国家标准《高质量数据集 格式要求》（征求意见稿）编制说明

一、工作简况

（一）任务来源

2025 年 12 月 31 日，根据《国家标准委关于下达 2025 年第十二批推荐性国家标准计划及相关标准外文版计划的通知》（国标委发〔2025〕76 号），国家标准《高质量数据集 格式要求》制定计划下达，计划号为 20256915-T-907。该标准由全国数据标准化技术委员会提出并归口，主管部门为国家数据局。

该标准的起草单位为中国电子技术标准化研究院华东分院、中国电子信息产业发展研究院、中国移动通信集团有限公司、中国电子技术标准化研究院、中国科学院计算技术研究所、国家数据发展研究院、中电数据产业集团有限公司、国务院国有资产监督管理委员会研究中心、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、浦江国家实验室、工业和信息化部电子第五研究所、中国联合网络通信集团有限公司、中国电信集团有限公司、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、商业信用中心、中国信息通信研究院、北京智网数科技术有限公司、石化盈科信息技术有限责任公司、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、中移互联网有限公司、国家电投集团数字科技有限公司、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、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、福建省大数据集团有限公司、湖北大数据集团数据开发有限公司、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、卡奥斯工业智能研究院（青岛）有限公司、杭州市临安区大数据管理服务中心、国网河南省电力公司经济技术研究院、软通智慧科技有限公司、烽火通信科技股份有限公司、太极计算机股份有限公司、复旦大学、同方知网数字科技有限公司、中移雄安信息通信科技有限公司、数据堂（北京）科技股份有限公司、《智慧中国》杂志社有限责任公司、河南金盾信安检测评估中心有限公司、河南省泛物网络科技有限公司、信阳市璀璨科技有限责任公

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、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、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、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、北京中数睿智科技有限公司、江苏中堃数据技术有限公司。

该标准的主要起草人为王为中、韩冰、张欢、吴坤、郭嘉丰、赵鹏飞、张群、廖华明、王超、温晓君、苏越阳、李天舒、时晓光、郭祎萍、连森、黄吉海、刘怡林、周艳芳、王亚沙、赵俊峰、马连韬、付文豪、谭瑾、李成博、蒋楠、刘陈宇、刘学忠、张黎明、周春雷、赵翔宇、贾蕾、李珍翔、程广明、李金夏、施晓辉、王锋、程健、骆意、张建中、武光城、谢卫军、罗腾、赵丽丽、王鑫、刘俊华、吴峥、李世奇、刘颖、刘广、杨二龙、邱泳钦、刘煜宏、肖邱勇、王雅琴、李荪、曹峰、姜辉、朱江涛、索寒生、严龙云、王晶、李亚楠、梁小涛、杨山、王海波、薛健、刘速、潘登、谭晓坤、胡力旗、邓成龙、汪睿棋、王兴旺、林云峰、张冲、袁芮、李娜、刘超、代勤、袁小乐、沈明辉、向海平、周志

华、邱会丽、孔亚文、蔡斯博、李静、孙晖、丁斌、张毅、周季峰、张锦辉、李凡、李科、黄宇恒、葛海龙、王培养、王宇、申中一、戴斌、王圆圆、林镇阳、陈刚、李佳忆、何震瀛、张庆国、杨彭年、齐红威、张挺、梁宏、何少英、马盼、郑随兵、王鹤、吴德亮、陈曦、邵志敏、范瑞、杨坤燚、张凯、邱旭东、鲁胜强、夏飞、王鹏飞、杨小红、王俊吉、李晓儿、刘云涛、严长春、庞俊奇、徐小传、彭荣、陈颖、温冬梅、韩涵、魏清。

起草单位、起草人及各自完成的工作如下：

王为中（中国电子技术标准化研究院华东分院）、韩冰（中国电子信息产业发展研究院）、张欢（中国电子技术标准化研究院）牵头制定高质量数据集格式要求技术框架，统筹标准主要章节内容、协调处理意见分歧等，负责各阶段的整体进度控制及内容审核。

