新一代数据标注产业对“人工智能+”范式创新的作用机理与实践路径研究

燕江依　李荪　樊威　曹峰

（中国信息通信研究院人工智能研究所，北京 100191）

摘要：数据标注作为人工智能产业基础层的关键环节，其发展质量直接影响人工智能算法模型性能与应用场景落地，是人工智能高质量数据集的核心生产力。系统梳理了数据标注产业的内涵界定、产业链结构、发展模式及政策环境，详细总结了数据标注产业赋能“人工智能+”重点行业应用实践情况，深入剖析了 DeepSeek 等大模型技术革新带来的产业变革，总结了当前数据标注产业存在的顶层设计缺乏、人才瓶颈、技术协同不足等核心问题。研究提出应通过强化国家级标注基地示范效应、提升数据标注技术水平、推进“人工智能+”行业应用水平、构建协同创新生态、完善标准体系、深化国际合作等路径，推动数据标注产业高质量发展。

关键词：人工智能+；数据标注；高质量数据集；行业应用

中图分类号：TP18；F49　　**文献标志码**：A

引用格式：燕江依，李荪，樊威，等．新一代数据标注产业对“人工智能+”范式创新的作用机理与实践路径研究[J]．信息通信技术与政策，2025，51(8)：26-34.

DOI：10.12267/j.issn.2096-5931.2025.08.004

0　引言

数据标注作为人工智能数据服务产业中的重要环节，其核心任务是对数据进行精准的分类、标记和描述，以确保数据资产在全生命周期管控中的准确性和可用性，涵盖数据的采集、存储、分析、流通、应用等各个阶段。数据标注是连接数据资源、算法模型与实际应用场景的关键桥梁，是挖掘数据要素价值的关键环节，是人工智能高质量数据集的核心生产力，在当今信息化、数字化、智能化的时代，数据标注服务产业已经成为推动“人工智能+”行动的重要环节。

1　人工智能数据标注产业界定

全球数据标注产业起源于 1984 年，旨在实现纸质内容电子化，1996 年，澳大利亚澳鹏公司（Appen Ltd.）诞生并布局数据服务领域业务。2007 年，李飞飞等[1]的 ImageNet 项目正式拉开数据标注行业序幕，该项目通过亚马逊公司的劳务众包平台 Amazon Mechanical Turk（AMT）来完成图片的标注和处理，得到的数据集供机器算法训练和学习。此后，全球开始涌现出众多的数据标注企业，数据标注行业也进入成长期。2022 年以后，生成式人工智能技术强势崛起，对高质量、大规模标注数据的需求呈指数级增长，数据标注产业由此步入爆发式增长阶段。

1.1　人工智能数据标注产业定义

从狭义角度来讲，数据标注产业是指对未经处理的原始数据添加说明、解释、分类或编码的过程，以便数据可以被人工智能算法所理解和使用。这一过程主

要是通过人工或半自动的方式，针对特定的数据集进行标注，以形成具有特定格式的结构化数据。通过高质量的数据标注，人工智能系统能够学习到更为丰富和真实的特征信息，进而提升其在各类应用场景中的表现力和泛化能力。狭义的数据标注旨在为人工智能提供标准化“教材”，助力机器实现更为精准和高效的处理与决策。

从广义角度来讲，数据标注产业是指以数据标注为核心的人工智能数据服务上中下游产业链，涵盖数据服务的全生命周期，具体包括数据采集、数据清洗、数据存储、数据标注、数据质测等多个环节。这些环节的协同发展推动了数据要素产业的持续健康发展，并为人工智能产业的快速发展提供了坚实的基础[2]。广义的数据标注产业超越了单一的数据处理环节，包括从原始数据到加工形成高质量数据集的数据基础服务全流程，它涉及到数字经济发展的发展战略和数据资源的整体规划。这一产业不仅承载着推动数据资源汇聚、提升数据质量和盘活数据要素价值的使命，更是数字经济体系中不可或缺的一环。通过加强顶层设计和布局，优化数据标注产业的发展环境，可以进一步释放其潜力，助力数字经济实现更快速、更可持续的发展。

总体而言，狭义的数据标注产业主要关注数据的标注过程和结果，不涉及数据收集、清洗等其他环节，它强调的是如何将人类知识转化为机器可理解的形式。广义的数据标注产业则不仅关注数据的标注本身，还涵盖了与之相关的整个人工智能数据服务产业链和生态系统，通过整合与优化各环节资源，推动人工智能技术的持续进步与广泛应用。

