人工智能公共语料库建设、共享与开放：困境、成因与政策路径*

王翔[①②] 何丹[③] 郑磊**[②④]

①广州南方学院公共管理学院

②复旦大学数字与移动治理实验室

③杭州市大数据管理服务中心

④复旦大学国际关系与公共事务学院

摘 要：共享开放的人工智能公共语料库有助于促进人工智能技术研发与应用普惠，被视为塑造未来数字竞争力的新型战略资源。当前，我国人工智能公共语料库建设、共享与开放仍处于起步探索阶段，在实践中面临诸多困境与挑战。以我国地方政府实践为研究样本，采用案例研究方法，系统探索我国人工智能公共语料库建设、共享与开放的发展现状、现实困境及其形成机制。研究发现，当前我国人工智能公共语料库的建设共享开放主要面临“不懂”“不愿”“不能”“不敢”四方面的困境，其成因分别为“认知不足与概念混乱、动力不足与激励缺乏、能力短板与协作障碍、安全风险与合规顾虑”，并形成一条“认知—动力—能力—责任”的因果递进关系，最终产生“锁定效应”。基于以上研究发现，建议从理念认知提升、激励机制创新、技术能力投入、安全合规保障等方面多管齐下，推动人工智能公共语料库的高质量建设、高效共享与有序开放。

关键词：人工智能；语料库；数据共享；数据开放；授权运营

DOI：10.16582/j.cnki.dzzw.2026.04.002

一、问题的提出

人工智能（Artificial Intelligence，简称AI）正成为引领新一轮科技革命和产业变革的战略力量，大语言模型的发展尤其受到关注。在驱动大模型发展的算法、算力和数据三要素中，算法的边际创新收益趋于递减，算力的普惠化趋势加快，而高质量语料数据的战略性地位却愈发显著。近年来，大规模预训练模型展示了前所未有的能力“涌现”效应，其性能跃升在很大程度上归功于训练语料数据规模和质量的指数级提升。可以说，高质量语料已成为提升AI模型性能的关键要素。各国科技企业和研究机构纷纷加速建设大型语料库，政府层面也投入大量资金和开放公共数据支持AI语料资源的建设。共享开放的AI公共语料库不仅能促进AI技术研发与应用普惠，也被视为塑造未来数字竞争力的新型战略资源。

我国政府高度重视数据要素对AI发展的支撑作用。2025年8月发布的《国务院关于深入实施“人工智能+”行动的意见（国发〔2025〕11号）》要求，“加强数据供给创新。以应用为导向，持续加强人工智能高质量数据集建设”。《生成式人工智能服务管理暂行办法》要求“推动公共数据分类分级有序开放，扩展高质量的公共训练数据资源”。各级地方政府也纷纷将语料数据供给纳入数字政府和AI产业规划，布局公共数据仓库、行业语料库等项目。

然而，尽管有政策指导和需求牵引，我国AI公共语料库建设仍处于起步阶段，各级公共管理和服务机构虽然积累了海量数据资源，但这些数据大多最初是为满足政务服务或管理需求而收集的，其侧重于解决具体问题的任务导向型数据结构在数据清洗、数据标注、数据更新等方面还不能满足大模型的需求，不完全符合AI语料的特性。另外，语料数据的共享开放程度仍然不高，公

*基金项目：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—17

共语料库的数据源复杂多样，呈现分散化、碎片化状态，缺乏统一的数据平台和语料协调机制，难以形成高质量、大规模语料共建共享的格局。

那么，当前我国各级地方政府在AI公共语料库建设、共享与开放中进展如何？遇到哪些困境？这些困境的成因及其作用机制如何，与传统公共数据治理与开发利用存在哪些差异？这些都是亟待关注和研究的问题。基于上述背景，将聚焦AI公共语料库这一新课题，以我国地方政府实践为研究样本，采用案例研究方法，梳理当前AI公共语料库建设的现状、问题与障碍及其形成机制，并探讨未来的政策路径，为我国AI公共语料库的建设、共享与开放提供参考。

