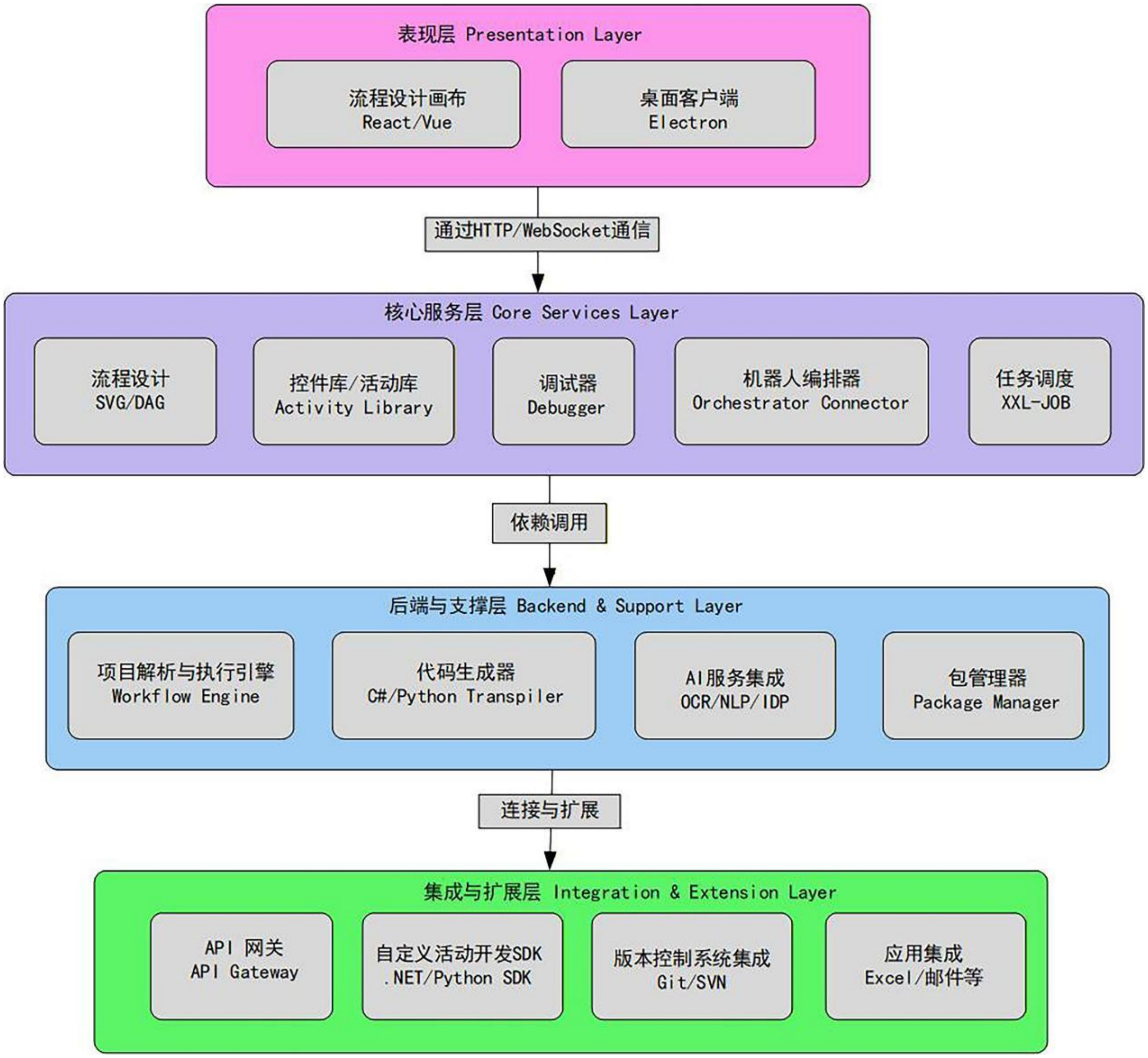


Figure 2 RPA Technical Architecture Diagram

3. OCR Contract Parsing and Settlement Rule Extraction Technical Solution

This solution builds an integrated intelligent parsing pipeline based on a unified "Perception–Cognition–Verification" architecture. The perception layer adopts a dynamic routing strategy, using a high-precision OCR engine for scanned documents and a PDF parsing library for digital files to ensure robust and accurate text extraction. The cognition layer deploys a Qwen-30B large language model, combined with carefully designed prompt engineering, to accurately extract key structured information such as contract amount, contracting parties, and payment terms. The verification layer introduces an innovative three-stage mechanism combining confidence scoring, rule-based validation, and human review to ensure output correctness and logical consistency. Finally, all results are standardized and exposed via API interfaces, providing reliable structured data support for downstream financial and settlement systems.

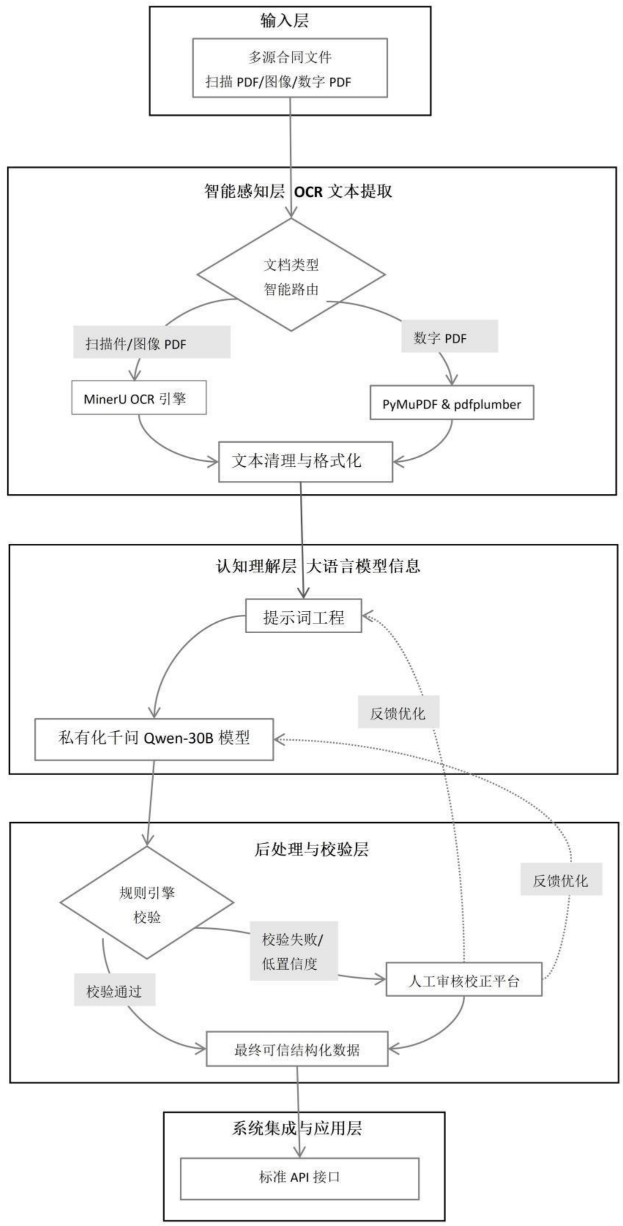


Figure 3 Contract Parsing and Rule-Based Architecture Diagram

4. Intelligent Q&A Technical Solution

This solution builds an intelligent financial Q&A system based on a RAG architecture. The data layer processes multimodal financial documents through document parsing tools, enabling intelligent chunking and vector-based representation. The retrieval layer adopts a multi-channel strategy combining semantic vector search and keyword matching, with a reranking model to select the most relevant content. The generation layer deploys a locally hosted Qwen-30B model, constrained by prompt engineering to ensure accurate and reliable responses. The application layer embeds the assistant into business systems and establishes a user feedback loop for continuous optimization, providing employees with real-time and precise financial policy and process guidance.

5. Intelligent Data Query (Text-to-Query) Technical Solution

This solution constructs an end-to-end intelligent data query and analysis system. A fine-tuned large language model is used to understand user intent and perform precise semantic parsing with the support of a financial knowledge base. The core adopts Text-to-DSL technology, converting natural language into structured query language to ensure high query accuracy. A distributed query engine is integrated to guarantee efficient data retrieval performance. In the reporting stage, RAG is used to retrieve analytical templates from the knowledge base, guiding the large model to generate professional data analysis reports. This forms a complete pipeline from natural language queries to actionable data insights, significantly lowering the barrier to data usage and empowering business users with self-service analytics.

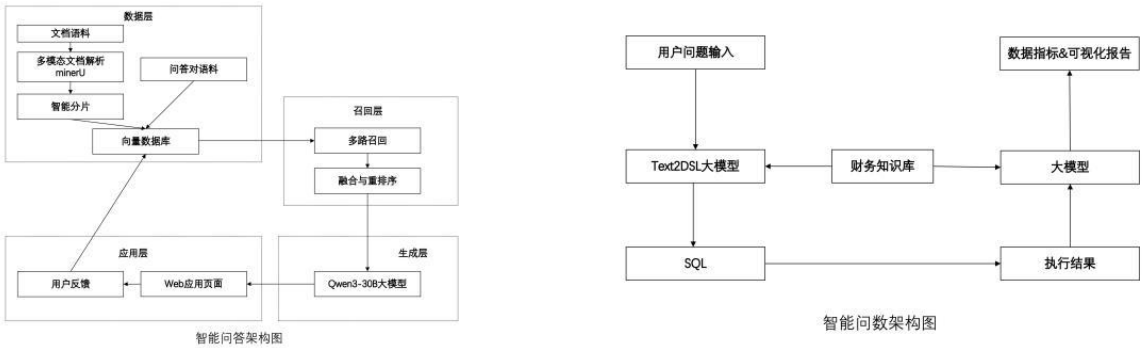


Figure 4 Intelligent Q&A and Intelligent Data Query Architecture Diagram

III. Project Highlights

(1) Key Technological Breakthroughs

1. Contract Parsing with a Dual "Cognition–Verification" Closed Loop

This project innovatively builds a "confidence–rule–human" three-level verification mechanism on top of the LLM cognition layer, enabling the system to perform structured reasoning and risk self-checking similar to financial experts. This achieves a fundamental leap from simple information extraction to intelligent contract review, significantly improving accuracy and compliance awareness.

2. AI-Assisted Intelligent Workflow Orchestration

By integrating RPA with AI agents, the system transforms workflow automation from fixed-script execution to intelligent collaborative orchestration. AI is used to generate workflow skeletons, while computer vision and NLP technologies automatically interpret and model business processes. As a result, workflow design efficiency is improved from weeks/months to days or even hours.

3. Hybrid Retrieval and Intelligent Reranking Mechanism

A novel "multi-channel retrieval + reranking" architecture is introduced, combining semantic vector search with keyword-based retrieval. This effectively overcomes the limitations of single retrieval methods, while the reranking model ensures higher precision, achieving a qualitative shift from "retrieving results" to "retrieving accurate results."

(2) Core Highlights

1. Deep Integration of Multiple AI Technologies to Address Core Financial Pain Points

The project breaks the limitations of isolated AI applications by constructing an end-to-end intelligent pipeline integrating AI + RPA + OCR, unifying perception, decision-making, execution, and validation. In automation scenarios, AI-assisted workflow generation combined with human configuration enables "digital workers," significantly improving workflow efficiency. In contract parsing scenarios, high-precision OCR combined with the "cognition–verification" closed loop ensures accuracy and logical consistency. In Q&A and analytics scenarios, a dual engine of RAG + Text2DSL, enhanced with user feedback mechanisms, continuously improves accuracy.

2. Full-stack Domestic Cloud-native Infrastructure for Security and Reliability

Built on Tianyi Cloud IaaS–PaaS infrastructure and integrated with domestic Qwen large models, the system ensures full autonomy and controllability. This architecture eliminates risks related to cross-border data flow and external computing dependency, while providing high availability, scalability, and compliance-grade security for critical financial operations.

3. Data Assetization and Low-code Enablement for Agile Delivery

An integrated data middle platform standardizes cleaned financial data into reusable API services, enabling "ready-to-use" data assets. Meanwhile, AI capabilities such as OCR, RPA, and LLM are modularized into components and deployed through low-code/no-code drag-and-drop configuration. This reduces deployment cycles from 3 months to about 4 weeks, enabling rapid reuse and agile innovation.

IV. Implementation Results

(1) Cost Reduction and Efficiency Improvement

- Audit efficiency improvement: Average processing time per document reduced from 5 hours to 0.6 hours (88% improvement), with ~10,000 documents processed monthly at 100% accuracy.
- Contract parsing efficiency: Key information extraction reduced from 30 minutes to under 5 minutes per contract (>75% improvement), supporting thousands of contracts annually.
- Q&A and analytics efficiency: Financial policy inquiries reduced from hours to seconds; ad-hoc data analysis reduced from days to minutes.
- Labor cost savings: RPA digital workers and automation systems save approximately 1.25 FTE annually, plus 1–2 data analysts' workload through intelligent query systems.

(2) Social Value

- Promotes digital transformation of financial management through a replicable "LLM + RPA + OCR" integrated framework, supporting national digital economy strategy.
- Enhances compliance and risk control via automated validation and intelligent anomaly detection (e.g., contract inconsistency, financial mismatches).
- Frees employees from repetitive tasks, enabling transition toward high-value financial analysis and risk management roles.

(3) Scalable Replication Framewor

The project summarizes a standardized four-step replication methodology:

1. Organizational governance first–Establish a triad team (Finance + IT + AI) to ensure alignment between business needs and model development.

