

企业高质量数据集建设运营路径研究

白玉真 杨靖世 袁博 王思佳 贾轩 王思源

(中国信息通信研究院人工智能研究所,北京 100191)

摘要:高质量数据集已成为支撑人工智能发展和行业智能化转型的关键基础。基于此,首先分析了高质量数据集的概念与发展现状;其次针对不同类型的数据集剖析了其建设需求与核心问题;再次对企业高质量数据集建设过程的关键环节和建设路径展开了研究,并针对医疗、能源、工业行业的高质量数据集建设路径进行了分析;最后探讨了高质量数据集的未来发展趋势。

关键词:高质量数据集;人工智能;数据工程

中图分类号:F49;TP308

文献标志码:A

引用格式:白玉真,杨靖世,袁博,等.企业高质量数据集建设运营路径研究[J].信息通信技术与政策,2026,52(5):10-15.

DOI:10.12267/j.issn.2096-5931.2026.05.002

0 引言

数据是国家基础性战略资源,而人工智能(Artificial Intelligence, AI)是当前赢得全球科技竞争主动权的重要抓手。近年来,随着大模型和智能体技术的持续突破,“以数据为中心的人工智能”正式进入深化发展阶段。其中,大规模高质量训练数据的持续投入,不仅是大模型性能跃升的关键支撑,更成为推动AI与实体经济深度融合的核心纽带。

从政策演进脉络来看,我国高质量数据集的发展可划分为3个阶段,各阶段政策层层递进,为国家数据要素市场建设工作提供了系统的顶层设计与落地路径。

大数据应用探索期(2015—2019年):2015年,国务院印发的《促进大数据发展行动纲要》提出要推动公共数据开放共享,为数据集概念落地搭建了政策框架;2016年,工业和信息化部印发的《大数据产业发展规划(2016—2020年)》进一步将数据定位为“国家基础性战略资源”,强调数据整合共享,间接奠定了数据

集建设的早期基础。

数据要素流通机制建设期(2020—2023年):2020年,数据被纳入生产要素范畴;2021年,“十四五”规划明确AI等前沿领域需依托专业化多模态数据集;2023年,国家数据局等17部门联合印发的《“数据要素×”三年行动计划(2024—2026年)》则将高质量数据集列为激活数据要素价值的重要基础,标志着国家数据要素市场建设工作进入政策落地阶段。

高质量数据集发展期(2024年至今):2024年,国家发展和改革委员会等6部门联合发文并明确“高质量数据集”概念^[1],将其定位为AI与实体经济融合的核心载体;2025年,国家数据局印发《数字中国建设2025年行动方案》,强调加强交通、医疗、制造等重点领域的标注工作^[2],建设行业高质量数据集;同年,国家数据局在2025中国国际大数据产业博览会上发布了104个高质量数据集典型案例,覆盖科学研究、工业制造、低空经济等多个领域^[3],标志着国家数据要素市场建设工作迈入“深度应用”新阶段。

本文从高质量数据集概念与发展现状、建设需求

与核心问题、建设运营路径和未来展望方面开展研究,期望为更好地促进高质量数据集供给、数据要素产业和 AI 纵深发展提供参考。

1 高质量数据集概念与发展现状

高质量数据集是指经过规范采集、清洗、标注等处理流程,可直接用于 AI 模型开发与训练,且能显著提升模型性能的数据集合。高质量数据集的核心特征紧密围绕 AI 模型训练需求构建:大规模体现在数据量需满足大模型从 GB 级向 TB 级的跃升需求,多模态涵盖文本、图片、音频、视频等多种格式以适配复杂场景,高质量要求数据具备准确性、完整性与一致性,易应用强调格式与接口标准化以降低使用门槛,安全合规覆盖数据全生命周期的合规管理,规避隐私泄露风险。