二、文献综述

（一）面向AI训练的公共语料库与传统数据集的区别

AI公共语料库一般是指由各级党政机关、企事业单位依法履职或提供公共服务过程中产生的，可用于训练、测试或优化AI模型的语言数据的集合。AI公共语料库是公共数据的一个子集，在国家数据局提出的“通识数据集”“行业通识数据集”和“行业专识数据集”三个分类中，AI公共语料库可以对应于第二类“行业通识数据集”。[1]

与传统数据的治理相比，语料数据的治理具有显著的区别。语料数据治理需要将数据从“能看、能查、能用”推进到“能训练、能泛化”，并注重语料的来源合法性、许可边界、清洗去重、标注一致性、分布代表性、版本可追溯等一整条可追溯、可问责的数据生命周期[2]。否则，语料即使看似可得，也难以真正进入可复用、可评测、可监管的训练与应用闭环。

进一步而言，语料数据的共享开放也与传统公共数据的共享开放有较大差异。从共享的角度来说，传统公共数据共享的核心是提升公共数据的可获取性、可交换性与可复用性，以支撑政务协同、公共服务和社会创新；而语料数据共享的核心，则是提升公共数据的可训练性、可解释性与可追责性，使其能够进入AI模型开发、评测与部署流程[3]。前者主要面向人和业务系统，后者直接面向模型能力形成，因此对数据的语义标注、质量筛选、来源说明和合规治理等提出了更高要求。从开放的角度来说，公共数据开放通常被定义为提升透明度、公共服务改进与社会创新的制度与技术组合，并强调目录编制、元数据标准、接口与平台建设等“可被外部主体稳定获取和再利用”的能力供给[4]。而语料数据的开放与传统公共数据开放相比，有两个主要的变化：一是开放目的的变化。公共数据更强调“可读、可用、可复用”，其价值实现路径依赖人类的解读、分析和开发；而语料更强调“可训练、可泛化、可控、可追溯”，其核心目标为赋能机器的“学习”与“推理”[5]。二是开放对象的变化。与传统的结构化数据的开放相比，语料往往是非结构化的，还包括多模态数据，其治理成本从传统的“编目与脱敏”跃迁到清洗去重、标注体系、分布均衡、版本迭代与回溯[6]。

（二）语料库建设、共享、开放的困境及应对

当前，国际主流的AI公共语料库在建设、共享和开放过程中普遍存在着个人隐私保护挑战、滥用与授权合规风险、偏见与歧视风险、治理成本居高不下四个方面的困境。

1.个人隐私保护挑战

语料数据往往包含各类能够识别个人身份的信息，如果在语料库建设、共享、开放过程中缺乏有效的脱敏技术和管理措施，极易造成个人隐私泄露，侵犯公民合法权益。然而在实际操作中，彻底剔除海量语料中的敏感信息极为困难。[7]常见做法是通过匿名化、替换等技术手段对明显的敏感字段进行处理，同时以许可协议方式限制数据用途。但匿名化无法保证百分之百有效，协议约束又缺乏对恶意违规者的强制力。[8]因此，即便许多公共语料库标榜已做脱敏，仍被发现残留个人信息隐患。隐私保护作为AI语料库治理的首要难题之一，既需要技术上的改进，更需要法律和管理手段并举，明确语料匿名化标准，建立违规使用惩戒机制，才能在共享开放与隐私安全之间取得平衡。

2.滥用与授权合规风险

公共语料数据共享和开放的初衷在于促进技术创新和学术研究，但现实中防范数据滥用极其困难。当语料

廉价甚至免费可得，总有机构会将其用于超出许可范围的商业用途甚至违法用途。例如，不少公开数据集声明仅限科研机构之间共享，但实际上缺乏有效监管，导致内部共享实际上变成了对外开放。有开放语料被不法分子用于训练生成虚假信息、深度伪造视频，造成社会危害。语料数据的开放程度越高，越难控制数据用途；而严格管控用途又与开放初衷相矛盾。[9]解决之道需要完善法治环境，如明确AI训练的数据使用边界、加强对敏感应用的监管，也需要行业自律与技术监测手段的结合。只有在法律上划定合理使用与侵权滥用的界限，并辅以一定的检测和追溯机制，才能在鼓励开放创新的同时降低数据被不当利用的风险。

