ICS 35.020
CCS L 80

DB21

辽　宁　省　地　方　标　准

DB21/T 4189—2025

生成式人工智能安全要求

Security requirements for generative artificial intelligence

2025-10-30 发布　　2025-11-30 实施

辽宁省市场监督管理局　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。

本文件由辽宁省工业和信息化厅提出并归口。

本文件起草单位：沈阳华睿博信息技术有限公司、国家计算机网络应急技术处理协调中心辽宁分中心、辽宁省烟草专卖局、东北大学、沈阳赛宝科技服务有限公司、辽宁省数据中心、沈阳航空航天大学。

本文件主要起草人：邵华、李凯、任志强、樊迪、王友民、李慧玲、王美琦、赵彬集、李雨、鲁凯、张晓晴、朱劭驰、王乐枭、唐文伟、石祥滨、刘芳、房琪。

本文件发布实施后，任何单位和个人如有问题和意见建议，均可以通过来电和来函等方式进行反馈，我们将及时答复并认真处理，根据实际情况依法进行评估及复审。

归口管理部门通讯地址：辽宁省沈阳市皇姑区北陵大街45-2号，联系电话：024-86913384。

标准起草单位通讯地址：辽宁省沈阳市和平区青年大街386号华阳国际大厦2396，联系电话：18698849086。

生成式人工智能安全要求

1 范围

本文件规定了生成式人工智能的安全基本要求、算力安全要求、数据安全要求、模型安全要求、服务安全要求、管理措施要求及安全评估要求。

本文件适用于开展生成式人工智能服务安全管理及评估等工作。

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 22239-2019 信息安全技术 网络安全等级保护基本要求

GB/T 25070-2019 信息安全技术 操作系统安全技术要求

GB/T 31168-2023 信息安全技术 云计算服务安全能力要求

GB/T 35273-2020 信息安全技术 个人信息安全规范

GB/T 35274-2023 数据安全技术 大数据服务安全能力要求

GB/T 42755-2023 人工智能 面向机器学习的数据标注规程

GB/T 43697-2024 数据安全技术 数据分类分级规则

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

GB 50174-2017 数据中心设计规范

3 术语和定义

下列术语和定义适用于本文件。

3.1

生成式人工智能 generative artificial intelligence

基于数据、算法、模型、规则，能够根据使用者提示，生成文本、图片、音频、视频等内容的人工智能。

3.2

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 基本要求

生成式人工智能安全应在满足GB/T 22239-2019中的要求的基础上，还应满足以下要求：

a) 保密性。采用技术和管理手段保证数据的保密性，并记录数据处理日志及确保日志的保密性，在数据存储和传输时，使用加密算法对人工智能数据进行数据存储和传输，如动态加密算法等，在检索增强生成场景下满足GB/T 35274-2023中的要求；

b) 完整性。优先在安全可控的网络环境（如局域网）中进行数据调取，并采用完整性校验的技术和管理手段，对人工智能数据的完整性进行监测；

c) 隐私性。严格保护个人信息和用户隐私，符合GB/T 35273-2020中5.5的规定；

d) 合规性。数据处理和模型应用应遵循法律法规及相关标准，通过模型对齐等手段确保输出内容合法合规，体现社会主义核心价值观，在意识形态、历史、民族、宗教等领域保持正确导向；
e) 分类分级。参照GB/T 43697-2024对所涉及数据进行分级分类，并设置相应安全防护等级；
f) 伦理安全。生成式人工智能数据全生命周期保障伦理安全，促进公平、公正、和谐，避免偏见、歧视、隐私和信息泄露等问题。

5 算力安全

5.1 基础设施安全

算力基础设施应满足以下要求：

a) 物理环境符合GB 50174-2017中第4章的规定，具备电力保障、防火、防雷、防静电、物理隔离等措施；
b) 云计算资源使用符合GB/T 31168-2023中第5章的规定；
c) 对部署在虚拟化平台的模型服务，采用强隔离策略，避免跨租户访问、信息泄露。

5.2 训练与运行环境安全

训练环境和推理部署环境的安全性直接影响算力任务的完整性与生成内容的可信度，应满足以下要求：

a) 系统安全加固：计算节点操作系统进行最小化安装、关闭无关服务，安装可信组件，采用GB/T 25070-2019中对应安全等级的配置要求；
b) 网络边界防护：训练平台、推理平台的网络边界部署入侵防御系统、防火墙、隔离网关等设备，并具备安全监测与溯源能力；
c) 训练软件合法合规：禁止使用未经许可的开源模型、非法下载的商业模型，优先使用可验证来源的软件包；
d) 安全隔离策略：在训练过程中采用资源隔离机制，防止不同训练任务之间的信息干扰；
e) 故障处理与回滚机制：针对可能出现的异常中断、数据损坏等故障情况，配置冗余计算节点、构建容灾系统，并建立训练状态回滚机制。

