TC609

全国数据标准化技术委员会技术文件

TC609-5-2025-03

高质量数据集 分类指南

High-quality dataset—Classification 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：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、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、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、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、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、北京智源人工智能研究院、上海人工智能创新中心、中电数据产业集团有限公司、中国质量认证中心有限公司、北京百度网讯科技有限公司、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、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、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、中兴通讯股份有限公司、浪潮电子信息产业股份有限公司、同方知网数字科技有限公司、烽火通信科技股份有限公司、蔚来汽车科技（安徽）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、辽宁省电子信息产品监督检验院、云基华海信息技术股份有限公司、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、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。

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，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，相应数据集需蕴含通用知识、行业领域通用知识、行业领域专业知识，然而，我国高质量数据集分类目前仍缺乏统一的标准规范。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，为组织机构开展高质量数据集分类工作提供指导，对于优化数据集供需匹配，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。

高质量数据集 分类指南

1 范围

本文件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。

本文件可为组织机构开展高质量数据集分类工作提供指导。

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：TC609-5-2025-01，3.1]

3.2

通用知识 general knowledge

面向社会公众的通用性知识，具有广泛性、基础性和常识性等特点。

注：通用知识主要包括基础概念、通用原理和典型事例等方面内容，无需专业背景即可理解和应用。

3.3

通识数据集 general knowledge dataset

蕴含通用知识的数据的集合。

3.4

行业领域通用知识 industry and field general knowledge

面向行业领域从业人员的通用性知识，在行业领域内部具有普适性和共识性等特点。

注：行业领域通用知识主要包括行业领域基础理论、通用技术和共性业务等方面内容，需要一定的专业背景方可理解和应用。

3.5

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。

3.6

行业领域专业知识 industry and field professional knowledge

面向行业领域机构内部业务人员的专业性知识，具有场景针对性、组织机构专属性和实践经验积累性等特点。

注：行业领域专业知识主要包括从研发、生产、管理、营销和服务等业务环节中产生和积累的知识，需要较深的行业背景和具体业务经验方可理解和应用。

3.7

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。

3.8

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]

4　类型划分

4.1　类型要素

高质量数据集可以分为通识、行业通识、行业专识三类，分别主要用于支撑通用、行业、场景模型落地应用。不同类型高质量数据集在多个类型要素方面的特征不同，其中，类型要素包括：

a）　知识内容：数据集中数据所蕴含知识的专业性、知识深度和目标受众。

b）　来源类型：数据集中数据的获取来源，如网络资源、文献类型、系统平台、组织机构等来源。

c）　时效性：数据集中数据的更新速度或有效期限。

d）　标注人员类型：对数据集中数据进行标注或审核的人员类型。

e）　敏感程度：数据集中数据公开后所产生的风险程度。

f）　模型类型：数据集中数据支持开发和训练人工智能模型的类型。

g）　主题范围：数据集所涉及通用知识领域、行业领域、业务场景等的范围。

4.2　类型特征

不同类型高质量数据集在多个类型要素方面的特征如表1所示。

表1　不同类型数据集的类型特征

数据集类型 / 类型要素	通识数据集	行业通识数据集	行业专识数据集
知识内容	面向社会普通公众，不需要专业背景即可理解	面向行业从业人员，需一定专业背景才能理解	面向内部业务人员，需较深专业背景才能理解
来源类型	来源不严格，主要来自百科、问答等互联网资源，综合性书籍等类型文献，以及生成数据	来源清晰，主要来自论文、报告、标准等专业文献，行业领域组织机构，以及生成数据	来源清晰，主要来自组织机构内部的业务系统、管理平台等系统平台，或文档图纸等文献
时效性	一般在较长时间内稳定，时效性要求较低	根据行业领域发展变化，时效性要求中等	根据业务场景需求变化，时效性要求较高
标注人员类型	普通标注人员	具备学科背景或从业经验的人员	行业领域专家
敏感程度	敏感程度较低	敏感程度较低	敏感程度较高
模型类型	通用模型、行业模型	通用模型、行业模型	场景模型
主题范围	不属于特定行业领域，主体范围较广	聚焦于特定行业领域，主体范围中等	聚焦于特定业务场景，主体范围较窄