3.偏见与歧视风险

语料数据不可避免地带有各种偏见，如果未经治理就用于训练模型，可能导致AI系统重现甚至放大这些偏见，产生算法歧视和不公正结果。偏见可能源自数据分布不均，也可能在标注过程中被标注人员引入。[10]这些偏见问题对AI伦理提出了挑战，但要消除偏见几乎不可能，只能在构建和使用数据集时尽量评估、减小和纠正偏差。常见做法包括：在数据集构建阶段，设定配额，保证不同人种、性别、地区样本的比例平衡；在数据集发布阶段，附带偏见分析报告和数据概况说明，让使用者知悉数据偏差并采取补救措施；在模型评估阶段，针对各亚群体分别测试性能，如发现明显不公，则回溯改进训练数据或模型算法。[11]

4.治理成本居高不下

完善的大规模语料库治理意味着高投入和复杂权衡。维护一个不断更新的语料库本身需要大量持续的人力物力投入。对包含亿万级数据的开放语料库，要逐条审查其中的版权、隐私、有害内容，几乎是不可能完成的任务。机器自动筛选不可能完全准确，人工复核又工作量惊人，远超学术机构或公益组织所能负担。[3]因此，很多公共语料库不得不采取折中策略，只做基本过滤（如明显的色情、暴恐内容），而对隐蔽的侵权或隐私问题鞭长莫及。同时，除了直接的经济成本，法律不确定性带来的决策成本也不可忽视。语料库运营方不得不时刻关注各国立法和判例动态，寻找风险最低的途径。这在无形中增加了运营难度和心理成本，降低了开放的积极性。长远来看，要缓解治理成本压力，需要弹性和前瞻性的治理策略：如优先处理高风险内容，动态监测争议数据；积极推动立法完善，争取更明确宽松的数据使用规则；研发更高效的技术工具以替代人工；依靠社区协作，共享治理经验、标准规范，降低单个机构的负担。[12]唯有如此，才能让AI公共语料库在可持续的投入下健康发展。

（三）文献评述

综合以上研究可以发现：第一，现有研究对于公共数据的治理、共享与开放问题已有较为系统的讨论，但相对缺乏对公共AI训练语料这一新型治理对象的专门研究，尤其对非结构化、多模态语料的关注不足。事实上，AI公共语料库既不是传统公共数据开放的简单延伸，也不是一般意义上的技术语料库建设，而是公共数据治理逻辑与AI训练数据治理逻辑叠加之后形成的新型治理议题，已有文献对于这一交叉地带的关注不够充分。第二，国外公共语料库建设主要由企业、科研机构和非政府组织牵头，现有研究对于语料数据治理的讨论也多置于“技术–产业”语境，其背景是欧美平台治理与科研开放情境，缺少基于中国制度环境的解释框架，对我国公共语料库建设与利用背景下的关键约束与行动逻辑尚未充分讨论。

基于此，本文拟在既有研究基础上，通过对中国AI公共语料库建设与共享开放的案例分析，揭示中国AI公共语料库建设的现实障碍及其成因机制，并据此提出针对性的政策建议。

三、研究方法

本研究采取案例研究法，选取东部某省会城市X市作为案例研究对象。该市是我国数字治理的领先城市之一，公共数据共享与开放工作起步早、基础好，数字经济和AI产业发达。该市公共管理和服务机构累积了海量多源数据，已初步开展政务大模型训练、行业语料库等探索，但也面临理念、技术、机制、安全等方面的挑战。因此，选择该市作为案例，有助于深入剖析地方政府推进AI公共语料库工作的现状和痛点，进而为全国其