吴坤（中国移动通信集团有限公司）、郭嘉丰（中国科学院计算技术研究所）牵头编制元数据属性章节等相关内容。

赵鹏飞、张群（中国电子技术标准化研究院）、廖华明（中国科学院计算技术研究所）、王超、温晓君（中国电子信息产业发展研究院）、苏越阳（中国科学院计算技术研究所）牵头编制数据集元数据章节中数据标识、关联数据标识、数据内容等相关内容。

李天舒、时晓光（国家数据发展研究院）、郭祎萍、连森（中电数据产业集团有限公司）、黄吉海（国务院国有资产监督管理委员会研究中心）、刘怡林、周艳芳（交通运输部公路科学研究

所）、王亚沙、赵俊峰、马连韬（北京大学）、付文豪（公安部第三研究所)牵头编制数据集元数据章节中标注信息、原始时间、最后修改时间等相关内容。

谭瑾（中国石油天然气集团有限公司）牵头编制数据集元数据章节中数据版本等相关内容。

李成博（中国电子技术标准化研究院）牵头编制标准范围、规范性引用文件等相关内容。

蒋楠、刘陈宇（中国石油化工集团有限公司）、刘学忠（中国交通建设集团有限公司）、张黎明（国家能源投资集团有限责任公司信息技术分公司）、周春雷（国家电网有限公司大数据中心）、赵翔宇（中国南方电网有限责任公司）、贾蕾（国家石油天然气管网集团有限公司)牵头编制数据集元数据章节中数据版本、授权类型等相关内容。

李珍翔（浦江国家实验室）、程广明（工业和信息化部电子第五研究所）、李金夏（中国联合网络通信集团有限公司）、施晓辉（中国电信集团有限公司）、王锋（中国质量认证中心有限公司）、程健、骆意、张建中、武光城（煤炭科学研究总院有限公司）牵头编制数据集元数据章节中来源类型、来源详情、生成数据标志等相关内容。

谢卫军、罗腾（中国稀土集团有限公司）、赵丽丽、王鑫（华为技术有限公司）、刘俊华、吴峥（科大讯飞股份有限公司）、李世奇（阿里巴巴（中国）有限公司）、刘颖、刘广（北京智源人工智能研究院）、杨二龙、邱泳钦（北京百度网讯科技有限公

司）、刘煜宏、肖邱勇（深圳市腾讯计算机系统有限公司）牵头编制数据内容元数据章节相关内容等相关内容。

王雅琴（商业信用中心）、李荪、曹峰（中国信息通信研究院）、姜辉、朱江涛（北京智网数科技术有限公司）、索寒生、严龙云（石化盈科信息技术有限责任公司）、王晶、李亚楠（中国交通信息科技集团有限公司）、梁小涛（中国移动通信集团有限公司）、杨山（中移动信息技术有限公司）、王海波（中移互联网有限公司）牵头编制标注信息元数据章节中相关内容。

薛健（国家电投集团数字科技有限公司）、刘速（中石油（北京）数智研究院有限公司）、潘登（联通数据智能有限公司）、谭晓坤（上海库帕思科技有限公司）、胡力旗（上海信投智能科技股份有限公司）参与编制元数据属性章节相关内容。

邓成龙、汪睿棋、王兴旺（中国电子技术标准化研究院）牵头编制术语定义、缩略语等相关章节内容。

林云峰、张冲（航天科工网络信息发展有限公司）、袁芮（中国邮政储蓄银行股份有限公司）、李娜（中电金信软件有限公司）、刘超（江苏省大数据管理中心）、代勤（内蒙古自治区大数据中心）、袁小乐（江西省大数据中心）、沈明辉、向海平（四川省卫生健康信息中心（四川省健康医疗大数据中心））参与编制数据集元数据章节中数据标识、关联数据标识、数据内容等相关内容。

