

构建企业级人工智能高质量数据集：方法与路径

姜春宇, 白玉真, 刘渊, 王超伦

中国信息通信研究院, 北京 100191

摘要

当前, 我国人工智能数据集面临质量评估方法缺失、能力建设体系不明确等挑战。梳理了人工智能数据集的构成和分类, 结合结构化数据质量评估, 提出一套人工智能数据集质量评估方法, 并基于产业实践, 提炼出企业高质量人工智能数据工程体系与能力建设路径。最后给出了我国建设高质量数据集的政策建议。

关键词

人工智能数据集; 数据质量评估; 人工智能数据工程

中图分类号: TP311.13

文献标志码: A

doi: 10.11959/j.issn.2096-0271.2025088

Building high-quality datasets for enterprise-level artificial intelligence: methods and pathways

JIANG Chunyu, BAI Yuzhen, LIU Yuan, WANG Chaolun

China Academy of Information and Communication Technology, Beijing 100191, China

Abstract

Currently, the artificial intelligence (AI) datasets in our country face challenges, including the lack of quality evaluation methods and an unclear capability-building system. This article reviews the composition and classification of AI datasets and combines structured data quality assessment to propose a set of evaluation methods for AI dataset quality. Based on industry practices, it distills a high-quality AI data engineering system, summarizes the pathways for enterprises to develop capability building, and provides policy recommendations for constructing high-quality datasets in our country.

Key words

artificial intelligence datasets, data quality assessment, AI data engineering

0 引言

在全球范围内，数据作为基础资源和创新引擎的重要性已获得广泛认可。作为新型生产力的重要元素，数据以其独特的增值方式推动了科技革命与产业变革，提高了全要素生产率。我国已将数据视为生产要素，美国、欧盟等主要经济体也高度重视数据价值，将其纳入国家战略规划中。随着国家数据局的成立，一系列文件（如《“数据要素×”三年行动计划（2024—2026年）》及《关于加快公共数据资源开发利用的意见》）相继发布，致力于实现数据“供得出”“流得动”“用得好”的目标，数据要素产业迎来蓬勃发展。

近年来，以 GPT4、DeepSeek R1、LLMA 为代表的大语言模型（简称大模型）技术取得了爆发性进展。这一波人工智能的发展趋势遵循所谓的“Scaling Law”，即通过独立增加数据量、模型参数规模或训练时间，持续提升模型性能。这表明，大模型已经成为释放数据价值的主要途径之一。以数据为中心的人工智能理论强调，优质的人工智能需要高质量、大规模且多样的数据集。从 GPT 系列的迭代过程中可以看出，其数据集规模从 GPT1 的 4.8 GB 增长至 GPT3 的 570 GB，其对数据质量的要求也在不断提高^[1]。

我国数据资源总量优势明显，《全国数据资源调查报告（2023 年）》显示，2023 年全国数据总量达 32.85 ZB，然而原始数据到高质量人工智能数据集还存在较大的转化鸿沟，面临开放不足、工具缺失、标准不完善、流通不畅通、方法论缺失等挑战^[2]。虽然开源大模型开始深入企业，但由于产业界尚未建立完善的人工智能数

据工程体系与治理体系，也缺少相关技术工具和方法论，限制了大模型在企业层面的进一步应用。为了解决这些问题，本文对人工智能数据集的构成与分类进行了分析，并提出一套评估数据集质量的方法。基于实际产业经验，本文将探讨构建和运营高质量人工智能数据集的工程体系，并梳理分步建设的路径，旨在凝聚共识，建立一套人工智能数据集建设与运营的通用方法，以指导各类企业、机构开展相关能力体系构建。

1 人工智能数据集的构成与分类

1.1 人工智能数据集的构成

人工智能数据集旨在为机器学习模型提供必要的训练资料，它包含了大量标注的图像、文本、语音等数据样本，这些样本用于训练人工智能系统识别和学习各种特征模式。具体来说，一个数据集由 4 个主要部分构成：特征、标签、元数据以及样本。