他地方提供参考。

本研究主要采用质性方法获取资料。一是文本分析法，收集国家法律法规、政策文件、行业报告等，了解X市建设AI公共语料库所面临的法律法规政策环境。二是焦点小组讨论，组织X市政府数据管理部门、行业主管部门、科技企业以及承担公共服务职能的企事业单位等进行焦点小组讨论共14场，了解相关主体对语料库建设的需求、举措与困难。三是参与式观察法，研究团队于2025年3—8月参与了X市AI公共语料库建设工作，在这一过程中，针对各方对AI公共语料库建设的认识、动机、需求、顾虑、投入的人力物力资源等方面进行了深入观察，收集了大量一手资料，积累了大量观察记录。之后，采用扎根理论方法分析资料，按照开放式编码、主轴编码、选择性编码三个阶段对通过以上方法采集到的数据进行了质性分析。

四、案例背景

为掌握AI公共语料库建设的基础和现状，研究团队对X市的相关工作进展进行了调研。近年来，X市在公共数据开发利用和智能应用方面走在全国前列，从政策、需求、实践等方面为研究提供了丰富的案例资料。

（一）政策驱动

X市AI公共语料库建设得到国家和省市政策的明确指引和要求。首先，国家层面的多个文件强调高质量数据集建设和公共数据授权开放，省里的公共数据条例对“公共数据开放与利用”进行了规范，成为市级工作的遵循依据。其次，该省还在全国率先开展数据知识产权登记试点，探索数据确权和价值转化的新路径，并于2025年6月获批国家数据要素综合试验区，计划在培育经营主体、繁荣壮大数据市场等方面开展先行先试，全面释放实体经济和数字经济融合效能。最后，市政府也在2025年政府工作报告中提出“加快推进公共数据开发利用和有序开放”，并将建设政务大模型训练场作为重点成果。这些政策要求赋予各部门开展语料库建设的驱动力。

（二）多元需求

调研发现，X市公共管理和服务机构以及科技企业对语料数据的需求十分迫切且多样。政府部门为建设数字政府、城市大脑等工程，需要海量政务文本（政策文件、办事指南、热线对话等）及多模态数据（监控视频、物联网感知数据等）来训练智能问答、智能审批、风险预警等模型，从而提升治理能力和服务水平。同时，国有企事业单位着力推进产业大模型和业务流程智能化改造，希望打通分散的数据“烟囱”，形成城市级数据底座，将历史业务数据转化为行业语料供AI模型学习。本地科技企业则渴求更大规模、更高质量的外部数据来提升其算法产品性能，表示“数据越多越好”，而当前自有的数据远不能满足模型训练需要。由此可见，在政策和应用双重驱动下，X市各主体对构建共享的AI语料库有普遍共识和内在动力。

（三）先行探索

在上级政策和自身需求的驱动下，X市各部门在AI公共语料库建设方面已取得一些阶段性成果。一是基于一体化智能化公共数据平台，初步汇聚了全市各部门的政务数据资源目录，并试点将部分非结构化文本纳入管理，为语料库建设打下基础。二是已归集全市70余个门户网站的文本、图片、视频等多模态数据并进行治理，拟在数据开放平台对外开放。三是已建成全国首个政务模型训练场，支持不少于五个单位同时进行模型训练任务，已开通训练账号900余个，已支持30多个部门训练任务。四是已完成知识数据、模型训练、智能编排、模型评测、应用发布、模型安全等六大中心的建设，基本构建形成训练推广一体、集约高效、监测熔断的垂域模型训练生产体系。五是各部门依靠现有语料数据建设了一批政务领域智能应用，市发改、公安、民政、卫健、医保等部门已经上线的AI智能体近20个，培育中的有30余个。

需要指出的是，上述成果多为各单位局部推进，缺乏全市层面的统筹，语料数据利用仍处于零散和初级阶段。X市有关部门负责人也意识到，尽管现在一些语料库建设已有成效，但面对更高层次的大模型训练需求，这些语料数据远远不够，还需更加系统、更大规模的建设与共享开放。整体而言，X市已迈出AI公共语料库建设的第一步，但整体仍处于起步阶段，还存在不少的困

