ICS 35.240
CCS L 70

中华人民共和国国家标准

GB/T XXXXX—XXXX

高质量数据集 质量评测规范

High-quality dataset—Specifications for quality evaluation and test

（征求意见稿）

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

国家市场监督管理总局
国家标准化管理委员会　发布

目　　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

本文件由全国数据标准化技术委员会（SAC/TC 609）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、中国科学院计算技术研究所、国家数据发展研究院、国务院国有资产监督管理委员会研究中心、中电数据产业集团有限公司、交通运输部公路科学研究所、北京大学、公安部第三研究所、中国移动通信集团有限公司、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、中国交通建设集团有限公司、国家能源投资集团有限责任公司信息技术分公司、国家电网有限公司大数据中心、中国南方电网有限责任公司、国家石油天然气管网集团有限公司、中国稀土集团有限公司、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、北京智源人工智能研究院、北京百度网讯科技有限公司、深圳市腾讯计算机系统有限公司、中国联合网络通信集团有限公司、中国电信集团有限公司、工业和信息化部电子第五研究所、中国信息通信研究院、商业信用中心、石化盈科信息技术有限责任公司、北京智网数科技术有限公司、浦江国家实验室、中国质量认证中心有限公司、煤炭科学研究总院有限公司、中国交通信息科技集团有限公司、中移动信息技术有限公司、国家电投集团数字科技有限公司、中电云计算技术有限公司、中石油（北京）数智研究院有限公司、联通数据智能有限公司、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、北京清竞数智科技有限公司、青岛国创智能家电研究院有限公司、成都高新愿景数字科技有限公司、每日互动股份有限公司、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、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、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、四川省卫生健康信息中心（四川省健康医疗大数据中心）、北京大学（天津滨海）新一代信息技术研究院、国家开放大学、杭州数美科技有限公司、中天钢铁集团（南通）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、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、广州数字健康科技有限公司、安徽飞数信息科技有限公司、湖北大数据集团数据开发有限公司、杭州市临安区大数据管理服务中心、软通智慧科技有限公司、同方知网数字科技有限公司、数据堂（北京）科技股份有限公司、睿尔曼智能科技（北京）有限公司、北京银河通用机器人股份有限公司、烽火通信科技股份有限公司、中兴通讯股份有限公司、浪潮电子信息产业股份有限公司、国网山东省电力公司、蔚来汽车科技（安徽）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、江苏省数据集团有限公司、四川数据集团有限公司、厦门赛西科技发展有限责任公司、云基华海信息技术股份有限公司、国网江苏省电力有限公司、广东省人民医院、辽宁省电子信息产品监督检验院、数字宁波科技有限公司、杭州景联文科技有限公司、北京腾云天下科技有限公司、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、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、北京八月瓜科技有限公司、北京中数睿智科技有限公司、江苏中堃数据技术有限公司。

本文件主要起草人：范科峰、张群、韩冰、郭嘉丰、赵鹏飞、王为中、廖华明、张欢、李成博、李冰、苏越阳、李天舒、时晓光、黄吉海、丁浩、李嘉宁、王超、温晓君、潘宗俊、陈英昊、王亚沙、赵俊峰、邓成龙、汪睿棋、初旭、吴坤、陈文萱、申端明、蒋楠、王奥、杨丽、王宇静、陈振宇、徐欢、宋一纯、李步伟、于海松、储培、林志强、李华杰、赵丽丽、王鑫、方昕、吴峥、李世奇、刘颖、刘广、杨二龙、邱泳钦、刘煜宏、董梁、杨瑞、万芊芊、程广明、樊威、李苏、袁星煜、赵岩、赵艺璇、王锋、程健、骆意、张建中、武光城、金柳、莫洋、梁小涛、王俊杰、薛健、冷家冰、魏如蓝、王昊、谭晓坤、

胡力旗、王吴越、王兴旺、申中一、桂志辉、陈亮、方毅、邵元勋、孙永泉、孟萦、王春涛、杜啸争、吴善鹏、崔连伟、张静、沈明辉、邓韧、李方平、谢魁、毛子卿、蔡斯博、余其竞、念灿华、符立峰、解凯、张锦辉、李凡、蒲逸凡、黄宇恒、葛海龙、胡环环、段先明、郑辉、林镇阳、薛德军、齐红威、郑随兵、曾辉、陈刚、吴德亮、陈曦、张闻彬、商泽坤、赵泓青昀、李玮燕、陈康、陈曦、曾良滨、鲁胜强、何金陵、马洲俊、梁会营、王俊吉、毛欢欢、刘云涛、张亚东、严长春、庞俊奇、刘杰、彭荣、陈颖、温冬梅、李长青、韩涵、魏清。