特征指的是数据集中的输入变量，它们描述了每个样本的具体属性。例如，在预测一个人的体重时，身高和年龄就是特征。根据不同的应用场景，特征可以分为数值型（如年龄、温度）、类别型（如性别、颜色分类）、文本型（如评论内容）和时间序列型（如股票价格随时间变化的数据）等多种类型。

标签则是监督学习中的目标变量，代表了模型试图预测的结果。比如在体重预测中，实际体重数值是标签；在图像分类任务中，“车”或“非车”的标识也是标签。标签为监督学习提供了“正确答案”，使模型能够通过学习特征与标签之间的映射关系来提升预测能力。传统上，标签需

要人工手动标注，但随着大模型技术的进步，利用大模型实现自动化标注正逐渐成为可行方案。

每个单独的数据记录被称为样本，它由一组特征向量和对应的标签组成。例如，在一个包含多个人员信息的数据集中，每个人的年龄、性别、身高等特征组合成一个特征向量，并配以该人的体重作为标签，共同构成了一个样本。

通过这样的结构，人工智能数据集不仅支持从简单到复杂的各种机器学习任务，而且促进了模型对特定领域知识的学习与应用。随着技术的发展，高效地构建高质量的数据集，特别是利用先进的自动化工具进行数据标注，已成为推动人工智能进步的关键因素之一。

1.2 人工智能数据集的分类

根据大模型的生成流程，人工智能数据集通常可以被分为3类：预训练数据集、指令微调数据集和评估数据集。预训练数据集指的是大模型在初始训练阶段使用的最大规模的数据集合，这类数据集主要由通用数据集和特定领域数据集构成。以大模型为例，通用数据集来源广泛，包括网页内容、书籍、学术文献、编程代码、平行语料库、社交媒体和百科全书等，旨在为模型提供丰富的语言素材和基础知识^[3]。特定领域数据集则是针对特定行业或主题量身定制的数据集，如增强金融领域性能的FinCorpus、FinGLM，以及提升医疗领域准确性的Medical-pt，目的是提高大模型在特定领域的专业能力和准确性。数据集的规模、多样性和质量会影响大模型的整体质量，其中多样性非常重要，多样化的数据来源能够为模型提供更全面的语言理解和知识基础，而网页数据由于其高可

访问性和丰富性，成为构建预训练数据集的重要资源^[3]。

指令微调数据集通常由一系列的问答对组成，问题一般是向大模型发出的请求或指令，答案一般是根据请求生成的响应。指令微调数据集用于进一步微调预训练的大模型，使大模型能够更好地理解和遵循人类指令，从而提高大模型的能力。指令微调数据集要么由人手动构建，要么利用模型来生成，或者直接利用现有的开源数据集。

评估数据集是一类精心设计的数据样本，主要目的是评估大模型在各种任务中的表现，为大模型提供性能测量的标准。通过评估数据集，研发人员可以衡量大模型的性能，识别大模型可以优化的方向，并间接衡量训练、指令、偏好数据集的质量。针对大模型的评估数据集通常覆盖考试、科目、自然语言理解、推理、长文本、代码、法律、医学、金融等20多个方面，表1整理了代码生成、考试、自然语言理解、数据与逻辑推理、多模态能力评估中的典型评估数据集。从趋势来看，越来越多的评估数据集尝试提供自动化的评估工具，以此来降低评估的门槛，同时支持动态的内容更新，不断提升评估的能力。

1.3 人工智能数据集的来源

通过梳理能够发现，人工智能数据集主要来自公开数据集、企业自有数据、外部机构数据和合成数据^[4]。公开数据集指的是通过网络爬取获得的数据集或互联网上的开源数据集，比如ImageNet、Librispeech等数据集。企业自有数据指的是企业在业务经营中积累的数据，分散在企业的业务系统、数据平台、知识库或物联网设备中。外部机构数据是与其他机构协作共享的数据或采购的数据，比如医疗