1.2 人工智能数据标注产业链

人工智能数据标注产业链是由资源提供方、数据标注核心服务方、配套支撑方三部分组成，其中资源提供方提供原始数据，同时又是数据标注业务的场景赋能对象。数据标注核心服务方提供数据标注技术服务、平台服务、交易服务和人力服务，有效提高数据价值，助力数据产业价值释放。配套支撑方从标准应用、人才培养、生态培育和安全保障 4 个方面赋能数据标注核心产业。具体产业链组成如表 1 所示。

2 数据标注产业发展现状

2.1 国内外数据标注企业呈多样化分布

从行业供给情况来看，全球数据标注行业企业主要分布在北美、欧洲、亚太等地区，但具有一定规模的企业数量相对较少[3]。北美地区主要集中在美国，数据标注企业较多，突出的特点是技术驱动导向，数据标注服务供给能力和质量较高，代表性企业有 Scale AI、Mighty AI、Mturk 等公司；欧洲地区代表性企业有 Mindy Support 等，但近些年欧洲地区的数据标注企业逐渐将业务转移到人力成本更低的亚太地区和非洲地区等。亚太地区的数据标注供给能力较为强劲，以中国、澳大利亚和印度为主，代表性的企业有海天瑞声（Speechocean）、澳鹏、Infolks、iMerit 等。中国地区的数据标注行业蓬勃发展，涌现出一批如海天瑞声、砺英数智、百度众包、云测数据、标贝科技、数据堂等人工智能基础数据服务企业。我国人工智能数据标注核心企业数量不断增长，产业链体系逐渐完善，呈现出井喷的趋势。预计在未来，随着人工智能产业的不断发展，数据标注相关企业数量将继续增长。

2.2 大模型高质量数据需求海量增长

近年来，大模型的训练数据规模呈现出显著增长趋势。据笔者统计，以 OpenAI 公司的 GPT 系列为例，2018 年发布的 GPT-1 模型，其训练数据量仅为 4.6 GB，而 2023 年的 GPT-4 模型的训练数据量已达到约 40 000 GB，总计包含 13 万亿个 token，这一数据规模的增长接近万倍，凸显了大模型对海量数据的依赖程度；谷歌公司的 PaLM2 模型在 2023 年使用了 3.6 万亿个 token 进行训练，而其 Gemini 模型的数据量也达到了 3.3 万亿个 token；2024 年，Meta 公司推出的 Llama 3 模型训练数据量提升至超过 15 万亿个 token。这些数据表明，大模型的训练数据规模正以惊人的速度增长。此外，大模型的高质量数据集来源也极为丰富，涵盖了文本、图片、音频、视频和多模态等多种形式，这些数据集包含海量的知识信息，涉及各类专业领域和多种语言。例如，ChatGPT、Claude、Llama 以及 DeepSeek 等大模型的训练数据，涵盖了互联网网页、文学作品、百科全书、论文专利、社交媒体以及学术文献等各类知识信息，这种多样化的数据来源，使得大语言模型具备了强大的通用能力和迁移能力，能够应对更广泛的任务和场景[4]。

表 1　数据标注产业上中下游具体组成

产业链位置	核心组成	主要作用
数据标注产业上游：数据资源提供和应用	公共数据	提供公共属性数据，例如政府或公共机构在履行职责或提供服务过程中产生的数据，可以被用于公共利益和社会发展等相关数据
	行业企业数据	提供各个行业、企业在运营过程中产生的数据，例如行业趋势、优化业务流程和提高效率等相关数据
	互联网数据	提供互联网平台和服务商在服务过程中收集的数据，例如用户行为、市场趋势和改进服务等相关数据
数据标注产业中游：高质量数据集开发和治理	算法服务	提供算法支持，帮助优化数据处理流程，提高数据标注的准确性和效率
	技术服务	提供技术支持，包括数据存储、处理和分析工具，以提高数据集的质量和可用性
	平台服务	提供数据标注和管理的平台，使数据标注工作更加系统化和高效
	交易服务	提供数据集的交易和管理，确保数据集的合法流通和使用
	人力服务	提供数据标注和处理所需的人力资源，包括培训和管理数据标注人员
	人才培养	提供培养和吸引数据科学、人工智能和数据标注领域的专业人才，以满足行业对专业技能的需求服务
数据标注产业下游：能力支持	生态培育	建立和维护一个健康的产业生态，包括支持创新、促进合作和提供必要的资源和服务
	数据安全	保护数据免受未授权访问、泄露、篡改或破坏，确保数据的安全性和完整性
	标准应用	制定和推广数据标注和处理的相关标准，以确保数据集的质量和一致性