高质量数据集是在传统数据集基础上,结合 AI 技术需求形成的。其核心包含 3 个关键要素:一是处理流程规范,覆盖数据全生命周期以保障质量可控;二是应用场景特定,专为 AI 模型设计,适配模型对数据格式、标注精度的要求;三是价值产出有效,能切实提升模型性能,如提高图像识别准确率、增强行业知识理解能力。目前,高质量数据集支撑的 AI 能力已广泛落地,其在基础能力层面可实现智能检索、目标检测等功能,在产业应用层面可渗透到工业质检、智能医疗、专业教育等领域,成为各行业数字化转型的重要技术支撑。

当前,我国高质量数据集建设工作已取得阶段性成效。截至 2025 年底,我国已建成的高质量数据集超 10 万个,总规模超 890 PB^[4]。这充分表明我国 AI 发展进入了快速增长阶段,应用场景不断深化,从能对话到能决策执行的智能体,AI 产业竞争力显著增强;同时,高质量数据集的大量供给,也促使数据要素价值不断加速释放,为 AI 创新发展进入良性互动阶段赋能。

2 高质量数据集建设需求与核心问题

2.1 高质量数据集建设需求分析

高质量数据集可以分为通识、行业通识和行业专识 3 类。当前,高质量数据集建设呈现“通识类成熟、行业类起步”的差异化格局,二者在供需模式、建设主体、规模上存在显著差异。

通识类高质量数据集建设已形成成熟的生态闭

环,以科技巨头和 AI 公司为核心需求方,以内部生产和公共供给为主要供给模式,通过开源社区平台实现高效对接。需求端的通识类高质量数据集主要用于基础模型训练,主要来源于科技巨头与头部 AI 公司。科技巨头自身拥有海量内部数据,如搜索日志、社交媒体内容、电商信息等。文心一言、通义千问等大模型均依赖内部数据构建核心语料库。供给端的通识类高质量数据集则以科技巨头自身积累为主,公共开源和定向商业补充为辅。科技巨头利用其强大的工程能力,自行爬取、整合并清洗全网数据,构建 TB 至 PB 级的私有语料库。例如,文心一言的训练数据词元数已超过万亿。在商业补充方面,海天瑞声等数据服务商提供数据采集、标注和清洗服务。

与通识类高质量数据集成熟的供给模式形成鲜明对比的是,行业类高质量数据集建设整体还停留在企业“私人作坊”的起步阶段,运营模式较不成熟,数据稀缺和“数据孤岛”成为阻碍行业大模型发展的关键瓶颈。从核心用途看,行业高质量数据集旨在为特定行业或垂直场景的 AI 模型提供精准判断和决策的能力,追求深度与专业性。从需求端来看,行业高质量数据集的需求主要是行业内企业,传统行业如医疗、金融、制造、能源等在 AI 快速发展趋势下模型应用需求激增,但每个行业乃至每个企业的具体需求又高度分散,专业化、个性化程度较高。从供给端来看,行业高质量数据集的供给也主要依赖企业自身。行业企业是数据的主要生产者和拥有者,其数据大多源于内部生产经营活动,而公共供给较少。究其原因,一是公开获取渠道少,数据是企业核心资产,关系核心竞争力,难以通过公共开源方式获取;二是商业化补充较匮乏,数据厂商多承担委托加工、标注工作,并非数据集的所有方,且行业专业壁垒较高、专业数据人才稀缺,进一步提高了数据集加工门槛;三是行业高质量数据集建设多以“一事一议”的项目制形式开展,供需双方需逐一谈判,涉及繁琐的保密协议签订过程;四是“数据孤岛”现象严重,数据被锁定在企业内部,即便归属同一行业,企业间数据也难以打通,极大浪费了数据价值,阻碍了行业整体模型性能的提升。

