# 国家标准《高质量数据集 分类指南》（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 分类指南》制定计划下达，计划号为 20256912-T-907，任务主管部门为国家数据局。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北京大学、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、中国石油化工集团有限公司、国家数据发展研究院、北京智源人工智能研究院、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、公安部第三研究所、中国石油天然气集团有限公司、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、中国移动通信集团有限公司、中国联合网络通信集团有限公司、中国电信集团有限公司、中国稀土集团有限公司、中电数据产业集团有限公司、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、北京百度网讯科技

有限公司、深圳市腾讯计算机系统有限公司、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、中国质量认证中心有限公司、煤炭科学研究总院有限公司、北京智网数科技术有限公司、中移动信息技术有限公司、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、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、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、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、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亚沙、赵鹏飞、韩冰、李成博、王为中、郭嘉丰、蒋楠、李睿童、赵俊峰、张欢、张群、李冰、王超、温晓君、廖华明、苏越阳、李天舒、时晓光、黄吉海、刘颖、刘广、初旭、刘冬梅、张悦、孙亚茹、方可、徐健、汲倩倩、李俊妮、冯勤宇、于涛、涂中英、吴坤、王皓天、周渭华、王有霖、余传健、顾延甲、刘彬彬、赵丽丽、王鑫、刘俊华、吴峥、李世奇、杨二龙、邱泳钦、刘煜宏、安德亮、程广明、燕江依、曹峰、孔佳、王锋、程健、杨培培、张建中、武光城、余海涛、李梁、王新明、张琳琳、谌湘临、刘晓遇、金柳、梁翠兰、薛健、杨博、赵兴华、山栋明、胡力旗、邓成龙、汪睿棋、王兴旺、申中一、周凯、白亚南、原国斌、王春涛、陈倩、刘瑛、崔连伟、孙杨、沈明辉、毛云鹏、李方平、马二燕、邱会丽、蔡斯博、程罡、何转琴、徐丹、张锦辉、李凡、廖晓玲、黄宇恒、葛海龙、谭昶、熊威、姜舰艇、林镇阳、张庆国、魏星华、郑随兵、张直政、周午杰、吴德亮、陈曦、朱璐、韩大鑫、张萌、顾永莉、李倩、郑

剑、鲁胜强、何金陵、马洲俊、李丹、王俊吉、卫学彬、刘云涛、严长春、庞俊奇、辛潇、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：

王亚沙（北京大学）、韩冰（中国电子信息产业发展研究院）、赵鹏飞、李成博、王为中（中国电子技术标准化研究院）、郭嘉丰（中国科学院计算技术研究所）牵头制定高质量数据集分类指南技术框架，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。

蒋楠、李睿童（中国石油化工集团有限公司）、赵俊峰（北京大学）牵头编制类型划分章节中类型要素相关内容。

张欢、张群、李冰（中国电子技术标准化研究院）、王超、温晓君（中国电子信息产业发展研究院）、廖华明、苏越阳（中国科学院计算技术研究所）牵头编制类型划分章节中类型特征相关内容。

李天舒、时晓光（国家数据发展研究院）、黄吉海（国务院国有资产监督管理委员会研究中心）、刘颖、刘广（北京智源人工智能研究院）、初旭（北京大学）、刘冬梅、张悦（交通运输部公路科学研究所）、孙亚茹（公安部第三研究所）牵头编制类型划分章节中分类方法相关内容。

方可（中国石油天然气集团有限公司）、徐健（中国交通建设集团有限公司）、汲倩倩（国家能源投资集团有限责任公司信息技术分公司）、李俊妮（国家电网有限公司大数据中心）、冯勤宇（中国南方电网有限责任公司）、于涛（国家石油天然气管

网集团有限公司）、涂中英（浦江国家实验室）、吴坤（中国移动通信集团有限公司）、王皓天（中国联合网络通信集团有限公司）、周渭华（中国电信集团有限公司）牵头编制类型划分章节中类型特征中行业通识数据集相关内容。

王有霖、余传健（中国稀土集团有限公司）、顾延甲、刘彬彬（中电数据产业集团有限公司）、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）、杨二龙、邱泳钦（北京百度网讯科技有限公司）、刘煜宏、安德亮（深圳市腾讯计算机系统有限公司）牵头编制类型划分章节中类型特征中通识数据集相关内容。

程广明（工业和信息化部电子第五研究所）、燕江依、曹峰（中国信息通信研究院）、孔佳（商业信用中心）、王锋（中国质量认证中心有限公司）牵头编制术语定义相关章节内容。

程健、杨培培、张建中、武光城（煤炭科学研究总院有限公司）、余海涛、李梁（北京智网数科技术有限公司）、王新明（中国移动通信集团有限公司）、张琳琳（中移动信息技术有限公司）、谌湘临、刘晓遇（石化盈科信息技术有限责任公司）牵头编制类型划分章节中类型特征中行业专识数据集相关内容。

金柳、梁翠兰（中国交通信息科技集团有限公司）、薛健（国家电投集团数字科技有限公司）、杨博（中石油（北京）数智研究院有限公司）、赵兴华（联通数据智能有限公司）、山栋明（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）参与编制类型划分章节中类型要素相关内容。