注：通用模型，指面向广泛领域、不依赖特定行业知识的人工智能模型；行业模型，指针对特定行业领域共性需求开发的人工智能模型；场景模型，指聚焦具体业务场景、依赖组织机构内部专业知识的人工智能模型。

4.2.1 通识数据集

通识数据集在多个类型要素方面所具备的特征如下：

a) 知识内容

数据集中数据所承载信息蕴含的知识面向社会公众，以基础概念、通用原理、典型事例等常识性知识为主，无需专业背景即可理解。

b) 来源类型

数据集中数据未严格溯源，主要来自百科、问答、互联网平台（短视频类、新闻门户类、用户内容生成类、视听资讯类、新闻机构类等）等网络资源，综合性书籍等类型文献，以及由模拟、合成等技术生成。

c) 时效性

数据集中数据一般在较长时间内保持稳定，时效性要求较低。如年度国民经济统计公报、重大事件新闻报道等，一经发布即可在较长时间内作为信息参考，较少出现频繁调整变更的情况。

d) 标注人员类型

数据集中数据通常由普通数据标注员标注，对行业领域专业背景和专业资质等级无特殊要求。

e) 敏感程度

数据集中数据的敏感程度较低，不涉及国家秘密、工作秘密、商业秘密、个人敏感信息等。

f) 模型类型

数据集中数据用于支持开展通用模型、行业模型的开发和训练。

g) 主题范围

数据集不属于特定行业领域，相比于行业通识数据集，主题范围较广。

4.2.2 行业通识数据集

行业通识数据集在多个类型要素方面所具备的特征如下：

a) 知识内容

数据集中数据所承载信息蕴含的知识聚焦于各行业领域的共性知识，以行业领域的基础理论、通用技术、共性业务为主，面向行业领域的从业人员，需要一定的专业背景才能理解和运用。

b) 来源类型

数据集中数据来源清晰，主要来自论文、报告、标准、专利、技术文档、非综合性书籍、官方文件等类型文献，行业领域的主管部门、协会、科研机构、企业、出版传媒机构等组织机构，以及由模拟、合成等技术生成。

c) 时效性

数据集中数据一般根据行业发展和管理需求变化，相对比较稳定，时效性要求高于通识数据集，低于行业专识数据集。

d) 标注人员类型

数据集中数据通常由行业领域内具备一定学科背景或相关从业经历的人员标注。

e) 敏感程度

数据集中数据的敏感程度较低，不涉及国家秘密、工作秘密、商业秘密、个人敏感信息等。

f) 模型类型

数据集中数据用于支持开展通用模型、行业模型的开发和训练。

g) 主题范围

数据集主要聚焦于行业领域内部，但也应考虑上下游及周边相关行业领域知识对于该行业知识的影响与补充，主题范围中等，介于通识数据集和行业专识数据集之间。

4.2.3 行业专识数据集

行业专识数据集在多个类型要素方面所具备的特征如下：

a) 知识内容

数据集中数据所承载信息蕴含的知识聚焦于各行业领域组织机构内部的专业知识，以组织机构自身业务中从研发、生产、管理、营销和服务等环节产生和积累的知识为主，面向内部业务人员，需要较深的专业背景和业务经验才能理解和运用。