周志华、邱会丽、孔亚文（北京大学（天津滨海）新一代信息技术研究院）、蔡斯博、李静（国家开放大学）参与编制数据

集元数据章节中标注信息、原始时间、最后修改时间等相关内容。

孙晖（杭州数美科技有限公司）、丁斌（福建省大数据集团有限公司）、张毅（湖北大数据集团数据开发有限公司）参与编制数据集元数据章节中数据版本、授权类型等相关内容。

周季峰（南京南瑞继保工程技术有限公司）、张锦辉（南京南瑞瑞中数据股份有限公司）、李凡（中通服网盈科技有限公司）、李科（北京海天瑞声科技股份有限公司）、黄宇恒、葛海龙（广州数字健康科技有限公司）、王培养（安徽飞数信息科技有限公司）、王宇（卡奥斯工业智能研究院（青岛）有限公司）参与编制数据集元数据章节中来源类型、来源详情、生成数据标志等相关内容。

申中一（中国电子技术标准化研究院）参与编制数据集元数据章节中生成数据标志等相关内容。

戴斌（杭州市临安区大数据管理服务中心）、王圆圆（国网河南省电力公司经济技术研究院）、林镇阳（软通智慧科技有限公司）、陈刚（烽火通信科技股份有限公司）、李佳忆（太极计算机股份有限公司）、何震瀛（复旦大学）参与编制数据内容元数据章节相关内容。

张庆国（同方知网数字科技有限公司）、杨彭年（中移雄安信息通信科技有限公司）参与编制标注信息元数据章节中相关内容。

齐红威（数据堂（北京）科技股份有限公司）、张挺（《智慧中国》杂志社有限责任公司）、梁宏（河南金盾信安检测评估

中心有限公司）、何少英（河南省泛物网络科技有限公司）、马盼（信阳市璀璨科技有限责任公司）、郑随兵（睿尔曼智能科技（北京）有限公司）、王鹤（北京银河通用机器人股份有限公司）参与标准内容的调研、研讨、试点验证等工作。

吴德亮（中兴通讯股份有限公司）、陈曦（浪潮电子信息产业股份有限公司）、邵志敏（国网山东省电力公司）、范瑞（蔚来汽车科技（安徽）有限公司）、杨坤燚（贵州大数据产业集团有限公司）、张凯（杭州市数据集团有限公司）、邱旭东（厦门赛西科技发展有限责任公司）、鲁胜强（云基华海信息技术股份有限公司）、夏飞、王鹏飞（国网江苏省电力有限公司）、杨小红（广东省人民医院）、王俊吉（辽宁省电子信息产品监督检验院）、李晓儿（数字宁波科技有限公司）、刘云涛（杭州景联文科技有限公司）、严长春（北京星河智源科技有限公司）、庞俊奇（山西集智数据服务有限公司）、徐小传（山东未来集团有限公司）、彭荣（广州维视达数字科技有限公司）、陈颖（厦门身份宝网络科技有限公司）、温冬梅（上海森栩医学科技有限公司）、韩涵（北京中数睿智科技有限公司）参与标准的试点验证、提供标准修改意见等工作。

魏清（江苏中堃数据技术有限公司）参与标准内容的调研、研讨等工作。

（二）制定背景及意义

人工智能作为引领新一轮科技革命和产业变革的战略性技术，深刻改变人类生产生活方式。随着人工智能技术快速发展，

研发重点正从“重点优化模型架构”转向“模型与数据协同优化”，其中，高质量数据的作用日益凸显。数据作为人工智能发展的三大核心要素之一，已成为人工智能模型开发和训练的核心要素资源。充分发挥标准的支撑和引领作用，加快高质量数据集规范化建设，对于推动人工智能赋能行业发展具有重要意义。

