

网络安全研究

赛迪研究院 主办

2025 年 7 月 15 日

第 2 期

总第 83 期

本期主题

□ 人工智能高质量数据集发展研究

赛迪智库

面向政府·服务决策

奋力建设国家高端智库

思想型智库 国家级平台 全科型团队
创新型机制 国际化品牌

《赛迪专报》《赛迪要报》《赛迪深度研究》《美国产业动态》《赛迪前瞻》
《赛迪译丛》《国际智库热点追踪周报》《工信舆情周报》《国际智库报告》
《新型工业化研究》《工业经济研究》《产业政策与法规研究》《工业和信息化研究》
《先进制造业研究》《科技与标准研究》《工信知识产权研究》《全球双碳动态分析》
《中小企业研究》《安全产业研究》《材料工业研究》《消费品工业研究》《电子信息研究》
《集成电路研究》《信息化与软件产业研究》《网络安全研究》《未来产业研究》

『所长导读』

数据集是人工智能模型训练的基础，高质量数据集是确保人工智能训练效果，进而形成新质生产力的重要保障。近年来，随着人工智能算力和算法的跨越式发展，数据集质量的重要性日益突出。国家、地方和企业高度重视人工智能数据集建设，加快推动高质量数据集赋能人工智能创新发展。

本期《人工智能高质量数据集发展研究》从高质量数据集概况、评价标准、国外发展情况、发展的难点堵点和意见建议五部分组成。系统梳理了高质量数据集的定义及内涵，明确了发展高质量数据集的意义和我国高质量数据集发展的整体概况。针对高质量数据集的评价体系，从质量、标注、格式三方面展开论述，阐述了高质量数据集标准评估体系框架。结合美国、欧盟等国外典型国家和地区发展高质量数据集的实践经验，分析梳理了国外发展人工智能与高质量数据集的思路与实施路径。总结了针对当前我国人工智能高质量数据集发展存在的问题和挑战并提出了针对性意见建议。

希望本期内容能为政府部门、企业机构和社会组织提供参考和借鉴，并欢迎各界读者不吝赐教。

赛迪研究院网络安全研究所所长 温晓君

2025年7月15日



目 录

CONTENTS

本期主题：人工智能高质量数据集发展研究

一、高质量数据集概况	1
（一）高质量数据集的定义及内涵	1
（二）发展高质量数据集的重要意义	1
（三）我国高质量数据集发展概况	2
二、高质量数据集评价标准	2
（一）高质量数据集质量标准	2
（二）高质量数据集标注标准	4
（三）高质量数据集格式规范	4
三、国外高质量数据集发展概况	4
（一）政策牵引数据开放加速高质量数据集培育	4
（二）需求推动多主体运营数据集产量大幅提升	5
（三）标准持续完善规范数据质量评估应用要求	6
四、高质量数据集发展难点堵点	6
（一）数据汇聚的产量小、质量低，数据资源共享流通不畅。	6
（二）数据集质量良莠不齐，缺失误差伪造问题严峻。	7
（三）算法偏见加剧数据遗失，数据要素价值挖掘短板明显。	7
五、高质量数据集发展意见建议	8
（一）加强数据质量源头发力，推动数据标注产业高质量发展。	8
（二）强化数据获取与共享，探索行业试点联合推进共建新模式。	8

- 
- (三) 完善质量与标准体系，推动建设重点行业数据集评价标准。.....9
 - (四) 加强数据隐私与安全保障，推动数据集安全评估能力建设。.....9
 - (五) 优化数据集运营模式，推动数据资源价值生态循环落地。.....9

本期主题：

人工智能高质量数据集发展研究

一、高质量数据集

（一）高质量数据集的定义及内涵

高质量数据集指已经过清洗、标注等数据处理，不涉及有关敏感信息，在格式、质量等方面符合相关要求，可直接用于开发和训练人工智能模型的数据集。按照建设类别，高质量数据集分为通识数据集、行业通识数据集和行业专识数据集。高质量数据集的标准从完整性、准确性、一致性、可用性、规范性、多样性、专业性、合规性、干净性、安全性、可溯性、可访问性、标注一致性、标注规范性等14方面定义。高质量数据集类型主要包含文本数据集、图像数据集、语音数据集、时序数据集、思维链数据集、多模态数据集等。

（二）发展高质量数据集的重要意义

国家高度重视高质量数据集建设，密集出台相关政策举措。国家先后出台《关于促进数据产业高质量发展的指导意见》《国家数据标准体系建设指南》《关于促进数据标注产业高质量发展的实施意见》等政策文件。2025年2月19日，国家数据局召开高质量数据集建设工作启动会，积极推进落实“人工智能+”行动，推动高质量数据集建设，高效赋能行业发展。

