TC609

全国数据标准化技术委员会技术文件

TC609-5-2025-01

高质量数据集 建设指南

High-quality dataset—Construction guidelines

2025-08-29 发布　　2025-08-29 实施

全国数据标准化技术委员会　发布

目　　次

前　言

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由全国数据标准化技术委员会（SAC/TC609）提出并归口。

本文件起草单位：中国电子技术标准化研究院、中国电子信息产业发展研究院、国家数据发展研究院、工业和信息化部电子第五研究所、中国信息通信研究院、国务院国有资产监督管理委员会研究中心、商业信用中心、中国科学院计算技术研究所、交通运输部公路科学研究所、北京大学、中国石油天然气集团有限公司、中国石油化工集团有限公司、石化盈科信息技术有限责任公司、中国交通建设集团有限公司、中国交通信息科技集团有限公司、国家能源集团信息技术公司、中国南方电网有限责任公司、中国电信集团有限公司、中国移动通信集团有限公司、中移动信息技术有限公司、中国联合网络通信集团有限公司、联通数据智能有限公司、中电数据产业集团有限公司、中国质量认证中心有限公司、国家石油天然气管网集团有限公司、国网山东省电力公司、国网江苏省电力有限公司、华为技术有限公司、科大讯飞股份有限公司、阿里巴巴（中国）有限公司、深圳市腾讯计算机系统有限公司、中兴通讯股份有限公司、浪潮电子信息产业股份有限公司、北京智源人工智能研究院、上海人工智能创新中心、北京百度网讯科技有限公司、上海库帕思科技有限公司、上海信投智能科技股份有限公司、南京南瑞继保工程技术有限公司、南京南瑞瑞中数据股份有限公司、杭州数梦工场科技有限公司、杭州市临安区大数据管理服务中心、安徽飞数信息科技有限公司、中通服网盈科技有限公司、北京海天瑞声科技股份有限公司、航天科工网络信息发展有限公司、中国邮政储蓄银行股份有限公司、江苏省大数据管理中心、内蒙古自治区大数据中心、江西省大数据中心、中国电子工程设计院股份有限公司、中电金信软件有限公司、软通智慧科技有限公司、厦门赛西科技发展有限责任公司、广东省医学科学院、烽火通信科技股份有限公司、同方知网数字科技有限公司、蔚来汽车科技（安徽）有限公司、江苏省数据集团有限公司、四川数据集团有限公司、贵州大数据产业集团有限公司、杭州市数据集团有限公司、睿尔曼智能科技（北京）有限公司、北京银河通用机器人有限公司、云基华海信息技术股份有限公司、辽宁省电子信息产品监督检验院、数字宁波科技有限公司、北京中数睿智科技有限公司、杭州景联文科技有限公司、北京星河智源科技有限公司、山西集智数据服务有限公司、山东未来集团有限公司、广州维视达数字科技有限公司、广州金域医学检验集团股份有限公司、厦门身份宝网络科技有限公司、上海森栩医学科技有限公司、苏州智行众维智能科技有限公司、北京八月瓜科技有限公司。

引　言

当前，随着新一代信息技术持续快速发展，人工智能正加速融入各行业领域，赋能实体经济高质量发展。高质量数据集是开发和训练人工智能模型的重要支撑，能够提高模型精度与可解释性、减少训练时长，已经成为人工智能发展的核心要素。目前，在我国高质量数据集建设推进过程中，仍存在数据集“如何建设”不明确、“建设方法论”缺失的问题。制定高质量数据集建设指南，为数据集建设全生命周期，包括数据需求、数据规划、数据采集、数据预处理、数据标注、模型验证等阶段，提供指导和建议，对于提升数据集优质供给，有力支持人工智能模型开发和训练，更好赋能经济社会发展至关重要。

高质量数据集 建设指南

1 范围

本文件提供了高质量数据集建设全生命周期的指导和建议，包括数据需求、数据规划、数据采集、数据预处理、数据标注、模型验证等阶段。

本文件适用于高质量数据集的规划、建设和维护工作。

2 规范性引用文件

本文件没有规范性引用文件。

3 术语和定义

下列术语和定义适用于本文件。

3.1

高质量数据集 high-quality dataset

经过采集、加工等数据处理，可直接用于开发和训练人工智能模型，能有效提升模型性能的数据的集合。

3.2

数据质量 data quality

在指定条件下使用时，数据的特性满足明确的和隐含的要求的程度。

[来源：GB/T 25000.24-2017，4.11]

3.3

数据质量模型 data quality model

已定义的特性集合，提供一个框架用于说明数据质量需求和评价数据质量。

[来源：GB/T 25000.24-2017，4.14]

3.4

数据标注 data labeling

给数据样本指定目标变量和赋值的过程。

[来源：GB/T 42755-2023，3.1]