2. Systematic data assetization–Unify financial data into domain models and enable rapid integration via ETL pipelines.

3. Low-code application enablement–Package OCR, RPA, and LLM capabilities into visual components for rapid assembly and deployment.

4. Continuous optimization loop–Build a feedback-driven system where user behavior and manual review continuously improve model performance and system accuracy.

Case 50

AI-Ready Platform for the Real Estate Industry Driven by Vehicle-Road Coordination and Multi-Agent Intelligence

Shanghai Shenyuan Intelligent Technology Co., Ltd.

I. Company Profile

Shanghai Shenyuan Intelligent Technology Co., Ltd. (“Shenyuan Intelligent”) is an AI technology enterprise specializing in digital transformation solutions for the real estate industry. Leveraging artificial intelligence technologies, the company empowers digital upgrading and the cultivation of new quality productive forces across the sector.

As an industry technology pioneer, Shenyuan Intelligent has developed a comprehensive AI technology framework built upon large-scale data and large language models (LLMs), covering land acquisition, real estate development, property management, and downstream business ecosystems.

The company has assembled a multidisciplinary R&D team comprising experts in AI, data analytics, and industry specialists with extensive experience in real estate. This enables the company to address key challenges associated with deploying large AI models in highly specialized industries, including domain adaptation, data security, and deep integration into business processes.

Based on independently developed “AI Primitive Layer” and extensive industry-specific knowledge repositories, Shenyuan Intelligent has established the AI-Ready platform, providing comprehensive AI-enabled solutions for real estate enterprises. Its products are now serving numerous leading developers and property management companies, becoming an important driver of intelligent productivity across China’s real estate industry.

II. Case Introduction

(1) Background Requirements

Under the national strategic guidance of the “AI+ Initiative” and the cultivation of new quality productive forces, the real estate industry—one of the pillars of the national economy—is entering a critical transformation phase from traditional growth models toward digital and intelligent development.

In particular, the Opinions of the CPC Central Committee and the State Council on Promoting High-Quality Urban Development issued on August 28, 2025, explicitly emphasize advancing smart and fine-grained urban governance and accelerating the construction of “good housing” that is safe, comfortable, green, and intelligent. This provides a clear direction for deep technology integration in the industry.

Market data further confirms this trend. According to CRIC research, 92% of leading real estate enterprises have prioritized large AI models as key frontier technologies, indicating that building an AI-native paradigm—which comprehensively reconstructs business processes, service delivery, and business models—has become an inevitable path to break through industry bottlenecks and align with national high-quality development goals.

Despite the growing consensus on digital transformation, structural challenges in the traditional real estate service system continue to constrain productivity release. The industry currently faces a dual challenge of data silos and professional barriers: massive heterogeneous data scattered across systems prevents unified analytics, while complex legal and financial expertise required in transactions leads to inconsistent service quality and steep onboarding difficulty.

Meanwhile, excessive reliance on manual data collection and report writing results in low efficiency, failing to meet the demands of the stock-driven market for high-frequency and high-precision services.

Against this backdrop, the core requirement of this project is to leverage AI to break information barriers, build a unified industry knowledge base and data infrastructure, and significantly improve the efficiency of housing analysis, market forecasting, and content generation through automation and intelligence—ultimately shifting from a labor-driven model to a computation-driven paradigm, empowering practitioners with expert-level capabilities.

(2) Implementation Plan Overview

The core concept of this project is “AI Native”, meaning AI is not merely an auxiliary tool but the foundation for reconstructing the entire business workflow and system architecture. The implementation follows a dual-layer structure:

- Four Core Infrastructure Foundations (Bottom Layer)
- Eight Industry Application Scenarios (Top Layer)
- Together, they form a tightly coupled, scalable, and intelligent ecosystem.

1. System Architecture and Four Core Foundations

The technical foundation consists of four integrated pillars: Database, Knowledge Base, Expert Base, and Engineering Capability Base. These serve as the “heart” and “brain” of the entire AI system.

(1) Database Foundation

The database layer serves as the system’s core infrastructure. Through a combination of manual research and automated data collection, the project deeply integrates CRIC’s 20-year proprietary real estate dataset, including:

- New housing project data
- Secondary housing transaction data
- Land transaction data
- Granular sub-district market analytics

By establishing unified data governance standards, the system performs multi-source data cleaning, integration, and standardization, eliminating information silos and forming a high-quality, fine-grained real estate data warehouse, which provides reliable “fuel” for upper-layer AI models.

(2) Knowledge Base Foundation

The knowledge base endows the AI system with industry cognition.

The team processes large volumes of unstructured data such as:

- Real estate policies and regulations
- Industry news and updates
- Professional research reports

These are cleaned, annotated, and transformed into vector embeddings stored in a domain-specific vector database.

The system supports hour-level and even minute-level updates, ensuring real-time synchronization with market dynamics and effectively solving the knowledge lag problem of general-purpose large models in vertical industries.

(3) Expert Base Foundation

The expert base represents the key differentiator of the project.

It focuses on externalizing and algorithmically modeling the implicit expertise of senior real estate analysts. By analyzing expert behavioral logs and conducting structured interviews, the system extracts:

- Market trend forecasting logic
- Investment valuation reasoning frameworks
- Analytical decision-making pathways

These are transformed into computable reasoning models, enabling the system to simulate expert-level thinking and generate insightful analytical reports rather than simple data aggregation.

(4) Engineering Capability Foundation

The engineering layer ensures system stability and scalability.

Built on a cloud-native architecture, it supports:

- High concurrency processing
- Elastic scaling

- Large model orchestration and scheduling
- Asynchronous task execution
- Edge computing support

It also integrates advanced capabilities such as digital human generation, image understanding, and multimodal content generation, providing rich interaction capabilities for upper-layer applications.

2. Business Logic and Application Ecosystem

Based on the four foundational layers, the project develops eight core AI application covering both B-end enterprises and C-end users, forming a full-chain real estate intelligent product ecosystem.

For professional institutions, the “CRIC2025” platform provides advanced capabilities such as:

- Automated monthly report generation
- Project evaluation systems
- Policy interpretation and research support

In terms of business logic, the system supports complex “data Q&A” and “industry knowledge Q&A”. Users can query multi-dimensional market data through natural language, and the system automatically generates:

- Analytical charts
- Structured insights
- Professional research reports

This enables a seamless transition from data retrieval to decision-grade intelligence.

Figure 1 CRIC2025 “Data Query & Automatic Interpretation” Scenario Example

For the sales scenario, the “AI Real Estate Sales Champion” is deeply embedded into the daily workflow of property consultants and real estate agents. By inputting community or unit-type parameters, the system automatically calls backend valuation models and competitor property databases to generate in real time a one-house-one-price list, comparative analysis of competing listings, and personalized marketing copy (such as Xiaohongshu-style posts and WeChat Moments promotional texts), significantly improving both customer acquisition efficiency and conversion rates.

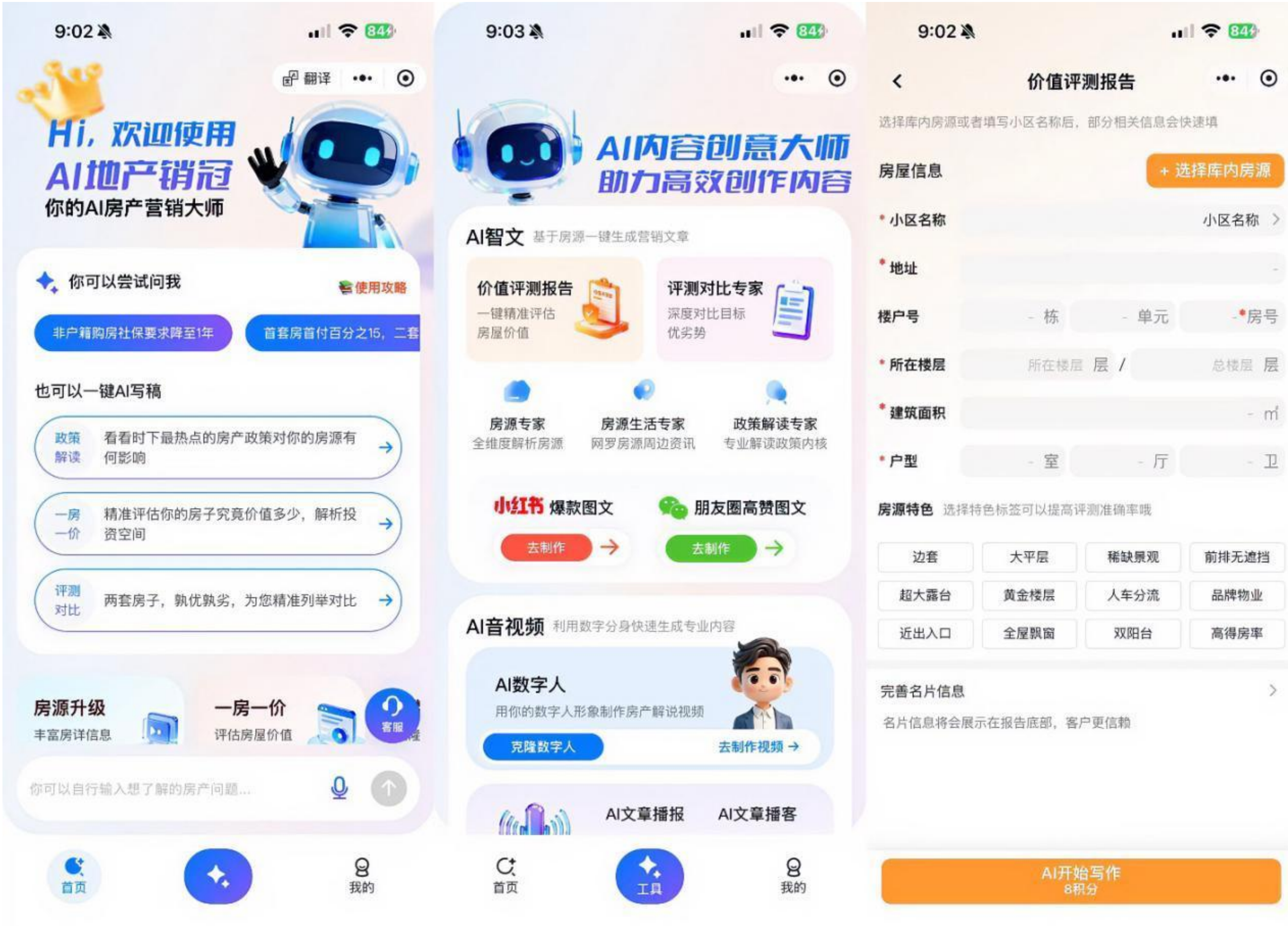


Figure 2 “AI Real Estate Sales Champion” Example of Property Valuation and Marketing Copy Generation

Additionally, for specific niche domains, the project has innovatively developed “Silver Age Digital Intelligence (银发数智)”, focusing on aging society research and the elderly industry chain, filling a market gap.

III. Project Highlights

This project has achieved significant breakthroughs in both technological innovation and scenario implementation. Its core highlight lies in the construction of a highly specialized real estate vertical cognitive system. Unlike general-purpose AI products on the market, the model in this project has been “immersed” in tens of billions of industry-specific corpora, forming a comprehensive real estate domain knowledge graph. As a result, the system not only “understands language” but also “understands business.” This deep vertical integration marks a shift of AI from a simple “tool attribute” toward an “expert attribute” with strong industry expertise.

One of the key technological breakthroughs is the intelligent multi-model routing and scheduling mechanism. For user requests of different complexity levels, the system does not simply rely on the most expensive large-scale model. Instead, it establishes an intelligent routing algorithm that dynamically selects the optimal cost-performance model—SLM (small model), MLM (medium model), or PLM (professional large model)—based on factors such as prompt length, numerical computation complexity, and whether extensive knowledge base retrieval is required. This mechanism maintains over 95% accuracy while significantly reducing inference costs, achieving a precise balance between performance and cost efficiency.

Another major highlight is the industry-specific vector engine and knowledge base optimization. General-purpose embedding models often struggle to accurately understand real estate terminology (e.g., the deeper meanings of “floor area ratio” and “unit efficiency”). This project adopts a domain-tuned embedding model trained on massive industry corpora, significantly improving understanding and retrieval accuracy for specialized terminology. Meanwhile, the knowledge base adopts multi-tenant isolation, HNSW+PQ hybrid hot-cold storage, and a mixed chunking strategy. In particular, the integrated PDF parser can recognize document hierarchy, footer noise, and table structures, and perform semantic decomposition. This enables the system to truly “understand” complex research report tables rather than merely extracting text, thereby providing high-quality long-document processing capabilities for downstream applications.

In terms of hallucination control and content quality assurance, the project establishes a strict “generation–review” dual-loop mechanism. Given the sensitivity of financial and real estate data, the system introduces token budget control and a three-stage validation process (content accuracy verification, logical consistency verification, and presentation compliance verification). More importantly, a human-in-the-loop review mechanism is implemented for critical outputs. Experimental results show that the system’s hallucination rate in handling complex real estate data is reduced by more than 50% compared to general-purpose large models, ensuring high reliability and compliance of outputs.