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。数据集是开发和训练人工智能模型的基础，开发和训练高质量的模型也对数据集质量提出越来越高的要求。对数据集进行质量评测是评判其是否“高质量”的基本路径，也是“以评促建”保障高质量数据集建设的重要手段，然而，我国高质量数据集质量评测目前仍缺乏统一的标准规范。制定高质量数据集质量评测规范，明确质量评测的指标要求和评测方法，为开展高质量数据集质量评测活动提供指导，对于提升数据集优质供给，促进数据集流通使用，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 质量评测规范

1 范围

本文件规范了高质量数据集的质量评测，明确了指标要求、评测方法。

本文件可为开展高质量数据集质量评测活动提供指导。

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。

[来源：20255407-T-907，3.3.27]

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。主要包括数据的完整性、真实性、可靠性、及时性、一致性、可追溯性和包容性。

[来源：20255407-T-907，3.1.24]

3.3

通识数据集 general knowledge dataset

蕴含通用知识的数据的集合。

[来源：20256912-T-907，3.3]

3.4

行业通识数据集 industry general knowledge dataset

蕴含行业领域通用知识的数据的集合。

[来源：20256912-T-907，3.5]

3.5

行业专识数据集 industry professional knowledge dataset

蕴含行业领域专业知识的数据的集合。

[来源：20256912-T-907，3.7]

3.6

数据标注 data labeling; data annotation

给数据样本指定目标变量和赋值的过程。

[来源：20255407-T-907，3.4.13]

3.7

数据记录 data record

一个单元化的相关数据项的集合。

[来源：GB/T 25000.24-2017，4.15]

3.8

无监督机器学习 unsupervised machine learning

仅用无标注数据实施训练的机器学习。

[来源：GB/T 41867-2022，3.2.32]

4 指标要求

4.1 评测指标要求

高质量数据集的质量要求应覆盖说明文档、数据质量及模型应用三个维度的指标。

a) 说明文档：数据集的说明文档应完整，包含基本信息、内容特征、建设过程及应用说明等；
b) 数据质量：数据集中数据的质量应满足人工智能模型开发和训练的基本要求；
c) 模型应用：数据集应能有效支撑目标人工智能模型的开发和训练。

4.2 说明文档指标要求

说明文档指标要求包括以下子指标。

a) 基本信息完整性：数据集说明文档应包含数据集规模（如样本数量、存储体积等）、格式规范、文件结构、访问渠道、技术支持方式等基本信息；
b) 内容特征完整性：数据集说明文档应包含模态类型、数据分布情况、标签类别统计、样本示例、局限性说明等内容特征；
c) 建设过程完整性：数据集说明文档应包含数据来源、采集方法、加工处理流程、标注规范、版本控制等建设过程；
d) 应用说明完整性：数据集说明文档应包含使用许可、目标应用场景、评估方法、基准测试结果、典型应用案例等应用说明。

4.3 数据质量指标要求

数据质量指标要求包括以下子指标。

a) 格式规范性：数据集中数据的格式应符合预定标准，可直接用于人工智能模型开发和训练；
b) 安全规范性：数据集中数据应符合人工智能模型开发和训练的安全要求，不包含违反社会主义核心价值观的内容、歧视性内容、商业违法违规、侵犯他人合法权益等非法内容；
c) 标注规范性：数据集中数据的标注应符合预定的标注规范，遵循预先设定的规范化流程；
d) 结构完整性：数据集中数据应填充完整，不包含缺失值或缺失值应在合理范围内；
e) 内容真实性：数据集中数据真实可追溯。非生成数据应具有明确的数据来源，并能提供详细的加工处理流程或记录；生成数据能追溯到生成算法和过程，且能符合目标场景真实数据的分布规律；
f) 内容一致性：数据集中相关联的数据间内容一致，能在语义和表达上保持匹配，包括不同模态数据间的一致性（如文本和图像之间、音频和视频之间在语义上对齐）和同模态数据间（如关联文本之间在语义上一致）的一致性；

g) 类型一致性：数据集中数据符合数据集类型要求，通识数据集中数据应蕴含通用知识，行业通识数据集中数据应蕴含行业领域通用知识，行业专识数据集中数据应蕴含行业领域专业知识；
h) 内容干净性：数据集中数据经过严格清洗处理，不包含脏数据（不同模态数据的内容干净性指标详见附录 A）。