制定该标准的必要性、重要性等主要体现在以下方面。

一是落实国家政策要求。国家高度重视高质量数据集建设工作，先后出台《国家数据局等部门关于印发〈“数据要素×”三年行动计划（2024—2026 年）〉的通知》（国数政策〔2023〕11 号）、《国家数据局等部门关于促进企业数据资源开发利用的意见》（国数资源〔2024〕125 号）、《国家发展改革委等部门关于促进数据标注产业高质量发展的实施意见》（发改数据〔2024〕1822 号）、《国家发展改革委等部门关于促进数据产业高质量发展的指导意见》（发改数据〔2024〕1836 号）、《国家发展改革委 国家数据局 工业和信息化部关于印发〈国家数据基础设施建设指引〉的通知》（发改数据〔2024〕1853 号）等多项政策文件，布局建设行业高质量数据集。标准在高质量数据集建设中可发挥规范和引领作用，《国家发展改革委等部门关于印发〈国家数据标准体系建设指南〉的通知》提出重点推进建设训练数据集采集处理标准，包括训练数据集格式要求、分类分级、采集性能、分析监测、质量要求等标准。

二是满足行业发展需求。当前，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，推动高质

量发展。高质量数据集是人工智能模型开发和训练的基础，是人工智能高效赋能行业发展的重要支撑。制定高质量数据集格式要求标准，明确其元数据及表示方法，有利于通过统一接口对数据集进行读取、使用，对于促进高质量数据集流通、应用，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。

（三）起草过程

2025 年 1 月：成立标准编制组，开展广泛调研和资料收集，明确工作思路和编制原则，讨论确定标准框架。

2025 年 2 月–5 月：编制组内部讨论并编制形成标准草案。

2025 年 6 月底：立项申报。

2025 年 6 月–8 月：全国数标委秘书处组织对该标准进行验证试点，共 33 家单位报名参与。各试点单位结合实际业务场景开展验证工作，编制组结合收集到的意见建议修改完善标准草案，进一步提升标准的科学性、适用性和先进性。

2025 年 9 月–12 月：组织开展研讨，不断完善标准草案。

2025 年 12 月底：国家标准委正式下达标准计划。

2026 年 1 月：持续修改完善标准草案，形成征求意见稿。

二、国家标准编制原则、主要内容及其确定依据

（一）编制原则

该标准的编制原则主要包含两个方面：

1. 该标准涉及相关方众多，鼓励高质量数据集相关建设主体、技术服务厂商、研究机构等广泛参与，以确保标准内容科学合理，具有普适性。

2. 该标准属于《国家数据标准体系建设指南》中的“C 数据资源−CE 训练数据集−CEA 训练数据集采集处理”标准，对人工智能数据产业发展具有重要的支撑作用。该标准应充分借鉴国际、国内相关先进研究成果，与国家相关政策导向相一致。

（二）编制依据

对于人工智能模型开发和训练，国内外均已发布不少高质量的数据集，国外有 Google 的 PaLM、Facebook 的 Wav2Letter、OpenAI 的 ImageNet-21K 等，国内有北京智源研究院的 WuDao 2.0、清华大学的 THUDataset、华为的 Aishell-1 等，具有各自不同的格式要求，然而，形成统一、规范的格式要求，是建设和复用高质量数据集的必然要求，也是未来的发展趋势。

该标准在编制过程中主要参考了 GB/T 18391.1-2009《信息技术 元数据注册系统（MDR） 第 1 部分：框架》、ECMA-404 The JSON data interchange syntax、IETF RFC 8259 The JavaScript Object Notation (JSON) Data Interchange Format、语义化版本（Semantic Versioning）等标准或文件。

（三）主要内容

该标准规定了高质量数据集的元数据及其表示方法，适用于指导建设、管理和加工高质量数据集。

高质量数据集的元数据包括数据标识、关联数据标识、数据内容、标注信息、原始时间、最后修改时间、数据版本、授权类型、来源类型、来源详情、生成数据标志等元数据，其中，数据内容元数据包括模态类型、内容等元数据；标注信息元数据包括