境与挑战。

五、公共语料库建设、共享与开放的困境机制及其成因

调研发现，X市AI公共语料库的建设、共享与开放主要面临“不懂”“不愿”“不能”“不敢”四个方面的困境，其成因分别为：认知不足与概念混乱、动力不足与激励缺乏、能力短板与协作障碍、安全风险与合规顾虑。四个方面困境的成因呈现出一条具有内在递进逻辑的因果链条并走向“锁定效应”，最终表现为一种整体性困境。

（一）“不懂”：认知不足与概念混乱

尽管经过多年的数字政府建设，X市各部门对公共数据资源的重要性已有共识，但对于AI语料这一全新事物仍缺乏充分认识。“过去我们认为只有结构化数据才有意义，所以花了大量精力做结构化，结果现在发现AI不一定需要结构化数据。”（访谈编码：ZF01[注1]）不少受访者坦言对语料库、数据集、知识库等概念理解模糊，甚至混为一谈，不清楚三者区别，“语料库和数据集实践中概念是有重叠的，大家一会儿说数据集一会儿说语料库”（访谈编码：ZF03）。很多业务人员认为日常工作中产生的文字记录只是留痕备查的“副产品”，没有意识到其可作为语料的价值。整体来看，各级干部对AI语料库在贯彻国家AI战略中的地位认识不足，缺乏主动推进的紧迫感。这种“对象不清、边界不清”的认知状态，导致一些部门对语料库建设的重要性认识不足，从而未能将其纳入日常工作规划，工作推进缺乏动力。

“不懂”的背后是认知不足与概念混乱，使得公共语料库建设、共享与开放的行动边界难以确立。公共管理与服务机构虽然已形成较成熟的数据治理话语体系与实践基础，但面对与语料库相关的新对象时，仍普遍存在概念混同、理解不一的问题。同时，一些单位仍将日常文本记录视为“留痕副产品”，尚未将语料库建设上升到支撑国家AI战略与数字政府能力跃迁的高度来理解与规划。这种“对象不清、边界不清”的认知状态，会导致对投入产出缺乏稳定预期，难以为后续协作与资源配置提供正当性基础，从源头上埋下动力不足的伏笔。

（二）“不愿”：动力不足与激励缺乏

语料数据治理涉及大量繁琐工作，如数据清洗、标注、更新等，需要业务部门和技术部门通力合作。但现实中各单位参与意愿不高。调研发现，许多部门没有明确考核压力去整理语料，主要靠领导推动和少数负责人的个人热情，缺乏长效激励机制。例如，某医院信息科反映，临床专家很少有动力进行数据标注，“标的人不懂，懂的人不标”（访谈编码：ZF06）。研究发现，“不愿”不只是简单的态度问题，而是在价值不确定与协作成本高企条件下的理性选择：在缺乏贡献溯源与激励安排的情况下，语料治理容易成为“高投入、低可见回报、强责任约束”的事务性负担。

由此造成公共数据开放力度不够，社会主体获取高质量语料数据仍然困难。有民营企业反映，有价值的语料数据大多需要通过授权运营方式使用，但现有的授权运营机制流程长、成本高，导致企业申请动力也不足，“数据价值随着时间流逝会耗损”，授权运营“三到六个月已经算快的了，有时要大半年甚至一年的时间”（访谈编码：QY03）。因此，“模型厂商需要重复采集和处理一些同质化数据，带来了资源的浪费”（访谈编码：QY01）。

（三）“不能”：能力短板与协作障碍

研究发现，X市一些公共管理和服务机构的语料库建设技术能力仍显不足，在语料数据编目、预处理、标注、更新、分布校准和多模态数据处理等方面还缺乏系统性能力。不少业务系统最初并未考虑到输出的数据将用于AI模型，缺乏规范标注，后续再进行治理代价高昂。各部门普遍反映缺少自动化的数据治理和标注工具，很多语料准备工作仍需人工完成，耗时费力且质量难保证。即使部门领导有心推进，也面临工具缺乏、人才不足的现实掣肘。智能问答与知识库需要随政策快速迭代，但目前缺乏及时、高效、便捷的更新方案，导致维护成本陡增，“上面一条政策更新后，我们业务部门和信息化部门就要忙上好几天”（访谈编码：ZF03）。同时，数据标准不统一的问题也较为突出，不同行业、