b) 来源类型

数据集中数据来源清晰，由组织机构日常生产经营生成和采集，主要来自组织机构内部的业务系统、管理平台等系统平台，或文档图纸等文献。

c) 时效性

数据集中数据一般根据业务场景需求变化，满足数据处理及时性要求，时效性要求较高。

d) 标注人员类型

数据集中数据通常由行业领域内具备深厚专业知识、丰富实践经验和高度权威性的专家标注。

e) 敏感程度

数据集中数据的敏感程度较高，仅供内部特定岗位人员使用，应用前需要明确权限和授权。

f) 模型类型

数据集中数据用于支撑开展场景模型的开发和训练。

g) 主题范围

数据集主要聚焦于业务场景，相比于行业通识数据集，主题范围较窄。

4.3 分类方法

如图1所示，高质量数据集的类型划分方法如下：

a) 分析数据集在 4.1 中各类型要素方面的特征；

b) 经综合判定，在整体上符合 4.2.3 中特征，数据集可确定为行业专识数据集；

c) 经综合判定，在整体上不符合 4.2.3 中特征，而符合 4.2.2 中特征，数据集可确定为行业通识数据集；

d) 经综合判定，在整体上不符合 4.2.2 中特征，数据集可确定为通识数据集。

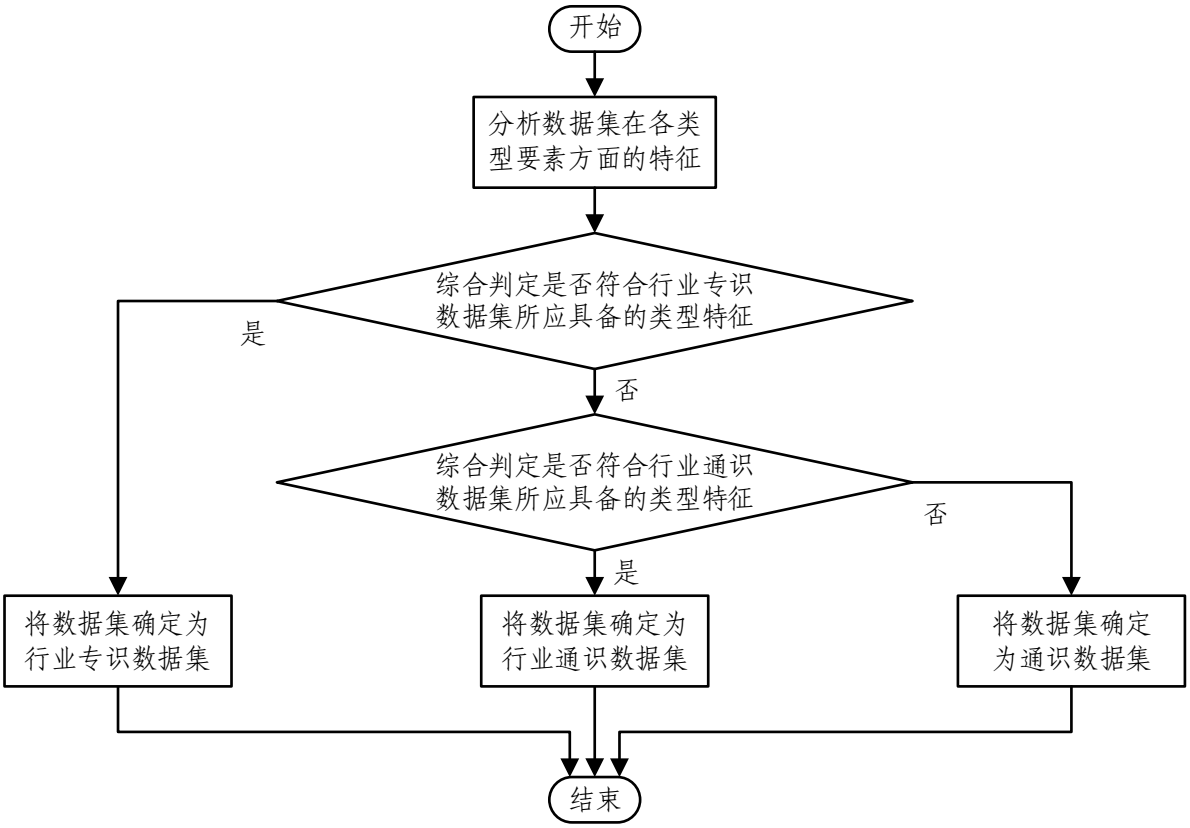

图1　高质量数据集类型划分方法

参 考 文 献

[1] GB/T 42755-2023 人工智能 面向机器学习的数据标注规程
[2] TC609-5-2025-01 高质量数据集 建设指南