

人工智能就绪的公共语料库：概念内涵、就绪框架与提升路径*

郑磊 杨涛**

复旦大学国际关系与公共事务学院 / 数字与移动治理实验室

摘要：在人工智能快速迭代发展的时代背景下，AI公共语料库已成为驱动智能涌现、赋能经济社会高质量发展的“新基建”。然而，既有研究多聚焦于传统语料库建设或结构化公共数据开发利用，对适配AI模型训练需求的公共语料库的概念及其治理关注不足。为此，研究系统构建了AI公共语料库的概念内涵与就绪框架。首先，界定了AI公共语料库的技术性内涵，厘清其与相关概念的关联和差异，进而从本体属性、权利属性、价值属性三个维度解析其公共性特征。在此基础上，构建了“技术性就绪-公共性就绪”双主线就绪框架。其中，技术性就绪涵盖可发现性、可访问性、可用性、可训练性与可信性五个核心要素，聚焦语料质量与技术适配性；公共性就绪包括公开性、普惠性、共创性与正当性四个关键维度，强调语料的公平供给与公共价值。此外，研究还探讨了提升AI公共语料库就绪度的实践路径。研究结论有助于深化智能时代背景下公共语料资源的理论认知，为我国构建AI就绪、公平可及、内容向善的公共语料基础设施提供实践参考。

关键词：人工智能；公共数据；语料库；AI就绪；高质量数据集

DOI：10.16582/j.cnki.dzzw.2026.04.001

一、问题提出

当前，以大语言模型为核心的生成式人工智能正引发一场深刻的范式革命。在此进程中，数据资源正跃升为驱动智能涌现、决定模型认知边界的基础性资源。^[1,2]大语言模型的性能遵循“缩放定律”（Scaling Law），其能力强弱关键取决于高质量训练数据的规模、多样性与语义深度。^[3]然而，当前全球人工智能（AI）发展正面临日益严峻的“数据瓶颈”。一方面，语料数据供给与需求之间形成显著的“剪刀差”。互联网高质量公开数据的自然增长速度，已远远滞后于大模型训练需求的指数级膨胀。有研究机构预测，全球高质量公网语言数据预计将在2028年前后趋于“枯竭”。^[4]而且，这类公开数据普遍存在信噪比偏低、知识碎片化、价值对齐模糊等问题^[5]，远未达到大模型训练所需的“AI就绪”（AI-Ready）要求。另一方面，数据资源的“壁垒化”态势日益加剧。头部商业平台纷纷筑起“数据护城河”，使得传统依赖低成本抓取实现大规模语料积累的模式已难以为继。

在此背景下，公共数据被视为破解语料供给梗阻、支撑大模型训练的重要战略资源。政府部门和公共机构在长期行政管理与公共服务实践中，沉淀了涵盖政策文书、执法记录、政务对话、城市感知等多模态海量数据。这些数据具有天然的权威性、规范性与逻辑性，且具备规模大、潜在价值高等特性，蕴含着与国计民生息息相关的规则、规律与专业知识，是训练政务智能体（Gov-Agent）和各类智能系统的优质语料素材。^[6]然而，这些存量数据大多是为适配传统业务流程生成的，与大模型所需的“AI就绪”语料之间，存在着显著的工程鸿沟。过去，政府共享开放的数据集多为结构化、机器可读的表格或关系型数据库，在数据提炼过程中，不可避免地剥离了丰富的背景信息与逻辑上下文。与此同时，仍有大量政务文档以非数字化格式留存，即便已完成数字化的部分数据，也存在语义标注精度不足、跨模态关联对齐水平偏低等问题，进一步加剧了语料适配难度。然而，公共部门对AI公共语料的建设与共享开放认

*项目基金：国家社会科学基金重大项目“面向数字化发展的公共数据开放利用体系与能力建设研究”（项目号：21&ZD337）。

**通讯作者

收稿日期：2026-02-27

修回日期：2026-03-25

知存在偏差,不仅对其核心内涵把握模糊,也未能充分掌握其价值规律,严重制约了公共语料的有效治理与利用。因此,盘活现有公共数据资源,建设适配本土语境、兼具可靠性与逻辑深度的公共语料库,破解语料供给难题,已是关乎国家长远发展与战略自主的重要议题。

放眼全球,推动“AI就绪”公共数据的共享开放,已成为加强训练语料供给、赋能AI发展的重要举措。美国先后发布《生成式人工智能与开放数据:指南与最佳实践》(Generative Artificial Intelligence and Open Data: Guidelines and Best Practices)、《赢得竞赛:美国人工智能行动计划》(Winning the Race: America's AI Action Plan)等政策文件,为政府机构准备和发布“AI就绪”数据集提供保障与指南。^[7,8]欧盟则通过《数据法案》(Data Act)及“欧洲公共数据空间”(Common European Data Spaces)建设等措施,强化政府高价值数据向AI可用训练资源的转化与治理。^[9]韩国发布了一系列面向AI发展的公共数据开放及利用计划,重点加强高价值“AI就绪”公共数据的治理和供给工作。^[10]

我国也在加速面向AI发展的公共数据战略布局。《中华人民共和国国民经济和社会发展第十五个五年规划纲要》强调要加强公共数据供给利用,加快建设人工智能语料库。《“数据要素×”三年行动计划(2024—2026年)》《国务院关于深入实施“人工智能+”行动的意见》《生成式人工智能服务管理暂行办法》等政策文件,也明确鼓励通过公共数据的开发利用,支撑和促进AI技术发展。在地方实践层面,上海、杭州、深圳、贵州等地正在积极探索扩大公共语料数据供给的机制创新。