注1：访谈编码中ZF为公共管理和服务机构，QY为企业。

不同部门的数据格式、标签体系不一致，导致语料难以融合互通，“同一个词在不同部门，概念可能是不一样的”“老百姓说的话跟我们的术语也不一样”（访谈编码：ZF08）。又如，X市方言口音较重，但当前语音识别模型对当地方言适配不佳，多模态融合也还存在技术瓶颈。

上述现象表明，X市各部门对于AI公共语料库的建设、共享与开放所需具备的关键能力还未形成体系化供给。“不能”并非单纯的技术短缺，而是由“低动力–低投入–低能力”的累积过程所固化：越缺乏激励机制与资源投入，越难以建设专业化工具与专业队伍；越缺乏关键能力支撑，语料治理和共享开放的边际成本也就越高，反过来又会强化“不愿”的理性基础，形成能力建设中的路径锁定。

（四）“不敢”：安全风险与合规顾虑

各单位在共享开放语料数据时顾虑重重，主要担心数据安全、隐私泄露和法律责任风险。首先，公共部门缺乏明确标准来界定哪些语料属于敏感或涉密，担心一旦共享开放出去如果引发信息安全事件，责任难以划分。多个政府部门反映，“对于内容敏感的定义很难把握，脱敏工作谁来做、怎么做，都是问题”（访谈编码：ZF09）。其次，企业也顾忌当前法规对AI训练用数据的合法性没有明晰界定，使用外部数据存在侵权风险，“我们可以去采购很多的数据，但是我们不清楚数据使用的边界，什么数据能用什么数据不能用很模糊，不敢买或者买了不敢用”（访谈编码：QY02）。数据授权与使用的边界模糊，法律法规对模型训练的版权责任豁免尚未出台，图书、期刊、报纸等版权内容市场化流通渠道匮乏，数据用于AI训练的合规风险较高。这种既“不敢开放”又“不敢使用”语料数据的氛围严重阻碍了语料价值的释放。

“不敢”并不只是因为主观上的“谨慎”，而是安全风险与合规顾虑下的合理防御行为。对政府部门而言，敏感或涉密语料缺乏清晰界定，合规审核与责任认定标准不统一，使得“内容敏感如何定义、脱敏谁来做、怎么做”等关键问题难以落地，敏感词也难以穷尽，进而放大“出事即问责”的心理预期。对企业而言，AI训练数据的授权与使用边界仍显模糊，侵权风险与合规不确定性抬升外部数据使用的谨慎程度，“合规边界不清晰，就像头上悬着一把剑”（访谈编码：QY02）；同时隐私计算等技术虽可实现但产业化成本高，也削弱了合规投入的可持续性。当缺乏自动化检测、可验证脱敏、全流程留痕与可追责的“技术–制度”组合时，即使主观上愿意探索，组织也很难形成“可被审计、可被证明、可被免责”的合规闭环，最终只能通过收缩共享开放范围、提高审批门槛、减少外溢风险来降低不确定性成本，由此出现“既不敢开放、也不敢使用”的双向抑制局面。

（五）小结

AI公共语料库建设推进的四方面困境，并非简单的四类问题并列叠加，而是呈现出一条具有内在递进逻辑的因果链条（参见图1）：认知不足导致目标、对象与方法边界不清，进而抬升语料治理与跨部门协作的组织成本并压低价值预期，诱发动力不足；动力不足进一步造成资源投入与制度供给不足，叠加既有信息化路径依赖与技术供给缺口，形成能力短板；能力短板又使脱敏、审核、监管等关键环节难以标准化与可验证，从而放大责任不确定性与合规风险感知，表现为风险规避的防御性行为选择；最终，风险规避又阻碍组织学习，形成负反馈回路，使这一线性链条走向“锁定效应”，