The uniqueness of scenario applications is reflected in the deep development of specialized AI agents. The project builds charting and mapping agents with multi-step reflection and code generation capabilities. For example, when generating a district value analysis map, the agent can automatically resolve text overlap issues in SVG graphics through an innovative repulsive-force algorithm, enabling high-quality, publication-grade automated visualization. In addition, the multimodal content generation engine integrates image-to-image generation, OCR, and voice cloning technologies to automatically produce real estate marketing posters, AI podcasts, and trending videos, greatly enriching content dissemination formats.

Another distinctive feature is the engineering-level prompt management system. The team has established a comprehensive prompt governance framework supporting variable injection, version control, and canary release. This allows prompt engineering for different scenarios (such as luxury housing sales versus affordable rental housing) to be managed and iterated like code, ensuring that AI outputs remain appropriate in tone and logically rigorous across different user groups.

IV. Implementation Outcomes

The deployment of this project not only validates the feasibility of vertical-domain large models at the technical level, but also delivers remarkable results in both economic and social dimensions, demonstrating the strong transformative power of digital intelligence technologies on traditional industries.

1. Economic Benefits and Operational Efficiency Gains

According to operational data and financial analysis, the system has delivered immediate and substantial efficiency improvements for pilot enterprises.

First, in knowledge retrieval and information access, the average time for employees to search historical materials and professional standards has been reduced from hours to seconds, improving retrieval efficiency by over 90%. This translates directly into massive time savings for thousands of employees, significantly reducing labor costs.

Second, in research report generation, the introduction of AI assistants has increased productivity by 300%. Reports that previously required overnight work can now be generated within minutes, with significantly improved formatting consistency and data accuracy.

In the core marketing process, the “Top Sales AI Assistant” has improved customer response speed by 40%, and standardized script adherence has reached 90%, directly driving higher conversion rates in sales scenarios.

In terms of direct financial return, the project generated over RMB 20 million in revenue in its first year of deployment and is expected to exceed RMB 100 million by 2026, demonstrating strong commercial scalability and return on investment.

2. Social Value and Industry Impact

Beyond economic benefits, the project has also generated profound social impact in service quality upgrading and talent structure optimization.

The implementation accelerates the digital transformation of professionals in the real estate industry. Through human-AI collaboration workflows, many traditional employees have acquired the ability to use AI tools, achieving signifcant skill upgrading.

Furthermore, as the first full-scenario AI-native implementation case in the industry, this project provides a replicable and scalable standardized blueprint for the digital transformation of the real estate sector. It demonstrates that by building vertical AI foundations, traditional industries can move away from extensive growth models toward a high-quality development paradigm driven by data and knowledge intensity. This represents an important benchmark for promoting deep integration between China’s digital economy and the real economy.

Case 51

An Intelligent Investment Research AI Analysis System Powered by Multi-Agent Collaboration and Knowledge-Driven Intelligence

Wind Information Co., Ltd.

I. Company Profile

Wind Information Co., Ltd. (hereinafter referred to as Wind) is one of China's leading providers of financial software, data, and information services. Headquartered in Shanghai's Lujiazui Financial Center, Wind operates 48 branches across China. It serves a broad range of financial institutions—including securities firms, fund management companies, insurance companies, banks, investment firms, and media organizations—and has become an important partner for more than 90% of China's financial institutions.

Wind provides clients with comprehensive financial data covering equities, bonds, funds, foreign exchange, commodities, macroeconomic indicators, and industry information. It also delivers professional software applications and integrated solutions for information retrieval, data analysis, investment research, and portfolio management, offering 24/7 reliable and timely financial information services to financial institutions, regulatory agencies, universities, research institutes, and corporate clients.

In response to the rapid development of artificial intelligence, Wind officially launched its proprietary large language model for the financial industry—Wind Alice. Built upon Wind's decades of expertise in financial data and investment research, Wind Alice serves as the core engine of the company's AI-powered financial services. It has already been deployed in more than 1,000 enterprise application scenarios, driving intelligent transformation across the investment research workflow.

II. Case Description

(I) Background

The financial investment industry has entered a critical stage of AI-driven transformation. Investment analysts face explosive growth in multi-source financial data and increasingly stringent requirements for research efficiency and analytical quality. Traditional investment research workflows rely heavily on manual data collection, information extraction, report interpretation, and logical reasoning. These labor-intensive processes consume a substantial portion of analysts' time, leaving limited capacity for higher-value research, strategic thinking, and innovation.

At the same time, financial expertise and research experience have long been difficult to standardize and replicate. Much of the industry's valuable knowledge remains embedded in individual experts, creating bottlenecks in knowledge transfer, collaboration, and organizational learning, while constraining further improvements in research efficiency and service quality.

(II) Solution

To address these challenges, Wind re-engineered the investment research workflow by integrating AI with professional financial expertise. The company systematically transformed analysts' research methodologies, logical reasoning processes, domain knowledge, and best practices into an intelligent knowledge system. Through knowledge engineering, rule extraction, and workflow restructuring, the Wind Alice Financial Large Model is capable of learning and applying the analytical thinking patterns of experienced financial professionals.

Built on this foundation, Wind developed a comprehensive intelligent solution centered on the Wind Alice Financial Large Model. Tasks that previously required significant manual effort—such as information retrieval, data aggregation, and preliminary analysis—can now be completed efficiently with AI assistance. More importantly, the system goes beyond automation by supporting deeper analytical reasoning and generating high-value research insights and recommendations.

1. Alice Chat: Intelligent Financial Q&A

Alice Chat is trained extensively on professional financial knowledge, enabling it to understand complex financial terminology and industry-specific concepts with exceptional accuracy

Rather than simply answering factual questions, Alice Chat is designed to explain the underlying rationale behind financial events and market developments. By analyzing historical cases and financial logic, it helps investment professionals move beyond asking "what happened" toward understanding "why it happened."

Leveraging deep knowledge of investment research methodologies, Alice Chat has mastered the complete analytical process used by professional researchers—from macroeconomic observation and industry analysis to company valuation and investment judgment. It can accurately identify key issues, automatically establish analytical frameworks, and transform raw financial data into structured investment insights.

Moreover, Alice Chat continuously updates its financial knowledge through ongoing learning from new market information, research reports, and expert interactions. As a result, the system not only answers existing questions effectively but also proactively incorporates the latest market developments into future analyses, providing forward-looking professional insights.

The following illustrates the solution architecture of Alice Chat intelligent Q&A:

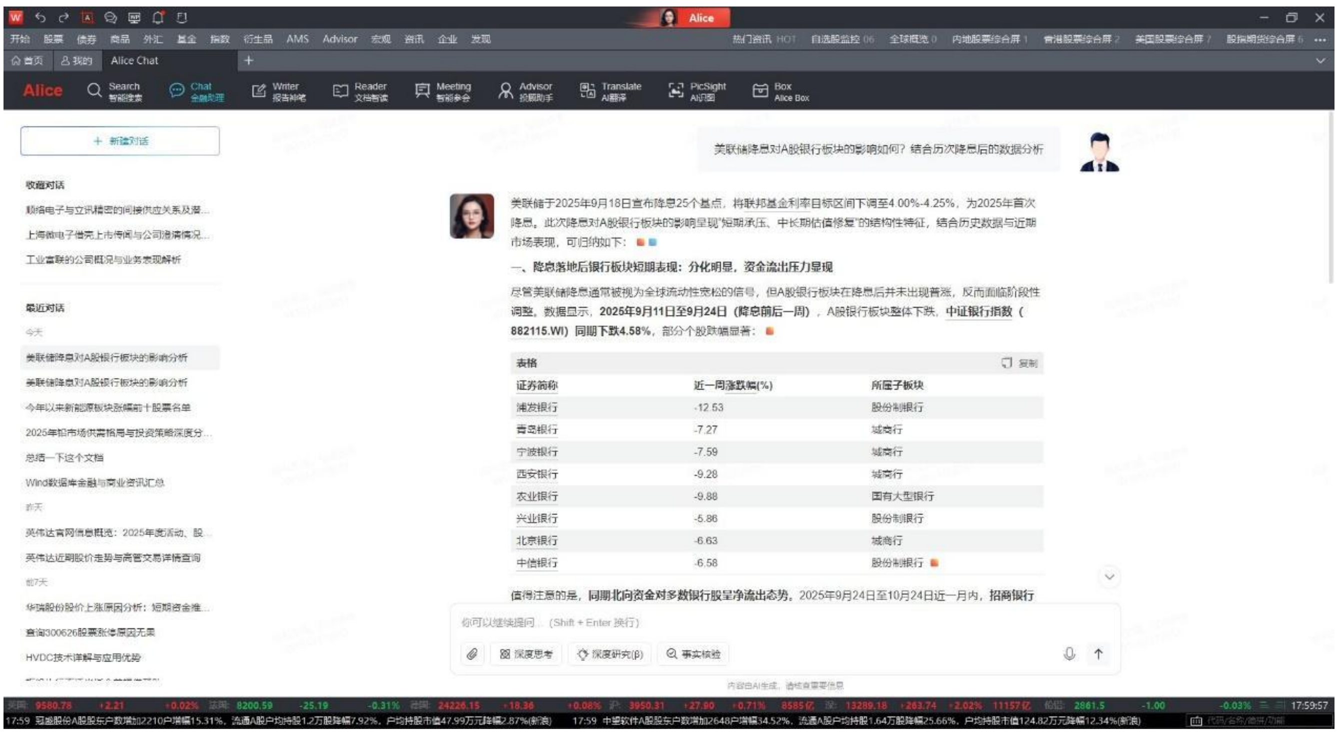


Figure 1 Alice Chat Enabling intelligent transformation from data to insights

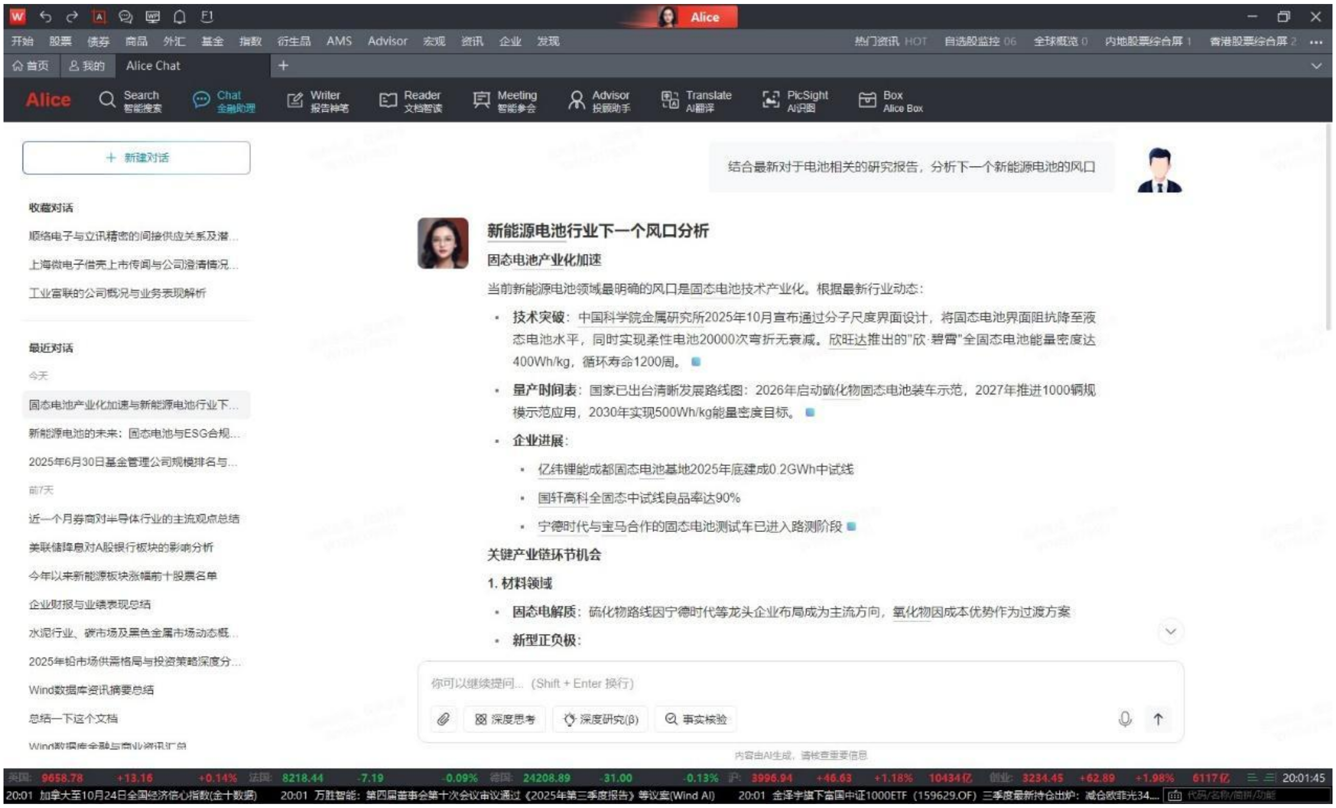


Figure 2 Alice Chat reasoning and analysis based on the latest market dynamics

2. Alice Writer: Intelligent Investment Research Report Generation

Through extensive learning from large volumes of high-quality investment research reports, Alice Writer has established a standardized knowledge base for financial writing. It has mastered the language conventions, professional writing styles, and analytical structures commonly used by experienced research analysts.

The system covers virtually the entire investment research workflow, including macroeconomic analysis, industry research, company studies, strategic analysis, fund evaluation, and thematic research. Drawing upon Wind's authoritative financial database, Alice Writer can automatically generate research reports that conform to professional investment research standards.

