大数据与决策研究

（政策与技术跟踪专题）

2025 年第 26 期（总第 323 期）

广西壮族自治区信息中心
广西壮族自治区大数据研究院　　2025 年 6 月 17 日

我区亟需加强高质量数据集建设 赋能人工智能发展

数据作为人工智能发展的三大核心要素之一，是模型训练与模型应用的基础和核心资源。加快建设人工智能高质量数据集，对于推动广西人工智能高质量发展有重要意义。

一、我国高质量数据集发展形势

近年来，大模型技术不断取得突破，一系列国内外大语

言模型展现出了高质量的智能水平，高质量数据集的投入在其中起到了关键性的作用。同时，人工智能与行业应用的深度融合也需要相关高质量数据集的支持，大模型对数据的需求量呈指数级增长，高质量、大规模和多样性的数据成为保障我国人工智能快速发展的根本要素。

（一）高质量数据集建设关键技术

高质量数据集的效能提升亟需突破并集成一系列关键技术：**一是数据采集与汇聚层面**，应用物联网等技术实时高效捕获多源异构数据，强化中文语料和特定行业场景数据的定向爬取能力。**二是数据处理与治理层面**，核心是发展基于规则引擎和机器学习的自动化清洗工具，同时要突破 AI 辅助标注技术，提升对多模态数据和中文语义、行业知识的标注效率与精度，形成“机器预标+人工精校”模式。**三是存储与管理层面**，依托云原生分布式存储架构和虚拟化存储池技术，可优化资源调度与成本。**四是安全与合规层面**，通过建立数据安全流动技术体系，推广利用联邦学习、多方计算、数据沙箱、区块链存证等技术，实现数据可用不可见。**五是服务与流通层面**，发展标准化 API 接口与数据资产登记技术，推动数据集安全可信流通，并构建自动化数据集构建流水线，提升高质量数据集生产的工程化水平。

（二）高质量数据集建设生态要素保证

一是政策的有效支持。通过顶层设计与专项规划明确数据采集、流通与安全标准，完善数据安全立法与隐私保护，

平衡市场化与权益保护。**二是商业模式的创新。**打造“数据即服务”等模式，将数据资产转化为收益，通过平台创新模式吸引多方协同，促进资源共享与标准化。同时，多元合作机制可分散风险、降低成本，并通过增值服务激活生态。**三是人工智能人才的培养。**复合型人才是突破技术瓶颈的核心支撑，需系统化学科重构培育跨界人才，通过产学研联动提升工程化能力。

（三）人工智能发展面临高质量数据集供需矛盾

在高质量数据集建设方面，我国目前仍存在以下的问题：**一是高质量数据集供给仍显不足。**中文数据质、量均落后于英文数据，公共数据开放利用程度有待提高，数据标准不统一，缺乏高质量行业数据集。**二是数据主体与商业模式尚不成熟。**缺乏高质量数据汇聚治理主体，具备大规模数据处理能力的公司不足，多领域公共数据授权运营主体仍待培育，相关商业模式有待完善。**三是相关规划政策亟待细化。**面向新一代人工智能的高质量数据集专项规划和支持政策尚未出台，建设、运营、流通、利用方面的举措需细化，尤其在数据采集标准规范、数据共享流通机制方面的不足限制了模型能力的快速提升。

二、我国建设高质量数据集政策布局规划

（一）顶层设计与战略规划

在由国家数据局等 17 部门联合印发的《“数据要素 ×”三年行动计划（2024—2026 年）》文件中明确提出：完善数

据资源体系，在科研、文化、交通运输等领域，推动科研机构、龙头企业建设行业共性数据资源库，打造高质量人工智能大模型训练数据集。发改委、国家数据局、工信部联合印发了《国家数据基础设施建设指引》文件，其中要求建立覆盖政府、行业、企业等主体及国家、省、市、县等层级的全国一体化的分布式数据目录，形成全国数据“一本账”，同时支持农业、工业、交通等 10 余个重点行业打造高质量数据集。

（二）分类建设高质量数据集标准体系

在由国家数据局推动制定的《高质量数据集建设指南（征求意见稿）》中，明确了三类数据集的建设标准:

一是推动通识类高质量数据集建设。由政府机构、科研机构等主导构建，覆盖自然语言处理等通用领域，能提供跨行业基准测试环境，并通过共享打破数据孤岛，降低企业基础数据成本。**二是推动行业通用类高质量数据集建设。**聚焦行业共性知识（标准术语、流程），使模型掌握领域内通用逻辑规则，既可作为垂直大模型预训练基底，又能支撑跨企业共性场景的快速落地。**三是推动行业专用类高质量数据集建设。**由企业根据核心业务场景需求构建绑定，支撑高精度决策场景（如医疗诊断、工业预测维护），通过专有数据持续优化模型，形成核心竞争力壁垒。

三、建设高质量数据集先进经验与模式

在推动高质量数据集建设过程中，国内先进省市通过机制创新与技术融合形成了可借鉴的实践经验。结合我区发展需求，现重点提炼出三种具有代表性的建设模式如下。

（一）省市一体化协同：构建纵向贯通的治理体系

通过建立省、市两级协同机制实现数据全域统筹与分级运营。**一是建立两级主体联动**：省级主体负责制定标准、搭建基座；市级主体对接本地需求开发特色产品（如**江苏**省级平台统一授权，市级开发医保风控）。**二是实施平台全域贯通**：依托省级平台建设市、县专区，实现目录统一管理、数据直达基层（如**贵州**三级联动枢纽）。**三是推行分级分类开放**：省级制定开放负面清单，市级按需申请高价值数据（如**山东济南**开放 22 类人口数据）。

（二）重点领域产业化开发：深耕垂直场景释放数据价值

聚焦产业需求打造行业级高质量数据集，推动资产化、产品化转型。**一是精准匹配需求**，省级部门联合龙头发布建设目录（如**湖北**推出 10 个行业数据集）。**二是构建产学研协同**，通过“知识库+智能标注”提升数据质量（如**苏州**推出丝绸纹样数据集）。**三是推动开源共享生态**，促进合规流通（如**北京**交易所发布 300 个数据集）。

（三）全域治理场景化应用：以用促建实现数据闭环迭代

以全域业务场景牵引治理，建立“供给—应用—优化”闭环。**一是场景化治理**，围绕高频场景制定标准（如**山东**构建一体化体系）。**二是业务反哺质量**，基于应用反馈动态校准数据（如**贵阳**“贵商易”优化模型）。**三是深化跨域融合**，打通公共与行业数据壁垒（如**济南**融合医保交通数据开发“健康出行指数”）。

一、广西提升高质量数据集供给能力的对策建议

（一）构建全域协同治理体系强化数据资源统筹效能

建立自治区级数据资源统筹主体，统一制定数据分级分类标准与安全规范；支持地市依托本地特色产业和民生需求开发垂直场景数据集，形成“自治区定标、地市落地”的协同架构，重点推动中国—东盟跨境贸易、特色农业等优势领域数据资源的定向汇聚与标准化治理。持续升级优化区级数据共享平台为全区数据共享提供核心枢纽，建立覆盖 14 个地市的分布式数据专区，实现目录统一管理、接口规范互通，优先打通交通物流、边贸通关等高频业务数据链条，支撑基层治理与跨境合作场景的实时数据调用。

（二）深化重点产业数据开发打造行业级核心数据资产

联合区内龙头企业发布特色产业数据集建设目录，构建涵盖机械制造、现代农业等领域的行业通用知识库。推动与区内各大高校、研究机构以及中国—东盟人工智能创新中心的深度协作，搭建“行业知识图谱+智能标注平台”双驱动体系，重点突破多语种（东盟语系）跨模态数据标注技术，降低各类特色场景的数据处理成本。

（三）创新场景驱动治理模式实现数据价值闭环迭代

围绕中国（广西）自由贸易试验区、西部陆海新通道等国家战略场景，制定电子口岸通关、跨境金融风控等专项数据治理规范，形成高频业务场景的数据质量评估体系。打通

公共领域数据壁垒，探索创新机制，试点开发各类融合应用，通过业务反馈持续优化相应业务的核心数据集精度。

（执笔人：黄子川）

编辑部地址：南宁市体强路 18 号广西信息中心 1412 号房

联系电话：0771-6113592

电子邮箱：dsjyjs@gxi.gov.cn

网　　址：http：//gxxxzx.gxzf.gov.cn/

扫描二维码获取
更多决策参考信息