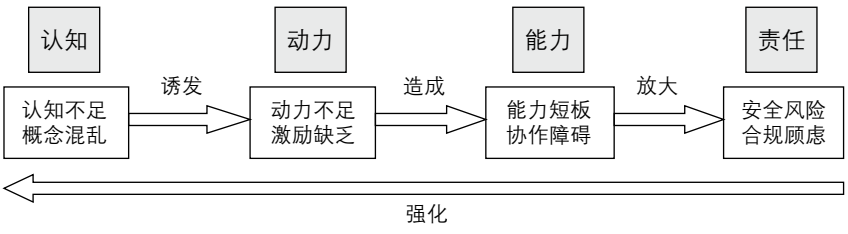

图1 AI公共语料库建设的因果递进关系

表现为整体性困境，即“认知不足—动力不够—能力不达—责任不小”，形成“不懂如何干、不愿主动干、没有条件干、不敢放手干”的局面。

六、结论与建议

（一）结论

本研究建构了我国地方政府在推进AI公共语料库建设、共享与开放过程中面临的“认知—动力—能力—责任”因果递进关系，揭示了认知偏差、激励缺位、能力约束与风险责任共同导致的“锁定效应”。与传统公共数据的治理、共享与开放相比，AI公共语料库建设、共享与开放过程中既存在老问题，也有新困境。老问题主要表现在语料的数据共享、开放动力不足等方面，而新困境突出表现在对语料数据及其重要性的认知不足，语料数据治理所需的技术能力不足，语料数据开放和利用面临的新型合规风险等方面，这些因素导致公共部门对推进该项工作的态度比传统公共数据的共享开放来得更加谨慎。

（二）建议

1.更新理念认知，增强语料意识

一方面，统一概念术语和标准。建议根据国家标准和业界共识，加快制定AI语料库相关名词术语指南，厘清语料库、数据集、知识库等概念边界，避免各自理解不一。同时，建立覆盖数据全生命周期的公共语料数据标准框架，统一各业务领域的数据格式、元数据和标签体系，消除跨部门的“数据语义差异”。另一方面，加强宣传培训。将AI应用和数据要素纳入教育内容，定期举办专题培训，提升领导干部和公务员的AI素养和语料意识。

2.创新工作机制，激发工作动力

一是高位协调推动。将AI公共语料库建设纳入政府重要议事日程，成立由地方主要领导牵头的工作专班或纳入现有工作机制，统筹协调语料库建设、共享、开放的推进落实。二是建立常态化工作机制。要求各单位在业务系统建设中同步规划语料数据的留存和输出，将“管项目”与“管数据”“管语料”相结合。探索将语料编目、归集、治理、开放等工作常规化、制度化，并纳入绩效考核。三是完善激励机制。设立专项资金，支持语料数据治理和标注等基础性工作投入。对贡献优质语料数据的部门和人员给予奖励，如在算力资源、数据使用权上倾斜支持。四是优化公共数据授权运营流程，通过减时间、减环节，提升公共数据授权运营“语料供给”质效。

3.完善技术平台，提升人员能力

一是提供语料数据公共服务能力。依托现有公共数据开放平台，增设语料数据专区，建立统一的语料资源目录，实现集中管理和动态更新。开发包括采集、清洗、标注、脱敏、合成、溯源于一体的语料处理工具箱，提供在线服务，避免各部门各自开发，降低技术门槛。同时制定语料质量评估指南，明确完整性、准确性、时效性等指标，并研发自动化的质量检测工具，保障语料库的数据品质。二是提升人员能力与素养。鼓励产学研合作，设立数据标注、语料治理方面的专项培训项目，提升从业人员技能。支持成立数据标注与语料服务联盟，整合社会化力量提供标准化、专业化的数据标注服务。通过联盟牵引，引进和培养一批熟悉AI与数据处理的复合型人才。鼓励高校、企业等多元主体联合攻关特定语料库建设任务，以成果共享的方式提高各方参与度，而非由政府一家包办。