2.2 高质量数据集建设的核心问题

当前,企业高质量数据集建设大多是以项目制方式为主,整体建设和运营能力尚不足,在制度、标准、技

术和生态等方面面临突出问题和挑战^[5-6]。

内外部制度待细化,制约数据集系统化建设。当前,企业在高质量数据集建设中普遍缺乏制度规范的约束。这些制度规范既包括跨主体的数据基础制度,也包括企业内部的配套制度和细则。数据所有权、使用权和收益分配等关键制度不清晰,缺乏明确的定价与利益分配机制,导致各企业对数据共享、流通交易普遍持审慎态度,数据资源获取和高质量数据集流通存在障碍。数据汇聚、共享机制缺失,尤其是涉及行业基础性、共性的数据汇聚、共享机制缺失时会导致部分数据集的重复建设,浪费大量人力、财力。另外,在企业内部以项目制方式推进数据集建设,尚未形成覆盖全环节的制度体系,影响高质量数据集的持续建设。

标准体系不完善,数据获取与共享困难。多数行业尚未形成覆盖数据采集、格式、标注和质量评估的标准体系,行业内不同企业间的数据结构、元数据、数据标注方式存在差异,造成数据的互通、数据标注结果和质量结果互认存在困难。例如,工业领域不同设备的协议、数据格式和接口不统一;交通领域时序数据的标注需结合统一的时空现状和物理规则;生态环境领域部分地方的生态环境开发管理系统数量过多,且各系统的数据标准和技术接口均不统一。

技术支撑能力不足,影响建设效率和效果。技术架构方面,尽管部分企业已搭建数据平台,具备数据标注、质量评估等工具,但数据采集、处理、标注、评估、应用全过程不打通,容易造成开发、标注等过程难以追溯、验证和迭代。数据处理方面,缺乏面向非结构化、多模态数据的智能数据处理工具,导致多模态数据的结构和特征提取能力不足,非结构化数据的联合建模可用性较差。数据标注方面,部分数据标注工具对建筑信息模型、影像等非结构化数据的标注效率较低,影响数据集的建设效率。质量评估方面,质量评估常作为事后环节开展,而非贯穿于数据采集、处理、标注等各阶段,导致高质量数据集开发系统无法及时发现与修正采集偏差、标注错误、样本分布失衡等问题。

协同生态未建立,阻碍数据集流通应用。目前,不少企业虽已开始构建企业或行业可信数据空间,但尚未形成成熟落地的运行模式。可信数据空间内的主体准入、应用成效、收益分配等运营机制和规则均未完全明确,导致开发的数据集很难被其他需求方使用,使用

反馈、再开发和迭代更是难以实现。另外,数据提供方、服务方与使用方之间缺乏以“数据即服务”为核心的商业模式,数据产品供给形式单一、价值实现路径模糊,进一步阻碍了高质量数据集的服务化推广和产业化落地。

3 高质量数据集建设运营路径研究

3.1 高质量数据集建设核心环节与核心技术

高质量数据集的建设是一个系统性工程,需通过数据需求分析、数据规划设计、数据采集汇聚、数据预处理、数据标注加工、模型验证优化六大核心环节层层推进,各环节依托专项技术工具支撑,最终形成“需求导向—技术落地—迭代优化”的生态闭环体系。

在核心环节方面,各环节聚焦不同目标。数据需求分析环节需结合应用场景明确数据规模、类型、质量等指标,为后续数据集建设提供方向;数据规划设计环节需围绕数据架构、存储方案、安全规范等制定整体框架,确保数据采集与处理的合规性和高效性;数据采集汇聚环节需整合多源异构数据,实现跨平台、跨格式的数据聚合;数据预处理环节是提升数据质量的关键,需通过数据清洗、转换、脱敏等操作解决数据缺失、冗余、不一致等问题;数据标注环节针对 AI 训练需求,对文本、图像、音频等数据进行结构化标注,为模型学习提供“标注样本”;模型验证优化环节则需通过将数据集投入模型训练,根据效果反馈调整数据质量参数,实现数据集与模型需求的动态适配。