2.3　全球数据标注市场规模蓬勃发展

数据标注行业作为人工智能领域的重要组成部分，其市场规模正在不断增长。市场咨询机构大观研究（Grand View Research）的报告显示，2022 年全球数据标注市场规模为 22.2 亿美元，预计 2023—2030 年将以 28.9%的年复合增长率增长[5]。近年来，中国数据标注行业发展迅速，规模实现了显著增长。华经产业研究院的报告显示，2023 年数据标注行业规模已经达到了 60.8 亿元，同比增长约 19.69%；2024 年，数据标注市场规模进一步扩大到 120 亿元以上，预计 2025 年可能达到 200～300 亿元[6]。这些数据表明数据标注行业正处于快速发展的阶段，并有望在未来继续保持增长势头。

2.4　数据标注产业政策体系初步形成

国外数据标注产业发展政策呈现多维度、市场化的特点。美国遵循“政府引导、企业参与、市场运作”的发展模式，通过《美国数据隐私和保护法案》等政策法规，不断完善数据要素市场法律体系，推进市场基础设施建设，投入大量资金用于数据采集、存储等环节的设施构建。同时，设立多个数据科学和技术中心，鼓励人才创新创业，并设立监管机构，建立风险评估机制，保障产业规范发展。欧盟遵循“数据一体化市场”战略，通过《通用数据保护条例》《数据法案》等法律法规，构建“欧洲共同数据空间”，整合多领域数据，推动数据自由流通。德国在 2024 年将数字和智能技术相关应用纳入政策重点，通过“制造-X”计划，构建数据空间，激发数据要素价值，推动制造业供应链数字化转型，提升产业竞争力[7]。

为抓住人工智能发展的重大机遇，构筑我国人工智能发展的数据先发优势，近年来我国国家政策利好

频出，针对激活数据要素潜能、加速释放人工智能技术红利做出新部署，政策中多次提及数据标注、确权、流通、共享、交换、审核、验证，为人工智能数据标注服务流程带来新的规范要求。2024 年 5 月，国家数据局提出开展数据标注基地试点，探索建设国家级数据标注基地，重点围绕技术创新、行业赋能、生态培育、标准应用、人才就业和数据安全 6 个方面推进国家级数据标注基地建设，推动数据要素价值释放和人工智能高质量发展，并于第七届数字中国建设峰会主论坛上发布了承担首批国家级数据标注基地建设任务的城市名单，分别是：四川省成都市、辽宁省沈阳市、安徽省合肥市、湖南省长沙市、海南省海口市、河北省保定市、山西省大同市。2024 年 12 月，国家发展和改革委员会、国家数据局、人力资源和社会保障部、财政部 4 个部门联合发布《关于促进数据标注产业高质量发展的实施意见》，旨在推动数据标注产业的高质量发展，为人工智能提供坚实基础，规范行业，促进就业和经济增长，提升国际竞争力，并推动区域经济平衡。

此外，我国各级地方政府也积极出台相关产业规划文件和扶持政策，以人工智能基础数据服务为切入点，寻求人工智能数据标注产业发展的参与机会。2024 年 12 月，山西省大同市印发《大同市数据产业发展三年行动计划（2024—2026 年）》，支持地区数据标注产业发展。2025 年 3 月，沈阳市数据局发布《沈阳市数据标注技术创新指导意见》，旨在进一步推动数据标注技术突破创新，培育壮大数据标注产业，强化数据标注技术对提升数据供给质量的支撑作用，助力沈阳在数字经济赛道上抢占先机。

3 数据标注驱动“人工智能+”行业实践

3.1 人工智能+工业领域

在工业领域，数据标注以覆盖生产全流程的多维度数据为处理核心，涉及设备运行实时参数、生产环境高清图像、产品质检多光谱数据、供应链物流信息等复杂数据类型。在标注过程中，通过结构化标签体系对数据进行层级分类，结合特征提取算法对关键生产指标进行量化标注，构建涵盖设计、加工、装配、检测等环节的标准化训练数据集。这些标注数据为人工智能视觉检测系统提供精准的缺陷识别基准，支撑工业机器人基于标注的装配路径进行自适应调整，同时为生产流程优化模型提供动态参数依据。通过数据标注实现的生产数据深度解析，推动制造环节从传统的事后质检向实时监控、从经验决策向数据驱动转型，不仅提升了生产线的柔性化程度与产品合格率，还促进了供应链上下游的数据协同，加速了工业互联网平台的构建与应用，为智能制造体系的全面落地奠定了数据基础。