然而,当前学界对这一新兴议题的理论回应仍较为有限。随着AI技术的爆发式发展,既有研究已关注到AI发展对公共数据生态系统的变革影响,以及对传统政府数据开放范式的“边界重构”挑战。^[11,12]相关研究指出了政府作为关键AI数据要素供给者的责任^[13,14],并围绕高质量训练数据的法律规制与技术规范展开了制度建构探讨^[15,16]。同时,部分研究也将语料库视为突破AI数据瓶颈、提升国家AI竞争力的新型基础设施。^[17-20]这些研究虽已勾勒出“AI公共语料库”这一议题的基本轮廓,但仍存在深层次局限。一方面,以往关于公共数

据开发利用的研究,多聚焦于结构化数据集或传统语料库,未能有效回应大语言模型对训练语料的特殊工程化需求与就绪性要求。另一方面,既有研究在探讨AI公共语料库技术性就绪的同时,忽视了其在价值层面的就绪性内涵,未能兼顾技术适配与价值导向的双重需求。鉴于此,本研究旨在系统界定AI公共语料库的概念内涵,构建复合视角的就绪框架,并探讨其价值实现路径。

二、AI公共语料库的概念内涵:“AI的”与“公共的”

要准确把握AI公共语料库这一核心概念,既需要厘清其作为AI基础设施的技术性内涵,明确其与普通数据资源的本质区别,还需要系统解析其公共属性,实现技术性与公共性的有机统一。

(一)“AI就绪”的语料库:技术性内涵及其重构

近年来,学界与实务界已广泛应用“数据集”“语料库”“训练数据”“AI就绪数据”“高质量数据集”等术语^[21-24],但彼此之间界限模糊。为厘清“AI语料库”的核心内涵,有必要对这些概念进行辨析。

数据集是一个组织形式概念,指按特定结构组织的数据集合,目的是方便计算机读取、处理和使用。它像一个“容器”,可泛指按照统一格式有序组织的数据集合,如结构化的表格、按类别整理的图片文件夹等。在AI语境下,数据集更多是指为特定任务构建的、带有明确“输入—输出”标签的“任务型数据集”(Task-Specific Datasets)。^[25]

语料库是一个内容类型概念,主要是指以语言数据为核心、经过系统化加工的资源集合。语料库这一术语发轫于语言学领域。20世纪60年代,相关研究将这种“为研究语言而系统收集的真实语言材料集合”定义为语料库。^[26]长期以来,其主要作用是供研究人员研究语言规律的“分析样本”。^[27]

然而,随着计算语言学的兴起,特别是大规模预训练模型的爆发,语料库的内涵发生深刻重构,从服务于人类认知的静态档案,演变为驱动机器认识世界的动态“教材”。在AI语境下,语料库特指用于支撑大模型预训练的自然语言资源集合,主要包含书面文本、口语转录或两者的组合。^[28,29]它通常以数据集的形式存在以便于模型

使用，但与任务型数据集存在本质区别：前者主要服务于模型的预训练阶段，使命是让模型通过海量、真实的语言材料建立对世界的通用认知^[30]，如同对人进行“通识教育”，让AI学会“听话、说话”（语言能力），追求规模、广度、多样性与上下文连贯性；后者主要服务于微调与测试等后训练环节，使命是让模型通过任务导向、结构化的标注数据习得特定技能^[31]，如同对人进行“专业教育”，让AI学会“做题、干活”（任务能力），关注精度、任务适配性与可量化评估。二者的主要差异见表1。

训练数据则是一个功能概念，可泛指用于训练AI模型的数据，无论其内容类型或组织形式如何，涵盖了语料库、任务型数据集等多种形态。语料库作为其中具有特定内容取向的子类，专注于语言数据，从而成为构建AI模型基础认知、常识体系与复杂逻辑不可替代的底层资源。而“AI就绪数据”和“高质量数据集”两个概念则多指向训练数据的“达标状态”^[24]，前者侧重AI训练适配的技术工程标准，而后者则兼具治理规范与应用价值视角。

需要指出的是，随着“预训练+微调”成为AI模型训练的主流范式，语料库的“任务化”与任务型数据集的“语言化”正在并行发展。不少传统语料库逐渐任务导向化，成为可同时支持预训练和多任务建模的“语料型数据集”；同时，许多经典数据集开始吸纳语料库构建原则，强调语言真实性与语料多样性，构成可计算和理解上下文的“数据型语料库”。尤其是在多模态人工智能的推动下，部分传统的文本语料库与任务型数据集逐步融合为一体化资源，二者功能日益交汇、界限趋于模糊。^[32,33]

这一趋势驱动传统语料库与任务型数据集从“分野”

表1 AI语境下的语料库和任务型数据集概念对比

对比维度	语料库	任务型数据集
主要作用	预训练	后训练
数据模态	以文本为主（含书面与口语转录）	文本、图像、音频、视频、传感器数据等
组织方式	按语言结构组织	按任务结构组织
标注要求	可无标注，或语言学标注（词性、句法等）	任务导向的精确标注（问答对、分类标签等）
主要特性	规模、广度、多样性、真实性	精度、标注质量、任务适配性

走向“同构”，催生了支撑机器复杂认知训练的“AI就绪语料库”新范式。一方面，从内容维度看，AI就绪语料库以语言数据为核心。因此，在信息表达中语言占据核心地位的数据，无论是否包含图像、视频等其他模态，均可纳入其范畴；而纯数值型数据及无语言模态的视觉数据则不属于。另一方面，从功能维度看，AI就绪语料库以支撑AI模型训练为基准，达到机器可理解层级、服务于模型训练的数据，即可纳入其范畴；而缺乏上下文语义信息的原始数据、仅供人类查阅的档案则不在此列。因此，AI就绪语料库也包含了经过语言化改造、保留了语言真实性的任务型数据集，即数据型语料库；而纯粹为特定任务构造、缺乏语言真实性的数据，则更宜归入一般任务型数据集。换言之，AI就绪语料库以语言数据为核心、以“语言真实性”为门槛，同时兼容经过语言化改造的多模态数据。