表1 典型评估数据集

名称	描述	类别
HumanEval	包含 164 个手工编写的编程题	代码生成
MMLU	包含 57 个子任务的英文评测数据集,涵盖初等数学、美国历史、计算机科学、法律等领域	考试
GLUE	包含了9个不同的自然语言理解任务,如文本蕴含、情感分析、语义相似性等	自然语言理解
GSM8K	包含了 8 500 道小学数学题,检验多步推理能力	数据与逻辑推理
MovieQA	视频问答 (VideoQA), 包含 15 000 个多项选择题,覆盖 400 多部电影的视觉和文本信息	多模态

机构间共享的疾病数据库、政府机构开放的公共数据等。合成数据指的是基于专用数学模型或算法生成的数据，一般可通过深度学习算法生成、统计模型生成、物理仿真模拟生成，具备隐私保护、可扩展数据规模、提升数据获取效率等优点。随着大模型的演进，合成数据正成为重要的数据来源。

2 人工智能数据集质量维度与评估方法

数据质量是对数据的定量和定性属性的全面特征描述与测量，根据应用场景的不同，数据质量的内涵不同。长期以来，围绕结构化数据的质量，产业界基本形成了一套比较成熟的质量度量维度和质量管理方法。从质量度量维度来看，结构化数据的质量主要考察六大维度，即规范性、完整性、准确性、一致性、时效性、可访问性^[5-6]。从质量管理来看，结构化数据质量管理形成了质量的需求梳理、质量评估、质量分析和质量提升4个主要环节。人工智能时代，数据集的质量愈发重要，其质量显著影响大模型的效果、公平性、稳定性和安全性，然而当前尚未对数据集质量定义形成统一共识，也没有建立一套比较科学的数据集质量管理框架。

2.1 人工智能数据集质量维度

相比结构化数据的质量管理，人工智能数据集质量维度更加复杂：一是因为数据的模态更多样，从结构化数据拓展到文本、图片、音视频等数据类型；二是因为要考虑场景和行业特点；三是因为要考虑数据集的内容安全性^[7-9]。在机器学习时代，数据质量维度通常考虑4个维度，即内在维度、情境维度、表示维度和可访问维度。具体来看，内在维度主要有正确性、冗余性、可信度3个指标；情境维度有均衡性、完整性、全面性、无偏差异性、多样性等指标；表示维度主要一致性和表示一致性两个指标；可访问性主要考虑可用性指标^[10]。在材料科学领域，面向机器学习的数据质量维度可参考9个指标，包括可追溯性、完整性、时效性、一致性、准确性、冗余性、均衡性、归一化、洞察力^[11]。对文本数据的质量评价，可以参考4种特定质量标准，包括写作风格、包含的事实数量、教育价值和理解文本需要的专业水平^[12]。

基于理论研究和实践探索，本文提出了一套人工智能数据集质量评价的维度，评估维度涵盖基础维度、安全维度、内容维度和应用维度四大类，包括准确性、合规性、全面性、均衡性等12个一级指标（如图1所示）。

从基础维度来看，主要包括数据集的规范性、准确性、完整性和一致性。规范性衡量数据集结构化程度，确保其遵循既定格式与内容准则。准确性衡量数据内容的正确性程度。完整性衡量数据集中应被赋值的数据属性的赋值程度。一致性衡量数据集在不同存储位置、不同时间、不同系统间的格式，以及格式转换时的一致性。

从安全维度来看，主要包括数据集的合规性、真实性。合规性衡量数据集不包含个人隐私信息、涉政、涉恐、数据投毒、数据后门等违规样本的程度。真实性衡量数据反映信息不含错误事实陈述的程度。

