《大模型安全护栏能力技术规范》

编制说明

标准起草小组

1. 标准范围

本文件规定了大模型安全护栏的技术要求、功能要求及评价方法，包括安全护栏的总体架构、能力要求、安全防护机制以及测试评价指标。标准内容覆盖输入安全防护、推理过程安全防护、输出安全防护、安全干预机制、日志与审计要求、安全护栏评价方法，以及相关安全风险类型示例和测试样本示例。

本文件适用于基于大语言模型及多模态基础模型构建的智能系统安全防护能力建设，可用于指导大模型应用系统在设计、开发、部署及运行阶段的安全防护能力建设与评估。标准适用于政务服务、公共管理、金融、医疗、教育等领域的大模型应用系统安全防护能力建设与评价，也可为大模型安全产品研发、系统集成和第三方安全测评提供参考。

本标准的主要用途包括:

（1）产品设计与开发：为大模型安全护栏产品、智能体安全防护组件及相关平台提供功能规划、能力建设和架构设计依据;

（2）产品选型与评估：为用户单位建设或采购具备大模型安全护栏能力的产品时，提供技术选型和能力评估依据;

（3）符合性测试与认证：为第三方测试机构对相关系统进行安全护栏能力测试、风险验证和符合性评价提供标准化依据。

2. 工作简况

随着大模型、智能体、检索增强生成、工具调用和多模态应用快速发展，大模型在政务、公共服务、企业运营和行业应用中的使用场景不断扩展。与此同时，提示词攻击、敏感信息泄露、违规内容生成、异常行为、多轮诱导攻击、外部知识污染、插件异常返回、工具调用异常、模型更新后安全能力下降以及运行环境安全缺陷等问题日益突出，迫切需要形成统一、可操作、可评估的大模型安全护栏技术规范。

在此背景下，标准起草组组织开展了大模型安全防护技术研究、应用场景分析、标准框架设计和文本起草工作，形成了标准草案。起草过程中，重点围绕安全护栏总体架构、功能组成、部署模式、输入/推理/输出全流程防护、安全干预机制、日志与审计要求、评价方法以及附录风险类型和测试样本等内容进行了系统梳理。

在专家征求意见阶段，重点吸收了以下意见：一是删除正文未实际使用的缩略语；二是补充模型训练阶段、系统运行阶段相关安全风险内容；三是增加安全护栏功能框架图；四是在部署模式中补充安全护栏与 AI 系统连接示意图；五是补充模型更新、版本留痕、异常调用监测等内容，并增强附录风险和测试样本对正文的支撑。经修改后，标准已形成较为完整的 V2.0 初稿。

3. 标准编制原则和确定标准主要内容

3.1标准编制原则

本文件依据 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》的规定编制。标准编制过程中，主要遵循以下原则：

一是科学性原则。围绕大模型安全护栏的关键能力，结合大模型、智能体、RAG、工具调用、多模态交互等实际应用场景，系统梳理安全防护需求。

二是实用性原则。突出输入、推理、输出、干预、日志审计和评价测试等可落地内容，增强标准对产品设计、系统建设和第三方测评的指导作用。

三是完整性原则。除交互环节外，兼顾训练数据、模型更新和运行环境相关风险，形成从风险识别、能力建设到测试评价的相对完整闭环。

四是可评估性原则。通过附录风险类型示例和测试样本示例，对标准中的能力要求提供支撑，便于开展验证、测试和评价。

3.2标准主要内容

本标准主要内容如下：

（1）基础定义与适用范围

标准明确了适用范围、规范性引用文件、术语和定义以及缩略语，为全文提供统一的概念基础。当前版本对缩略语

进行了收敛，重点保留与正文内容密切相关的术语。

（2）总体架构与功能组成

标准提出了大模型安全护栏总体架构，明确了功能框架示意图，并将功能组成划分为输入安全检测、推理过程安全监测、输出安全检测、安全干预、安全日志与审计、安全策略管理与版本管理、人工复核与反馈优化等模块，形成较完整的能力体系。

（3）部署模式与连接关系

标准对模型前置部署、模型内嵌部署、模型后置部署、全流程部署和外围协同部署等模式进行了规定，并增加了“安全护栏与 AI 系统连接示意图（输入输出视角）”，明确安全护栏在输入侧、模型服务侧、输出侧及全流程链路中的接入位置和作用目标。

（4）输入安全防护要求

标准规定了输入安全检测、提示词攻击防护、敏感信息检测、恶意内容识别、输入安全策略管理及适用边界与处置要求，并明确输入内容除用户直接输入外，还包括检索增强生成场景中的外部知识片段、工具调用返回结果、插件回传内容及其他进入模型上下文的外部信息。