参考国际标准化组织（ISO）和国际电工委员会（IEC）提出的SMART模型^[34]，以及我国的机器可读标准（GB/T 45508—2025）^[35]，以机器对信息不同层次的理解和执行能力为标尺，AI就绪语料库的内涵及其发展态势可如图1所示。

综上，本文将AI就绪语料库界定为：为支撑人工智能模型训练，遵循机器认知规律，经系统性加工处理而构建的，以语言数据为核心、兼容多模态数据，具备规模化、高知识密度与迭代进化能力的数据资源集合。该集合构成一个连续谱系，涵盖从基础性预训练语料库到语言化任务型数据集这一区间。

（二）“公共的”AI语料库：公共属性的解析

在技术属性的“AI就绪”之外，为深入解析AI公共语料库的本质，还需重点回答三个核心问题：它是什么性质，为什么是公共的，为谁服务。

1.本体属性：“数智公物”与新型公共基础设施

首先，需要回答AI公共语料库是“什么性质的物品”。从物品属性来看，公共数据具有非竞争性、非排他性（也不应该有排他性），需要且应该通过各种行政性法律性强制规定来开放与共享。^[36]作为公共数据的组成部分，AI公共语料库同样具备显著的公共物品属性，关乎公共利益与公共价值，应在制度上明确其公共资源定位。

基于这一本体属性，AI公共语料库应同传统的水、

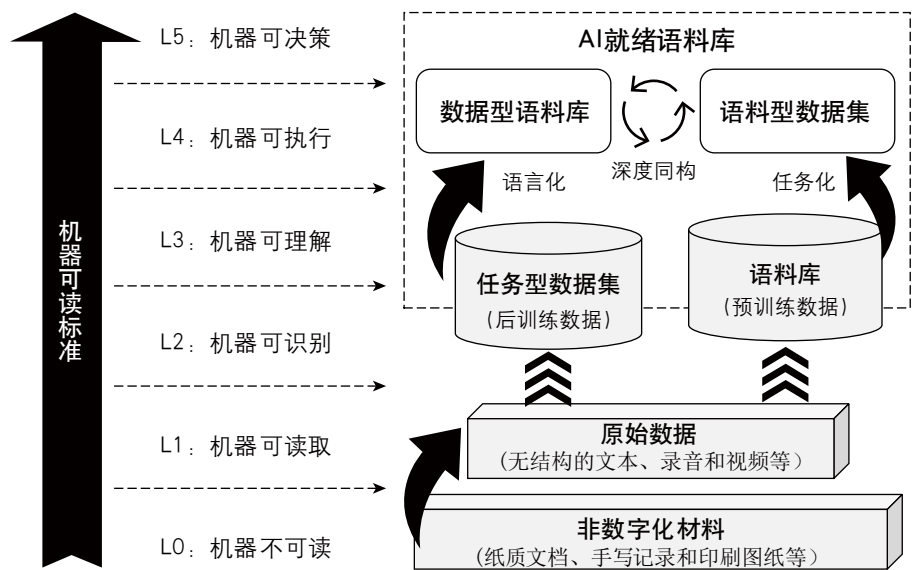


图1 AI就绪语料库的内涵及其发展态势

电、道路、网络等公共基础设施一样，视为由公共部门主导建设运营的数智公共基础设施。公共语料蕴含着社会各方的公共利益和基础信息，具有强大的正外部性、规模效应与网络效应，结合各类社会场景后有望产生巨大价值，其发展水平关乎国家AI产业整体竞争力与发展安全。^[20]同时，也正因其建设周期长、投资规模大、战略意义重大，且绝大部分“建材”源自公共部门，使其成为超越了单个市场主体的能力和责任边界的基础性资源。这种基础性、外部性和战略性，决定了它作为“数智公物”的根本定位，即一种服务于全社会的新型公共基础设施。

2. 权利属性：公共财政与公共信托

在明确了AI公共语料库的本体性质之后，还需要追问它“为什么是公共的”。公共语料库的权利属性，取决于其生成与存续的法理基础。从权利来源看，这类资源之所以是“公共的”，根源在于其原料是公权力运行的“副产品”。各级党政机关、国有企事业单位在依法履职过程中形成的数据资产，其产生并非源于市场行为，而是基于法律授权，以履行公共管理与服务职能为目的。这一资源生成机制，使其从源头便被打上了“公共”的烙印，决定了其权利归属不同于私有财产。

从权利支撑看，这类资源的生成与存续均依赖公共财政的投入。公共数据依托财政经费采集存储，后续的加工同样消耗公共资金，即纳税人的资源。^[37]这意味着，公民实际上已经通过纳税方式完成了对公共语料的

“预付”，任何主体均无绝对的支配权。因此，依据“取之于民，用之于民”的公共财政原则与公共信托理念，政府作为受托方管理和开发这些资源，所形成资产的终极产权应归属于社会公众。^[38]将沉淀的公共数据资源开发为高价值的AI公共语料并实现共享开放，已超出了对“部门资产”的简单处置逻辑，而是实现公共资产价值“返还”与升华的行为。

需要澄清的是，社会主体依托公共数据自行建设的语料库并不等同于“公共的”语料库，仅属于“基于公共数据建的”语料库。反之，“公共的”语料库也不等同于“公共部门的”语料库，其并不排斥对社会数据的吸纳与整合。从AI发展和社会现实需求来看，科研学术资源、网络公开数据、产业行业数据，及各类社会主体建设的语料库，均可通过协议授权、共享开放等方式纳入公共语料库，以此丰富语料资源的多样性与应用价值。