从内容维度来看，主要包括内容的全面性、多样性、独特性、时效性。全面性衡量数据在类别、时间跨度满足真实情况或业务需求全面程度的比例。多样性衡量数据集在目标领域或场景对涉及类别丰富的程度。独特性衡量数据集内相同或近似样本规模的比例。时效性衡量信息或数据是否最新，并确保其及时更新，衡量数据反映的信息是否处于前沿，以及更新频率是否符合具体场景需求。

从应用维度来看，主要侧重数据集的均衡性和有效性。均衡性衡量各类别样本数量分布与实际使用场景类别分布的适配度。有效性衡量数据中有效、高价值信息的比例，衡量在应用中满足应用需求的程度。

2.2 人工智能数据集质量评估

依据基本的质量维度，企业可以开展人工智能数据集的质量评估工作，具体的质量评估流程如下。

• 建立评估团队：由企业组织待评估数据集相关的开发、维护和管理人员，形

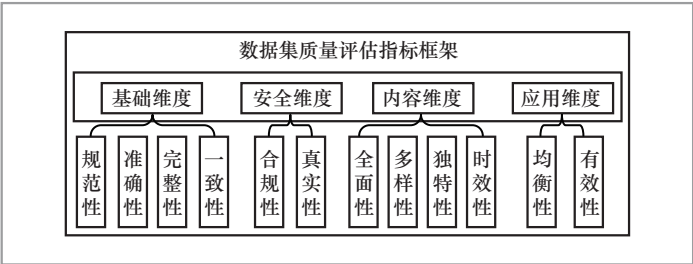


图1 数据集质量评价维度

成评估团队或引入外部第三方。

• 选取评估指标：对数据集进行分析，根据分析结果选择相应的一级指标（如完整性、准确性、及时性、一致性等）和对应的二级指标（如必填字段缺失率、业务规则校验通过率、合规审计通过率等），设置各指标权重和计算方式^[13]。

• 执行评估：选取需要评估的数据集及其涉及的关键表、核心字段或业务实体，将采集到的样本数据放入评估工具或计算环境中进行预处理，如处理异常值、格式转换等，确保指标计算时具备可操作性。

• 统计执行结果：指标项评分是对数据集作进一步评估等级划分的依据，数据集应在通过符合性审查的基础上，使用已定义好的方法，对每一条指标进行统计和计算，根据各指标项打分，利用加权方法计算总体得分。

当前数据集质量的评估是很多企业关注的热点议题，评估工作的工具化、智能化是业界努力的方向。同时，要意识到质量评估只完成了质量问题的发现，后续数据集质量的提升以及构建数据集全生命周期的管理能力才是关键。

3 企业人工智能数据集能力体系建设

企业构建人工智能数据集是一项体系

性的工程，涉及组织、制度、技术和人员能力的建设，面临不少挑战。一是跨团队，这项工作涉及智能、数据和业务部门等团队，需要构建专业、高效的协作机制来共同推进；二是缺方法，企业端数据集建设的实践刚刚开始，可以参考的经验不多，尚未建立较好的方法论；三是缺工具，技术工具是这项工作的具体承载物，然而当前业界缺少成熟的管理工具，阻碍了企业快速开展这项工作。

3.1 人工智能数据工程参考框架

训练数据集的处理通常有5个环节，包括数据收集、数据过滤、数据去重、数据标准化和数据审查^[3]。结合通用方法，笔者团队联合40多家机构编写技术标准《人工智能数据工程能力要求》，旨在建立人工智能数据集建设的目标、内容和规范化流程体系，指导企业开展能力建设。《人工智能数据工程能力要求》详细定义了数据集建设的五大能力，即研发管理、交付管理、运维管理、运营管理和技术工具，共18个能力子域（如图2所示）。