3.2 人工智能+医疗领域

在医疗领域，数据标注聚焦医疗数据的专业化与精准化处理，针对医学影像、结构化电子病历、非结构化临床文本、生理信号等多元数据展开深度标注。在标注过程中，需结合医学专业知识进行语义层面的特征标识，如对影像数据中的病灶边界进行像素级标注、对病历文本中的症状描述进行实体关系抽取、对生理信号中的异常波形进行时序特征标注，形成符合临床规范的训练数据集。经过专业标注的数据集为人工智能辅助诊断系统提供了高可信度的参考样本，使其能够精准识别影像中的病变特征；为疾病风险预测模型提供了纵向的临床数据支撑，提升了慢性病早期预警的准确性；同时为个性化治疗方案生成工具提供了患者特征与治疗效果的关联数据，推动治疗方案从标准化向个体化转变[8]。这种基于数据标注的医疗数据价值挖掘，不仅提升了基层医疗机构的诊疗能力，还促进了医疗资源的跨区域协同，为医疗服务的均质化与高效化提供了数据驱动的技术支撑。

3.3 人工智能+教育领域

在教育领域，数据标注围绕教学交互与学习过程的全场景数据展开系统性处理，涵盖课程资源（视频、文本、习题等）、学习行为（登录时长、点击轨迹、讨论发言等）、评测数据（答题结果、错误类型、得分分布等）、教学反馈（教师评价、学生互评等）等多维度信息。在标注过程中，通过对课程内容进行知识图谱构建式标注、对学习行为进行时序模式识别标注、对评测数据进行能力维度映射标注，形成覆盖教与学全流程的结构化训练数据集。这些标注数据支撑智能教学系统实现课程内容的精准推送，根据学生的知识薄弱点动态调整学习路径；支撑学习效果评估模型进行多维度能力画像，突破传统分数评价的局限性；同时为教学策略优化提供实时反馈，帮助教师针对性改进教学方法。通过数据标注实现的教育数据深度应用，推动教育模式从“教师中心”向“学生中心”转型，促进了教育

资源的个性化分配与高效利用,不仅提升了学生的学习主动性与效果,还为教育公平的实现提供了技术路径,加速了教育数字化转型的进程。

3.4 人工智能+交通领域

在交通领域,数据标注以交通场景与运行状态的全要素数据为处理对象,包括道路环境图像/视频(含天气、光照、路面状况等)、车辆运行数据(轨迹、速度、姿态等)、交通参与者信息(行人、非机动车、其他车辆等)、交通设施状态(信号灯、标志标线、护栏等)、交通流数据(流量、密度、延误等)等。在标注过程中,通过目标检测标注(如车辆类型、行人动作)、语义分割标注(如道路区域、车道线)、时序状态标注(如信号相位、车辆变道意图)等方式,构建动态更新的交通场景训练数据集。这些标注数据为自动驾驶系统的环境感知模块提供了丰富的场景理解样本,使其能够精准识别复杂路况并做出决策;为智能交通管控平台的流量调度算法提供了实时参数输入,提升了信号配时的合理性与路网通行效率;同时为出行服务平台的路径规划模型提供了历史与实时数据支撑,优化了出行方案的时效性。基于数据标注的交通数据价值挖掘,不仅推动了自动驾驶技术的迭代升级,还促进了交通管理从被动应对向主动预判转变,为构建高效、安全、绿色的综合交通运输体系提供了数据驱动的解决方案。

3.5 人工智能+金融领域

在金融领域,数据标注覆盖金融业务全链条的多模态数据,包括交易数据(金额、时间、对手方、渠道等)、客户数据(基本信息、风险偏好、行为特征等)、市场数据(利率、汇率、股价、新闻舆情等)、风控数据(违约记录、担保信息、合规报告等)等。在标注过程中,通过对交易数据进行风险特征提取标注、对客户数据进行信用等级映射标注、对市场数据进行趋势模式识别标注,形成符合金融业务逻辑的训练数据集。这些标注数据支撑智能风控模型实现欺诈交易的实时识别与拦截,提升了风险响应速度;支撑信用评估系统进行多维度信用画像,突破了传统信用评价的局限;支撑量化交易模型进行市场趋势预测与策略优化,提高了投资决策的科学性;同时为客户服务系统提供精准的需求画像,实现金融产品的个性化推荐。基于数据标注的金融数据深度应用,不仅提升了金融机构的运营效率与风险抵御能力,还促进了金融服务从线下向线上、从标准化向个性化转型,为普惠金融的推进与金融科技的创新发展提供了数据基础。