Compared with traditional report-writing processes, the improvement in efficiency is remarkable. Reports that previously required one to two days of manual preparation can now be completed within approximately five minutes. Beyond generating narrative text, Alice Writer automatically retrieves relevant financial data, produces charts and visualizations, structures analytical frameworks, and transforms fragmented information into coherent, professional research reports.

Furthermore, every AI-generated report undergoes multiple layers of validation to ensure factual accuracy, logical consistency, and regulatory compliance. This mechanism effectively minimizes hallucinations and factual errors while ensuring that the generated content meets the rigorous quality standards required by financial institutions.

More importantly, Alice Writer is not intended to replace professional analysts. Instead, it functions as an intelligent research assistant, enabling analysts to devote more time to high-value activities such as investment judgment, strategic thinking, and client engagement, thereby substantially enhancing overall research productivity.

Figure 3 Alice Writer report generation system framework

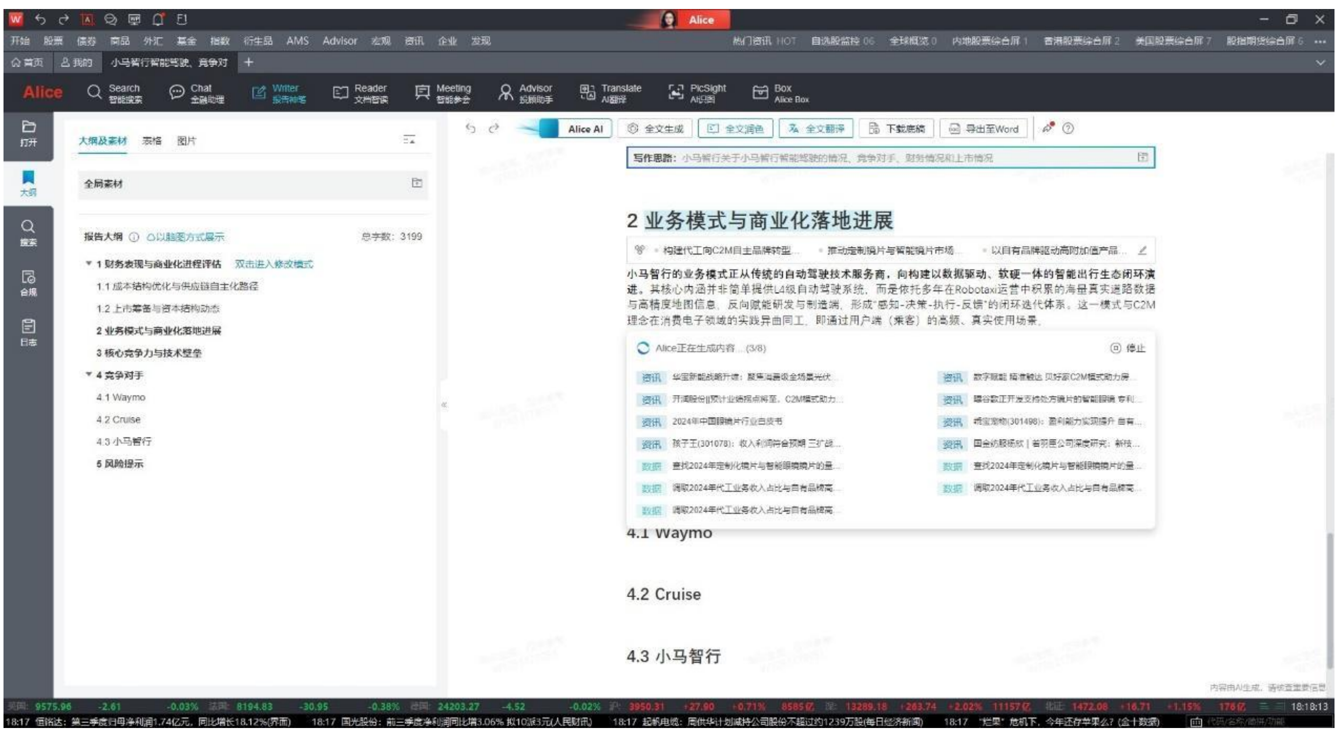


Figure 4 Alice Writer automated generation of outlines, logical structuring, research collection, and content drafting

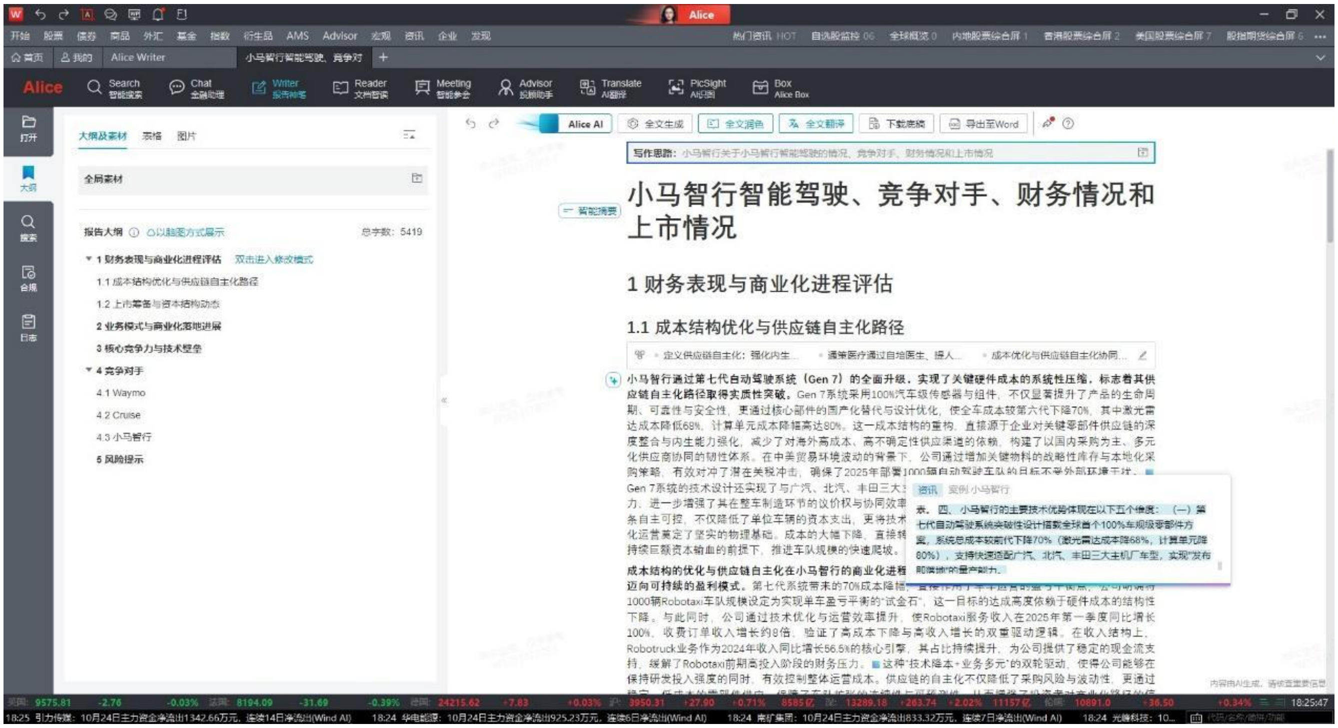


Figure 5 Writing content citation and reference standards

3. Agent Workflow Automation

WIND (Wind Information) conducts in-depth research on the workflows of financial experts and decomposes and models the working methodology of senior analysts into intelligent systems. Each Agent functions like a professional assistant analyst equipped with specialized expertise, rather than a simple task executor—it learns expert thinking patterns and decision-making logic.

Through an intelligent multi-Agent collaboration architecture, different Agents are assigned dedicated responsibilities such as information gathering, data analysis, logical reasoning, and report generation, faithfully simulating the collaborative workflow of an investment research team.

Each Agent transforms high-frequency and complex investment research tasks into knowledge-driven intelligent processes. They are embedded with domain expertise, analytical frameworks, and decision rules, enabling them to make professional judgments like experienced analysts. Agents understand the essence of tasks, determine when to retrieve data, select appropriate analytical methods, and structure output effectively.

For users, only a natural language description of the task objective is required, with no technical background needed. The Agents can automatically interpret requirements, intelligently plan execution paths, allocate necessary resources, and complete end-to-end professional analysis.

This knowledge-driven Agent collaboration mechanism not only significantly improves efficiency but also standardizes and replicates the working methodologies of top analysts, ensuring consistency and stability in output quality. Through Agent workflow automation, every investment researcher is empowered with expert-level intelligent support, truly enabling the intelligent transformation of financial research work.

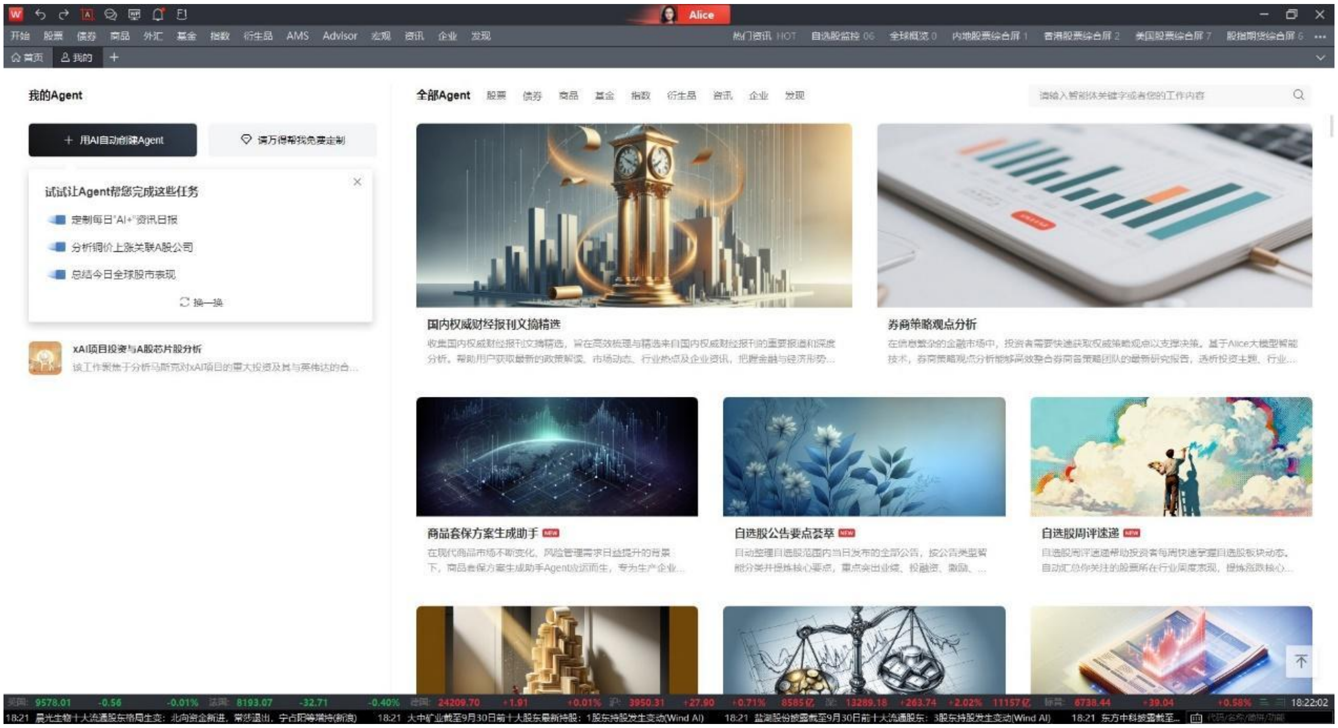


Figure 6 Alice Agent Marketplace

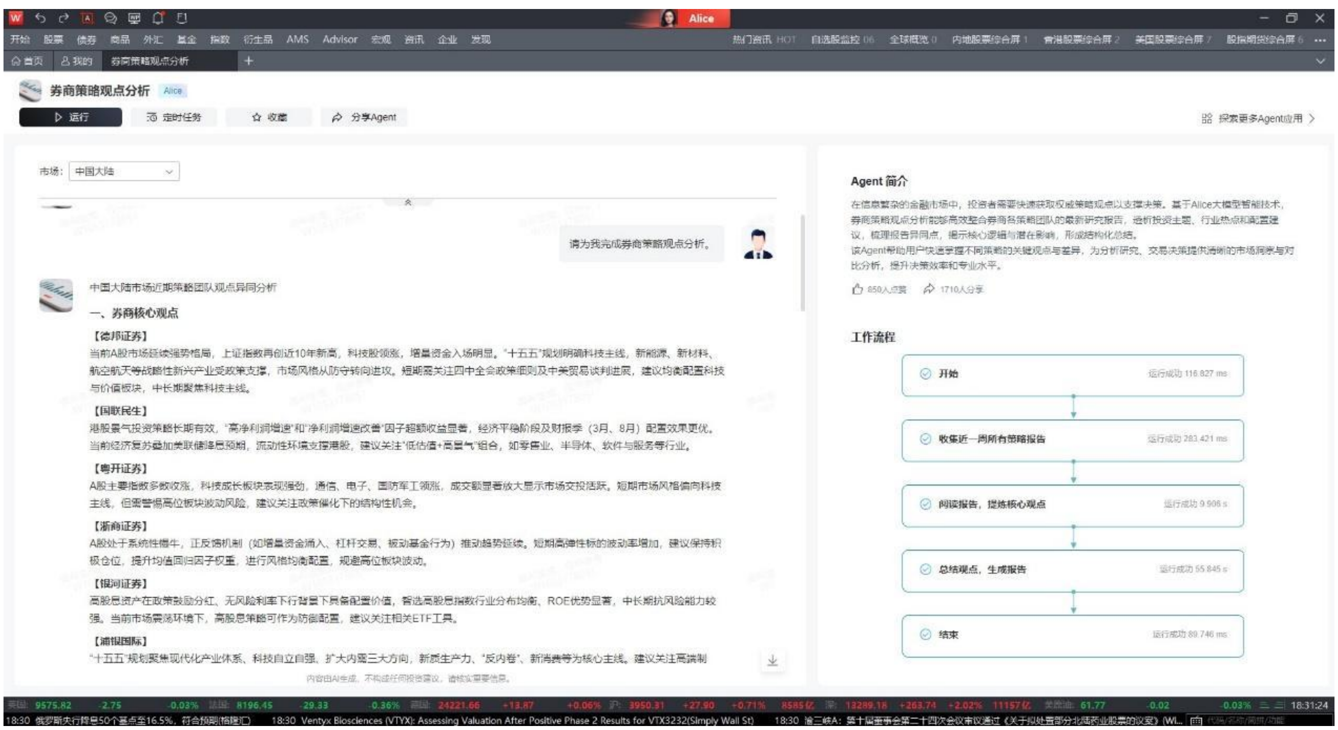


Figure 7 Workflow automation

4. Alice Reader and Alice Meeting: Intelligent Document and Meeting Understanding

Alice ' s document and meeting intelligence capabilities are built on a deep understanding of financial experts ' information processing methodologies. The system not only masters text recognition technologies but also learns the information screening logic and value judgment criteria used in investment research analysts. Just like an experienced analyst who reads reports, it pays special attention to key signals such as abnormal fluctuations, structural changes, and hidden risks, thereby developing a professional sense of information sensitivity and judgment.