展开来看，研发管理是对数据集的生

成流程进行管控，其中需求管理的目的是明确人工智能团队和业务部门对数据集有哪些需求，规范化需求的描述，确认其优先级和合理性；设计管理的目的是构建数据集标准、质量、安全、合规的规范体系，先立规矩再开展具体的开发行为；数据加工管理的目的是梳理数据集加工的整体流程，明确对应的技术能力和管理要求。交付管理面向数据集的交付过程开展管控活动，其中测试管理的主要目的是对数据集上线前的质量进行把关，开展质量和合规性验证；数据集发布管理的目的是建立发布流程体系，对发布后的数据集版本进行生成、对比、回滚等管理活动。运维管理关注数据集的日常监控和维护，对过程中涉及的数据、计算、存储资源进行管理。运营管理关注数据集在用户端的使用情况，衡量投入产出的收益，关注数据集长期的质量提升。

人工智能数据工程顺利推行，离不开相应的技术工具的支持，框架中列举了核心的7项技术能力，涉及数据采集、数据处理、数据管理、数据标注、数据合成、数据质检工具、数据服务运营，这些技术能力可以集成在一个或多个技术工具中。

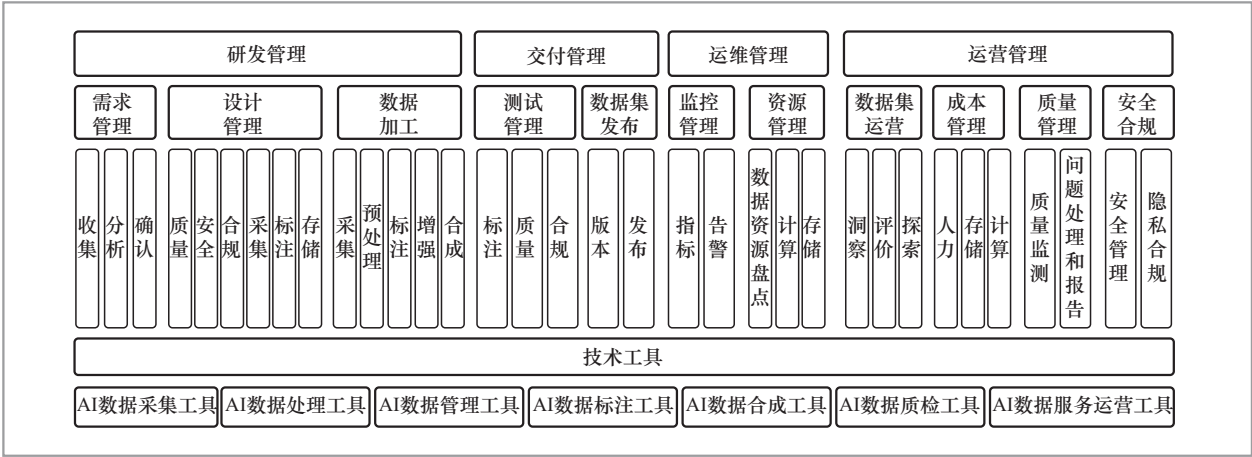


图2 人工智能数据工程能力要求

3.2 企业人工智能数据集建设实践

企业级的人工智能数据集需要考虑4个要素：一是围绕企业大模型训练和企业人工智能应用场景，全面收集数据；二是构建企业内统一的处理规范和流程，保证数据集生产流水线的规范性；三是需要统筹部署数据集处理和管理工具，实现工具的集约建设；四是建立数据集持续运营体系，开展目录整理、检索、质量评价、版本更新维护等工作，优化数据集的生产环节，提供更好的服务能力。