3.5

数据架构 data architecture

对企业业务主要数据类型和来源、逻辑数据资产、物理数据资产和数据管理资源结构和交互的描述。

[来源：ISO TR 21965:2019,3.2.6]

3.6

有监督机器学习 supervised machine learning

仅用标注数据进行训练的机器学习。

[来源：GB/T 41867-2022，3.2.37]

4 建设方法

高质量数据集建设应按照生命周期有序展开，包括数据需求、数据规划、数据采集、数据预处理、数据标注、模型验证等阶段。各阶段主要按以上顺序逐步开展，同时会对其他阶段进行反馈，或者会在其他阶段反馈下进行迭代优化。

高质量数据集建设方法如图1所示。

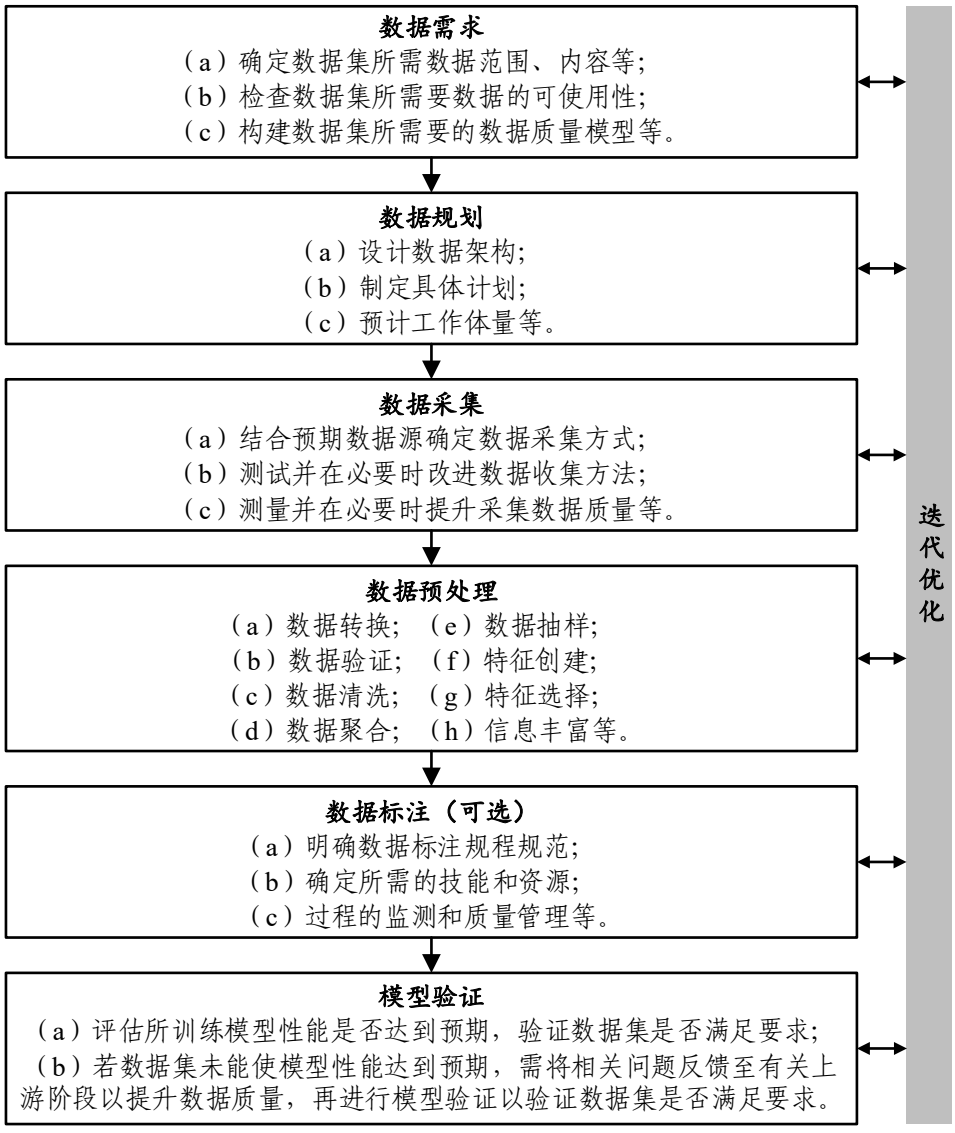

图1 高质量数据集建设方法

5 数据需求

数据需求阶段主要是确定人工智能应用对数据的需求，即根据预期人工智能用途明确数据集在数据范围、内容、可用、质量等方面的需求。该阶段包括但不限于：

a) 确定数据集所需数据范围、内容等，即根据预期人工智能应用，明确具体需要哪些数据，包括数据格式、统计特性和可分性（即训练、验证、测试数据等子集的有效划分说明，包括每个子