在核心技术层面,各环节依托不同的专项技术支撑。比如,数据采集汇聚环节依赖多源异构数据融合技术、边缘计算辅助采集技术及生成式 AI 合成技术;数据预处理环节运用自动化清洗工具、机器学习智能补全技术及隐私计算脱敏技术;数据标注环节采用半自动化标注工具、分布式众包管理技术及主动学习标注策略;模型验证优化环节则通过数据质量评估算法、样本均衡性优化等技术,确保数据集适配模型训练需求。

上述核心环节与核心技术共同构成高质量数据集从“需求定义”到“落地应用”的全流程支撑体系,为高质量数据集的质量可靠性与场景适配性提供保障。

3.2 高质量数据集建设路径分析

高质量数据集建设运营需从体系管理、工程建设、

生态运营三维协同发力,使之各司其职又相互支撑,共同夯实数据根基、释放数据价值^[7]。

体系管理聚焦“由谁来做、应该怎么做”,旨在明确数据集建设的组织与流程规范。具体来说,企业需搭建专门负责数据集建设的组织架构与项目团队,确保责任主体清晰;同时,需制定相关工作流程、标准规范,使数据集建设过程有据可依(如,明确数据采集、标注、管理的统一标准,保障数据集建设的系统性与合规性)。

工程建设围绕“用什么工具、实际怎么做”,覆盖数据集全生命周期的技术工具与安全保障。从数据集全生命周期来看,需具备数据采集、清洗、标注、质检、运营管理等环节的技术工具或平台,实现数据处理的高效化;同时,要保障数据集全生命周期的数据安全与应用安全,通过加密、权限管理、安全审计等技术手段,防范数据泄露、篡改等风险,为数据集的可靠应用筑牢技术防线。

生态运营关注“做得怎么样、还能做什么”,聚焦数据集的价值评估与生态拓展。具体来说,企业一方面要评估数据集覆盖的应用场景及应用效果,明确其在业务场景中的价值贡献(如衡量数据集在模型训练、业务决策支持等场景中的精度与效率);另一方面要推动数据共建共享,扩大数据生态范围,通过构建数据交易、合作共享机制,使其在更广泛的生态中流转,持续释放潜在价值,形成“数据—应用—生态”的正向循环。

上述3个维度应相互协同、层层递进。体系管理为工程建设与生态运营提供组织和制度保障,工程建设是体系管理落地的技术支撑,生态运营则是前二者价值的最终体现与持续延伸,三者合力推动高质量数据集从“建设”到“运营”再到“价值变现”的全链路发展,共同为企业数智化转型提供坚实的数据底座。

3.3 典型行业高质量数据集建设的路径分析

本文基于医疗、能源、工业制造三大行业的数字化基础、场景需求与核心痛点,围绕“打造完善的体系规划能力、全面提升数据工程建设能力、建立可持续运营的管理机制”三大重点目标,构建了分行业、全流程的高质量数据集建设路径。

医疗行业高质量数据集建设以“标准统筹+技术破壁”为核心路径,通过体系规划、工程建设与运营机

制的协同推进,破解“数据孤岛”与质量难题。在体系规划方面,以医院信息科为核心,联合临床、质控、AI团队成立专项组,明确“业务需求—技术落地—质控考核”分工,同时参考国家医疗数据规范,构建“术语—格式—质量”三级标准体系,为数据整合与质量管控奠定基础。在工程建设方面,采用数据编织技术构建虚拟数据平面,通过关联跨系统患者数据并同步脱敏保障合规,以“自然语言处理+AI辅助标注”实现数据加工。在运营机制方面,建立“全流程质控+动态优化”体系,定期审计数据质量并根据模型反馈调整规则,同时探索“院内科研+外部合规合作”模式,输出脱敏数据集用于AI模型训练并按贡献度分配收益,激发数据流通动力,最终实现医疗数据从“分散治理”到“体系化价值释放”的转变。