4.4 模型应用指标要求

模型应用指标要求包括以下子指标。

a) 内容多样性：数据集的数据分布全面程度应满足目标应用场景人工智能模型开发和训练的要求；
b) 规模完整性：数据集的规模满足目标应用场景人工智能模型开发和训练的要求；
c) 内容时效性：数据集中数据的采集时间和更新状态满足目标应用场景人工智能模型开发和训练的要求；
d) 标注准确性：数据集中数据的标注能精准标记出目标应用场景人工智能模型开发和训练所需的所有信息；
e) 模型适配性：数据集应能有效提升目标应用场景人工智能模型的性能。

5 评测方法

5.1 评测方式要求

数据集同时满足说明文档指标、数据质量指标和模型应用指标的要求，视为符合高质量数据集的质量要求。具体要求为：说明文档指标得分达到 90 分及以上，数据质量指标得分达到 90 分及以上，模型应用指标得分达到 90 分及以上。

5.2 说明文档指标评测

5.2.1 说明文档指标计算方法

说明文档指标的计算方法如表 1 所示。

表1 说明文档指标的计算方法

指标编号	子指标名称	目标数据形式	指标描述	计算方法
0101	基本信息完整性	数据集	数据集说明文档满足基本信息完整性要求的方面比例 **注：** 基本信息完整性所要求的方面包括数据集规模、格式规范、文件结构、访问渠道、技术支持方式等。	$X = A/B$ 式中： A——数据集说明文档满足基本信息完整性要求的方面数量； B——数据集说明文档需满足基本信息完整性要求的方面总数
0102	内容特征完整性	数据集	数据集说明文档满足内容特征完整性要求的方面比例 **注：** 内容特征完整性所要求的方面包括模态类型、数据分布情况、标签类别统计、样本示例、局限性说明等。	$X = A/B$ 式中： A——数据集说明文档满足内容特征完整性要求的方面数量； B——数据集说明文档需满足内容特征完整性要求的方面总数

表1　说明文档指标的计算方法（续）

指标编号	子指标名称	目标数据形式	指标描述	计算方法
0103	建设过程完整性	数据集	数据集说明文档满足建设过程完整性要求的方面比例 **注：** 建设过程完整性所要求的方面包括数据来源、采集方法、加工处理流程、标注规范、版本控制记录等。	$X = A/B$ 式中： A——数据集说明文档满足建设过程完整性要求的方面数量； B——数据集说明文档需满足建设过程完整性要求的方面总数
0104	应用说明完整性	数据集	数据集说明文档满足应用说明完整性要求的方面比例 **注：** 应用说明完整性所要求的方面包括使用许可、目标应用场景、评估方法、基准测试结果、典型应用案例等。	$X = A/B$ 式中： A——数据集说明文档满足场景说明完整性要求的方面数量； B——数据集说明文档需满足场景说明完整性要求的方面总数

注：数据形式是指相应评价指标基于哪一种数据单位，主要包括数据记录、数据集等类型。

5.2.2　说明文档指标评分方法

说明文档指标的得分确定可基于以下公式。

$$P = \sum_{1}^{n} w_i \times s_i \times 100 \cdots\cdots (1)$$

其中，P为数据集的说明文档指标得分，i为说明文档指标中子指标的索引号，n为说明文档指标中子指标的数量，w_i为第i项子指标的计算权重，s_i为数据集在第i项子指标上的取值。

说明文档指标评分采用百分制，当得分达到 90 分及以上时，视为符合高质量数据集的说明文档指标要求。各子指标权重可根据实际应用场景和评测需求调整，建议各子指标权重相等时取值为$1/n$。

5.3　数据质量指标评测

5.3.1　数据质量指标计算方法

数据质量指标的计算方法如表 2 所示。

表2　数据质量指标的计算方法

指标编号	子指标名称	目标数据形式	指标描述	计算方法
0201	格式规范性	数据记录	数据集中格式符合预定标准的数据记录比例 **注：** 预定标准是指相关标准、惯例或其他用户自定义规则。	$X = A/B$ 式中： A——数据集中格式符合预定标准的数据记录数量； B——数据集中的数据记录总数
0202	安全规范性	数据记录	数据集中符合人工智能模型开发和训练安全要求的数据记录的比例	$X = A/B$ 式中： A——数据集中符合人工智能模型开发和训练安全要求的数据