3.6 人工智能+能源领域

在能源行业,数据标注聚焦能源生产、传输、消费全周期的关键数据处理,涵盖发电设备运行数据(温度、压力、振动等参数)、能源传输网络数据(管网压力、线路负荷、损耗率等)、终端消费数据(用户用量、时段分布、设备类型等)、环境数据(风速、光照、地质条件等)等。在标注过程中,通过对设备数据进行故障特征提取标注、对传输网络数据进行状态评估标注、对消费数据进行需求模式识别标注,构建覆盖能源系统各环节的训练数据集。这些标注数据为发电设备故障预警模型提供了早期异常识别依据,降低了非计划停机概率;为能源传输网络优化模型提供了负荷分布与损耗特征数据,提升了传输效率与稳定性;为能源消费预测模型提供了用户行为与需求关联数据,支撑精准的供需平衡调度;同时为可再生能源并网模型提供了环境变量与发电效率的关联数据,促进清洁能源的高效利用。基于数据标注的能源数据价值挖掘,推动了能源行业从传统的经验式管理向数据驱动的智能化管理转型,不仅提升了能源系统的运行效率与安全性,还为能源结构调整与“双碳”目标的实现提供了技术支撑。

4 数据标注产业发展趋势

4.1 技术创新成为产业发展核心动力

随着机器学习、深度学习以及大模型算法的不断进步,自动化标注、智能审核及合成数据等新兴技术正逐步走向成熟并广泛应用于实际场景。这些技术通过自动对数据进行分类和标注,显著提升了标注效率与准确性,同时大幅减少了人工工作量。此外,数据标注工具也在不断进化,从单一的人工标注模式向人工标注与人工智能辅助标注相结合的半自动化模式转变,人工智能模型对数据进行预处理后,标注人员在此基础上进行校正,进一步提升了标注效率与质量。

4.2 逐步转向专业化与细分化方向发展

随着数据标注行业的不断发展,其产业结构正在由粗放式扩张向专业化、精细化方向演进。一方面,面对复杂多样的应用场景,企业开始聚焦特定领域,形成专业化的数据标注服务能力。例如,在自动驾驶领域

注重三维点云标注与运动轨迹追踪，在医疗影像识别中强调精准的器官边界划分与病理特征标注。另一方面，标注流程本身也在不断细化，从项目管理、质量控制到人员培训、数据清洗等环节均趋向标准化与专业化，以满足客户对标注结果一致性和准确性的高要求。此外，行业内的分工协作日趋明确，出现了专注于工具开发、平台建设、数据采集或质量审核等不同职能的企业与机构，形成了完整的产业链生态[9]。这种专业化与细分化趋势有助于提升整体行业效率与服务质量，也为从业者提供了更清晰的职业发展路径。

4.3 数据安全与隐私保护愈发重要

在数据标注产业蓬勃发展的同时，数据安全与隐私保护问题愈发凸显，成为产业发展中不容忽视的关键因素。数据标注过程中涉及大量的敏感信息，如个人隐私数据、商业机密等，一旦泄露将给企业和用户带来严重的损失。当前，越来越多的企业开始采用数据脱敏、加密传输、访问控制等技术手段来加强数据安全管理，并通过建立完善的数据生命周期管理体系，确保数据采集、存储、处理与销毁各环节符合相关法律法规要求。同时，随着《中华人民共和国个人信息保护法》《中华人民共和国数据安全法》等政策法规的出台，监管力度不断加大，行业标准逐步完善，进一步推动企业在数据治理方面加大投入。未来，构建可信、透明、可追溯的数据标注环境将成为行业发展的关键方向。

4.4 高素质专业型人才需求不断增大

随着人工智能数据标注产业向专业化、智能化方向发展，对高素质专业型人才的需求日益增大。一方面，产业的技术创新需要具备深厚技术功底的人才，他们能够熟练掌握机器学习、深度学习等相关技术，开发和优化数据标注工具与算法。另一方面，各行业对专业化数据标注的需求，要求标注人员不仅具备数据标注技能，还需掌握相应行业的专业知识，如医疗、金融、交通等领域的专业术语和业务流程等，以便更好地理解标注对象并提升标注质量。此外，随着人工智能辅助标注技术的发展，标注人员还需具备一定的编程能力与平台操作经验，以适应新型工作流程。