集的相关统计特性、代表性、合适规模大小等）等；

b） 检查数据集所需要数据的可使用性，即确认用于特定人工智能应用的数据是否可获取并使用；

c） 构建数据集所需要的数据质量模型，即实例化一个具有相关数据质量特性（例如，完整性、准确性、一致性等）的数据质量模型。

6 数据规划

数据规划阶段主要是确保所用数据满足数据需求阶段的要求，同时为使用这些数据完成人工智能应用的目标提供支持。该阶段包括但不限于：

a） 设计数据架构，即界定所需数据的全部属性、来源、范围等，以及如何使用这些数据；

b） 制定具体计划，即制定涵盖数据采集、数据预处理、数据标注、模型验证等阶段的具体计划，包括各阶段实施计划、数据质量计划等，以满足数据规范等方面要求；

c） 预计工作体量，即预估获得和准备数据以支持特定人工智能应用所需的工作量，可能包括任何必要的数据重组、传输或收集的时间，以及为特定人工智能应用构建数据质量模型的时间。

7 数据采集

数据采集阶段主要是收集用于特定人工智能应用的数据，即从数据规划阶段所确定的数据源收集的实时和历史数据。该阶段包括但不限于：

a） 结合预期数据源确定数据采集方式，即根据所需数据是否已存在并可直接再利用、是否可转化现有数据来满足要求、是否可通过购买或许可获得数据、是否可以生成数据、是否需要采集新数据等情况，确定是以获取和组合现有数据集、生成数据（如模拟数据、合成数据等）、收集数据（如传感器采集、手动输入等）等之中何种方式采集数据；

b） 测试并在必要时改进数据收集方法，即如需收集新数据，则要测试数据收集方法，在必要时调整相关配置和参数设置、操作条件、传感器规格和安装位置等，以满足相关数据收集规范要求；

c） 测量并在必要时提升采集数据质量，以降低采集阶段数据质量问题引入下游阶段的风险，避免为下游阶段增加不必要的工作量。

8 数据预处理

数据预处理阶段主要是将所收集到的数据处理成可供数据标注等后续阶段使用的形式。该阶段涉及以下可选过程：

a） 数据转换，以最小的内容损失，将数据从一种表示或空间转换为另一种表示或空间；

b） 数据验证，根据验证正确性、有意义、安全性、隐私性等数据质量特性，确保数据是正确的；

c） 数据清洗，检测错误、重复或缺失数据，并通过替换、修改、输入或删除等方式修正数据；

d） 数据聚合，将两个或多个数据集以汇总的形式合并为一个数据集；

e） 数据抽样，从数据集中选择数据，抽样可以替换或非替换方式进行；

f） 特征创建，创建比原始特征更能有效捕捉数据中主要信息的新特征；

g） 特征选择，使用可用特征的子集来降低数据的维数；

h） 信息丰富，链接各类数据源，并为数据增加额外的上下文语境。

9 数据标注

数据标注阶段为可选阶段，主要是针对有监督机器学习，其训练、验证和测试数据需要对单个或多个目标变量赋值。该阶段包括但不限于：

a） 明确数据标注规程规范；

b） 确定所需的技能和资源，如标注技能、工具、平台等；

c） 对数据标注过程进行监测和质量管理。

10 模型验证

模型验证阶段主要是将所准备的数据用于人工智能模型开发和训练，对模型性能是否达到预期进行评估，以验证数据集是否满足要求。若模型性能达到预期，则表明数据集已满足要求。若模型性能未达到预期，则可采取以下步骤：

a） 对于人工智能模型，确定数据集相比于算法，是否为致使模型性能未达到预期的根本原因；

b） 对模型验证阶段所发现的数据质量问题进行分析，将对模型性能产生不利影响的数据质量问题反馈给上游阶段，以改进相关阶段的数据质量；

c） 重复数据需求、数据规划、数据采集、数据预处理、数据标注等阶段以提升数据质量；

d） 重建人工智能模型，对模型性能进行评估。

参 考 文 献

[1] GB/T 25000.24-2017 系统与软件工程 系统与软件质量要求和评价（SQuaRE） 第24部分：数据质量测量（ISO/IEC 25024:2015，MOD）

[2] GB/T 41867-2022 信息技术 人工智能 术语

[3] GB/T 42755-2023 人工智能 面向机器学习的数据标注规程

[4] ISO/IEC 5259-1:2024 Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 1: Overview, terminology, and examples

[5] ISO/IEC 5259-3:2024 Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 3: Data quality management requirements and guidelines

[6] ISO/TR 21965:2019, Information and documentation - Records management in enterprise

[7] ISO/IEC 23053:2022 Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)