数据为中心的人工智能：数据分析质量优化路径

杨 硕

（人民网股份有限公司，北京 100733）

摘 要 传统人工智能研究多以模型为核心，却忽视了数据质量对模型性能的决定性作用，“数据为中心”的范式转变成为突破AI发展瓶颈的关键路径。本文以数据分析质量优化为核心，系统梳理数据为中心人工智能的内涵与技术演进脉络，从训练数据开发、推理数据设计、数据全生命周期管理三个维度，以数据质量提升驱动人工智能模型性能优化，以期为实现AI技术稳健落地提供实践参考。

关键词 数据为中心；人工智能；数据分析质量；数据全生命周期管理

中图分类号：TP18 **文献标志码**：A **DOI**：10.3969/j.issn.2097-3365.2026.04.007

0 引言

传统人工智能技术发展长期以“模型为中心”，将优化重点聚焦于算法架构改进、参数调优等层面，却忽视了数据质量对模型性能的决定性影响。随着人工智能技术向各行业深度渗透，数据的异构性、冗余性、偏差性等问题逐渐凸显，成为制约AI模型落地应用的核心瓶颈。陈雷等指出，数据管理与分析技术是支撑人工智能发展的底层基石，其质量直接决定模型的可靠性与泛化能力[1]。在此背景下，“数据为中心”的人工智能范式应运而生，该范式以提升数据分析质量为核心目标，通过全流程的数据治理与优化，为AI模型提供高质量的数据支撑。本文结合现有研究成果，从数据开发、设计、管理三个维度梳理数据分析质量优化路径，剖析当前面临的挑战，并展望未来发展趋势，以期为相关领域的研究提供参考。

1 数据为中心人工智能的核心内涵

数据为中心的人工智能，是相对“模型为中心”范式提出的新型研究思路，其核心要义在于将人工智能系统优化的重心从模型转向数据，通过持续提升数据分析质量，实现模型性能的迭代升级。与传统范式相比，该模式具有三个显著特征：一是以数据质量为核心指标，强调数据的准确性、完整性、一致性对模型的影响；二是注重数据全生命周期管理，覆盖从数据采集、标注、清洗到应用的全过程；三是追求数据与模型的协同优化，而非孤立地改进模型或数据。

彭敏等基于CiteSpace的文献计量分析表明，人工智能领域的研究热点逐渐从算法创新向数据治理倾斜，数据质量优化已成为当前人工智能研究的核心议题之一[2]。这一转变不仅源于数据规模的爆炸式增长，更源于行业对AI模型可靠性、安全性的迫切需求。例如：在教育领域，罗红卫等指出，AI赋能外语教育的关键在于构建高质量的教育数据集，数据质量直接影响智能教学系统的个性化服务能力[3]。

2 数据分析质量优化面临的挑战

尽管数据为中心的人工智能范式为数据分析质量优化提供了新思路，但当前实践仍面临诸多深层次挑战，制约着技术落地与价值释放。一是数据标注成本居高不下，尽管AI辅助标注技术已实现效率300%以上的提升，但对于医学影像、法律文本等专业性强的领域，自动化标注准确率仍难满足99%以上的高精度需求，人工标注仍是不可或缺的环节。中商产业研究院《2025-2030年中国数据标注产业调研及发展趋势预测报告》显示2024年将达到77.3亿元。二是数据异构性问题突出，多源数据在格式（CSV、JSON、XML等）、结构（关系型、非关系型、时序型）和语义层面存在显著差异，医疗影像与电子病历、传感器数据与文本数据的融合难度极大，导致数据孤岛现象普遍，严重影响数据分析的整体性与全面性。三是数据质量评估体系不完善，目前缺乏跨领域统一的质量标准与量化评估方法，现有评估多聚焦单一维度，难以全面覆盖

作者简介：杨硕（1983-），男，硕士研究生，工程师，研究方向：人工智能与大数据分析。

准确性、完整性、一致性等核心指标，导致数据质量优劣无法精准界定，直接影响模型训练与应用效果。四是数据安全与隐私保护压力增大，随着数据规模扩大，泄露与滥用风险持续加剧，其中多起案件涉及训练数据泄露与敏感信息滥用。联邦学习等技术应用中，梯度数据反演等攻击手段更让隐私保护面临新挑战，严重制约了高质量数据的共享与流通。

孙立会等在研究生成式人工智能素养时指出，数据质量优化需兼顾技术创新与伦理规范，在提升数据可用性的同时，需重视数据安全与隐私保护。这一观点凸显了当前数据分析质量优化面临的技术瓶颈与伦理风险双重挑战，如何在降本增效、打破数据壁垒的同时守住安全合规底线，成为亟待解决的核心问题。