6 数据安全

训练数据作为算力执行的重要输入，应保证合法、完整、无毒、无害，严禁携带侵权、违法、敏感或虚假内容。训练数据安全在满足GB/T 45654-2025第4章中的要求的基础上，还应满足以下要求：

a) 数据来源合法合规，并获得必要的授权许可，符合GB/T 35274-2023中6.1的规定；
b) 涉及自然人信息的数据，进行脱敏处理或取得用户明示同意，符合GB/T 35273-2020中第5章的规定；
c) 建立训练数据输入日志机制，记录数据来源、处理过程、入库记录及调用操作，确保数据可审计、可回溯。

7 模型安全

7.1 基本要求

模型安全基本要求应符合GB/T 45654—2025中第5章的规定。

7.2 模型对齐

为确保生成式人工智能模型的输出与人类合法意图、社会主义核心价值观、法律法规及伦理准则一致，应满足以下要求：

a) 目标对齐设计：在模型架构设计阶段，明确对齐目标，如符合主流价值观、规避歧视性输出、遵守行业合规要求；
b) 训练过程对齐：使用包含合规性、伦理正向性标注的数据开展对齐训练，针对政治敏感、伦理道德等关键领域，构建专项对齐训练数据集，降低模型输出偏离目标的风险；
c) 对齐验证与迭代：将对齐效果纳入模型安全评估体系，采用对齐准确率和违规输出率等量化指标，定期或不定期验证模型对齐状态；对验证中发现的对齐偏差，通过模型微调、规则优化等方式迭代修正；
d) 人机协同对齐机制：建立模型自动过滤+人工审核复核的双层对齐机制，对高风险场景，强制触发人工审核流程，确保对齐结果的可靠性；
e) 对齐日志管理：记录模型对齐训练的数据来源、目标参数、验证结果及修正过程，日志满足保密性要求，确保对齐过程可审计、可回溯。

8 服务安全

8.1 用户信息保护

生成式人工智能的安全措施要求涵盖了用户信息保护，具体要求至少应包括：

a) 收集与使用。仅在提供生成式人工智能服务所必需的范围内收集用户信息，并明确告知用户信息的收集、使用目的和范围，不得非法留存能够推断出用户身份的输入信息和使用记录，避免用户隐私泄露；
b) 信息安全。采取加密技术和其他必要的安全措施，确保用户信息在存储和传输过程中的安全性，定期和不定期对存储的用户信息进行安全检查和备份，以防止数据丢失或损坏；
c) 用户权利。允许用户随时查询、更正、删除自己的信息。在收到用户关于用户信息处理的请求时，在规定时间内及时响应并处理。

8.2 服务透明度与可控性

生成式人工智能的安全措施要求涵盖了服务透明度与可控性，具体要求至少应包括：

a) 明确并公开服务的适用人群、场合、用途等信息，以便用户了解并合理使用服务；
b) 在交互界面或说明文档中公开服务的局限性、所使用的模型、算法等方面的概要信息；
c) 提供用户控制选项，允许用户根据需要调整服务的使用方式，如设置隐私权限、选择服务内容等；
d) 允许用户随时停止使用服务，并保障用户在停止服务后的个人信息和数据的安全。

8.3 投诉举报与应急响应

生成式人工智能的安全措施要求涵盖了投诉举报与应急响应，具体要求至少应包括：

a) 建立用户投诉接收处理机制，设置便捷的投诉、举报入口，公布处理流程和反馈时限，及时受理、处理公众投诉举报并反馈处理结果，确保用户权益得到保障；
b) 制定应急预案，对可能发生的安全事件进行预防和应对，在发生安全事件时，及时采取措施解决安全问题，并向相关主管部门报告；
c) 定期和不定期对服务进行安全评估和合规性检查，接受政府和相关主管部门的监管和指导，确保服务的安全性和合规性。