4.5 DeepSeek 开启大模型训练数据开发利用的新范式

DeepSeek-R1 模型在后训练阶段使用了强化学习技术，在仅有极少数据的情况下，将数据标注视为提升模型性能的核心因素之一，深入到数据标注的每一个环节，确保每一条数据的精准和高效，极大提升了模型推理能力。其对数据开发利用的独特性具体体现在三方面。一是自动生成高质量数据集，减少传统数据标注需求。DeepSeek 模型训练采用自动化推理和数据生成技术，大幅提升自动化数据标注技术方式占比，传统数据标注需求减少。二是“数据蒸馏+人类协同”技术提升数据标注质量和效率。DeepSeek 通过数据蒸馏技术，从低质量数据中高效提炼生成高质量训练数据，同时采用自动化筛选和人类专家标注反馈机制保障数据标注质量，大幅提升数据标注质量和效率。三是提出强化学习新范式，聚焦高质量推理型数据集。DeepSeek 聚焦高质量推理数据，收集了大约 60 万条推理相关训练样本和 20 万条非推理训练样本，推理型数据与非推理型数据配比约 3∶1[10]，推理训练监督微调数据占比大幅减少。

5 数据标注产业发展面临的挑战

5.1 顶层设计有待完善

当前数据标注产业过程管理和质量控制缺少统一标准，头部数据标注企业主要提供定制化数据标注服务，数据标注结果存在各成体系的现象，企业间数据标注规范难以自发实现统一，数据流通存在门槛。随着人工智能开发中心不断向专业应用拓展，定制化服务占据市场需求主体。据笔者统计，2023 年我国数据标注市场中定制化服务的占比已达 86%，标准化的数据集产品仅占 13%。此外，不同行业对数据标注需求和标准存在差异，这些差异影响了整个数据标注产业的标准化进程。比如医疗行业对数据标注的精度要求极高，任何标注错误都可能导致严重后果；在社交媒体分析中，标注的灵活性和适应性则更加重要，这些行业特定的需求增加了标准化工作的难度，也表明在制定统一的标准体系中需要充分考虑行业的差异性和特殊性。

5.2 技术创新能力有待提升

首先，数据标注技术的研发和市场推广之间存在一定脱节，技术成果未能及时转化为实际应用，导致技术价值未能充分释放。其次，标注技术本身仍存在一些技术瓶颈和算法局限性，例如在某些特定领域或复杂场景下，标注技术受到场景数据质量、标注工具等因

素的限制,准确性和效率仍有待提升[11]。此外,在市场竞争激烈的环境下,不同标注企业往往以保护自身利益为出发点,难以形成合力进行技术协同攻关,并且标注技术复杂性、标准不统一等问题也严重阻碍了企业间的标注技术协同创新,这些因素共同制约了标注技术的广泛应用和协同发展。

5.3 高水平人才缺口较大

随着大模型的发展,高质量数据集的评判标准变得更加复杂,要求标注者必须具备更深层次的理解和分析能力,以及更高的逻辑思维和专业知识体系要求。同时,在处理复杂、多模态数据时,专业技能和学术素养变得尤为重要,导致部分项目高水平数据标注人才短缺。此外,行业场景的多样化促使数据需求量长期持续增长,对数据标注人员的需求进一步扩大。猎聘大数据研究院研究数据显示,2024 年数据标注岗位数量增长速度较 2023 年大幅提高,但高质量数据集的高要求与低产能成为数据标注企业发展的痛点。

5.4 专业化平台能力不足

当前数据标注平台面临多重技术挑战与生态适配困境,制约行业高质量发展。首先,平台基础设施建设能力薄弱,受限于行业发展周期短及资源约束,多数企业在数据采集、处理、标注及流通环节存在显著技术短板,自建智能化处理平台能力不足,核心算法研发与高质量数据集平台化处理水平亟待提升,尤其在应对大规模数据时普遍存在性能“瓶颈”与智能化辅助功能缺失问题。其次,平台功能体系与可靠性存在缺陷,现有系统在高并发场景下易出现响应延迟或服务中断,严重影响标注效率与连续性。此外,信创生态适配能力不足问题突出,平台对硬件架构、操作系统及数据库的兼容性研发投入不足,未能有效整合技术生态资源以优化系统性能,制约了技术迭代与稳定性提升[12]。

6 数据标注产业发展建议

6.1 提升数据标注技术创新能力

鼓励各地区与行业头部企业联手共建数据标注技术创新联合实验室,持续加大在数据标注工具与机器学习等智能算法融合方面的研究力度,致力于提升标注工具在效率、质量、精度和稳定性等多方面的性能指标。同时,积极开展产学研合作,与高校、科研机构携手共同开展前沿技术研究,加速科技成果向实际应用的转化,持续推动数据标注技术的创新与发展,为产业升级注入源源不断的动力。