表2 数据质量指标的计算方法（续）

指标编号	子指标名称	目标数据形式	指标描述	计算方法
				记录数量； B——数据集中数据记录总数
0203	标注规范性	数据记录	数据集中符合预定标注规范的数据记录比例 **注：** 预定标注规范指数据标注方面相关标准、惯例或其他用户自定义规则；无监督机器学习等不需标注的应用场景不适用。	$X = A/B$ 式中： A——数据集中符合预定标注规范的数据记录数量； B——数据集中的数据记录总数
0204	结构完整性	数据记录	数据集中缺失值在合理范围内的数据记录比例	$X = A/B$ 式中： A——数据集中缺失值在合理范围内的数据记录数量； B——数据集中的数据记录总数
0205	内容真实性	数据记录	数据集中真实可追溯的数据记录比例	$X = A/B$ 式中： A——数据集中真实可追溯的数据记录数量； B——数据集中的数据记录总数
0206	内容一致性	数据记录	数据集中内容一致的数据记录比例 **注：** 包括不同模态数据间内容一致和同模态数据间内容一致。	$X = A/B$ 式中： A——数据集中内容一致的数据记录数量； B——数据集中的数据记录总数
0207	类型一致性	数据记录	数据集中符合其所属数据集类型要求的数据记录比例	$X = A/B$ 式中： A——数据集中符合其所属数据集类型要求的数据记录数量； B——数据集中的数据记录总数
0208	内容干净性	数据记录	数据集中内容干净的数据记录比例	$X = A/B$ 式中： A——数据集中内容干净的数据记录数量； B——数据集中的数据记录总数

5.3.2 数据质量指标评分方法

数据质量指标的得分确定可基于以下公式。

$$P = \sum_{1}^{n} w_i \times s_i \times 100 \quad (2)$$

其中，P为数据集的数据质量指标得分，i为数据质量指标中子指标的索引号，n为数据质量指标中子指标的数量，w_i为第i项子指标的计算权重，s_i为数据集在第i项子指标上的取值。

数据质量指标评分采用百分制，当得分达到 90 分及以上时，视为符合高质量数据集的数据质量指标要求。各子指标权重可根据实际应用场景和评测需求调整，建议各子指标权重相等时取值为$1/n$。

5.4 模型应用指标评测

5.4.1 模型应用指标计算方法

模型应用指标的计算方法如表 3 所示。

表3 模型应用指标的计算方法

指标编号	子指标名称	目标数据形式	指标描述	计算方法
0301	内容多样性	数据集	数据集的数据分布全面程度满足目标应用场景人工智能模型开发和训练需求的比例	$X = A/B$ 式中： A——数据集分布覆盖范围； B——目标应用场景人工智能模型开发和训练所需的数据集分布覆盖范围
0302	规模完整性	数据集	数据集规模满足目标应用场景人工智能模型开发和训练所需规模的比例	$X = A/B$ 式中： A——数据集规模； B——目标应用场景人工智能模型开发和训练所需数据集规模
0303	内容时效性	数据记录	数据集中满足目标应用场景人工智能模型开发和训练对采集时间和更新状态要求的数据记录比例	$X = A/B$ 式中： A——数据集中满足目标应用场景对采集时间和更新状态要求的数据记录数量； B——数据集中的数据记录总数
0304	标注准确性	数据记录	数据集中能精准标记出目标应用场景人工智能模型开发和训练所需所有信息的数据记录比例 **注：** 无监督机器学习等不需标注的应用场景不适用。	$X = A/B$ 式中： A——数据集能精准标记出目标应用场景人工智能模型开发和训练所需所有信息的数据记录数量； B——数据集中的数据记录总数
0305	模型适配性	数据集	利用数据集训练的人工智能模型在目标应用场景中达到预期性能水平的程度 **注：** 使用数据集开发和训练模型，通过对比模型实际性能与预期性能的差异，衡量数据集的模型适配性。	$X = A/B$或$X = B/A$ 式中： A——使用数据集开发和训练的人工智能模型在目标应用场景的性能； B——目标应用场景人工智能模型的预期性能

表3 模型应用指标的计算方法（续）

指标编号	子指标名称	目标数据形式	指标描述	计算方法
				注：X的计算方式取决于指标的选择，对于期望值越高越好的指标用A/B计算；对于期望值越低越好的指标用B/A计算。