能源行业高质量数据集建设以“统筹协同+平台支撑”为核心路径,通过体系规划、工程建设与运营机制的深度联动,破解数据分散与安全流通难题。在体系规划方面,企业组建由业务、技术、安全等人员构成的高质量数据集建设团队,绘制行业数据集全景图谱,同时联合科研机构共同统一数据采集频率与标注规范,为全行业数据集建设锚定方向。在工程建设方面,部署边缘设备采集多源数据,以“湖仓一体”架构存储数据并进行加密保护,同步打造行业协同平台,集成数据自动化清洗、特征提取等工具,支持产业链上下游企业在可信环境中联合建模。在运营机制方面,按数据敏感度进行分级授权,通过数据水印与审计日志追溯流向,同时建立“贡献度量化+收益共享”模式,依托行业数据交易所或行业可信数据空间推出标准化数据产品,推动数据价值向产业应用转化,最终实现能源数据从“分散建设”到“全链条协同运营”的升级。

工业制造行业高质量数据集建设以“场景突破+技术攻关”为核心路径,通过体系规划、工程建设与运营机制的协同发力,破解生产场景数据采集、人工经验沉淀难题。在体系规划方面,以生产部门为核心,优先选择冶炼、矿山等数据基础好、价值密度高的场景开展试点,明确各场景数据需求,同时构建“基础通用+场景专用”的标准体系,适配金属冶炼、汽车制造等行业分支的工艺差异。在工程建设方面,采用高清相机与传感器采集跨模态数据,通过边缘设备预处理保障数据有效性,同步研发定制化标注工具,构建多任务模型

挖掘工艺参数间的隐性关联。在运营机制方面,按生产一线反馈动态调整数据采集频率与标注规则,形成从应用场景到数据的迭代闭环,同时将试点场景的建设经验形成可复制模板向业内推广,依托数据流通利用平台共享非敏感数据,解决单个企业数据量不足的问题,最终实现工业制造数据从“经验驱动”到“数据与技术双轮驱动”的转型。

4 高质量数据集发展展望

企业高质量数据集建设是关乎国家数字竞争力、产业安全与现代化产业体系构建的战略工程。为推动高质量数据集建设,需充分发挥企业支撑的作用,使其超越责任边界,明确自身定位,承担起更宏大的国家使命;在供给侧,成为国家关键领域高价值、高可信数据的核心供给者;在生态侧,成为融通产业链、创新链、价值链的数据生态重要参与者。未来,行业高质量数据集建设将呈现标准化、智能化与生态化的发展趋势。

在制度突破与标准先行层面,数据基础制度、治理体系和安全监管将不断完善,将为数据集建设、流通和使用提供更清晰的边界与运行规则;与此同时,行业数据规范、标签体系与质量评估体系将逐步统一,跨企业、跨场景数据可用性将显著提升。企业可在内部率先探索数据资产确权、内部核算与收益分享机制;行业主管部门可联合头部企业,在基础较好的领域,率先研制并发布行业高质量数据集建设标准(含数据格式、标注规范、质量评估等),为跨企业数据互通奠定基础。

在技术攻坚与平台赋能层面,多模态数据处理、数据合成等技术将逐步成熟,将促使数据生产方式更加智能高效。企业可进一步加大数据工程技术投入,研发适配行业特点的数据智能标注、质量评估、动态监控工具;大型企业可建设集团级数据资产运营平台,将分散的项目能力沉淀为可复用、可共享的平台服务,降低下属企业的建设门槛。

在生态构建与价值释放层面,基于行业可信数据空间、区域数据平台和跨主体协同机制的新型生态将加速形成,将推动企业高质量数据集在更大范围内流通利用。有条件的企业可牵头建设行业可信数据空间,在保障安全与权益的前提下,探索数据“可用不可见”的合作模式,完善空间内相关制度规则;将自身高