3. 价值属性：公益导向与观念载体

最后，还需要回答AI公共语料库是“为谁服务”的问题。公共语料库的价值属性，重点体现在其公益导向与服务宗旨上。公共语料库以增进公共利益为根本价值取向，因服务社会发展需求而存在。“公共”是“全社会公有共用”之意^[39]，即面向全社会，服务于所有人，而非特定主体的私人利益，这正是公共语料库区别于私有语料库的根本特征。其价值定位指向社会共同利益，是全社会共有共享的重要资源，服务于满足民生需求、促进经济发展

与社会创新、支撑科学研究等多元公共利益诉求，在坚持公共利益优先的同时，包容并赋能多种价值追求。

与此同时，AI公共语料库还承担着承载公共价值观的独特功能。与一般公共资源不同，语料不仅是信息载体，更是价值观的塑造媒介。AI模型的价值取向与认知框架均从语料中习得，而其所习得的不仅是语言规律，更包括语料所蕴含的世界观、思维方式与价值偏好。^[18]公共语料库在内容层面深层次关联特定的社会文化传统与价值体系，语料中蕴含的价值取向将直接塑造AI系统的输出逻辑，进而影响其服务于人的方式。这一本质凸显了公共语料区别于其他公共资源的独特之处：其不仅具备“资源”与“利益”意义上的公共性，更承载着“文化”与“认同”意义上的公共性。

综上，AI公共语料库可界定为：以增进公共利益和公共价值为导向，由公共部门依托公共权力与公共财政主导建设、吸纳社会力量参与，面向公共用途进行共享与开放的AI模型训练语料资源集合。其公共性源于本体、权利与价值三类属性的有机统一。

三、AI公共语料库的就绪框架：技术性就绪与公共性就绪

AI公共语料库的双重内涵，也对其建设与治理提出了实现“技术性”就绪与“公共性”就绪的双重要求。

本研究在审视现有数据治理和语料库建设原则及其局限的基础上，构建了一个适配AI公共语料库内涵与特性的就绪框架。

（一）既有框架及其局限性

围绕数据管理与AI训练，国际国内已发展出多个具有影响力的原则、标准与评估模型（参见表2），但尚不能完全适用于指导“AI公共语料库”的建设与治理。

其一，重“人用”而轻“机用”，面向AI的技术性就绪程度不足。以FAIR原则^[40]、政府数据开放原则^[41]为代表的传统数据管理框架，以及部分传统语料库建设原则，为数据流通利用建立了“通用护照”。然而，这些准则的核心逻辑是保障“人”能够找到、拿到、看懂和用好数据，虽奠定了开放获取的基本原则与技术要求，但并未适配AI模型对训练数据的特殊要求。^[42]例如，大模型预训练需要海量连贯文本，指令微调要求精准的“指令—输出”配对，思维链训练需要逻辑步骤标注。^[43]由于这些面向机器认知学习的细粒度要求在传统开放原则中付之阙如，实践中开放的数据多为机器可读的原材料，缺乏支撑深度语义理解与复杂推理任务所需的AI就绪语料。对此，科学数据领域已开始在FAIR原则中纳入AI就绪维度^[23]，但公共数据领域在这方面还相对滞后。

其二，重“好用”而轻“用对”，公共性就绪程度

表2 数据管理和语料库建设相关框架（示例）

类别	相关框架
数据管理原则	FAIR（Wilkinson等，2016）、FAIR-R（Verhulst等，2025）、FAIRsFAIR（Huber，2025）、FAIR×FAIR（沈志宏等，2026）；CARE原则（Carroll等，2020）；《数据伦理框架草案》（美国联邦总务署，2020）；我国《国家数据标准体系建设指南》和《数据管理能力成熟度评估模型》（GB/T 36073-2025）
数据开放原则	政府数据开放8项基本原则（开放政府工作小组，2007）；《开放数据宪章》、开放数据五星标准；纽约大学治理实验室和世界银行相关标准等
AI就绪数据	AI-Ready Data（美国商务部，2025；世界银行，2025；开放数据研究所，2025；韩国行政安全部，2026）；AI-ready data framework（Majithia等，2026）；AIDRIN（Hiniduma等，2024）
高质量数据集	《人工智能高质量数据集建设指引》（国家数据局、中国信息通信研究院等，2025）、《高质量数据集建设指南》（TC609-5-2025-01）
AI数据质量标准	国际标准化组织ISO/IEC 5259系列标准、《人工智能数据集质量评价指标》（T/CQAE18003-2025）、《高质量数据集质量评测规范》（TC609-5-2025-04）
语料库技术规范	《人工智能语料库建设导则》（T/SAIAS）系列团体标准（上海市人工智能行业协会，2024）、《IEEE Standard for Energy Market Corpus Construction Framework for Large Language Models Training》（IEEE Std 3834-2025）、《语料库通用技术规范》（ZYF001-2018）、《双语平行语料加工服务基本要求》（GB/T40035-2021）