5.4.2 模型应用指标评分方法

模型应用指标的得分确定可基于以下公式。

$$F_m = \begin{cases} 1 & s_m \geq 1; \\ e^{s_m - 1} & s_m < 1 \end{cases} \qquad (3)$$

$$P = F_m \sum_{1}^{n-1} w_i \times s_i \times 100 \qquad (4)$$

其中，P为数据集的模型应用指标得分，i为模型应用指标中子指标的索引号（不包括模型适配性），n为模型应用指标中子指标的数量，w_i为第i项子指标的计算权重，s_i为数据集在第i项子指标上的取值；F_m为模型适配性指示函数，s_m为模型适配性子指标的取值。

模型应用指标评分采用百分制，当得分达到 90 分及以上时，视为符合高质量数据集的模型应用指标要求。各子指标权重可根据实际应用场景和评测需求进行调整，建议各子指标权重相等时取值为$1/(n-1)$。对于无监督机器学习，“标注准确性”子指标不适用，权重取值为 0，其余子指标权重需进行适当调整。

附 录 A
（资料性）
不同模态数据的内容干净性指标

表A.1 文本数据的内容干净性指标

指标编号	指标名称	指标描述
02080101	文本困惑程度（PPL）	文本语言流畅度水平。困惑度数值越小表示文本内容越通顺自然，越符合语言表达规律。例如，可通过预训练语言模型（如BERT系列模型等）计算得出
02080102	知识信息密度	文本中包含有效知识信息的丰富程度
02080103	重复内容程度	文档、段落、句子之间存在重复内容的程度
02080104	文本完整程度	文本内容过短或不完整的程度
02080105	信息缺失程度	标题、句子、段落等关键结构要素的缺失程度
02080106	文本纯净程度	广告、页码、特殊符号、乱码等干扰内容的存在程度
02080107	文本连贯程度	主题与内容的一致性以及是否存在异常截断、错别字等问题

表A.2 图像数据的内容干净性指标

指标编号	指标名称	指标描述
02080201	图像分辨率	图像的像素尺寸大小，通常以宽×高像素表示。分辨率越高，视频的细节和清晰度越好
02080202	图像重复度	数据集中相同或高度相似图像的重复程度
02080203	图像信噪比	图像中有效信号与噪声的比值，数值越高表示图像质量越好，噪声干扰越少
02080204	图像清晰度	图像的视觉清晰程度，不应存在水印、抖动、字幕干扰、条纹等问题

表A.3 视频数据的内容干净性指标

指标编号	指标名称	指标描述
02080301	视频分辨率	视频的像素尺寸大小，通常以宽×高像素表示。分辨率越高，视频的细节和清晰度越好
02080302	视频重复度	数据集中相同或高度相似视频的重复程度
02080303	视频帧率	视频每秒播放的帧数，帧率越高，视频播放越流畅
02080304	视频时长	视频文件的播放时间长度，时长过短可能导致内容不完整或信息量不足
02080305	视频清晰度	视频的视觉清晰程度，不应存在水印、抖动、字幕遮挡、条纹等问题
02080306	视频动态范围	视频可以表现的亮度和暗度的范围。HDR（高动态范围）视频的动态范围比SDR（标准动态范围）视频更大

表A.4 音频数据内容干净性指标

指标编号	指标名称	指标描述
02080401	音频信噪比	音频信号与背景噪声的比值，信噪比越高表示音频质量越好，噪声干扰越少
02080402	信号失真比	音频中目标信号与噪声、干扰等信号的强度比值，数值越高表示音频质量越好
02080403	音频采样率	音频数据的采样频率，采样率越高，音频还原度越好
02080404	音频位深度	音频数据的量化精度，位深度越高，音频的动态范围和分辨率越好
02080405	音频比特率	音频数据的传输速率，比特率越高，音频质量通常越好
02080406	音频时长	音频文件的播放时间长度，时长过短可能导致内容不完整或信息量不足

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 36344-2018 信息技术 数据质量评价指标

[3] GB/T 41867-2022 信息技术 人工智能 术语

[4] 20255407-T-907 数据 基础术语

[5] 20256912-T-907 高质量数据集 分类指南

[6] YD/T 4522-2023 面向机器学习的电信数据规范 数据质量

[7] TR-REC-064 数据质量评测方法与指标体系