邓成龙、汪睿棋、王兴旺、申中一（中国电子技术标准化研究院）牵头编制标准范围、规范性引用文件相关章节内容。

周凯、白亚南（航天科工网络信息发展有限公司）、原国斌（中国邮政储蓄银行股份有限公司）、王春涛（中国电子工程设计院股份有限公司）、陈倩（中电金信软件有限公司）、刘瑛（江苏省大数据管理中心）、崔连伟（内蒙古自治区大数据中心）参与编制类型划分章节中类型特征相关内容。

孙杨（江西省大数据中心）、沈明辉、毛云鹏（四川省卫生健康信息中心（四川省健康医疗大数据中心））、李方平、马二燕、邱会丽（北京大学（天津滨海）新一代信息技术研究院）、蔡斯博、程罡（国家开放大学）参与编制类型划分章节中分类方法相关内容。

何转琴（杭州数美科技有限公司）、徐丹（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）、廖晓玲（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）、谭昶（安徽飞数信息科技有限公司）、熊威（湖北大数据集团数据开发有限公司）、姜舰艇（杭州市临安区大数据管理服务中心）、林镇阳（软通智慧科技有限公司）、张庆国（同方知网数字科技有限公司）、魏星华（数据堂（北京）科技股份有限公司）、郑随兵（睿尔曼智能科技（北京）有限公司）、张直政（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。

周午杰（烽火通信科技股份有限公司）、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）、朱璐（国网山东省电力公司）、韩大鑫（蔚来汽车科技（安徽）有限公司）、张萌（贵州大数据产业集团有限公司）、顾永莉（杭州市数据集团有限公司）、李倩（四川数据集团有限公司）、郑剑（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）、何金陵、马洲俊（国网江苏省电力有限公司）、李丹（广东省人民医院）、王俊吉（辽宁省电子信息产品监督检验院）、卫学彬（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）、辛潇（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）负责标准的试点验证、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。**人工智能模型对数据集的需求正从“通用知识”向“专业知识”延伸拓展，“分类建设”高质量数据集，才能有效支撑通用模型、行业模型、场景模型等落地应用。制定高质量数据集分类指南，明确类型划分的类型要素、类型特征、分类方法，为开展高质量数据集分类活动提供指导，对于提

升数据集供需匹配，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。

（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

通用模型、行业模型、场景模型等不同类型模型需要不同类型的数据集，相应数据集需要蕴含通用知识、行业领域通用知识、行业领域专业知识，鉴于从数据集“所蕴含知识深度”确定其类型较为困难，该标准主要根据不同类型数据集在知识内容、来源类型、标注人员类型等不同方面所具备的特征不同，来确定数据集的类型。

（三）主要内容

该标准规定了高质量数据集的类型划分，给出了类型要素、类型特征、分类方法，可为开展高质量数据集分类活动提供指导。

高质量数据集可分为通识、行业通识、行业专识等类型。不同类型的高质量数据集在知识内容、来源类型、时效性、标注人员类型、敏感程度、模型类型、主题范围等类型要素方面的特征不同。在确定高质量数据集类型时，可以按行业专识、行业通识、通识的顺序，根据数据集是否满足相关特征条件，将数据集的类型确定为行业专识、行业通识或通识数据集。

**三、试验验证的分析**

该标准所规定的内容经过森翎医学、腾讯、杭州数据集团、中国移动、云基华海、公路院、贵州大数据集团、联通数智、同方知网数科、国网山东电力、山东未来、中国联通、维视达数科、中数睿智、四川数据集团、数字宁波、国家能源、浪潮电子、蔚来汽车、杭州景联文、国家管网、广东医科院、中兴通讯、中国

石油、中国石化、中国交建、国网江苏电力、集智数据、星河智源、中电数产、辽宁电子信息院、厦门身份宝、烽火通信等 33 家企事业单位验证，已被证明确实可行，对数据集相关建设方、技术服务方等开展分类活动具有实际指导价值。

**四、与国际、国外同类标准技术内容的对比情况**

国际方面，尚无数据集分类标准；ISO/IEC 20547-3:2020《信息技术 大数据参考体系结构 第 3 部分：参考体系结构》有涉及“数据分类”。总体上看，国外目前尚无针对“高质量数据集”的分类标准。

**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**

高质量数据集可以从多个维度进行分类，在模态方面，可以分为文本、图片、音频、视频等类型；在来源方面，可以分为公开、专有、合成等类型；在用途方面，可以分为训练、验证、测试、基准等类型；在应用领域方面，可以分为计算机视觉、自然语言处理、语音识别、自动驾驶等类型；……。目前，在国内高质量数据集建设中，国内相关企事业单位从“支撑通用、行业、场景等不同层次类型模型开发和训练”的维度开展数据集分类工作的实践仍相对较少。

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基础，本标准的发布必定有助于提升数据集优质供给，促进数据集流通应用，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数据集分类建设、管理等，构建能够有效“分类”支撑通用、行业、场景等不同类型模型落地应用的高质量数据集；有关技术服务方可基于标准，为建设主体提供在类型方面符合要求的数据集建设、管理相关服务。

**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类

型划分，在数据集分类建设、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类型划分，对数据集分类建设、管理等富有成效的项目或企业给予资金等方面激励。推动数据集分类与行业监管、产业扶持政策精准衔接，将数据集分类建设、管理等作为衡量数据集项目建设成效的重要参考，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的类型要素、分类特征、分类方法，开展数据集分类评测，建立健全高质量数据集分类评测能力。同时，推动政府与市场建立多方采信机制，推动分类评测结果在数据集运营、流通中发挥作用，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 分类指南》

编制工作组

2026-01-30