不足。为回应AI训练的需求，“AI就绪”数据框架及相关质量标准应运而生。如开放数据研究所（ODI）、美国商务部指南、韩国行政安全部、世界银行等提出的“AI就绪”（AI Ready）数据框架，强调数据需经过准备、质量可控、文档完备^[44]；国内《高质量数据集建设指引》等文件亦明确了AI训练数据在准确性、完整性、一致性、多样性等方面的具体要求。这类框架推动AI数据准备从“可用”向“好用”演进，并发展出相应的质量标准。^[45,46]而以《人工智能语料库建设导则》为代表的原则为适配AI发展的语料资源加工处理贡献了技术标准。然而，这些原则标准的制定多基于效率与最优逻辑，旨在最大化AI模型的性能指标和经济价值，但若直接应用于公共语料资源的治理，其效率优先导向的内在局限便会凸显。例如，对“价值对齐”的要求多停留于负面内容过滤，缺乏对服务公共利益的主动建构与正面引导，对“公平普惠”“公益导向”的要求则缺乏深度整合。因此，这类标准虽然回答了“如何让数据更好用”的技术性问题，但未能系统回答“数据为谁服务、体现何种价值”等公共性问题。

可见，现有框架虽从不同侧面为AI公共语料库的建设与治理提供了参考，但也各有其解释盲区：传统数据管理与开放原则以及传统语料库标准，未能回应AI模型训练对语料“可理解”“可执行”等深层技术性就绪需求；而部分面向AI模型训练的数据管理原则与语料库建设标准，则对公共语料的公共性规范价值回应不足。因此，现有框架还缺乏对AI公共语料库技术适配与价值导向的整体关照，难以为其建设治理与价值实现提供完整指引。

（二）AI公共语料库的就绪框架

AI公共语料库的智能与价值“涌现”，需要将技术判断和价值判断有机整合。基于上述分析，本研究将AI公共语料库的“就绪”分为两个相互关联的层面：技术性就绪和公共性就绪。技术性就绪聚焦公共语料库面向AI模型训练“是否好用”，而公共性就绪则关乎其在价值层面“是否用对”，前者属于技术判断，后者侧重价值判断，两者共同构成AI公共语料库的就绪框架（参见图2）。

1. 技术性就绪

AI公共语料库的技术性就绪包含可发现性、可访问性、可用性、可训练性和可信性五个层层递进的维度。

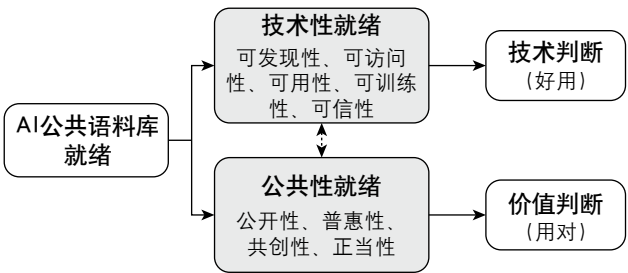


图2 AI公共语料库双主线就绪框架

（1）可发现性：能被“看见”

可发现性是指AI公共语料资源能被用户（含人类与AI系统）高效、准确地检索、定位与识别，是技术性就绪的起点。它呼应FAIR原则中的“可发现”要求，强调语料具备完备的元数据描述（如主题、规模、格式、来源、更新频率）和统一的标识体系，使潜在使用者不仅能找到数据实体，还能理解其内容主题、背景语境与潜在应用价值。^[40,42]对于AI训练而言，可发现性还意味着元数据应当采用标准化的格式和词汇，支持机器自动读取和解析，以便于通过自动化工具进行语料检索与筛选。

（2）可访问性：能被获取

可访问性关乎AI公共语料资源在权限、接口、协议等技术层面上的可获取程度。其要义是语料资源以适当的方式发布和托管，确保目标用户能够便捷地获取并使用。^[40]通常需要具有提供访问、持续更新和长期存留的能力，使数据在未来依然可被调取用于新的用途。同时，可访问性要求提供完善的文档、易用的工具和持续的运营支持，以降低获取门槛。

（3）可用性：能被处理

可用性指AI公共语料资源能够被AI工具读取和进行基础处理的便利程度，其核心在于确立语料的“形式质量”。相关研究表明，解决数据质量问题往往占据数据分析项目的绝大部分时间，如果语料包含大量噪音、错误或不一致，将严重影响后续模型的可靠训练。^[47]因此，语料库构建流程中必须设置专门的预处理阶段。这不仅包括统一文件格式、纠正明显错误，更关键的是大幅提升数据的“信噪比”，滤除无意义的乱码与重复内容、进行基础性脱敏处理等。可用性原则还要求提高数据的互操作性，即采用统一的格式标准与接口规范，使多源、多模态语料能够被无缝集成与协同处理。

(4)可训练性：能被学习

可训练性是AI公共语料库区别于传统数据资源的核心特性。相较于可用性，可训练性更关注语义精度与逻辑一致性，直接指向语料的“内容质量”。它不仅要求数据在技术格式上达标，更需在规模性、多样性、准确性、时效性、逻辑性以及任务适配性等方面支撑AI训练的要求。^[31]例如，预训练语料需具备充足规模并丰富多样性，而微调语料则应拥有高精度任务标注与清晰逻辑链条。由于AI模型的实际性能受制于底层训练数据的质量，数据残缺、误差与适用性不足等问题，均会造成模型输出不可靠、决策结果存在偏差等问题。^[48]这一原则强调对数据质量进行持续监控、评估和改进，更需要语言学、数据领域专家与AI工程师的深度协同。

(5)可信性：能被安全合规使用

可信性是保障AI公共语料资源在来源、版权、隐私、安全等方面可被验证、可被信任的关键，是技术就绪的底线保障。随着大规模语料被用于训练模型，数据来源不透明和许可不当引发了紧迫的法律与合规关切。有研究表明，当前超过70%的数据集缺失许可证信息，面临数据集许可属性的“危机”。^[49]对此，语料可信性需要在工程层面夯实两项基础：一是来源可溯与许可可验，即为语料库配置统一的开放许可协议，并将其嵌入元数据供程序自动校验，让版权信息可被机器识别、来源链条清晰可查；二是隐私保护，通过自动化脱敏算法识别并擦除个人敏感信息，从源头守住隐私底线。此外，面对大模型训练中日益严峻的“数据投毒”威胁，还有必要延伸至数据完整性校验、内容防伪与抗污染鲁棒性，确保语料从源头到使用全程可追溯、可验证。