国内高质量数据集快速培育发展。一是数据资源开放牵引高质量数据集发展。2024年，全国地市级以上地方公共数据开放平台数量增长7.5%，开放的数据量增长7.1%，高质量数据集数量同比增长27.4%。二是高质量数据集标准体系加速构建。在国家数据局指导下，推动制定《高质量数据集标注规范》《高质量数据集质量规范》《高质

量数据集格式规范》等多项标准，构建起高质量数据集标准的“四梁八柱”。

（三）我国高质量数据集发展概况

我国高质量数据集建设已进入规模化、规范化的发展阶段。国家数据局统筹推进数据资源体系建设，全国7个数据标注基地已构建医疗、工业、教育等领域的335个高质量数据集，标注总规模达17282TB，支撑121个国产大模型的研发，带动数据标注相关产业产值超83亿元。2025年4月29日召开的第八届数字中国建设峰会上，国务院国资委发布了涵盖智慧能源、工业制造、绿色低碳、金融服务等10余个行业30项人工智能行业高质量数据集优秀建设成果。贵阳大数据交易所累计发布的939个多模态数据集覆盖了金融、工业、医疗、商贸等关键领域，为大模型厂商提供了丰富且优质的数据资源。广州市政务服务和数据管理局对标“12218”现代化产业体系组织开展了高质量数据集征集工作。经公开征集并遴选，选出第一批16家企业

28个高质量数据集。深化数据资源开发利用和开放共享，以高质量数据集建设赋能大模型训练、助力高价值场景落地，为人工智能高质量发展筑牢数据底座。

二、高质量数据集评价标准

（一）高质量数据集质量标准

一是充分考虑多个数据集的质量指标。高质量数据集的质量要求应满足人工智能应用场景的需求，充分考虑多个质量指标，包括完整性、准确性、一致性、可用性、规范性、多样性、专业性、合规性、干净性、安全性、可溯性、可访问性、标注一致性、标注规范性等。其中，完整性指数数据集中数据所具有属性和所记录信息是否完整；准确性指数数据集中数据是否真实反映所代表的信息；一致性指数数据集中数据是否在所有来源和记录中完全一致；可用性指数数据集中数据是否可直接用于人工智能模型的开发和训练；规范性指数数据集中数据是否格式标准、文档完备。多样性指数数据集中数据是否满足相关人工智能任务对数据分布全面程度的要求。专业性指数数据集中数据是否蕴含深层次行业领域知识。

合规性指数数据集中数据是否符合数据安全等方面相关法律法规要求。干净性指数数据集中数据是否包含重复信息和噪声信息。未篡改性指数数据集中数据是否被篡改，包括数据投毒、数据后门等恶意篡改。可溯性指数数据集中数据是否在来源和处理过程上可溯源。可访问性指数数据集是否容易被所授权用户查询、获取、使用和共享。标注一致性指数数据集中数据的标注结果在不同标注者之间是否一致。标注规范性指数

据集中数据的标注过程和结果是否符合预定标注规范和指导准则。二是通过质量指标的必要性，判断不同类别数据集是否满足高质量数据集的质量要求。数据集按照多个特征（知识内容、来源类型、时效性、标注人员类型、敏感程度、模型类型、主题范围、信息含量等）的不同表现，分为：通识数据集、行业领域通识数据集、行业领域专识数据集。不同质量指标对于不同类别数据集的重要程度具体如下表：

表 1 不同质量指标对于不同类别数据集的重要程度

数据集类型 不同质量指标	通识数据集	行业领域通识数据集	行业领域专识数据集
完整性	必要	必要	必要
准确性	非必要	必要	必要
一致性	必要	必要	必要
可用性	必要	必要	必要
规范性	必要	必要	必要
多样性	必要	非必要	非必要
专业性	非必要	必要	必要
合规性	必要	必要	必要
未篡改性	必要	必要	必要
干净性	必要	必要	必要
可溯性	必要	必要	必要
可访问性	必要	必要	非必要
标注一致性	必要	必要	必要
标注规范性	必要	必要	必要

（二）高质量数据集标注标准

一是流程要求。标注前期准备阶段包括需求确认、需求分析、数据获取、数据试标、规则细化、专项培训、计划制定和资源分配等环节。标注任务实施阶段包括任务创建、任务分发、任务执行、标注审核、任务回收等环节。标注任务交付验收阶段包括初步验收、任务交付、最终验收、后期维护、任务总结等环节。