6.2 推进“人工智能+”行业应用水平

高质量行业数据集为传统产业的数字化、智能化转型提供了坚实支撑,有力推动了行业整体发展水平的提升。为了实现这一目标,应深入挖掘“人工智能+各个行业”的数据标注需求,支持公共数据在“人工智能+多领域”的标注与开发利用,并积极推动数据标注服务纳入政府采购范围。同时,鼓励企业加大对数据的开发利用力度,激发企业释放更多的数据标注需求,共同建设高质量的行业数据集,为人工智能技术在多领域的应用赋能。此外,数据标注企业应与各行业开展深度合作,推动标注数据在新型工业化、智慧教育、智能诊断、金融风险评估等具体场景中的应用,助力企业优化业务流程、增强市场竞争力,加速实现“人工智能+”智能化转型[13]。

6.3 构建数据标注生态体系

加速构建数据标注生态,通过实施“龙头引领+中小微孵化”双轮驱动策略,加速构建完善的产业链、价值链和生态系统。一方面,集中资源培育和引进数据标注龙头企业,发挥其在技术、资金和市场方面的优势,引领产业方向,制定行业标准,推动数据标注技术的创新与应用。另一方面,通过税收优惠、资金扶持和创业空间等为中小微企业提供良好的孵化环境,激发中小企业的创新活力,形成产业链上下游的协同发展。此外,支持龙头企业与中小企业建立紧密的合作关系,促进资源共享与优势互补,共同开展项目研发和业务合作,实现互利共赢。

6.4 推动数据标注标准应用

积极推动数据标注标准编制和应用,鼓励数据标注头部企业积极参与数据标准产业标准的制定,构建涵盖技术、质量、流程等多维度的标准框架体系,加快制定国家标准与行业标准,为数据标注提供明确规范。同时,推动标准在实际标注过程中的广泛应用,通过实践不断检验和完善标准体系,促进数据标注产业的规范化与高质量发展[14]。此外,建立健全标准实施与监督机制,强化对数据标注企业和项目的监督检查,确保标准有效执行。

6.5 培养数据标注高水平人才

加强数据标注人才培育力度。通过设立实训基

地、举办职业技能大赛等多种形式，推动产教融合发展，培育高端标注人才队伍，形成对就业的带动效应。此外，支持高校和职业院校开设数据标注相关专业和课程，结合产业需求更新教学内容，培养适应数据标注产业发展的专业人才。鼓励行业联盟、高校、科研院所与企业建立长期合作机制，共同开展科研项目和人才培养，实现资源共享、优势互补，推动数据标注技术的创新和应用。

6.6 保障数据标注安全可靠

持续加强数据安全防护力度，搭建数据标注安全溯源机制，推动数据标注安全生产环境建设，开展数据合规认证，建立完善的数据安全管理体系，加强数据在采集、传输、存储、处理等全生命周期的安全防护，采用加密、权限管理等技术手段，防止数据泄露、篡改和滥用[15]。此外，加强员工的数据安全培训，提高安全意识，定期开展安全审计和风险评估，及时发现和整改安全隐患，确保数据标注过程的安全可靠。

6.7 促进数据标注国际合作

依托我国数据基础设施优势，鼓励国内企业承接数据标注国际业务，深化数据标注领域技术及产业合作，推动我国数据标注企业逐步走向国际市场，拓展海外业务，为国际供给一批符合我国社会主义核心价值观的高质量数据集。同时，开展数据标注科技人才国际交流，培养一批具有国际视野的数据标注人才，加速人才链与产业链的有效国际对接，显著增强我国在全球数据标注产业中的话语权和影响力。此外，支持企事业单位牵头制定数据标注国际标准，主导形成国际统一的数据标注标准和共享机制，促进数据标注产业高质量、国际化发展。

7 结束语

当前，我国数据标注产业已迈入以规模应用反哺技术跃升、以高质量数据驱动“人工智能+”场景落地的新阶段。数据标注作为连接数据资源、算法模型与“人工智能+”实际应用场景的关键桥梁，已成为各国科技竞争的关键要素。加快研发多模态、跨领域、人机协同的智能化标注技术和工具，培育高水平、专业化的数据标注人才，构建可信、可控、可流通的高质量数据集供给体系，打造贯通“数据资源—标注服务—算法训练—场景应用”的完整产业生态，有利于加速人工智能赋能千行百业，促进我国人工智能与数据要素产业高质量蓬勃发展。

参考文献

[1] 潘剑宜．整数智能：人工智能行业的数据合伙人[J]．杭州科技，2025，56(2)：30-33.