2.公共性就绪

AI公共语料库的公共性就绪则包含公开性、普惠性、共创性与正当性四个互为支撑的维度。

(1)公开性

公开性特指AI公共语料库治理与运营过程的透明度，这是公共性就绪的基础。它并不是指AI公共语料资源本身是否向社会无条件公开，而是强调建设运营过程的程序透明。具体而言，谁参与管理决策、谁提供服务、依据何种标准、经过哪些环节、最终实施结果如

何，这些关键信息均应向社会公开。其核心在于构建贯穿公共语料库建设运营全流程的信息披露与溯源机制，确保AI公共语料库“运行在阳光下”。具体来看，在决策层面，语料库的立项依据、建设目标、资金来源、责任主体应公开可追溯；在过程层面，数据来源、标注规范、质量评估、伦理审查等关键加工环节应有据可查；在产出层面，资源使用情况、主要利用主体和应用领域等信息应定期披露，确保公共投入的使用效果可监督可检验。^[15]只有将公共语料库的运营逻辑与过程向社会适度敞开，方能有效防范公共资源的暗箱操作，切实维护公共利益。

(2)普惠性

普惠性特指AI公共语料库供给结果的公平可及。它强调将公共语料库作为公共资源向社会进行普惠性供给；若其供给仅面向少数部门或企业的利益，其公共性便无从体现。其核心要求包括：一是供给的普遍性，依托多元化渠道和便捷可及的服务方式，拓宽公共语料供给的覆盖范围；二是获取的公平性，杜绝公共语料资源的供给异化为谋取部门利益或市场盈利的工具，严防用纳税人的资金“补贴”特定群体的开发利用需求；三是结果的包容性，在保障机会平等的基础上，进一步关注结果公平，保障“长尾需求”和弱势群体公平获取、利用公共语料的实际成效，确保“数据红利”无偏无倚、不厚此薄彼。

(3)共创性

共创性强调治理结构的开放性与价值创造过程的互动性，是AI公共语料库实现多元协同发展的核心支撑机制。公共语料库不应成为部门自建自用、封闭排他的“知识小作坊”，更不能沦为一次性交付、以营利为导向的数据仓库，而应打造为多元主体共建、共治、共享的公共性基座。^[50]大模型能力的持续突破，离不开公共部门、社会企业、开源社区等多元主体的数据资源融合，更需要整合各领域的资源、能力和知识，实现协同训练、共建提升。^[18]这就要求构建权责清晰的治理体系，建立多元协商、价值共创的合作治理机制。政府需发挥“召集者”“守门人”与要素供给者的核心作用，市场主体贡献工程技术能力，学术机构提供专业理论支撑，社会公众依法行使监督权利。更为关键的是，使用者同时也是反馈者，其在开发应用中产生的新需求需反

向驱动语料迭代优化，形成“需求牵引供给、供给适配需求”的互动闭环。唯有各方主体协同投入、共享受益，才能打破封闭化的“知识壁垒”，构建能伴随技术与社会持续生长的AI公共语料训练基座。

(4)正当性

正当性特指公共语料库内容与法律、伦理及社会价值的一致性要求，是构成AI公共语料库的底线保障。这一原则要求构筑双重规范防线：一是坚守法律合规底线，依托制度手段保障语料来源清晰、版权规范、隐私保护到位，从源头确保语料内容不侵犯用户合法权益；二是强化伦理价值对齐与偏见防控。有研究指出，若不对原始语料进行有效管控，模型可能会像“随机鹦鹉”（Stochastic parrot）一样，机械放大人类社会的偏见、歧视与仇恨。^[51]这正是学界和实务界大量探讨“负责任的AI”的深层动因。因此，公共语料需主动规避可能隐含的性别歧视、种族偏见、仇恨言论等不良内容，从源头防范结构性偏见被模型习得与传播。

四、结论与启示

在AI范式下，传统语料库的内涵已发生深刻变革：从服务于人类语言研究的“静态样本”，演进为驱动机器智能涌现的“动态要素”。面对这一时代背景，本研究系统构建了AI公共语料库的概念内涵与就绪框架，为后续相关理论研究与实践探索提供基础性支撑。

首先，界定了AI公共语料库的技术性内涵，厘清其与任务型数据集、训练数据、高质量数据集等相关概念的关联与差异，进而从本体属性、权利属性、价值属性三个维度解析其公共性特征。在此基础上，构建了“技术性就绪—公共性就绪”双主线框架，其中，技术性就绪涵盖可发现性、可访问性、可用性、可训练性与可信性五个核心要素，聚焦语料质量与技术适配性；公共性就绪包括公开性、普惠性、共创性与正当性四个关键维度，强调语料的公平供给与公共价值。两条主线相互支撑，共同促进公共语料资源的高质量开发利用与价值实现。

基于上述框架，建议围绕技术性就绪与公共性就绪两条路径来系统提升我国的AI公共语料库就绪度。为达成技术性就绪，需通过体系化的语料治理，将原始、异

构的公共数据加工为符合AI模型训练要求的优质语料资源，这一过程需实现制度保障与技术手段的深度融合，切实筑牢语料质量根基，为AI模型训练提供可靠支撑。为实现公共性就绪，需通过多元化的开发利用路径，确保语料公平、合规地赋能经济社会高质量发展。囤积状态下的AI公共语料库难以发挥其核心价值，其价值实现关键在于多元化开发利用。可结合语料的敏感性、价值密度与服务目标，将公共性就绪所要求的公开透明、公平可及、多元参与、内容向善等原则落到实处。具体可通过四条路径“内部共享以赋能智能政务高效运转、普惠开放以激发社会创新动能、授权运营以依托市场机制激活要素价值、生态共创以推动多方主体协同共治”推进，从而形成全方位、多层次的公共语料开发利用格局。

