ICS 35.030
CCS L 80

中华人民共和国国家标准

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

Cybersecurity technology—Basic security requirements for generative artificial intelligence service

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：中国电子技术标准化研究院、国家计算机网络应急技术处理协调中心、浙江大学、北京中关村实验室、上海人工智能创新中心、复旦大学、北京百度网讯科技有限公司、阿里云计算有限公司、北京快手科技有限公司、华为云计算技术有限公司、北京航空航天大学、联想(北京)有限公司、蚂蚁科技集团股份有限公司、科大讯飞股份有限公司、北京大学、中国网络安全审查认证和市场监管大数据中心、北京深度求索人工智能基础技术研究有限公司、北京奇虎科技有限公司、中国科学院自动化研究所、河南科技大学、中国政法大学、上海交通大学、清华大学、中国科学院软件研究所、OPPO广东移动通信有限公司、中国移动通信集团有限公司、深信服科技股份有限公司、北京面壁智能科技有限责任公司、北京瑞莱智慧科技有限公司、国家工业信息安全发展研究中心、公安部第三研究所、国家信息中心、上海燧原科技股份有限公司、深圳陆兮科技有限公司、杭州网易智企科技有限公司、贝壳找房(北京)科技有限公司、北京天融信网络安全技术有限公司、北京零一万物科技有限公司、上海稀宇科技有限公司、广州市动悦信息技术有限公司、天翼安全科技有限公司。

本文件主要起草人：姚相振、郝春亮、张妍婷、张震、任奎、刘勇、杨珉、秦湛、胡影、夏文辉、陈钟、王迎春、贺敏、张凌寒、许晓耕、刘建伟、落红卫、王凤娇、徐恪、陈洋、张向征、包沉浮、谢安明、彭骏涛、谷晨、郑子木、吴少卿、王姣、王秉政、郭建领、孟令宇、徐甲、杨子祺、王庆龙、邱锡鹏、黄晴、石琳、张宗洋、边松、张志勇、张谧、洪赓、潘旭东、胡永启、林冠辰、刘俊华、乔玉平、梅敬青、贾开、赵静、张严、权高原、谭知行、杨光、姚龙、李琦、王晖、朱贵波、周芃、安勍、沈俊成、赵睿斌、刘栋、马梦娜、王俊、张立尧、贾雨萌、王海棠、彭韬、李根、邱勤、江为强、徐阳、游建舟、周呈辉、刘楠、丁治国、王荣仕、李大海、朱晓芳、王雨晨、薛智慧、肖博峰、危嘉祺。

引　言

当前，生成式人工智能技术不断发展，相关服务已广泛应用，为社会生产生活等各方面提供便利，但也带来大量网络安全新风险、新挑战，亟需标准规范设立安全基线。

本文件重点面向具有舆论属性或者社会动员能力的生成式人工智能服务，支撑备案管理、检测评估等方面工作开展。重点关注数据标注安全时，本文件可与GB/T 45674《网络安全技术　生成式人工智能数据标注安全规范》结合使用；重点关注预训练和优化训练数据安全时，本文件可与GB/T 45652《网络安全技术　生成式人工智能预训练和优化训练数据安全规范》结合使用。

网络安全技术　生成式人工智能服务
安全基本要求

1　范围

本文件规定了生成式人工智能服务在训练数据安全、模型安全、安全措施等方面的要求。

本文件适用于服务提供者开展生成式人工智能服务相关活动，也为相关主管部门以及第三方评估机构提供参考。

注：训练数据及生成内容涉及的主要安全风险见附录A，生成式人工智能服务安全评估参考方法见附录B。

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 25069　信息安全技术　术语

3　术语和定义

GB/T 25069界定的以及下列术语和定义适用于本文件。

3.1

生成式人工智能服务　generative artificial intelligence service

利用生成式人工智能技术向公众提供生成文本、图片、音频、视频等内容的服务。

3.2

服务提供者　service provider

以交互界面、可编程接口等形式提供生成式人工智能服务的组织或个人。

3.3

分类模型　classification model

对给定输入数据，输出其所属的一个或多个类别的机器学习模型。

[来源：GB/T 41867—2022，3.2.6]

3.4

训练数据　training data

所有直接作为模型训练输入的数据。

注：包括预训练数据和优化训练数据。

3.5

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”。