9 管理措施

9.1 组织机构安全

应建立与生成式人工智能安全服务相匹配的组织架构，并满足以下要求：

a) 明确领导责任。成立由主要负责人牵头的生成式人工智能安全工作领导小组，明确其在安全战略、投入保障、应急指挥等方面的职责；
b) 设立专职部门。设立专门的生成式人工智能安全管理部门，或明确现有相关部门的安全管理职责，负责具体安全措施的规划、实施、运营和监测；
c) 建立协同机制。建立安全协同与信息共享机制，确保安全要求贯穿业务全流程；
d) 授权与决策。明确关键安全事务（如重大数据泄露处置、模型重大风险下线、重要安全资源投入）的决策机构和审批流程。

9.2 岗位安全

应设置关键安全岗位，并明确其职责与权限，以满足以下要求：

a) 岗位设置。至少设置系统管理员、安全管理员、安全审计员等关键岗位；
b) 职责分离。遵循最小权限和职责分离原则，系统管理、安全管理和安全审计的权限相互独立、相互制约，避免出现权限过度集中的岗位；
c) 岗位说明。制定明确的岗位说明书，定义每个安全岗位的职责、权限、任职要求和工作要求。

9.3 制度安全

应制定覆盖生成式人工智能服务全生命周期的安全管理制度体系，并满足以下要求：

a) 制度体系。建立包括但不限于算力安全、数据安全、模型安全、服务安全、安全评估及人员管理等方面的制度体系；
b) 评审与更新。定期或在重大变更发生、重大安全事件发生后，定期和不定期对安全管理制度进行评审和更新，以确保其适用性和有效性；
c) 发布与执行。正式发布安全管理制度，确保相关员工易于获取、充分理解并严格遵循；
d) 数据标注伦理规范。依据GB/T 42755-2023的要求建立数据标注的伦理规范和质量核查制度，确保标注过程可控、结果可信，避免引入偏见和歧视。

9.4 人员安全

应对所有相关人员，特别是关键岗位人员，进行严格管理和安全教育，并满足以下要求：

a) 背景审查。在录用关键岗位人员前，对其进行必要的背景审查，了解其专业能力、信用状况和守法记录；
b) 安全协议。与所有接触敏感数据、核心算法和模型权限的人员签订保密协议，明确其安全责任和保密义务；
c) 安全教育与培训。制定年度安全培训计划，定期对全体员工进行生成式人工智能安全的培训与考核。关键岗位人员接受专项技能培训和伦理培训；
d) 意识提升。通过宣传、演练等方式，持续提升全体人员的网络安全意识和风险防范能力；
e) 离岗管理。建立人员离岗流程，及时终止离岗人员的所有访问权限，并收回其持有的所有敏感资料和设备。

10 安全评估

10.1 法律法规与价值观导向

10.1.1 生成式人工智能系统应符合 GB/T 35273-2020 中第 4 章的规定，保障系统安全和用户合法权益。

10.1.2 坚持社会主义核心价值观导向，应建立意识形态安全防护机制，防止生成内容偏离主流价值观和社会公序良俗。

10.1.3 在模型设计与训练过程中，应将政治敏感性、法律合规性和伦理道德纳入测试集，构建多层次风险识别及屏蔽机制，有效降低不合规内容生成风险。

10.2 安全评估体系

10.2.1 应建立涵盖算力安全、数据安全、模型安全、服务安全及管理措施等方面的系统化安全评估体系，定期和不定期评估模型是否符合安全评估体系的要求。

10.2.2 应采用合格率、拒答率、误拒率、伦理偏离率等量化指标进行评估，以确保评估结果科学、客观。

10.2.3 应配备关键词库、多模态测试题库及内容风险分类模型等技术支撑工具，提升评估的覆盖度和自动化水平。

10.2.4 对于具体的评估内容，应对所需的关键词库、生成内容测试题库、拒答测试题库和分类模型等安全评估基础设施提出具体清晰的建设规范，形成可操作的安全评估标准，测试题库需覆盖多模态生成内容(如图文混合问答)，并满足 GB/T 42755-2023 中的要求。

10.3 合规响应与治理机制

10.3.1 应对检测到的不合规生成内容应实施日志溯源管理，确保问题可追溯、可追责。

10.3.2 应采取模型纠偏、内容屏蔽、封禁等措施，及时阻断和修正违规内容传播。

10.3.3 应建立数据源头黑名单管理机制，防止不合规数据进入训练流程。

10.3.4 应定期和不定期检查及更新公共风险词库和风险识别规则，动态适应法规政策和社会环境变化，形成闭环风险防控体系。

参 考 文 献

[1] 《人工智能安全治理框架》，全国网络安全标准化技术委员会，2025