[2] 刘瑜．数据标注是人工智能的基石工程[N]．西宁晚报，2025-03-29(A03)．

[3] 王峰，张天意，朱方昊，等．数据标注技术在人工智能领域的研究和应用[J]．信息技术与标准化，2024(12)：22-26.

[4] 黄丽．生成式人工智能训练数据的软硬法协同治理研究[J]．宁夏大学学报(人文社会科学版)，2024，46(1)：112-121.

[5] Grand View Research. Data collection & labeling market size, share & trends analysis report, 2023-2030 [R], 2023.

[6] 华经产业研究院．2024—2030 年中国数据标注行业全景调研及发展趋势研究报告[R]，2024.

[7] 陈兵，傅小鸥．生成式人工智能数据训练的法治基调及展开[J]．辽宁师范大学学报(社会科学版)，2024，47(3)：1-10.

[8] 高宏旭，曹大军．人工智能中数据集的分类，获取与处理[J]．科学大众：科技创新，2020(5)：62-63.

[9] 黄熙．人工智能数据集标注工具研究与开发请补充卷期和页码．中国宽带，2024，20(6)：67-69.

[10] DeepSeek-AI. DeepSeek LLM：scaling open-source language models with longtermism[J]. arXiv Preprint, arXiv：2401. 02954, 2024.

[11] 马俊．人工智能幻觉，怎么破[N]．环球时报，2025-06-13(008)．

[12] 陈俊秀，徐玉琴．人工智能大语言模型引发的数据污染风险及其规制路径[J]．大连理工大学学报(社会科学版)，2025，46(4)：55-62.

[13] 苏德悦．数据标注产业乘风起航加速发展[N]．人民邮电，2025-06-16(03)．

[14] 叶菁．高质量数据集驱动 AI 模型突破与创新[N]．通信信息报，2025-06-11(02)．

[15] 李雨泽．大数据时代背景下人工智能(AI)的创新与应用趋势研究[C]//《中国招标》期刊有限公司．新质生产力驱动第二产业发展与招标采购创新论坛论文集(二)．北京：《中国招标》期刊有限公司，2025：70-71.

作者简介：

燕江依 中国信息通信研究院人工智能研究所工程师，主要从事人工智能数据质量与模型性能闭环反馈机制与方法、人工智能数据集质量评估体系和工具平台研发、人工智能高质量数据集建设路径以及人工智能高质量数据集标准体系设计等方面的研究工作

李荪 通信作者。中国信息通信研究院人工智能研究所高级工程师，主要从事人工智能政策、标准、产业研究，涵盖机器学习、语音感知认知技术以及产品融合应用等方面的研究工作

樊威 中国信息通信研究院人工智能研究所高级工程师，主要从事人工智能高质量数据集建设及数据标注等方面的研究工作

曹峰 中国信息通信研究院人工智能研究所高级工程师，人工智能关键技术和应用评测工业和信息化部重点实验室副主任，主要负责牵头可信 AI 人工智能评测标准体系和能力建设，以及工程化能力等相关评估规范的研制与评测工作

Research on the mechanism of action and practical path of the new generation of data annotation industry on the innovation of the "AI+" paradigm

YAN Jiangyi, LI Sun, FAN Wei, CAO Feng

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

Abstract: Data annotation, as a key link in the foundational layer of the artificial intelligence industry, directly affects the performance of artificial intelligence (AI) algorithm models and the implementation of application scenarios, and is the core productive force of high-quality AI datasets. This paper systematically reviews the connotation definition, industrial chain structure, development model and policy environment of the data annotation industry, presents a detailed summary of the application practices regarding how the data annotation industry empowers the "AI+" initiative across key sectors, deeply analyzes the industrial transformation brought about by the technological innovation of large models such as DeepSeek, and summarizes the core problems existing at present, such as the lack of top-level design, talent bottlenecks, and insufficient technological collaboration. This study proposes that the high-quality development of the data annotation industry should be promoted through paths such as strengthening the demonstration effect of national-level annotation bases, improving the technical level of data annotation, continuously advancing the application of "AI+" in key industries, constructing a collaborative innovation ecosystem, improving the standard system, and deepening international cooperation.

Keywords: AI+; data annotation; high-quality dataset; industry application

(收稿日期：2025-06-30)