标签、标注方式、标注工具、标注人员类型等元数据。每个元数据用 7 个属性描述，包括中文名称、英文名称、定义、数据类型、值域、数据填充约束、备注等。

三、试验验证的分析

该标准所规定的内容经过贵州大数据集团、杭州数据集团、腾讯、国网江苏电力、森翎医学、公路院、云基华海、国家能源、中数睿智、国家管网、维视达数科、浪潮电子、国投数科、数字宁波、广东医科院、蔚来汽车、山东未来、联通数智、同方知网数科、杭州景联文、集智数据、国网山东电力、中兴通讯、星河智源、厦门身份宝、烽火通信、中国石油、中国石化、中国交建、中国联通、辽宁电子信息院、中电数产、中国移动等 33 家企事业单位验证，已被证明确实可行，对相关建设方、技术服务方等开展数据集建设、管理、加工等具有实际指导价值。

四、与国际、国外同类标准技术内容的对比情况

目前，国外尚无与数据集格式强相关的标准在研或发布。

五、产业化情况、推广应用论证和预期达到的经济效果、社会效益和生态效益

在国内，对于通用领域，主要结合国外优质开源数据集和有限的中文数据形成高质量数据集，其中，中文数据主要来自百科、问答等互联网公开数据，其他类型优质数据来源仍相对较少；对于行业领域，尽管行业高质量数据集供给仍略显不足，但已取得一定积极进展，目前已建成高质量数据集超 10 万个，规模超 890PB。目前，在通用和行业领域的高质量数据集建设中，国内

相关企事业单位在格式要求方面的实践不一，亟需通过标准进行规范、统一。

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。训练数据集是开发和训练人工智能模型的基础，本标准的发布必定有助于促进训练数据集流通应用，推动人工智能高效赋能行业发展。在推广应用方面，在本标准发布之后，有关建设主体可依据标准进行训练数据集建设，构建能够方便地进行流通应用的训练数据集；有关技术服务方可基于标准，对建设主体提供在格式方面符合规范的数据集建设、管理和加工相关服务。

六、是否合规引用或者采用国际国外标准

该标准未引用或者采用国际国外标准。

七、与现行相关法律、法规、规章及相关标准的协调性

该标准与现行相关法律、法规、规章及相关标准协调一致。

八、重大分歧意见的处理经过和依据

该标准研制过程中未涉及重大分歧意见。

九、涉及知识产权或专利的情况说明

该标准不涉及知识产权或专利。

十、实施国家标准的要求

建议作为推荐性国家标准，在标准报批阶段及正式发布后，同步开展标准宣贯培训与应用示范工作。建议标准发布6个月后正式实施。

十一、贯彻标准的要求和措施建议

1. 加强政府引导与宣传推广。标准发布后，在国家数据局指导下，由全国数据标准化技术委员会组织开展标准宣贯活动，在高质量数据集相关产业链和应用领域加强宣传，提升标准的宣传权威性和受众针对性；

2. 完善配套政策与激励措施。建议相关部门结合本标准中数据集格式要求，在数据集建设、运营、流通等方面研究出台配套政策措施。鼓励各类企事业单位依据标准开展数据集建设，对符合格式要求的数据集建设项目给予资金等方面激励。推动数据集格式要求与市场准入、产业扶持政策精准衔接，将格式规范性作为衡量数据集是否“高质量”的重要指标，切实增强企业应用标准的积极性；

3. 推动第三方评测与生态协同。鼓励第三方机构依据本标准所规定的元数据及表示方法，开展数据集格式符合性测试，建立健全高质量数据集格式评测能力。同时，推动建立政府与市场协同的多方采信机制，促进格式符合性测试结果在数据集流通与价值实现中发挥作用，提升标准的实施效力与行业认可度。

十二、替代或废止现行相关标准的建议

无。

十三、公平竞争审查结论

该标准已完成公平竞争审查，并填写了《公平竞争审查表》。该标准起草过程中无限制或变相限制市场准入和退出、商品要素自由流动等情况，未对经营者生产经营成本、生产经营行为造成不利影响，不存在违反《公平竞争审查条例》规定的情况，符合

公平竞争审查标准。

十四、其它应予说明的事项

无。

国家标准《高质量数据集 格式要求》

编制工作组

2026-01-30