（5）推理过程安全防护要求

标准规定了推理安全监测机制、安全探针机制、异常行为识别、推理风险拦截、推理安全策略管理及适用边界与处

置要求，强调对工具调用链、插件服务、外部接口、知识库访问以及相关运行链路的异常监测。

（6）输出安全防护要求

标准规定了输出安全检测机制、违规内容识别、敏感信息过滤、输出风险控制、输出安全策略管理及适用边界与处置要求，用于规范模型生成内容的检测、过滤、替换和风险控制。

（7）安全干预机制

标准规定了风险识别触发机制、自动化安全干预、输出替换机制、人工审核机制、安全告警机制和人工复核与回退要求，支撑风险拦截、安全替代输出、人工复核和异常处置闭环。

（8）日志与审计要求

标准规定了安全日志记录、审计追溯能力、安全事件记录、日志安全管理以及日志留痕与版本追溯要求。当前版本进一步补充了训练数据版本、标注规则版本、模型版本、模型发布时间、发布方式、回滚时间以及模型更新、版本切换、异常调用等关键日志记录要求。

（9）安全护栏评价方法

标准建立了评价原则、评价对象、评价指标、判定规则以及输入安全测试、推理安全测试、输出安全测试、安全干预测试等方法，并补充了训练数据治理、模型更新管理、运

行环境安全和日志审计相关的配套能力评价要求。

（10）附录风险与测试样本

附录 A 给出了大模型安全风险类型示例，当前版本已扩展至训练数据安全风险、模型训练与微调过程安全风险、系统运行环境安全风险等内容；附录 B 给出了大模型安全测试样本示例，为提示词攻击、敏感信息、多轮对话攻击、对抗样本以及工具调用链异常、外部知识污染、版本切换/策略回滚一致性等测试提供支撑。

4. 主要试验（或验证）的分析、综述报告

本标准当前阶段主要基于大模型安全防护场景分析、能力结构设计、风险类型梳理和测试样本构造开展编制，形成了可用于输入安全、推理安全、输出安全、安全干预以及日志与审计能力验证的测试方法框架。标准附录 B 已给出典型测试样本示例，可作为后续开展符合性验证和能力测评的基础。

目前，标准文本中尚未单列形成独立的试验综述报告。后续在标准实施或验证阶段，可进一步结合典型应用场景、测试环境和样本库建设情况，补充形成相关验证材料。

5. 标准在起草过程中遇到的问题及解决办法；重大分歧意见的处理经过和依据；有无重要技术问题需要说明

本标准在起草过程中，主要面临以下问题：一是安全护栏概念在不同产品和系统中的边界不完全一致；二是标准原

始版本更偏重输入/输出交互防护，对训练数据、模型更新和运行环境相关内容覆盖不足；三是部署模式和安全能力关系在文本表述上不够直观。对此，起草组通过补充功能框架图、补充安全护栏与 AI 系统连接示意图、扩展附录 A 和附录 B、完善日志追溯和评价测试等方式进行了修订。

在征求意见与修订过程中，未出现重大原则性分歧。专家意见主要集中在缩略语精简、训练及运行阶段风险补充、图示表达增强、测试样本完善等方面，相关意见已在当前版本中吸收。当前标准不存在尚未解决的重大技术问题。

6. 与国外标准的关系：包括采用国际标准和国外先进标准的程度，与国外标准主要技术内容的差异

目前本项目未直接采用国际标准或国外先进标准文本。标准主要依据国内大模型应用安全防护实际需求，围绕大模型安全护栏能力建设、部署模式、测试评价及风险类型构建技术内容，体现了面向国内政务、公共管理及行业场景的大模型安全防护需求特点。

7. 修订标准时，说明与标准前一版本的重大技术变化，并列出所涉及的新、旧版本的有关章条；废止/代替现行有关标准的建议

不涉及

8. 说明标准与其他标准或文件的关系，特别是与有关现行法律、法规和强制性国家标准的关系

本标准与 GB/T 22239—2019《信息安全技术 网络安全等级保护基本要求》、GB/T 25069—2022《信息安全技术 术语》、GB/T 28448—2019《信息安全技术 网络安全等级保护测评要求》、GB/T 35273—2020《信息安全技术 个人信息安全规范》、GB/T 41867—2022《信息技术 人工智能 术语》等文件相衔接，为大模型安全护栏能力建设与评价提供补充性、专门化的技术依据。

本标准内容符合现行法律、法规和国家有关网络安全、数据安全、个人信息保护等要求，不与现行强制性国家标准相冲突。

9. 标准作为强制性标准或推荐性标准的建议

建议本文件作为推荐性团体标准发布实施。

10.贯彻标准的要求和措施建议；标准发布后，对国内外业界可能产生的影响

建议本文件作为推荐性团体标准发布实施。实施过程中，可结合典型场景开展标准宣贯、能力评估、测试样本建设和第三方测评应用，推动大模型安全护栏相关产品和系统的规范化建设。

同时，建议围绕政务服务、公共管理、金融、医疗、教育等重点场景，逐步形成与本标准配套的测试样本库、评价

方法细则和实施指南，提高标准落地性和可操作性。

标准实施后，有助于统一大模型安全护栏能力建设的技术基线和评价口径，提升大模型在重点行业应用中的安全防护水平、部署一致性和测评规范性，也可为后续标准扩展、行业落地及生态协同提供基础支撑。

11.标准是否涉及知识产权的情况说明；如标准中含有自主知识产权，说明产品研发程度、产业化基础及进程

目前未发现本文件明确涉及必须披露的知识产权内容。如后续在标准实施、测试工具、样本库或安全策略机制中涉及自主知识产权成果，可另行补充说明。

12.其他应予说明的事项

无。