By constructing domain-specific financial knowledge, Alice is able to understand the business meaning and investment implications behind information. When processing documents, Alice Reader builds a multi-dimensional analytical framework similar to that of a professional analyst, including vertical comparisons of financial indicators, horizontal analysis of business changes, and in-depth mining of risk factors. For meeting content, Alice Meeting not only extracts spoken key points but also identifies implicit information within discussions and captures subtle shifts in management sentiment—an embodiment of the experience accumulated by seasoned investors over many years.

Alice's intelligent analytical capability is based on a deep learning of expert methodologies. It understands which combinations of financial indicators may signal potential risks and which statements from management deserve deeper attention.

It can also identify which trends in industry reports are forward-looking and possess predictive value. This kind of information-driven processing enables what previously required hours of manual reading and analysis to be completed within minutes, delivering professional-level deep insights almost instantly.

Figure 8 Alice document information processing

(III) Problems Addressed

Wind Alice deeply integrates financial experts' methodologies with AI technologies, fundamentally resolving efficiency bottlenecks and quality challenges in investment research work.

First, Alice enables a shift from experience dependency to knowledge empowerment. In traditional investment research models, professional knowledge and working experience are often concentrated in senior analysts, making them difficult to transfer and replicate. Wind Alice transforms this valuable expert know-how into a reusable AI system, enabling users to quickly access the working methods of senior analysts. This standardizes knowledge transfer and enables large-scale application, systematically enhancing the research capabilities of the entire organization.

Second, in terms of information processing, Alice completely alleviates the problem of analysts being overwhelmed by massive volumes of data. Data retrieval time is reduced from tens of minutes to near real-time response in seconds, allowing analysts to access required information instantly. Facing the daily influx of massive research reports, Alice can automatically extract key insights to ensure that critical information is not missed. Even complex documents exceeding 100 pages can be deeply analyzed within 5 minutes, with precise identification of key information. This intelligent filtering capability allows researchers to truly control information rather than be controlled by it.

Finally, in terms of productivity improvement, Alice fully liberates researchers from repetitive tasks. Report writing that previously required 2 hours can now be completed in 5 minutes. Routine monitoring tasks operate automatically 24/7. This enables researchers to dedicate their time and energy to high-value work such as deep analysis and strategy development, truly returning investment research to its core essence.

Finally, through an intelligent Agent collaboration mechanism, the traditional individual working model is upgraded into an intelligent team-based collaborative model. Multiple intelligent Agents can simulate the collaboration patterns of a professional team, automatically conducting cross-dimensional analysis and multi-source information integration.

The entire investment research workflow is transformed into a standardized, automated, and intelligent process. This not only significantly improves work efficiency but also ensures stability and consistency in research quality.

With this comprehensive solution, Wind Alice truly achieves the intelligent transformation of financial investment research work, making AI a powerful assistant for every finance professional.

II. Project Highlights

(I) Key Technological Breakthroughs

1. Financial Industry Large Model Innovation

Wind Alice has achieved multiple groundbreaking advances in financial large model technology. Through deep learning techniques, the system successfully transforms the implicit knowledge and expertise of financial analysts into a computable intelligent model, breaking through the traditional limitation that AI can only handle explicit rule-based problems.

In natural language understanding, the system not only recognizes financial terminology and concepts, but also understands complex business logic and causal relationships, achieving a leap from "language understanding" to "logical reasoning." By constructing dynamically evolving financial knowledge representations, the system can automatically discover correlations between different types of information, forming association-based reasoning capabilities similar to human experts.

These technological breakthroughs enable AI to truly understand the essence of financial business for the first time, rather than merely performing pattern matching.

2. Integration of Financial Big Data and AI

In the financial domain, data accuracy and real-time performance are critical. Wind Alice leverages a vector database to deeply capture semantic relationships within unstructured financial texts. With the support of a knowledge graph, it accurately maps dynamic relationships among companies, securities, individuals, policies, and events, enabling structured understanding of complex financial ecosystems.

Through a distributed large-scale information retrieval engine, the system achieves millisecond-level financial information retrieval. Combined with a distributed stream-processing engine, it ensures that transaction, market, and risk data can be computed and responded to in real time.

By deeply integrating these four core capabilities with a financial large model, Wind Alice achieves intelligent deployment in highly complex data environments. It establishes a trustworthy, precise, and interpretable intelligent financial hub, equipped with both semantic reasoning and real-time decision-support capabilities.

(II) Scenario Specificity

Financial investment research scenarios are highly professional and complex, which brings unique challenges and opportunities for AI applications. Unlike general domains, financial research requires processing multi-source heterogeneous data, conducting cross-temporal and multi-dimensional analysis, and making balanced quantitative judgments, where every decision may involve significant economic impact.

Wind Alice deeply aligns with these characteristics by tightly integrating AI capabilities with financial domain expertise, forming a distinctive "AI + Expert Intelligence" model. The system is not merely automating existing workflows; instead, it redefines the investment research process itself, enabling AI to act as an intelligent partner with professional judgment capabilities.

This scenario-level deep customization allows the system to demonstrate capabilities far beyond general-purpose AI in the financial domain.

IV. Implementation Results

(I) Economic Benefits

Wind Alice creates significant economic value for financial institutions through a triple mechanism of "cost reduction, efficiency improvement, and risk control." Currently, Wind Alice has a usage penetration rate exceeding 90% among thousands of clients, which itself demonstrates strong economic effectiveness.

1. Cost Reduction: Compressing Research Costs

According to in-depth interviews with clients including major banks and research institutions, the efficiency gains brought by Wind Alice are reflected in several key areas:

- Significant reduction in data organization cycles:

In-depth research tasks involving data collection and structuring have been shortened from the original 2 people working for 24 hours to just 2 people working for 4 hours, substantially reducing labor and time costs.

- Improved research output efficiency:

Report drafting time has been reduced by at least 50%, with production cycles shortened from one week to three days, significantly improving analyst productivity and return on labor investment.

2. Value Enhancement: Driving Investment Returns

- Decision-making efficiency advantage:

Enables rapid processing of large volumes of information. During the earnings reporting season for listed companies, the time required to process annual and quarterly report information is reduced from one day to just 3 hours, helping institutions seize decision-making opportunities and identify market opportunities ahead of competitors.

- Research quality assurance:

Standardized analytical frameworks ensure the stability and accuracy of research quality, enabling more precise identification of investment opportunities and providing strong support for portfolio return growth.

3. Loss Control: Reducing Potential Risks

- Compliance automation:

By strengthening automated compliance checking mechanisms, the system proactively identifies regulatory risks from the source, significantly reducing economic losses and reputational damage caused by non-compliant corporate behavior.

- Enhanced risk forecasting capability:

Strengthens risk prediction capabilities and reduces losses caused by investment misjudgments, thereby protecting asset safety.

Overall, Wind Alice, through the synergistic effects of "cost reduction, efficiency improvement, and risk control," significantly enhances institutional profitability and risk resilience. With a customer adoption rate exceeding 90%, these tangible economic benefits have been fully validated, making it an essential tool for improving competitiveness in the financial industry.

(II) Technological Innovation Points

1. Deep Integration of Large Models and Professional Databases

The system innovatively integrates a large-scale parameter LLM with Wind's professional financial database, constructing a "trusted AI" framework in the financial domain.

This integration not only preserves the powerful language understanding and generation capabilities of large models, but more importantly, ensures output accuracy and timeliness through real-time data integration. It effectively addresses the "hallucination" problem of general-purpose large models in financial applications.

Each piece of data and each conclusion can be traced back to its source, enabling true verifiability and transparency in AI-generated financial insights.

2. Intelligent Financial Workflow

Wind Alice has achieved innovative breakthroughs across multiple technical dimensions.

In terms of knowledge enhancement, it continuously absorbs the latest market information and expert experience through a continuous learning mechanism, forming a dynamically evolving knowledge system.

In intelligent retrieval, the system not only accurately locates information, but also performs intent-based reasoning and correlation analysis. A multi-agent collaborative architecture enables automatic decomposition and parallel processing of complex tasks, where each Agent has domain-specific expertise and works collaboratively to complete end-to-end complex analytical workflows.

The system also innovatively integrates natural language interaction with professional financial analysis, allowing users to obtain deep insights as if they were conversing with experts. These innovations collectively enable the system to truly possess "reasoning" capabilities.

(III) Social Value

1. Driving Intelligent Transformation in the Financial Industry

The Wind financial terminal is not merely a tool, but a key driver of intelligent transformation in the financial industry.

Through successful applications with financial clients, it provides proven AI implementation cases for the sector and accelerates the intelligent upgrading of more financial institutions. This demonstration effect is helping the entire financial industry shift from a traditional labor-intensive model to a technology-driven model.

2. Enhancing Financial Services for the Real Economy

By improving research efficiency and quality, financial institutions are able to provide broader research coverage and investment banking services to more enterprises.

In particular, for small and medium-sized enterprises (SMEs) and emerging industries, AI technology lowers the research barrier, enabling them to gain greater attention in capital markets. This inclusive improvement directly enhances the breadth and depth of financial services supporting the real economy.

3. Promoting the Application of AI Technology in Vertical Domains

The successful implementation of Wind Alice has provided valuable experience for applying AI technology in industry-specific fields. By deeply integrating general-purpose AI technologies with domain knowledge, it demonstrates the tremendous potential of AI in vertical industries.

4. Enhancing the International Competitiveness of China's FinTech

As a fully self-developed financial AI product, Wind Alice represents China's innovation capability in financial technology. In the context of increasingly intense global FinTech competition, this improvement in independent innovation capability not only serves domestic financial institutions but also lays a technological foundation for China's FinTech sector to compete in international markets.

(IV) Industry Impact

Wind Alice's intelligent investment research capabilities have been widely adopted across multiple financial institutions, including securities firms, mutual funds, insurance companies, banks, and government regulatory departments, earning broad recognition from users.

The product has not only transformed traditional investment research workflows but, more importantly, has driven a paradigm shift in the industry's perception of AI applications. From initial skepticism to active adoption, financial institutions have gradually realized that AI is not meant to replace professionals, but to empower them—enabling financial practitioners to create greater value.

This conceptual transformation is accelerating the financial industry's transition into a new stage of intelligent development.

Case 52

Industrial Circulating Water System Control Optimization Technology Empowering Traditional Manufacturing with Energy Savings, Cost Reduction, and Efficiency Improvement

HaiRun Zhiyun (Shanghai) Data Technology Co., Ltd.

I. Company Profile

HaiRun Zhiyun Technology Co., Ltd. was established in 2015 as the diversified second-growth business of Haigang Group, one of China's leading enterprises in the apparel industry. Headquartered in Jingjiang, Jiangsu Province, the company has progressively expanded from the Yangtze River Delta to the national market.

Focusing on energy efficiency, HaiRun Zhiyun helps industrial enterprises improve quality and efficiency, reduce operating costs, promote intelligent upgrading, accelerate digital transformation, and achieve greener development. To strengthen its industrial influence and attract high-caliber talent, the company established HaiRun Zhiyun (Shanghai) Data Technology Co., Ltd., located in Shanghai's Yangpu District. The company has been recognized as a National High-Tech Enterprise, a Shanghai Specialized, Sophisticated, Distinctive and Innovative SME, and a Yangpu District Technology Little Giant Enterprise, among other honors.

Leveraging the group's diversified product portfolio, extensive market resources, and deep industrial expertise, HaiRun Zhiyun has developed an integrated cloud–edge–device comprehensive energy management solution. Through data acquisition, analytics, diagnostics, simulation modeling, and artificial intelligence algorithms, the company optimizes customers' energy structures, enables energy-saving transformation projects, and delivers intelligent operation and maintenance services.