从建设路径来看，企业可以依据“盘、建、研、管、用”五步走的方法开展高质量数据集体系建设。“盘”指的是理清内外部数据资源的来源、规模、质量、类别情况，明确应用与采集的格式、质量、内容要求，主要输出数据资源地图和需求列表，建立采集的模板。“建”指的是提前制定数据集的质量、安全、运管要求，构建标准化工作流程，并开展工具平台的引入和定制化开发。“研”指的是依托前期建立的流程和规范，开展数据集的研发工作，生成数据集。“管”指的是对数据集质量、安全建立常态化管理体系，包括质量的评价、安全和隐私的检测。“用”指的依托前序工作成果，跟踪、维护、优化数据集的评价、使用、成本和价值。每个步骤都涉及相应的输入物、工作活动以及输出物，同时要注意实施过程的进度管理。

企业实践方面，某通信运营商大力投入高质量数据集建设工作，围绕工具平台、数据标注、质量评测、数据生产、数据集管理运营构建了全面的能力体系：在盘点环节，聚焦集团6类通用大模型的训练需求、30类场景需求，形成统一全面的数据集建设需求列表；在建设环节，构建了集团、省公司两级部署的高质量数据供给和管理平台，支持数据集的存储、汇聚、处

理、标注、合成、评测、管理，联合高校构建数据标注团队，定义标注规范；在研发环节，构建了基础数据集、合成数据、应用数据三大生产链路，以及一个数据治理流程，目前已经形成覆盖32个行业领域，涉及文学、教育151个方向，26TB的行业高质量数据集；在管理环节，形成了针对文本数据、音频数据、图像数据、视频数据的200多项质量评测指标，每类数据均形成了基础评测、内容质量评测、标注质量评测、数据可用性评测等系列工具。

某出版社着力构建教育大模型，同步形成系统的高质量数据集：在盘点环节，面向高等教育和职业教育两大方向，分别形成训练数据集、微调数据集、测试数据集需求清单；在建设环节，构建了统一的数据资源中心，覆盖图像、视频、音频、文本等类型，部署了数据存储、预处理、清洗、标注全链路工具；在研发环节，形成了数据集生产链路体系，目前形成了13个学科门类、19个专业大类的数据集体系，打造了高质量课程、高质量期刊、高质量题库等应用方向；在管理环节，具备数据分类分级、数据集检索、预览、版本控制等能力，并对数据质量进行定期评价和提升。

4 结束语

从机器学习到深度学习，再到现在的大模型时代，人工智能的基本范式始终是让机器从数据中学习。近年来，不少人工智能领域的学者提倡“以数据为中心的人工智能”理念，强调提升数据的规模、质量和丰富度可以显著提升模型的效果。人工智能数据集由文本、图像、音视频等数

据类型构成，主要用于模型训练、微调和评测等环节，覆盖互联网、图书、学术材料、代码库、企业知识库等多个来源。在训练数据集处理中通常采用基于规则的过滤技术来提升数据质量，例如去除终端标记、检测句子重复以及强制设定长度限制，以排除低质量数据^[14-15]。在指令微调阶段，可以通过基于指标体系、基于可训练的大模型、基于大模型、小模型等方法来判断数据集的效果，从而进行数据集的筛选^[16-18]。相比结构化数据，人工智能数据集的质量维度和质量评价更复杂，除了传统的七性维度，还要关注内容、应用等维度，本文提出了一种人工智能数据集质量评价的维度和评价方法，可以作为评价高质量数据集的参考框架。同时，在前期标准化的基础上，围绕数据集生命周期，本文提出了一套企业人工智能数据工程的体系，覆盖了研发、交付、运维、运营4项管理活动和7项技术能力，参考数据工程体系，企业可以通过“盘、建、研、管、用”5个步骤来开展能力体系建设。

未来，国家层面推动高质量数据集建设可以考虑4个方面的内容。一是做好顶层规划和统筹，整体推进各行业数据集能力建设，包括开展行业数据集的摸底工作，建立行业数据集的地图，征集优秀建设案例，鼓励经验共享。二是建立标准和规范体系，围绕质量评价、建设运营能力、数据处理流程、工具平台等方面开展标准研制工作，体系性构建方法论，选取头部机构，推动标准的验证落地。三是推进生态体系建设，汇聚数据源企业、数据标注企业、版权机构和模型训练企业，梳理数据集需求清单和数据集地图，精准匹配模型训练企业和数据集供应企业，鼓励更多主体参与到数据集开放、开源和流通体系建设中^[19]。四是研究版权和安全等问题，出版行业可以