本研究聚焦公共数据资源与AI训练语料的交叉领域，旨在回应智能时代公共部门向社会供给高质量AI公共语料资源，这一新型公共服务形态的前沿议题，填补当前该领域研究的薄弱环节。研究结论有助于深化智能时代对公共语料资源的理论认知，也为构建AI就绪且公平可及的公共语料基础设施提供实践参考。受限于该议题的前沿性与复杂性，本研究仅对AI公共语料库开展了初步理论探索，相关领域仍存在诸多可深入挖掘的研究空间。未来研究可进一步加强实证分析，深入调研AI公共语料资源的社会需求、现实挑战与典型实践案例；同时，可加强跨国比较研究，借鉴国际实践经验，提炼适配中国情境的AI公共语料库发展模式；此外，还可聚焦AI公共语料库的治理规则、激励机制、市场化运作等关键议题，为破解公共语料资源“供不出、流不动、用不好”等现实困境，提供更具针对性的理论支撑与可操作的实践对策。

参考文献：

[1]Kusumegi K, Yang X, Ginsparg P, et al. Scientific production in the era of large language models[J]. Science, 2025, 390(6779): 1240–1243.
[2]施锦诚，刘洋，朱凌. 人工智能驱动创新管理：理论演进、研究框架及未来展望[J]. 管理世界，2025，41(12): 198–224.

- [3]Xiao C, Cai J, Zhao W, et al. Densing law of LLMs[J]. Nature Machine Intelligence, 2025, 7(11): 1-11.
- [4]Villalobos P, Ho A, Sevilla J, et al. Will we run out of data? Limits of LLM scaling based on human-generated data[EB/OL]. [2026-03-25]. <https://doi.org/10.48550/arXiv.2211.04325>.
- [5]Shumailov I, Shumaylov Z, Zhao Y, et al. AI models collapse when trained on recursively generated data[J]. Nature, 2024, 631(8022): 755-759.
- [6]林韬. 我国AI训练数据生产流通的制约因素与应对策略研究[J]. 中国科学院院刊, 2025, 40(04): 672-680.
- [7]U. S. Department of Commerce. Generative Artificial Intelligence and Open Data: Guidelines and Best Practices[R/OL]. [2026-03-25]. <https://www.commerce.gov/news/blog/2025/01/generative-artificial-intelligence-and-open-data-guidelines-and-best-practices>.
- [8]The White House. Winning the AI race: America's AI action plan[R]. [2026-03-25]. <https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf>.
- [9]夏义堃, 宋佳, 纪昌秀. 高价值数据集开放与复用的内涵边界与实践进阶分析[J]. 电子政务, 2024 (06): 54-68.
- [10]Ministry of the Interior and Safety (MOIS). MOIS Selects TOP 100 High-Value Public Data and Deliberates the 5th Basic Plan for Public Data [EB/OL]. [2026-02-08]. https://www.mois.go.kr/frt/bbs/type010/commonSelectBoardArticle.do?bbsId=BBSMSTR_000000000008&nttId=122913.
- [11]Nikiforova A, Lnenicka M, Luterek M, et al. Theorizing the evolution of public data ecosystems: An empirically grounded multi-generational model and future research agenda[J]. Government Information Quarterly, 2025, 42(03): <https://doi.org/10.1016/j.giq.2025.102062>.
- [12]吕文增, 杨正军. 边界重构: 生成式人工智能对政府数据开放的新挑战[J]. 信息通信技术与政策, 2025, 51(12): 33-39.
- [13]赵精武. 论人工智能训练数据高质量供给的制度建构[J]. 中国法律评论, 2025(01): 92-104.
- [14]苏宇. 人工智能要素的政府供给: 法理基础与制度因应[J]. 财经法学, 2025(02): 77-93.
- [15]孙清白. 论人工智能大模型训练数据风险治理的规范构建[J]. 电子政务, 2024(12): 41-52.
- [16]刘霜, 张潇月. 生成式人工智能数据风险的法律保护与规制研究——以ChatGPT潜在数据风险为例[J]. 贵州大学学报: 社会科学版, 2023, 41(05): 87-97.
- [17]陈强, 邢窃窃. 面向生成式人工智能的语料库建设: 现状、挑战与未来策略[J]. 社会科学, 2025(06): 177-192.
- [18]张凌寒. 加快建设人工智能大模型中文训练数据语料库[J]. 人民论坛·学术前沿, 2024(13): 57-71.
- [19]Domagalski M J, Lu Y, Pilozi A, et al. Preparing clinical research data for artificial intelligence readiness: Insights from the National Institute of Diabetes and Digestive and Kidney Diseases data centric challenge[J]. Journal of the American Medical Informatics Association, 2025, 32(10): 1609-1616.
- [20]李兴腾, 冯锋, 黄鹂强. 突破人工智能大模型的“数据瓶颈”——构建国家级语料库运营平台的思考[J]. 中国科学院院刊, 2025, 40(03): 522-529.
- [21]Hiniduma K, Byna S, Bez J L. Data readiness for AI: A 360-degree survey[J]. ACM Computing Surveys, 2025, 57(09): 1-39.
- [22]Citron D T, Ginsparg P. Patterns of text reuse in a scientific corpus[J]. Proceedings of the National Academy of Sciences, 2015, 112(01): 25-30.
- [23]沈志宏, 朱小杰, 朱国亮. FAIR×FAIR: AI时代科学数据的共享与利用[J]. 数据分析与知识发现, 2026, 10(01): 31-47.
- [24]张晓林. “人工智能+”背景下的高质量数据集建设: 图书馆的机遇与挑战[J]. 中国图书馆学报, 2025, 51(06): 4-17.
- [25]Yang J, Tan Y, Hu W, et al. Are Task-Specific Datasets the Be-All and End-All for LLM Performance?[C]//2025 International Joint Conference on Neural Networks (IJCNN). IEEE, 2025: 1-8.
- [26]Sinclair J. Corpus, Concordance, Collocation[M]. Oxford: Oxford University Press, 1991: 171.
- [27]McEnery T, Hardie A. Corpus linguistics: Method, theory and practice[M]. Cambridge: Cambridge University Press, 2012: 1-3.
- [28]丁富登, 毛文蔚, 陈媛, 等. 基于大语言模型的政务智能化应用: 技术现状、挑战与未来展望[J]. 工程科学学报, 2026, 48(01): 98-113.
- [29]Liu Y, Cao J, Liu C, et al. Datasets for large language models: A comprehensive survey[J]. Artificial Intelligence Review, 2025, 58(12): 403.