II. Case Introduction

(I) Background Requirements

Against the backdrop of actively promoting green transformation and intelligent upgrading in the manufacturing industry, the energy system is undergoing a critical transition toward digitalization and intelligence.

Fine-grained control and energy efficiency improvement have become key tasks. As an important component supporting stable operation in traditional manufacturing production systems, circulating water systems still widely rely on manual scheduling, suffer from imbalance in water-energy matching, weak system coordination, lack of intelligent analytical capabilities, and other common issues. These problems are specifically reflected in unstable water temperature control, significant energy waste during intermittent production, large fluctuations in inlet/outlet water temperature differences, reliance on personnel experience in multi-variable system operation, and insufficient fault warning capabilities.

Therefore, there is a need to build a smart water system management platform with full perception, intelligent analysis, and closed-loop optimization capabilities. Such a platform has become a necessary path for steel enterprises to achieve energy saving, emission reduction, and efficient low-carbon operation.

(II) Implementation Plan Introduction

This project is a typical solution for intelligent upgrading of water systems in the steel industry within traditional manufacturing. It aims to comprehensively retrofit circulating water systems and related power equipment to achieve intelligent optimization of water supply and usage, thereby reducing energy consumption.

The project constructs an integrated control system based on computer networks, integrating key technologies such as intelligent water monitoring, system flow optimization, hydraulic balance, and thermal balance dynamic adjustment, enabling centralized monitoring and fine-grained management of circulating water systems.

The system covers the entire lifecycle from energy data collection, production process monitoring, online analysis, evaluation, and optimization control. Relying on existing informatization foundations, it forms a closed-loop optimization mechanism of energy-production coordination. While ensuring production safety, it significantly improves system efficiency.

In terms of implementation, the project builds a complete water system control network. Through flow sensors installed at water supply terminals, precise flow control is achieved. Valve optimization is carried out for water distribution, and real-time collection of temperature, pressure, and operating status data of key equipment is enabled. This realizes full-process monitoring and intelligent control from water supply to usage. The main implementation content includes infrastructure control, acquisition equipment upgrades, system control network construction, and multi-variable control system development, ensuring efficient, stable, and reliable operation of the water system.

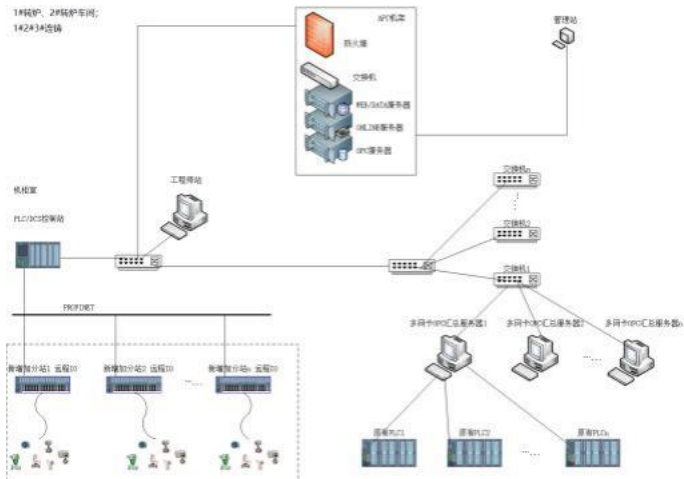


Figure 1 Overall System Design

1. System Control Network Construction

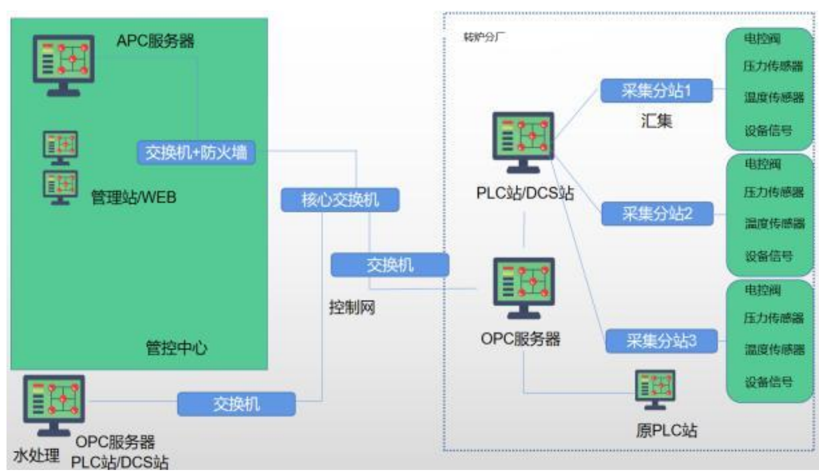


Figure 2 System Control Network

The project adds collection substations, PLC substations, OPC servers, and switches to build a water system control network, establishing data links and control loops between the water treatment system and the steelmaking sub-plant, and enabling real-time online regulation through an APC server.

A new PLC system is deployed in the water treatment sub-plant to collect signals such as temperature, pressure, valve status, and frequency converter data. It communicates with the original converter pump station system to enable data acquisition and control of the existing system. The system also communicates with the sub-plant's OPC server to collect process parameters at water points. A PID-based control loop is implemented, applying strategies such as constant pressure, constant temperature, and variable flow according to process requirements from the user side, while operating within safety constraints. This replaces the original manual control logic and enables automatic pump switching, as well as coordinated control of liquid level and fan operations. The retrofit involves upgrades to the original system control logic and front-end configuration, along with a fiber-optic network deployed at the sub-plant level.

In the steelmaking sub-plant, additional collection stations are installed to aggregate temperature, pressure, and valve signals via distributed I/O modules. A newly built PLC station communicates with each acquisition node through an industrial Ethernet network and interacts with the existing secondary systems, collecting key equipment signals to support decision-making. This PLC station is integrated into the water system control network, exchanges data with the APC server, and implements single-loop PID control based on temperature differentials for certain water-use circuit valves, thereby adjusting branch water flow.

2. Multivariable Control System

In traditional industrial circulating water systems, end users are numerous and complex, and it is difficult to unify requirements for pressure and flow at different water points. To meet the demands of downstream users, operators often increase the pressure of the main supply pipeline. However, this leads to a chain reaction in the network: excessive flow, overly high velocity, and increased pipeline pressure loss, while also causing impact on certain valves.

In multi-user water distribution networks, if only temperature difference regulation is applied at the main pipeline, temperature deviations may occur in certain branches. Conversely, if regulation is applied only at branch lines, flow demand cannot be effectively coordinated with energy supply, resulting in limited energy-saving benefits. Therefore, introducing a multivariable control system is essential.

The project designs and deploys an independent APC (Advanced Process Control) server to run multivariable control programs. The APC server is integrated into the converter water treatment system and connected to the control network via optical fiber.

The multivariable predictive controller is implemented using dedicated software and algorithms, independent of the PLC system. Due to its high computational complexity, it is typically deployed on a standalone server acting as the upper-level control system, communicating with PLCs via standard OPC protocols. Since the controller is independent from the PLC, logic switching programs must be configured within the PLC to enable seamless interaction between APC and PLC systems, ensuring smooth switching between advanced and conventional control.

(1) Configuration Content

To achieve seamless switching between advanced control and conventional control systems, auxiliary variable tags are created within the PLC system (including custom digital, analog, and integer variables), without modifying the original PLC configuration.

These tags mainly serve two functions:

1) Communication verification

PLC communication protection signals.

2) Control system switching

Advanced control system tags include control switches, controlled variables (CV) related signals (upper/lower limits, status bits), and manipulated variables (MV) signals (output value, constraints, limits, etc.).

(2) Communication Monitoring Program

To ensure system safety, when communication failure or system crash occurs between the APC host and PLC, the advanced control system must be automatically disengaged. This is achieved through a communication monitoring program in the PLC.

Mechanism: Under normal operation, the APC host periodically writes zero to a monitoring tag in the PLC. The PLC continuously accumulates this value. If the reset signal is not received for a long time and the accumulated value exceeds a threshold, it indicates a communication failure. The PLC then automatically disables the advanced control system.

(3) Switching Program

The switching program ensures seamless transition between advanced control and conventional control systems. Each control loop can be switched independently. It includes safety start/stop of controllers, protection of variable limits, communication status checks, and loop status verification. It ensures safe operation of the control system.

(4) Operation Interface

The interface is the platform for operators to access and manage the advanced control system. It enables switching of control modes, adjustment of process parameters, and management of system functions such as global switches, subsystem switches, and upper/lower limits. Design principles emphasize usability, safety, and practicality.

III. Project Highlights

The company developed advanced control systems including APC for air separation and water systems based on industrial big data platforms. These systems break traditional PID, cascade, and ratio control methods. Their core technologies include system simulation modeling and multivariable predictive control, reducing operator workload and optimizing process control for cost reduction and efficiency improvement.

(I) System Simulation Modeling Technology

Our company relies on the world-leading CFD commercial software ANSYS Fluent. Through collaboration with top domestic and international industry experts and hundreds of enterprise on-site experiments and validations, we have developed industrial energy-consumption mathematical models applicable to fluid dynamics, heat transfer, and chemical reaction processes.

These models can fully reproduce real production processes in a computer-based environment by collecting actual production data, process parameters, and equipment information from enterprises via the Haiyan Intelligent Cloud Platform. The system evaluates energy demand across different process units and accurately identifies energy waste points, thereby providing targeted energy-saving solutions and significantly improving overall energy utilization efficiency while delivering strong economic benefits.

The system is built based on pipeline network configurations (including pipe length, diameter, and valve states), physical parameters (such as ambient pressure, temperature, water head elevation, fluid density, and pipe material characteristics), and node demand data (such as pressure, flow rate, temperature differences, etc.). Using built-in mathematical formulations, a system-level simulation model is constructed to accurately replicate real operating conditions.

The model dynamically represents flow direction, flow rate, velocity, and pipeline resistance across the network, reflecting demand and actual operating conditions at each node, including flow and pressure distribution. It also provides real-time visualization of operational parameters such as real-time flow and water replenishment demand versus actual demand, enabling full-system simulation and analysis.

The true value of the system lies in its ability to design and simulate various energy-saving schemes based on the model. These solutions are safe, reliable, and practical, and do not introduce operational risks to production. The predictive capability of the model can estimate future behavior of controlled objects and determine control strategies while adapting to system variability. Meanwhile, feedback correction mechanisms use measurable data to adjust predictions, reducing errors and improving robustness.

In addition, rolling optimization selects the optimal sequence for each control cycle, typically implementing only the first step while continuously updating based on real-time feedback to achieve optimal performance under uncertainty.

After implementation, and based on real-time monitoring of multiple data sources on the Haiyan Intelligent Cloud Platform, the system not only achieves significant energy-saving effects but also elevates the enterprise's safety management level, ensuring safer and more stable production operations.

(II) Multivariable Predictive Control System

Traditional water system control methods include temperature difference control, pressure control, flow control, cooling tower control, and water level balance control. These systems are typically based on independent PID loops with closed-loop operation. However, there are complex interdependencies among subsystems, which leads to limitations in practical applications.

Because control variables are highly coupled, interactions between subsystems can create strong coupling effects, making unified optimization difficult for closed-loop systems and often resulting in unpredictable outcomes.

Therefore, the introduction of a multivariable control system is essential for energy-saving optimization in water systems. The system establishes a complete multivariable input/output control framework. Input signals include water temperature, pressure, flow rate, ambient temperature, liquid level, and valve positions. Output signals include pump speed regulation, frequency control, valve actuation, and switching commands.

The core multivariable controller consists of PLC/DSP, upper-level control computer, frequency converters, and electric control valves.

Energy-Saving End (Heat Exchange System) Control

In the heat exchange system, the control strategy focuses on flow balance and thermal balance:

Flow balance control: The model considers pump head and minimum pump operating conditions to determine the optimal water delivery path. It also ensures that the minimum flow requirements inside the heat exchanger are met, in order to determine inlet and outlet pressure differences and flow distribution.

Thermal balance control: Based on equipment heat load and allowable temperature rise, the system determines the minimum required flow distribution for each water point.

During the modeling stage, flow rate, temperature, and pressure constraints are defined. Real-time valve adjustments are used to maintain equipment operation within optimal ranges, ensuring stable and economical system performance.

Cooling End (Cooling Tower) Control: In cooling tower operation, the system performs coordinated control of water distribution and fan operation. The system dynamically adjusts water distribution to cooling towers and fan switching based on operating conditions. To prevent temperature fluctuations caused by sudden fan activation or shutdown (which may cause water temperature instability), the system adjusts return water valves and fan frequency in coordination. This ensures balanced water distribution and stable outlet water temperature, achieving precise control of the entire cooling process.

IV. Implementation Results

(I) Direct Economic Value

Through the deployment of an intelligent industrial water system, a special steel enterprise in Hubei Province completed intelligent optimization and energy-saving retrofitting for its water supply and converter water systems. The project focuses on circulating water systems and upgrades power equipment such as pumps and valves. It started in July 2023 and passed acceptance in December 2024.

According to data provided by the Party A, electricity consumption per ton of steel decreased from 16.47 kWh/t before the retrofit to 14.5 kWh/, a reduction of 11.96%. Based on annual steel production, electricity savings are approximately 5.91 million kWh/year, saving about 3.84 million RMB/year.