二是技术要求。一方面是总体要求。所有标注需遵循统一标准，确保不同标注员的结果一致，避免因主观差异导致的标注偏差；标注结果需准确无误，确保数据可用于高质量模型训练，避免因标注错误影响模型性能；标注过程需记录详细信息，包括修改时间、修改记录等，便于问题追溯和质量控制。另一方面，是针对不同类型数据如图像数据、文本数据、语音数据、时序数据、点云数据、多模态数据、思维链数据的专项技术要求。

三是管理要求。标注相关人员需分工明确、遵守标注规则，把控标注质量，及时完成标注任务，相

互监督等；标注工具需满足标注工作中各项要求，保证标注工作顺利、高质量完成；在安全管理方面，需满足数据合规、数据安全、管理验证等要求。

（三）高质量数据集格式规范

一是规范标注文件存储格式。

目标检测、图像分割、人脸识别等人工智能应用场景下的数据集，需符合通用的标注文件存储格式标准。**二是规范标注文件元数据格式。**不同应用领域需符合统一的标注文件元数据格式，包括但不限于命名方式、属性丰富度等方面。**三是规范数据集文件组织结构。**不同应用领域定制数据集文件的目录结构和命名规则，需符合统一的文件组织格式标准。

三、国外高质量数据集发展概况

（一）政策牵引数据开放加速高质量数据集培育

近年来，美国、欧盟等全球主要经济体通过系统化制定政策，促进数据要素价值释放，加快培育高质量数据集。美国自2009年启动《开放政府指令》以来，持续完善

数据开放制度体系，先后出台《开放数据政策》《政府信息公开和机器可读行政命令》等政策，构建了覆盖联邦、州及市三级政府的数据开放网络，截至 2024 年已累计发布 29 万个多领域数据集；2016 年，美国《国家人工智能研究和发展战略计划》确立数据集建设战略地位；2022 年美颁布《人工智能权利法案蓝图》，提出五大治理原则；2023 年新版战略计划强调构建可访问的共享数据环境，形成“基础建设—应用规范—生态优化”的递进式发展路径。欧盟通过制度创新构建数据治理新格局，2020 年，《欧洲数据战略》提出九大领域公共数据空间建设框架；2024 年 6 月，《数据治理法》进一步确立统一的数据获取与共享规则；欧盟《人工智能法案》将数据集质量要求细化为可用性、完整性、无偏性等可操作标准，形成“战略规划—法律约束—技术规范”的三维保障体系。

（二）需求推动多主体运营数据集产量大幅提升

一是政府机构或公共部门基于公共利益或政策需求，将收集或

制作的数据集向公众开放并服务于公共服务优化、科研探索及产业创新。美国政府开放数据平台 Data.gov 整合了农业、气候、教育等领域的 19 万余个数据集，通过统一门户实现跨部门数据共享；欧盟开放数据平台则汇聚了欧盟机构及成员国的农业、能源等跨领域数据资源，形成区域性数据协作网络。二是社区驱动的数据集建设模式，以 Kaggle、Hugging Face 等开源平台为代表，数据来源涵盖学术机构、企业和个人开发者，通过众包协作机制推动数据集持续更新与多样化发展。三是科研机构主导的学术数据集建设，如加州大学欧文分校维护的机器学习存储库（UCI Machine Learning Repository）、剑桥大学主导的 CORD-19 医学数据集、普林斯顿大学开发的 WordNet 语义网络等，多通过学术期刊、机构知识库等渠道实现开放共享。四是企业自建数据集的商业化运营，如谷歌、亚马逊等科技企业基于业务积累数据，经脱敏、标注等处理后形成标准化数据产品，通过 API 接口或订阅服务实现数据资产的价值转化。

（三）标准持续完善规范数据质量评估应用要求

一方面，在知识产权保护政策方面做出调整。各国通过制度创新平衡技术创新与权利人利益。如美国采用“例示列举+一般要件”的合理使用制度，其《版权法》第107条将批评、评论、学术研究等六类行为纳入合理使用范畴，但该制度的适用仍需结合案件具体情况判断，存在一定的解释弹性。日本2023年修订的《著作权法》虽未新增AI生成内容相关条款，但政府明确表态不保护AIGC模型训练阶段使用的既有作品，并创设“非以欣赏作品价值为目的”的新型合理使用场景。欧盟通过《单一数字市场版权指令》，为科研目的的文本数据挖掘设置双重例外条款，既允许非商业研究的数据挖掘，又为商业性应用提供法定许可机制。另一方面，各国加强技术标准建设。国际标准化组织（ISO）SC42分委员会系统构建了涵盖数据质量评估、算法可信赖性、治理框架等维度的人工智能标准体系，其制定的ISO/IEC 24026标准专门规范了数据全

