ICS 35.030
CCS L 80

中华人民共和国国家标准

GB/T 45674—2025

网络安全技术
生成式人工智能数据标注安全规范

Cybersecurity technology—Generative artificial intelligence data annotation security specification

2025-04-25 发布　　2025-11-01 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。

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。

本文件由全国网络安全标准化技术委员会(SAC/TC 260)提出并归口。

本文件起草单位：国家计算机网络应急技术处理协调中心、中国电子技术标准化研究院、北京中关村实验室、北京快手科技有限公司、北京百度网讯科技有限公司、北京天融信网络安全技术有限公司、阿里云计算有限公司、北京大学、国家计算机网络应急技术处理协调中心江苏分中心、公安部第三研究所、清华大学、上海人工智能创新中心、北京市公安局人工智能安全研究中心、西安邮电大学、浙江大学、中国科学院信息工程研究所、中国移动通信集团有限公司、小米科技有限责任公司、蚂蚁科技集团股份有限公司、华为云计算技术有限公司、北京数安行科技有限公司、北京晴数智慧科技有限公司、北京零一万物科技有限公司、北京奇虎科技有限公司、科大讯飞股份有限公司、联想(北京)有限公司、启明星辰信息技术集团股份有限公司、亚信科技(成都)有限公司、杭州萤石软件有限公司、北京东方通网信科技有限公司、广东省信息安全测评中心、厦门美柚股份有限公司、北京瑞莱智慧科技有限公司、天翼安全科技有限公司、北京远鉴信息技术有限公司、上海商汤智能科技有限公司、苏州核数聚信息科技有限公司、南京领行科技股份有限公司、江苏满运软件科技有限公司、长安通信科技有限责任公司、OPPO广东移动通信有限公司。

本文件主要起草人：张震、谭知行、张妍婷、贺敏、刘勇、孙旭东、徐恪、陈钟、杜金浩、郝春亮、任奎、刘楠、落红卫、叶晓俊、安勍、胡影、王龑、姚龙、谢安明、嵇程、江为强、丁治国、雷晓锋、戴娇、谷晨、张晴晴、郭建领、张勇、罗磊、刘玉红、廖双晓、蒋慧、赵云、张峰、许晓耕、王文宇、陈洋、张夏、彭骏涛、包沉浮、王海棠、孟凡芹、赵丽丽、刘俊华、李家锟、崔婷婷、余瀚洋、李峰风、臧娇娇、林冠辰、丁欣、王士进、韩晗、张向征、胡嵩智、徐怡悦、管铭、张天奕、黄喆、刘俊、周雪、郑榕、刘栋、罗旭鹏、郑鸿咚、蒋发群、马梦娜、田伟丽、胡月、黄鹏华、张小敏、张中维、周城、李根、李笑如、张秉晟、王和俊、刘洞宾。

引　言

数据标注是生成式人工智能的关键活动，直接决定了训练数据以及生成内容的质量和安全水平，但由于标注规则不完善、人员管理不规范、核验标准不明确等原因，在数据标注过程中也可能为生成式人工智能引入新的风险隐患，亟需标准规范用于提高数据标注的安全水平。为加强生成式人工智能数据标注活动的安全管理，采取有效措施防范和处置相关风险，编制本文件，旨在帮助服务提供者、数据标注组织方以及数据需求方明确数据标注的安全基线、提高服务安全水平。

网络安全技术
生成式人工智能数据标注安全规范

1 范围

本文件规定了生成式人工智能训练的数据标注平台或工具安全要求、数据标注规则安全要求、数据标注人员要求、数据标注核验要求，描述了数据标注安全评价方法。

本文件适用于生成式人工智能数据标注组织方开展训练数据标注活动，并为生成式人工智能数据需求方对于数据标注进行检查、验收或第三方机构对数据标注进行安全性评估提供参考。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本(包括所有的修改单)适用于本文件。

GB/T 42755—2023 人工智能 面向机器学习的数据标注规程

GB/T 45654—2025 网络安全技术 生成式人工智能服务安全基本要求

3 术语和定义

下列术语和定义适用于本文件。

3.1

提示信息 prompt

引导生成式人工智能模型完成特定任务并提供合理输出内容的输入信息。

3.2

响应信息 response

在生成式人工智能数据标注中，按照提示信息要求形成的符合人类认知的应答信息，用于训练模型形成对提示信息输出相应内容、模式或风格的响应的能力。

3.3

生成式人工智能数据标注 generative artificial intelligence data annotation

通过人工操作或使用自动化技术机制，基于对提示信息的响应信息内容，将特定信息如标签、类别或属性添加到文本、图片、音频、视频或者其他数据样本的过程。

注：以下简称“数据标注”。

[来源：GB/T 45654—2025，3.5]

3.4

功能性数据标注 functional data annotation

用于训练生成式人工智能模型具备完成特定任务能力的数据标注。

[来源：GB/T 45654—2025，3.6]

3.5

安全性数据标注 security data annotation

用于训练生成式人工智能模型提升输出响应信息安全性的数据标注。

[来源：GB/T 45654—2025，3.7]