For converter and hot-rolling workshop systems, comprehensive electricity consumption decreased from 0.2453 kWh/m^3 to 0.1849 kWh/m^3, a reduction of 24.62%, saving approximately 4.23 million kWh/year and 2.75 million RMB/year.

In total, the project achieves:

- Total annual electricity savings: 10.13 million kWh
- Total annual economic benefit: 6.59 million RMB

Total annual electricity savings	Total annual economic benefit
10.13 million kWh	6.59 million RMB

(II) Indirect Economic Value

Centralized control: The system enables full-process centralized monitoring, control, and data analysis of the water system. Through a single platform, managers can understand the overall operating status of the entire water plant system, while data and reports can be automatically generated and pushed to relevant personnel.

Automatic control: Under normal production conditions, the system enables automatic control, requiring operators to only handle abnormal equipment failures.

Energy-saving and consumption reduction: By integrating a unique hydraulic and thermal balance analysis system, the system enables optimized allocation of water supply, intelligent control of pumps and fans, and achieves energy savings.

Technology upgrade: Operators are freed from on-site manual operations, while the system provides simulated operations and real-time simulation results, helping maintenance personnel make more reasonable parameter adjustments. This reduces operational difficulty and human workload while enabling continuous system optimization.

Safety enhancement: By deploying an intelligent industrial water system, sensors related to key equipment, network nodes, and heat exchange devices are installed. These critical system data are updated in real time (per second), and the system model performs real-time analysis based on the data to predict potential risks or failures in advance. Potential safety hazards are monitored and managed proactively, achieving "prevention before failure" and ensuring safe production.

Maintenance cost reduction: Through the application of the intelligent industrial water system, mathematical models are used to analyze pressure fluctuations and flow bottlenecks across the system, enabling continuous optimization. High pressure is a major cause of wear and damage to valves, pipelines, and mechanical seals. By optimizing system parameters, equivalent theoretical models are applied to eliminate damage risks caused by excessive pressure at pump outlets. At the same time, improving low-flow areas reduces sedimentation and prevents heat exchanger blockage, effectively extending equipment service life.

Case 53

An Intelligent NL2SQL Query System Powered by Large Language Models

Shanghai Urban Construction Information Technology Co., Ltd.

I. Company Profile

Shanghai Urban Construction Information Technology Co., Ltd. (hereinafter referred to as SUCI Info) is controlled by Shanghai Urban Construction Group, China's largest state-owned listed enterprise in the municipal engineering sector, which has ranked first nationwide for 17 consecutive years. Operating under a fully market-oriented model, the company is a high-tech enterprise dedicated to providing one-stop digital information services for urban infrastructure.

As a key subsidiary supporting the digital transformation of Shanghai Urban Construction Group, SUCI Info focuses on the full lifecycle of infrastructure projects. It provides digital solutions including engineering information services, engineering big data analytics, AI-enabled integrated applications, and intelligent operation platforms.

The company brings together professionals from architecture, information technology, artificial intelligence, and other interdisciplinary fields. Employees aged 25–40 account for 94% of the workforce, while 92% hold bachelor's degrees or above. Its projects have received more than 40 awards, including the prestigious Tianyou Award and various municipal excellence awards.

Among its flagship achievements is the digital twin project for the Yangpu Bridge, the first "Metaverse" application in Shanghai's transportation sector. Through 1:1 digital reconstruction, the project integrates visualization, management, and safety monitoring into a unified platform, promoting digital transformation across the industry.

II. Case Introduction

(I) Project Overview

Stec AutoBI is an intelligent data query system designed for the brokerage stock trading and market research scenario. The system uses a fully automated workflow as its core, which includes:

"Natural language question → AI large model executing standardized NL2SQL → local secure execution of queries → result-driven business output."

It introduces the MCP (Model Context Protocol) as a standardized bridge between large models and external tools, and leverages multi-agent collaborative orchestration to decompose and understand query tasks.

The system performs intent understanding, query matching, SQL generation, execution control, and exception retry.

(II) Background Requirements

With the continuous expansion of group business, the market department needs to manage bidding and tender data for thousands of projects. These core business assets are stored in multiple data platforms.

The existing query system has the following issues:

Monthly report query cycles are long; from requirement submission to result delivery takes on average 1–2 working days, which cannot meet real-time or ad-hoc needs.

- Query conditions rely on predefined templates, resulting in low flexibility.
- Personalized queries require manual SQL writing by database engineers, leading to high cross-department communication costs.

Therefore, there is a strong need to develop an intelligent data query system to solve these problems. However, at the early stage of system redesign, several implementation challenges arise due to complex business rules and enterprise governance constraints.

From a business logic perspective, the system involves multiple table joins and complex logical relationships in the data warehouse. Traditional monolithic architectures suffer from limited concurrency, low flexibility, and restricted scalability.

From an information security perspective, bidding and tender data are highly sensitive core assets. Data is strictly prohibited from being exported outside the company network or uploaded to external cloud AI systems. Additionally, due to budget constraints, large-scale proprietary model deployment is not feasible, while small models struggle with complex NL2SQL reasoning tasks.

(III) Implementation Plan

In response to the above requirements, the project takes system construction content and objectives as the starting point, and proceeds from both system architecture and business logic to design and implement technical and business solutions. Finally, the application is deployed through system integration and workflow tuning.

1. Construction Content and Objectives

Functional objectives: Shift from “fixed monthly report queries” to “natural language free-form queries,” eliminating dependence on technical staff and enabling frontline business users to access data directly. The query scope is expanded from “monthly reporting fields” to “data warehouse data items,” with results returned in an automated loop execution process. Data flow is strictly isolated: query results are not fed into large models, and third-party database cloud services are not relied upon.

Efficiency objectives: Improve efficiency by compressing data query turnaround time from 1–2 working days to just a few seconds; reduce costs by decreasing reliance on database engineers for cross-department communication; and support decision-making by providing timely and more comprehensive data to market strategies.

2. System Architecture

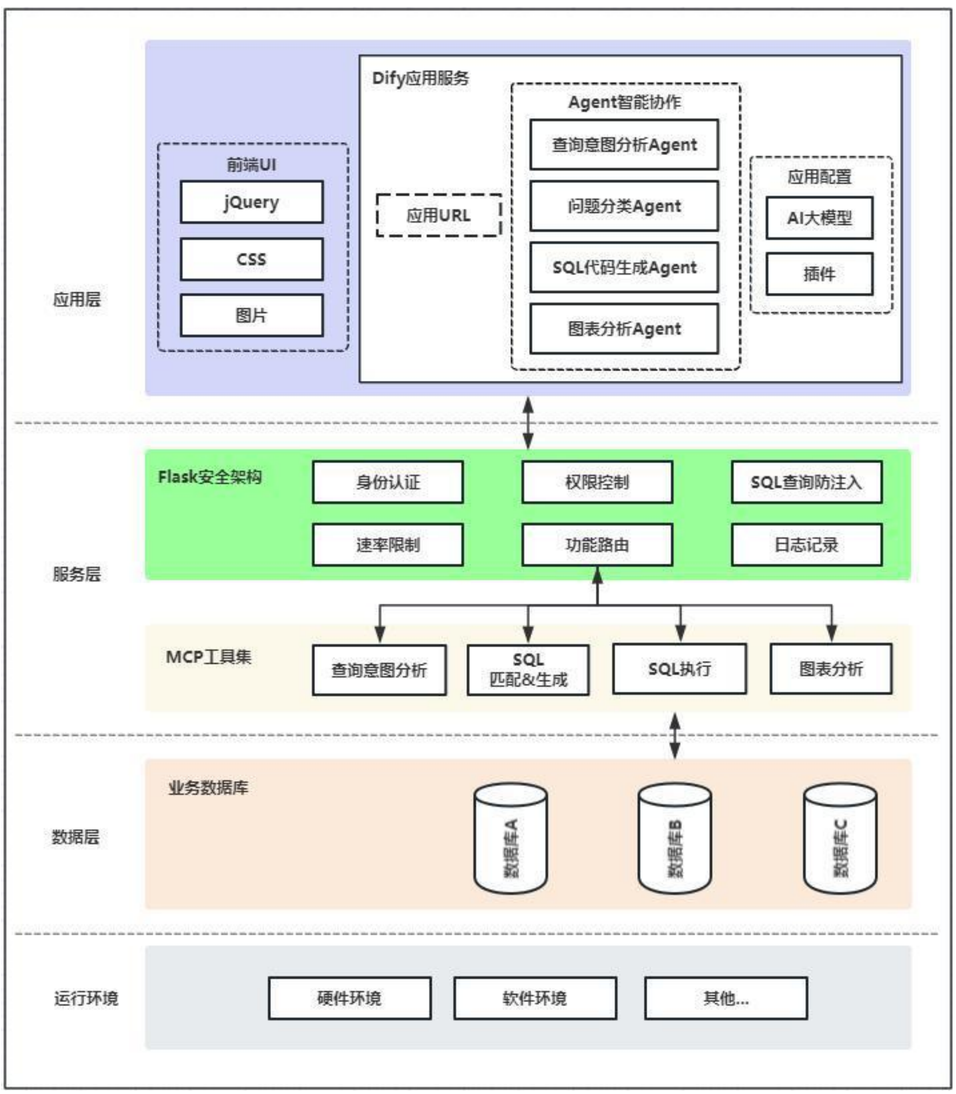


Figure 1 AutoBI System Architecture Diagram

We define the technical roadmap as “Dify open platform + domestic third-party online large models + MCP protocol + multi-agent architecture.” On this basis, we build an isolated execution loop with the principle of “no data leaving the intranet and no model backflow.”

In addition, the system adopts a layered microservice architecture to ensure clear responsibilities and low coupling. The main layers are as follows:

Application Layer

As the core of user interaction and intelligent processing, this layer mainly includes a user-friendly Web UI at the frontend, supporting jQuery-based interaction, CSS styling, multimedia display, etc.

Through Dify application service workflow orchestration, it integrates frontend UI, intelligent Agent collaboration, and application configuration management. The Agent subsystem consists of multiple specialized intelligent agents responsible for tasks such as query intent analysis, question classification, SQL generation, and chart analysis. Through multi-agent collaboration, complex AI processing workflows are implemented. The application configuration module manages large models and plugin settings, enabling flexible invocation of different AI capabilities and extensibility of the system.

Service Layer

As the core for business logic processing and security assurance, this layer mainly includes Flask-based security architecture, providing comprehensive protection such as identity authentication, permission control, SQL injection prevention, rate limiting, functional routing, and logging to ensure system reliability.

It also includes the MCP toolset based on the Model Context Protocol (MCP), providing standardized tool invocation interfaces, including query intent analysis, SQL generation, SQL execution, and chart analysis.

Data Layer

As the foundation of data storage and management, this layer maintains multiple business-related database instances and supports classification storage and management of different types of business data, ensuring data organization and maintainability.

Runtime Environment Layer

As the infrastructure support layer, it includes:

- Hardware resources: computing, storage, and network infrastructure
- Software environment: operating systems, middleware, runtime environments
- Other environment configurations: network configuration, monitoring tools, and other essential operational components

3. Business Analysis and Logical Implementation

Business requirement clarification: The application boundaries, functional requirements, and database/table dependencies are clearly defined with business users. The final scope covers 2 core requirement categories, 10 business modules, 1 database, and 16 tables.

Business scenario analysis: Based on monthly sample queries, the system identifies frequently occurring user question types, abstracts them into reusable semantic patterns, and builds template combinations.

Data dependency analysis: For each question category, relevant data tables and external key relationships are identified, including time dimensions, primary keys, and data fields.

Query logic analysis: Business logic and data extraction methods are matched based on data fields. The system considers that identical fields may have different query logic across tables or time dimensions, and thus builds a unified rule library (e.g., different rules for cumulative quarterly values vs. incremental values).

Technology stack selection: The system evaluates budget constraints and NL2SQL performance to select the optimal solution. Dify (open-source, flexible, highly compatible, and ecosystem-friendly) is used as the application development platform, while domestic third-party large models are used as foundational models (balancing cost and performance).

Testing and tuning: Following the principle of "minimal template coverage for maximal query logic," AB testing and template compression are conducted for each sub-module. In the optimization stage, approximately 10 – 20 rule templates are reduced to 5 high-quality templates. A refined data table schema description is used as NL2SQL grounding to improve accuracy and stability of business logic execution.

Information security enhancement: Security protections such as XSS/SQL injection/SSL encryption are implemented at the network layer. At the MCP layer, access control and data isolation are strengthened. Full-chain auditing and rate limiting are applied to ensure controllability within an intranet environment.