生命周期管理要求。OECD《人工智能原则》、IEEE《人工智能伦理设计准则》等国际框架，也为数据治理和算法伦理提供了可参考的规范模板。

四、高质量数据集发展难点堵点

当前，我国加速推动高质量数据集创新发展，但对于高质量数据集突出存在数据产量低、数据质量低、数据利用效率低三方面问题。

（一）数据汇聚的产量小、质量低，数据资源共享流通不畅。

一是高质量数据储备量低，中文数据规模较小。美国加州大学研究表明，应用于人工智能的数据集可能会在2026-2032年间耗尽所有高质量语言数据。此外，国际主流大模型数据集主要以英文为主，最流行的Common Crawl中文数据只占据4.8%。二是数据流通开放力度不足，公共数据获取渠道不畅。中国科学院院士陈松蹊表示，近十年来，网络下载数据的通道并不稳定，数据发布格式时有变化，易造成数据缺失，缺乏公共数据的有效获取途径。三是数据标注自动化程度不

足，数据集产量与数据增速不匹配。2024 年，全国年度数据生产总量 1 达 41.06 泽字节(ZB)。然而，据调研，当前我国数据标注自动化、智能化程度较低，专业标注人员队伍数量缺口较大，数据集产量小，部分专业数据集无法规模化生产，难以满足专业场景需求。

(二) 数据集质量良莠不齐，缺失误差伪造问题严峻。

一是数据集存在缺失、尺度不一问题。据北京大学中国社会科学调查中心调研发现，真实数据来源通常通过试验采集，抽样、问卷设计、访员行为、受访者行为等，受制于数据采集、加工过程中各类误差、工具手段稳定性等影响，数据集普遍存在分布偏差、颗粒度不一致、采集缺失甚至错误数据等问题。二是数据集标准不一，规范性参差不齐。据调研，我国数据标注人才专业性不足，加之标注行业监管规范欠缺，部分企业存在层层外包，低价竞争，导致数据标注不规范、不专业，数据集质量良莠不齐。三是数据集伪造、混用造成影响训练效果。部分企业和研究机构为追求

模型训练效果，混用、篡改数据集，甚至进行数据投毒，严重影响模型训练质量。高瓴人工智能学院和伊利诺伊大学研究表明，在大模型训练的基准测试中相关数据被用于模型训练的情况变得越来越常见，导致部分测试分数虚高，泛化能力下降，不相关任务表现骤降等问题。

(三) 算法偏见加剧数据遗失，数据要素价值挖掘短板明显。

一是算法缺陷导致原始数据遗失率高。在大模型训练过程中，数据呈现长尾分布，为提高训练成功率，多数大模型算法采用“去尾”方法，即训练过程中对原始数据进行选择性“忽略”，因而导致数据遗失问题，甚至造成对原始数据的破坏。二是数据使用率较低。据统计，2023 年，在我国存储的数据中，一年未使用的数据占比约 4 成，企业一年未使用的数据占比为超过 30%，大量数据被存储后便不再被读取和复用，成为“死”数据。三是数据价值挖掘不足。数据加工能力不足导致大量数据价值被低估、难以挖掘复用。据统计，2023 年，全国数据产存转化率为 2.9%，海量数

据源头即弃。在开展数字化转型的大企业中，实现数据复用增值的仅有 8.3%，数据价值挖掘效率极低。此外，高质量数据集的价值实现路径不清晰也引发企业运营建设积极性降低。

五、高质量数据集发展意见建议

随着人工智能算法模型准确度不断提升和算力成本不断降低，围绕数据服务领域的智能化、高效标注技术将持续高速发展，高质量数据集和数据标注服务将成为企业竞争发展的重点，垂直领域应用场景和一体化高质量数据服务生态体系将成为区域数据标注产业发展核心。应当从供给、标准、安全、价值四方面发力，推动高质量数据集建设赋能。

（一）加强数据质量源头发力，推动数据标注产业高质量发展。

供需双向发力带动数据标注产业发展，从“四类供给，两类需求”发力，培育壮大数据标注产业，强化高质量数据集供给应用。供给端：以技术创新提高数据供给质量能力，面向市场持续供给数据产品

服务。发展健全产业链条，促进数据标注产业生态繁荣。完善数据集标准评价体系，规范数据产品质量服务，保障数据集供给质量。人才引育强化数据标注产业供给支撑，构建高水平人才发展体系。需求端：推动公共领域数据开放共享，释放重点领域数据集需求。深化企业数智发展，挖掘人工智能应用场景需求。