制定AI版权管理办法，确定出版社、创作者作品被用于AI训练的知情和收益权，并构建数据调用的收费模式和收费标准，可以尝试按年整体付费、按接口调用条数、按作者贡献度等多种收费方式，构建版权协作的机制和安全保护的基线^[20]。

参考文献：

[1] ZHA D C, BHAT Z P, LAI K H, et al. Data-centric artificial intelligence: a survey[J]. ACM Computing Surveys, 2025, 57(5): 1-42.

[2] 程乐. 我国高质量场景数据集的供给现状与发展策略[J]. 人民论坛, 2025(5): 68-72.
CHENG L. Supply status and development strategies of high-quality scene datasets in China[J]. People's Tribune, 2025(5): 68-72.

[3] LIU Y, CAO J H, LIU C Y, et al. Datasets for large language models: a comprehensive survey[EB]. arXiv preprint, 2024, arXiv:2402.18041.

[4] 杜晓明, 赵晔蕾, 张媛玥, 等. 大模型数据集现状与对电信运营商的启示[J]. 互联网天地, 2023(9): 53-56.
DU X M, ZHAO Y L, ZHANG Y Y, et al. Present situation of large model data set and its enlightenment to telecom operators[J]. China Internet, 2023(9): 53-56.

[5] 穆晓君, 吴桐, 加雄伟, 等. 大模型时代: 高质量数据的构建、治理与未来展望[J]. 信息通信技术, 2024, 18(3): 63-68.
MU X J, WU T, JIA X W, et al. The age of big models: building, governing and exploring of high-quality data[J]. Information and Communications Technologies, 2024, 18(3): 63-68.

[6] 国家市场监督管理总局 国家标准化管理委员会. 信息技术 数据质量评价指标: GB/T 36344-2018[S]. 北京: 中国标准出版社,

- 2018.
- Standardization Administration of the People's Republic of China. Information technology—Evaluation indicators for data quality: GB/T 36344—2018[S]. Beijing: Standards Press of China, 2018.
- [7] WHANG S E, ROH Y, SONG H, et al. Data collection and quality challenges in deep learning: a data-centric AI perspective[J]. The VLDB Journal, 2023, 32(4): 791–813.
- [8] GUPTA N, MUJUMDAR S, PATEL H, et al. Data quality for machine learning tasks[C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. New York: ACM, 2021: 4040–4041.
- [9] CHEN H H, CHEN J P, DING J H. Data evaluation and enhancement for quality improvement of machine learning[C]//Proceedings of the 2020 IEEE 20th International Conference on Software Quality, Reliability and Security. Piscataway: IEEE Press, 2020: 13–13.
- [10] ZHOU Y H, TU F J, SHA K W, et al. A survey on data quality dimensions and tools for machine learning invited paper [C]//Proceedings of the 2024 IEEE International Conference on Artificial Intelligence Testing. Piscataway: IEEE Press, 2024: 120–131.
- [11] LIU Y, YANG Z W, ZOU X X, et al. A general framework to govern machine learning oriented materials data quality [J]. Materials Science and Engineering: R: Reports, 2025, 166: 101050.
- [12] WETTIG A, GUPTA A, MALIK S, et al. QuRating: selecting high-quality data for training language models[EB]. arXiv preprint, 2024, arXiv:2402.09739.
- [13] TAE K H, ROH Y, OH Y H, et al. Data cleaning for accurate, fair, and robust models: a big data – AI integration approach[C]//Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning. New York: ACM, 2019: 1–4.
- [14] LIU F Z, ZHOU W D, LI B B, et al. QuaDMix: quality-diversity balanced data selection for efficient LLM pre-training[EB]. arXiv preprint, 2025, arXiv: 2504.16511.
- [15] LONGPRE S, YAUNEY G, REIF E, et al. A pretrainer's guide to training data: measuring the effects of data age, domain coverage, quality, & toxicity[EB]. arXiv preprint, 2023, arXiv: 2305.13169.
- [16] QIN Y L, YANG Y C, GUO P C, et al. Unleashing the power of data Tsunami: a comprehensive survey on data assessment and selection for instruction tuning of language models[EB]. arXiv preprint, 2024, arXiv: 2408.02085.
- [17] ZHANG B L, WANG J H, DU Q L, et al. A survey on data selection for LLM instruction tuning[EB]. arXiv preprint, 2024, arXiv: 2402.05123.
- [18] LI M, ZHANG Y, LI Z, et al. From quantity to quality: boosting LLM performance with self-guided data selection for instruction tuning[EB]. arXiv preprint, 2023, arXiv: 2308.12032.
- [19] 张文娟, 邓辉, 艾政阳, 等. 我国AI大模型数据集建设发展刍议[J]. 人工智能, 2024, 11(3): 85–95.
- ZHANG W J, DENG H, AI Z Y, et al. On the construction and development of AI large model dataset in China[J]. AI-View, 2024, 11(3): 85–95.
- [20] 种晓明, 孔正毅. AIGC数据供给侧视角下我国出版业版权资产的开发[J]. 出版发行研究, 2023(8): 29–36.
- CHONG X M, KONG Z Y. On development of copyright assets in China's publishing Industry from the perspective of AIGC data supply side[J]. Publishing Research, 2023(8): 29–36.