4. Workflow Refinement and Key Adjustments

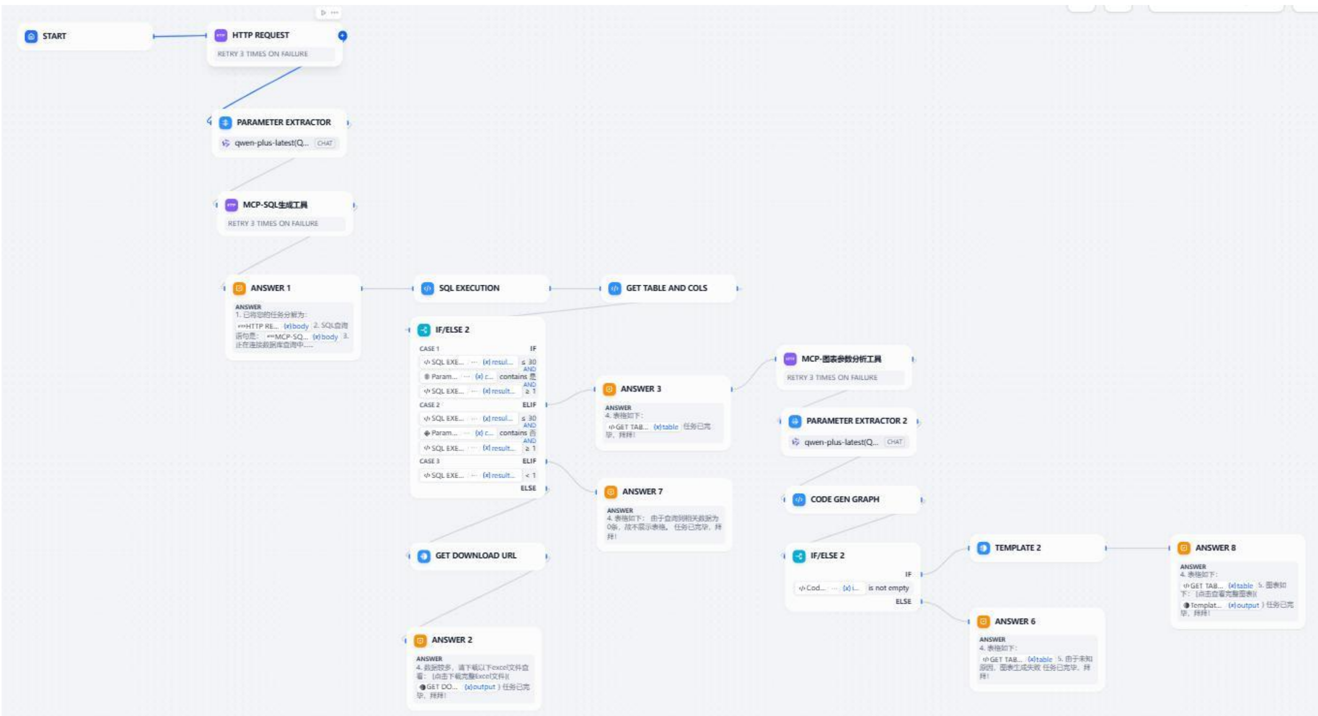


Figure 2 Agent + MCP Workflow Interface Diagram (Single-Branch Scenario)

The project initially adopted a RAG-based approach, storing SQL query templates in the Dify knowledge base. SQL was generated by using the Agent to invoke MCP tools and retrieve templates. However, as the number of templates increased, the precision and controllability of retrieval gradually decreased, and the latency of NL2SQL increased from seconds to minutes.

Therefore, the solution was refactored into a pure MCP-based SQL generation approach, embedding templates into Agent prompts and converting them into a single MCP SQL generation tool. This reduced NL2SQL generation time to within 15 seconds (or even just a few seconds in most cases). On this basis, the project introduced an Agent-based intelligent question classification mechanism, which categorizes logically conflicting query types into different MCP tool invocation branches, with each branch corresponding to a dedicated MCP SQL generation tool, thereby avoiding logical interference.

At the same time, to meet the requirements of the second version, a parallel chart generation workflow was implemented. The system adopts an Agent + MCP + local CODE node hybrid architecture for data security processing. After query execution, the CODE node first extracts data field information (without exposing raw data). The AI model then analyzes only field names and initial query intent, generates chart parameters, and finally merges these parameters with actual query results in a local CODE node to generate charts using a Python library.

Ultimately, the system implements a complete workflow in Dify Chatflow, including:

"User query → Agent intelligent analysis → query intent parsing and formatting → MCP standardized tool invocation → CODE node local secure database query → Agent-based chart parameter analysis → CODE node secure chart rendering."

This end-to-end pipeline integrates MCP protocol and Agent technology while strictly ensuring data security compliance.

III. Key Highlights

(I) Key Technological Breakthroughs: A Five-Pronged Innovative System

A new secure querying paradigm for "data never leaves the perimeter" is created for data access. It establishes a dual mechanism of "model participates in reasoning, data remains computed locally," achieving "no data enters the model, no data leaves the network, and full process auditability." This provides a safe and reliable online AI intelligent query experience under strict compliance requirements.

The engineering capabilities standardized on MCP are reused to build a unified "model–tool" interaction protocol, breaking the traditional bottleneck of AI applications being "hard to develop once and difficult to migrate."

Through prompt engineering and template compression, a two-layer strategy of "selected templates + schema dynamic injection" is adopted. Prompt length is reduced by 40%, while maintaining an accuracy rate above 85%. Average response time is stabilized at 2–3 seconds, and token consumption is reduced by 40%.

A domestic full-stack technical verification is implemented based on a combination of domestic OS (Kylin v10 SP3), an open platform (Dify), and domestic large models, completing end-to-end localization validation from infrastructure to application layer and breaking the perception that "domestic solutions are inherently less capable."

A safe pipeline for intelligent chart generation is designed by continuing the "data separation" principle. Semantic metadata and real data are decoupled; chart parameter configuration is processed by large-model analysis, while visualization rendering is completed locally via Python, achieving a secure "parameters enter the model, data does not enter the model" generation process.

(II) Scenario Characteristics: Architectural Innovation Under a "Trilemma"

The project breaks through the "impossible trilemma" of cost–accuracy–security constraints through a threefold trade-off mechanism. Analysis of constraints:

Cost constraint: Large models capable of meeting high accuracy requirements (85%+) typically require GPU cluster deployment, with monthly costs exceeding tens of thousands. Even smaller deployable models (7B–13B) only achieve around 60% accuracy in complex query scenarios, which cannot meet production requirements.

Security constraint: While third-party large-model APIs are relatively low-cost, data governance requirements prohibit exposing real business data to external services.

Based on this, the project proposes an innovative architecture called an "intelligent partitioned execution layer." In this design, SQL statements themselves are treated as abstract expressions of query logic that do not contain real business data (sensitive information only appears in execution results), allowing security boundaries to be enforced at key points in the data flow.

Capability offloading: The "natural language → SQL" transformation is handled by third-party large models. By sacrificing some marginal cost, the system obtains reasoning capability equivalent to GPU cluster deployment while improving accuracy from ~60% to over 85%.

Data-in / data-out separation: Database query execution and result rendering are encapsulated within local CODE nodes. Sensitive data never leaves the system throughout the process.

(III) Core Value: From Architectural Innovation to Full-Stack Engineering Implementation

The project introduces the architectural concept of an "intelligent partitioned execution layer," achieving conceptual innovation at the architectural level (breaking the trilemma), while also achieving deep technological breakthroughs at the engineering implementation level (from security compliance to domestic innovation verification).

This dual-driven approach—“top-level design + bottom-level implementation”—not only solves specific business problems, but also establishes a reproducible, scalable, and continuously evolvable technical path for AI applications in regulated scenarios.

IV. Implementation Outcomes

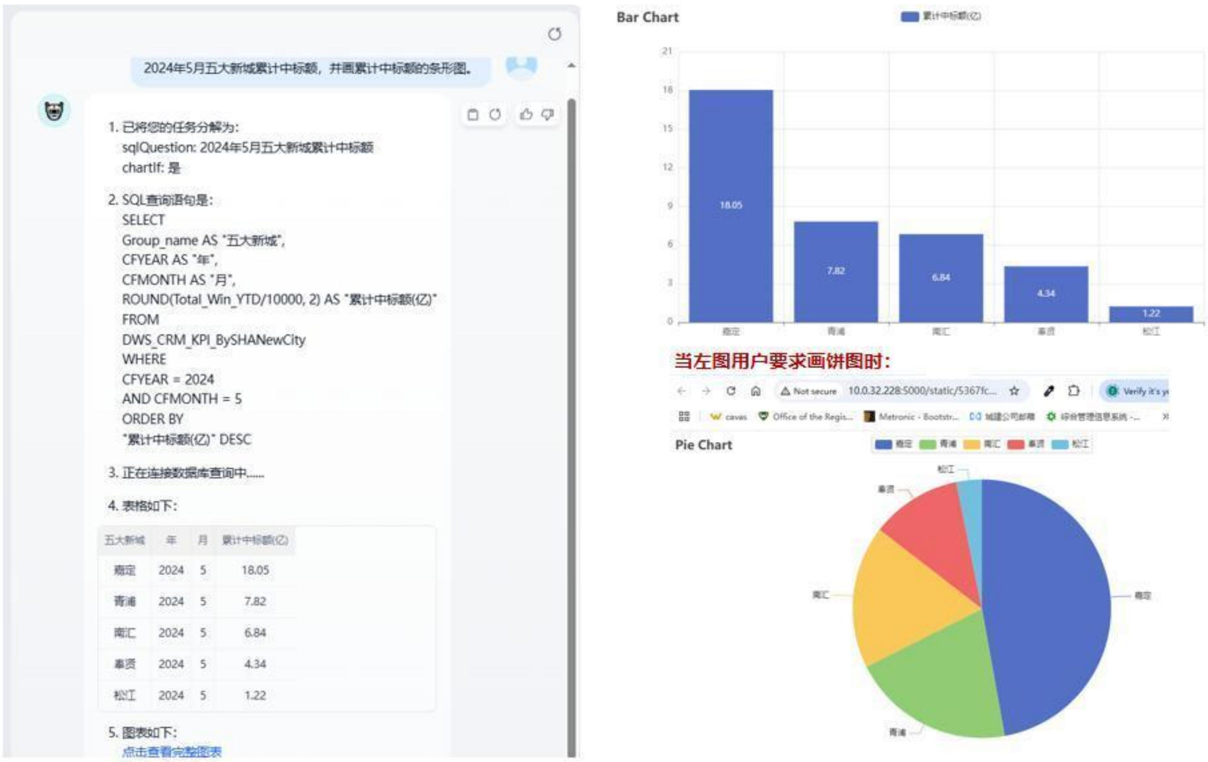


Figure 2 Query Results Example (Test Data Used to Ensure Data Security)

(I) Economic Benefits

Market application outcomes: The data query process duration was reduced from 1–2 working days to an average of 5–15 seconds, with response speed improved by 99%. The average number of daily queries increased from 3–5 times to over 30 times, significantly improving efficiency.

Business department replication effect: Based on rapid reuse of mature modules from the Market Department to the “Corporate AI Strategic Analysis Module” (capital preservation / financing modules), code reuse rate exceeds 70%. The time to launch new scenarios has been shortened from quarterly/seasonal cycles to monthly/weekly cycles, with replication cycles reduced by 75% and development cycles reduced by 50%. The project team size decreased from 10 people to 4 people.

(II) Social Value

Market application outcomes: The query scope expanded from monthly report keywords to a data platform, with coverage increased by 2–3 times and data granularity refined to a more detailed level. Non-technical users’ query learning time was reduced from 1–2 days to 2–3 hours. “Real-time Q&A for urgent issues and fast answers to complex problems” was achieved, with query accuracy stabilized above 90%. Within 6 months of system launch, active users increased from 3–5 to over 30, and average daily queries rose from about 200 to over 1,000.

(III) Ecosystem and Replication Effects

A standardized workflow node template library with more than 5 templates has been established, with over 60% template reuse. New business scenario onboarding time has been reduced from 1–2 months to 2–3 weeks. Future rollout plans: modular design supports cross-scenario replication, expected to cover modules such as cost management and supply chain management, forming a unified group-level “AI + Data” foundation serving more than 10 business domains.

(IV) Risk Control and Compliance

Under the constraint of keeping data inside the internal network, the system provides usable, reliable, and auditable AI capabilities that meet compliance requirements. Audit coverage reaches 100%, with zero security incidents. All query operations are fully traceable, with 100% log completeness, and data access permissions are controlled down to table-level and field-level.

Afterword

The compilation of this casebook brings together the practical experience of Shanghai's small and medium-sized enterprises at the forefront of AI adoption, as well as the insights of industry experts and research institutions. The cases are organized into two principal themes:

AI Empowerment. This section broadens the industrial perspective from individual applications to the vision of "AI for Thousands of Industries." It is dedicated to exploring a comprehensive methodology for realizing AI value, examining in depth how AI technologies can be deeply integrated with industry-specific know-how to foster new ecosystems and business models, reshape industrial value chains, and demonstrate AI's transformative power as a general-purpose technology.

AI Applications. This section focuses on the intelligent transformation of traditional enterprises and organizations. Serving as a practical implementation guide, it systematically illustrates the entire process—from strategic planning, scenario identification, and technology selection to value assessment. Through these cases, readers can see how small and medium-sized enterprises overcome resource constraints to deploy AI successfully, achieving both cost reduction, efficiency gains, and business innovation.

We hope that this casebook will serve as a valuable reference for AI practitioners, entrepreneurs, policymakers, researchers, and all those interested in artificial intelligence in Shanghai and across China. It is our aspiration that readers will gain a deeper understanding of the opportunities presented by the AI era, better appreciate the underlying principles of AI-enabled industrial transformation, and engage with greater confidence in the practical journey toward intelligent transformation.

The successful completion of this casebook would not have been possible without the dedicated efforts of the Shanghai Municipal Office for Promoting the Development of Small and Medium-sized Enterprises, the Shanghai SME Service Center, the Shanghai Artificial Intelligence Industry Association, and many other organizations that participated in its planning and editorial work. We are also sincerely grateful to the numerous enterprises and institutions that generously contributed the valuable case materials featured in this volume.

Shanghai Artificial Intelligence Industry Association

December 2025