（二）强化数据获取与共享，探索行业试点联合推进共建新模式。

一是加强物联网等数据接口开放，广泛汇聚高质量数据，提高原始数据直连比率。建立高质量数据集汇聚平台，推动重点行业高质量中文数据集、思维链数据集和主流价值数据集建设，支持行业专业机构深度参与数据集建设、训练、应用全流程。二是推动公共高质量数据集开放共享，搭建数据集共享平台，加快构建安全数字底座，支持由专业机构配合全流程数据开放合规工作，推动数据集高效安全开放共享。三是鼓励各地因地制宜出台指导意见，探索建立委托授权、模型训练知识产权保护豁免机制，试

点行业间、地区间联合共建数据共享开放交流机制，逐步提升数据流通共享效率。

（三）完善质量与标准体系，推动建设重点行业数据集评价标准。

一是建立数据集质量评估标准，有机融入《国家数据标准体系建设指南》体系。加快研究制定《高质量数据集质量评测规范》等行业高质量数据集质量评估相关标准，建立安全风险、有害内容评估专业数据集，全生命周期把控数据集质量水平。**二是**制定重点行业、主流价值数据标注评估标准，规范数据集接口标准。加快研究制定《高质量数据集数据标注规范》，规范面向人工智能模型训练的高质量数据集数据标注流程。制定合成数据使用标准，平衡好合成数据与原始数据应用的“度”，助力共同发挥最佳作用。**三是**建立数据集流通应用质量评估标准。规范数据集使用、流通范围，明确数据集提供方、使用方、服务方权利义务，建设数据集应用效率评估体系，指导动态分配数据采集、标注资源，提升数据资源利用效率。

（四）加强数据隐私与安全保障，推动数据集安全评估能力建设。

一是强化数据集安全保障技术水平。加强数据伦理、风险评估监管判断技术工具研发，推动构建数据集隔离仓库、原始数据资源池、数据安全屋等措施，加强真实数据保护管理能力。**二是**建立对合成数据集的持续监控评估机制，加强多模态数据融合技术鉴伪能力，建设深度合成鉴伪检测平台，支持联邦学习、差分隐私、可信数据交换等AI安全技术工具发展。**三是**建立跨行业产学研合作平台，加强研究人员、数据工程师、行业专家多方紧密合作，增强算法与数据的匹配度。提高模型算法水平，在数据集处理全过程加入数据可靠性评估分析，提升数据资源利用效能。

（五）优化数据集运营模式，推动数据资源价值生态循环落地。

一是搭建全国一体化的行业高质量数据集供需对接机制和平台，建立数据集资源地图，促进高质量数据集供需对接，推动数据集的流通和共享。**二是**加强政策引导，完善数据集定价和收益分配机制，鼓

励企业探索商业模式创新，实现数据集的可持续发展和应用。加快形成面向高质量数据集的价值循环体系，打造数据集产业生态。三是因地制宜挖掘优势产业，分类开展行业高质量数据集的建设运营及应用工作，以试点先行，逐步推广方式，推动实现区域、行业数据标注产业

和数据集建设生态有序发展。发布高质量数据集建设典型案例，为行业和地方开展高质量数据集建设提供靶向支撑。

(网安所：邓攀科 黄莹莹 韩冰 陈熠祺 胡景朝 王超 温晓君)

联系人电话：18811443518

**思想，还是思想，才使我们与众不同
研究，还是研究，才使我们见微知著**

新型工业化研究所（工业和信息化部新型工业化研究中心）
政策法规研究所（工业和信息化部法律服务中心）
规划研究所
产业政策研究所（先进制造业研究中心）
科技与标准研究所
知识产权研究所
工业经济研究所（工业和信息化部经济运行研究中心）
中小企业研究所
节能与环保研究所（工业和信息化部碳达峰碳中和研究中心）
安全产业研究所
材料工业研究所
消费品工业研究所
军民融合研究所
电子信息研究所
集成电路研究所
信息化与软件产业研究所
网络安全研究所
无线电管理研究所（未来产业研究中心）
世界工业研究所（国际合作研究中心）

通讯地址：北京市海淀区万寿路27号院8号楼1201 邮政编码：100846

联系人：王 乐

联系电话：010-68200552 13701083941

传 真：010-68209616

电子邮件：wangle@ccidgroup.com

赛迪研究院

《网络安全研究》编辑部

编辑部：赛迪研究院

通讯地址：北京市海淀区万寿路27号院8号楼12层

邮政编码：100846

联系人：王 乐

联系电话：010-68200552 13701083941

传 真：0086-10-68209616

网 址：www.ccidthinktank.com

电子邮件：wangle@ccidgroup.com