作者简介



姜春宇（1987–），男，中国信息通信研究院高级工程师，大数据与智能化部主任，主要研究方向为大数据与人工智能技术、大模型、数据要素流通、数据资产管理、数据安全。参与起草多个国家文件，包括《促进国家大数据发展行动纲要》《大数据产业发展规划（2016—2020年）》。建立了国内首个大数据人工智能产品能力评测体系，制定了50多项数据标准规范，牵头编写《数据资产管理实践白皮书》《数据安全治理实践指南》《数据要素白皮书》《智能中台实践指南》等多本研究报告。



白玉真（1990–），女，博士，中国信息通信研究院工程师，主要研究方向为数据要素流通、高质量数据集、隐私计算等。参与构建可信数据流通技术标准与评测体系，制定多方安全计算、联邦学习、隐私计算互联互通、公共数据等20多项数据标准规范，牵头和参与编写《可信隐私计算研究报告》《可信数据流通网络（TDN）白皮书》《隐私计算白皮书》等多本研究报告。



刘渊（1994–），女，中国信息通信研究院中级工程师，主要研究方向为数据基础设施、人工智能技术和应用、数据智能技术和应用、数据管理和治理等。牵头《集成化大数据平台能力分级要求》《流批一体大数据平台技术要求》《数据智能平台技术要求》等团体标准的制定工作；主持执行国内大数据产品测试70余项；主持《大数据技术发展成熟度研究》等课题的研究工作和研究报告编写。



王超伦（1990–），男，中国信息通信研究院中级工程师，主要研究方向为大数据与人工智能技术及行业应用。参与起草多个国家文件，包括《促进国家大数据发展行动纲要》《大数据产业发展规划（2016—2020年）》。参与编写了《检索增强生成技术要求》《智能体平台技术要求》《大模型应用开发平台技术要求》《大模型驱动的智能数据分析工具技术要求》《大模型一体机能力要求》《知识中台能力成熟度模型》《人工智能数据工程能力要求》《数据智能白皮书》等多项标准及研究报告，并开展相关评测工作。

收稿日期: 2025-08-25
通信作者: 姜春宇,jiangchunyul@caict.ac.cn