4.强化合规保障，营造包容环境

一方面，推动合规保障体系建设。建立试错容错机制，支持各地区各部门在语料数据安全治理中就法律、行政法规未禁止的事项先行先试，在不违背公序良俗、危害国家和公共安全的前提下，充分审视动机、过程和结果，合理界定行政责任。另一方面，强化安全治理技术支撑。探索建立数据“避风港”，在确定规则红线的基础上，分级分类接受各类主体加入，开展数据流通利用机制创新。探索语料库流通“专有域”模式，在指定授权运营域内进行模型部署及训练推理，确保“原始数据不出域、数据可用不可见、数据可控可计量”。推动先进、高效、安全、合规的一体化流通利用基础设施体系建设，深化密态计算、区块链、数据沙箱等技术应用，探索建设支撑数据可信流通的密态计算基础设施，促进语料数据合规高效流通使用。

总之，我国AI公共语料库建设需要政府牵头、多方协同，在认识、机制、技术、监管各层面同步发力。只有综合施策，破解当前“不懂、不愿、不能、不敢”的难题，才能加快建立起高质量、可持续、负责任的公共语料资源体系，支撑我国AI及相关产业的健康发展。

参考文献：

[1]国家数据局：分三类建设高质量数据集赋能AI发展[EB/OL].(2025-04-30)[2026-03-13]. https://www.nda.gov.cn/sjj/swdt/mtsy/0430/20250430115203906544518_pc.html.

[2]Bender E M, Friedman B. Data statements for natural language processing: Toward mitigating system bias and enabling better science[J]. Transactions of the Association for Computational Linguistics, 2018(06): 587-604.

[3]Longpre S, Mahari R, Chen A, et al. A large-scale audit of dataset licensing and attribution in AI[J]. Nature Machine Intelligence, 2024, 6: 975-987.

[4]Janssen M, Charalabidis Y, Zuiderwijk A. Benefits, adoption barriers and myths of open data and open government[J]. Information Systems Management, 2012, 29(04): 258-268.

[5]饶高琦，胡星雨，易子琳. 语言资源视角下的大规模语言模型治理[J]. 语言战略研究，2023，8(04): 19-29.

[6]Zha D, Bhat Z P, Lai K H, et al. Data-centric artificial intelligence: A survey[J]. ACM Computing Surveys, 2025, 57(05): 1-42.

[7]Kushida C A, Nichols D A, Jadrnicek R, et al. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies[J]. Medical Care, 2012, 50(07): S82-S101.

[8]Chevrier R, Foufi V, Gaudet-Blavignac C, et al. Use and understanding of anonymization and de-identification in the biomedical literature: Scoping review[J]. Journal of Medical Internet Research, 2019, 21(05): doi: 10.2196/13484.

[9]Torrance A W, Tomlinson B. Training is everything: Artificial intelligence, copyright, and "fair training"[J]. Dickinson Law Review, 2023, 128(01): 233.

[10]Mehrabi N, Morstatter F, Saxena N, et al. A survey on bias and fairness in machine learning[J]. ACM Computing Surveys, 2021, 54(06): 1-35.

[11]Quy T L, Roy A, Iosifidis V, et al. A survey on datasets for fairness-aware machine learning[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2022, 12(03): https://doi.org/10.1002/widm.1452.

[12]Gebru T, Morgenstern J, Vecchione B, et al. Datasheets for datasets[J]. Communications of the ACM, 2021, 64(12): 86-92.

作者简介：

王翔，博士，广州南方学院公共管理学院讲师，复旦大学数字与移动治理实验室研究员，主要研究方向为数字政府与数字治理、都市人类学。

何丹，硕士，杭州市大数据管理服务中心高级工程师，副科长，主要研究方向为公共数据全生命周期管理。

郑磊，博士，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，主要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。