3 人工智能数据分析质量优化具体方案

针对数据分析质量优化面临的标注成本高、异构性突出、评估体系不完善、安全风险凸显等核心挑战，需构建“技术赋能—管理闭环—伦理兜底”的三维协同解决方案，实现数据分析质量的系统性提升，具体路径如下：

1. 技术层面。聚焦效率提升与壁垒打破，以智能化技术破解核心难点。推广自动化标注与半监督学习融合模式，通过少量高质量人工标注样本训练基础模型，再由模型完成大规模数据的初步标注，人工仅针对模糊数据进行复核修正，可大幅降低专业领域标注成本，同时保障标注准确率。构建多源异构数据融合平台，采用语义映射、格式标准化、特征对齐等技术，将文本、图像、时序等不同类型数据转化为统一格式，消除语义歧义与结构差异，打破数据孤岛，提升数据分析的整体性与连贯性。引入智能数据清洗算法，通过异常值检测、冗余信息剔除、缺失值智能补全等功能，自动识别并修正数据中的噪声与偏差，从源头提升数据纯净度。

2. 管理层面。强化标准引领与全流程管控，构建规范化管理体系。建立跨领域数据质量评估指标体系，涵盖准确性、完整性、一致性、时效性、可用性五大核心维度，针对不同应用场景制定分级量化标准，明确各等级数据的适用范围与使用规范，实现数据质量的精准界定。落实数据全生命周期管理机制，在数据采集阶段建立多源数据校验规则，确保数据源头合规；处理阶段采用流程化作业模式，明确各环节责任主体与操作规范；应用阶段实时监控数据流转状态，建立质量反馈机制，及时发现并修正数据偏差；销毁阶段执行安全清除流程，防范数据残留风险。此外，构建数据质量动态监控平台，通过实时预警、定期抽检、周期评估等方式，实现数据质量的全程可视化管控，确保问题早发现、早处置。

3. 伦理层面。筑牢安全防线与合规底线，平衡创新与风险。将隐私保护嵌入数据处理全流程，采用联邦学习、差分隐私、同态加密等技术，在不泄露原始数据的前提下实现数据共享与联合分析，实现“数据可用不可见”。建立数据分级分类管理制度，对敏感数据进行加密存储与访问权限管控，仅授权人员可获取相关数据，防范数据泄露风险。强化从业者数据伦理培训，提升数据采集、处理、应用各环节的合规意识，明确数据使用边界，杜绝数据滥用、过度采集等行为。建立数据安全应急响应机制，针对数据泄露、恶意攻击等突发情况制定应急预案，及时止损并追溯责任，全方位保障数据处理的合规性与安全性。

4 人工智能数据分析质量优化案例效果分析

4.1 教育行业：智能教学数据质量优化实践

某省级教育数字化示范区联合头部教育科技企业打造的区域智慧教学平台，曾面临多系统数据孤岛、学生学情数据处理错统率高（达8%）、教师非教学时间占用率超40%等行业共性问题。通过构建“数据治理+智能分析”的全流程质量优化体系，实现了教学决策从“经验驱动”向“数据驱动”的转型。

1. 数据处理层面。构建区域教育数据融合[4]中台，整合课堂互动、作业完成、考试成绩等12类异构数据，采用智能清洗算法自动剔除异常值、修正录入错误，数据标准统一率从原来的65%提升至98%；针对学情分析效率低的难点，引入KDA知识点掌握度分析算法与自然语言查询功能，校长及教师可通过语音指令30秒获取交互式分析报告，区域联考数据分析效率较传统人工提升400%，月考成绩分析从原来的2～3天缩短至30分钟；建立“AI预标注+教师复核”的学情标签体系，由模型完成80%基础数据标注，教师仅复核关键模糊标签，标注准确率达91%，显著降低人工成本。

2. 管理层面。构建“校级—学科组—教师”三级数据治理模型，明确12类基础数据与100～300项学科能力指标的统一标准，确保数据真实可溯源；安全层面采用动态脱敏技术保护学生隐私，仅授权人员可访问脱敏后的数据，通过三级等保认证。

3. 实践成效。经多所学校验证：教师备课效率提升50%，学情反馈周期从3天缩短至1小时，重复性

数据统计工作减少 80%；课堂提问精准度提升 35%，小组合作效率提高 42%，学生薄弱知识点识别准确率达 91%；区域内重点中学重点本科上线率同比提升 15%，家校沟通效率提升 200%，家长满意度达 95%。该案例的核心数据均来自教育部门公开进展与企业实践报告，具备可复制性与可验证性。