- [30]Silberztein M. The limitations of corpus-based methods in NLP[M]//Linguistic resources for natural language processing: On the necessity of using linguistic methods to develop NLP software. Cham: Springer Nature Switzerland, 2024: 3–24.
- [31]Zhang Y, Sun S, Galley M, et al. DialoGPT: Large-scale generative pre-training for conversational response generation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 270–278.
- [32]Incelli E. Exploring the Future of Corpus Linguistics: Innovations in AI and Social Impact[J]. International Journal of Mass Communication, 2025, 3: 1–10.
- [33]Schuhmann C, Beaumont R, Vencu R, et al. Laion-5b: An open large-scale dataset for training next generation image-text models[J]. Advances in Neural Information Processing Systems, 2022, 35: 25278–25294.
- [34]ISO. ISO-IEC/ISO SMART[EB/OL]. (2022-09-16)[2026-02-24]. <https://www.iso.org/smart>.
- [35]国家市场监督管理总局, 国家标准化管理委员会. 机器可读标准能力等级模型与评估[EB/OL]. (2025-03-28)[2026-02-27]. <https://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=A2637679FA8163227FD4C0AF476DE5EC>.
- [36]江小涓. 数智时代的秩序重构与治理合作：合理合意双重目标[J]. 管理世界, 2025, 41(05): 1–14, 58, 241.
- [37]支振锋. 公共数据资源开发利用促进及其体系化制度构建[J]. 行政法学研究, 2025(04): 55–69.
- [38]王锡锌, 王融. 公共数据概念的扩张及其检讨[J]. 华东政法大学学报, 2023, 26(04): 17–27.
- [39]夏义堃. 数字环境下公共数据的内涵、边界与划分原则分析[J]. 中国图书馆学报, 2024, 50(02): 100–114.
- [40]Wilkinson M D, Dumontier M, Aalbersberg I J, et al. The FAIR Guiding Principles for scientific data management and stewardship[J]. Scientific Data, 2016, 3(01): <https://doi.org/10.1038/sdata.2016.18>.
- [41]郑磊. 开放不等于公开、共享和交易：政府数据开放与相近概念的界定与辨析[J]. 南京社会科学, 2018(09): 83–91.
- [42]Raza S, Ghuge S, Ding C, et al. FAIR enough: Develop and assess a FAIR-compliant dataset for large language model training[J]. Data Intelligence, 2024, 6(02): 559–585.
- [43]Guo D, Yang D, Zhang H, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning[J]. Nature, 2025, 645(8081): 633–638.
- [44]Open Data Institute. A Framework for AI-Ready Data[EB/OL]. [2026-02-26]. <https://theodi.org/insights/reports/a-framework-for-ai-ready-data/>.
- [45]Hiniduma K, Byna S, Bez J L, et al. AI data readiness inspector (AIDRIN) for quantitative assessment of data readiness for AI[C]//Proceedings of the 36th International Conference on Scientific and Statistical Database Management, 2024: 1–12.
- [46]Majithia N, Carey-Wilson T, Simperl E, et al. An actionable framework for AI - ready data[J]. AI Magazine, 2026, 47(01): <https://doi.org/10.1002/aaai.70054>.
- [47]Dasu T, Johnson T. Exploratory Data Mining and Data Cleaning[M]. Hoboken, NJ: John Wiley & Sons, 2003: 25–26.
- [48]Mohammed S, Budach L, Feuerpfeil M, et al. The effects of data quality on machine learning performance on tabular data[J]. Information Systems, 2025, 132: <https://doi.org/10.1016/j.is.2025.102549>.
- [49]Longpré S, Mahari R, Chen A, et al. A large-scale audit of dataset licensing and attribution in AI[J]. Nature Machine Intelligence, 2024, 6(09): 975–987.
- [50]何哲, 曾润喜, 郑磊, 等. ChatGPT等新一代人工智能技术的社会影响及其治理[J]. 电子政务, 2023(04): 2–24.
- [51]Bender E M, Gebru T, McMillan-Major A, et al. On the dangers of stochastic parrots: Can language models be too big?[C]//Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021: 610–623.

作者简介:

郑磊, 复旦大学国际关系与公共事务学院教授, 博士生导师, 数字与移动治理实验室主任, 研究方向包括数字政府与数字治理、公共数据开放利用、移动政务服务、城市数字治理等。

杨涛, 复旦大学国际关系与公共事务学院博士研究生, 复旦大学数字与移动治理实验室研究助理, 研究方向为公共数据开发利用、数字公共服务、人工智能治理等。