质量数据集作为“锚点”,吸引上下游企业、科研机构贡献数据、共同开发,将数据优势转化为产业生态领导力。

5 结束语

综上所述,通过在标准、技术与生态领域的持续投入,企业将具备更稳定的数据资源供给能力、更强大的国际竞争力和生态组织能力,成为推动我国数据要素市场繁荣的关键力量。通过不断系统性地破解制度、技术与生态层面的难题,企业不仅能筑牢自身智能化转型的基石,更将有力提升我国重点行业整体的数字化水平,为深化落实“人工智能+”专项行动、壮大国家数据要素市场、发展新质生产力提供坚实支撑。

参考文献

- [1] 国家发展和改革委员会,国家数据局,教育部,等.关于促进数据产业高质量发展的指导意见[EB/OL]. 2024[2026-04-01]. https://www.gov.cn/zhengce/2020-04/09/content_5500622.htm.
- [2] 国家数据局.《数字中国建设2025年行动方案》近日印发[EB/OL]. 2025[2026-04-01]. https://www.nda.gov.cn/sjj/ywpc/sjzg/0516/20250516160504500412691_pc.html.
- [3] 国家数据局.高质量数据集典型案例名单发布[EB/OL]. 2025[2026-04-01]. https://www.nda.gov.cn/sjj/zwgk/tzgg/0912/20250912130433113171778_pc.html.
- [4] 新华社.我国建成高质量数据集超10万个,规模超890PB[EB/OL]. (2025-12-30)[2026-04-01]. https://www.gov.cn/lianbo/202512/content_7053257.htm.
- [5] 中国通信标准化协会大数据技术标准推进委员会.高质量数据集实践指南(1.0)[R], 2025.
- [6] 中电数据产业集团有限公司,中国信息通信研究院.中央企业高质量数据集建设研究报告[R], 2025.
- [7] 姜春宇,白玉真,刘渊,等.构建企业级人工智能高质量数据集:方法与路径[J].大数据,2025,11(6):47-56.

作者简介:

白玉真 中国信息通信研究院人工智能研究所高级工程师,主要从事数据流通、高质量数据集方面的研究工作

- | | | |
|--|--|---|
| <p>杨靖世 中国信息通信研究院人工智能研究所工程师,主要从事可信数据空间、高质量数据集方面的研究工作</p> <p>袁博 中国信息通信研究院人工智能研究所大数据与智能化部副主任,高级工程师,主要从事数字经济、大数据、数据流通、数据基础设施方面的研究工作</p> <p>王思佳 中国信息通信研究院人工智能研究所工程</p> | <p>贾轩 中国信息通信研究院人工智能研究所高级工程师,主要从事数据基础设施、数据流通方面的研究工作</p> <p>王思源 中国信息通信研究院人工智能研究所工程师,主要从事隐私计算、数据基础设施方面的研究工作</p> | <p>师,主要从事人工智能、高质量数据集、数据要素、数据应用方面的研究工作</p> |
|--|--|---|

Research on the construction and operation path of enterprise-level high-quality datasets

BAI Yuzhen, YANG Jingshi, YUAN Bo, WANG Sijia, JIA Xuan, WANG Siyuan

(Artificial Intelligence Institute, China Academy of Information and Communications Technology,
Beijing 100191, China)

Abstract: High-quality datasets have become a critical foundation for supporting the development of artificial intelligence and the intelligent transformation of industries. Based on this, the study first analyzes the concept and current development status of high-quality datasets. Secondly, it examines the construction requirements and core issues of different types of datasets. Thirdly, it investigates key stages and pathways in the process of building enterprise-level high-quality datasets, analyzing the construction approaches for high-quality datasets in the healthcare, energy, and industrial sectors. Finally, it explores future development trends of high-quality datasets.

Keywords: high-quality datasets; AI; data engineering

(收稿日期:2026-04-07)