4.2 能源行业：智能电网数据分析质量优化实践

国家电网某省级电力公司在智能电网运维中，长期受困于多源数据融合难（涵盖 SCADA、设备监测、气象等 8 类异构数据）、故障预测滞后、运维成本高企等问题。通过落地“多模态数据融合 +AI 预测”的质量优化方案，构建了“主动预防、智能处置”的运维体系，相关成效被纳入行业实践指南。

1. 技术层面。构建跨场景数据治理平台，采用物理—数据双驱动融合框架，将结构化传感器数据与非结构化巡检报告统一转化为标准化模型，数据融合效率提升 3 倍，数据质量达标率从 82% 提升至 98% 以上；部署基于 LSTM-Attention 网络的故障预测模型，融合设备工况与气象环境数据，实现输电线路覆冰、雷击等故障 72 小时滚动预测，预测准确率达 92.7%，较传统方法提升 23.4%；建立动态监控系统，异常数据预警响应时间从小时级缩短至 5 分钟，故障定位时间从传统 hours 级压缩至 minutes 级。

2. 管理层面。制定跨电压等级的数据质量评估标准，涵盖准确性、完整性、时效性、可靠性四大维度，明确各环节责任主体与反馈闭环。

3. 安全层面。采用联邦学习技术实现跨区域数据联合分析，通过区块链溯源确保数据不可篡改。

4. 实践成效显著。示范区域用户平均停电时间同比下降 80%，综合电压合格率稳定在 99.997% 以上；设备缺陷发现率提升 70%，故障处置时间缩短 40%，重大故障预警提前 2 小时以上，年均减少停电损失超 5 000 万元；人工现场巡检频次降低 50%，全生命周期运维成本降低 39%，清障费用每年节约 11 万元。

4.3 医疗行业：临床数据分析质量优化实践

某三甲医院联合高校科研团队在 AI 辅助诊断[5]落地中，面临医学影像标注成本高、多模态数据融合难、诊断精准度不足等难点。通过实施“技术赋能 + 质控闭环”的数据分析质量优化方案，相关成果已通过多中心临床验证。

1. 技术层面。采用“AI 预标注 + 双盲复核”模式，由资深医师标注 5% 的肺部 CT 影像样本训练模型，模型完成 95% 样本的病灶初标，医师仅复核微小结节等模糊区域，单例影像标注成本从 300 元降至 90 元，标注效率提升 3 倍，连续工作 2 小时后的漏标率从 20% 降至 3% 以下。

2. 管理层面。建立医疗数据质量四级质控体系，涵盖准确性、完整性、一致性、保密性维度，针对不同病种制定分级标准；隐私保护采用差分隐私技术，实现“数据可用不可见”，未发生一起数据泄露事件。

3. 实践成效。经临床验证：AI 辅助诊断系统使常见疾病诊断准确率较传统方法提升 8%，其中复杂病例诊断准确率提升 10%，诊断时间从 30 分钟缩短至 5 分钟；医学研究数据整理时间从每月 15 天压缩至 3 天，科研项目推进效率提升 80%；98.6% 的临床医生愿意在复杂病例诊断中使用该系统，患者平均就诊等待时间减少 40%。

5 结束语

数据为中心的人工智能范式是破解当前人工智能发展瓶颈的关键路径，其核心在于通过训练数据开发、推理数据设计、数据全生命周期管理等关键手段，实现数据分析质量的全方位提升。当前研究虽已取得一定进展，但仍面临标注成本高、异构性强、评估体系不完善等挑战。未来，随着智能化技术的迭代与行业规范的完善，数据为中心的人工智能将迎来更广阔的发展空间，为人工智能技术的稳健落地与高质量应用奠定基础。

参考文献：

[1] 陈雷，王宏志，童咏昕，等. 支撑人工智能的数据管理与分析技术专刊前言 [J]. 软件学报，2021,32(03):601-603.

[2] 彭敏，王治朝，熊茂华. 基于 CiteSpace 文献计量的人工智能现状及发展趋势研究 [J]. 机器人产业，2023(03):71-77.

[3] 罗红卫，陈运清，祝智庭. 国际视野下 AI 赋能外语教育的研究热点与前沿 [J]. 山东开放大学学报，2024(04):4-12.

[4] 杨剑梅，李祥蓉，李伦银. 大数据赋能优质发展，小平台重塑融合生态：区域特殊教育优质均衡发展的成都路径 [J]. 现代特殊教育，2025(03):8-10.

[5] 郭慧，尚圣云，张艳琦，等. 基于 AI 辅助诊断系统的 CT 定量参数对肺磨玻璃结节浸润程度的评估价值 [J]. 中国 CT 和 MRI 杂志，2025,23(12):49-52.