《大模型服务与应用安全评测技术规范》编制说明

标准起草小组

1. 标准范围

本文件规定了大模型服务与应用安全评测的评测对象、评测环境与测试准备、安全风险类型、安全检测方法、评价指标、评估结果判定规则以及测试报告要求。

本文件适用于对大语言模型服务平台、大模型应用系统、智能体系统以及集成大模型能力的应用程序接口（API）服务开展安全评测。本文件重点适用于文本交互、代码生成、多轮对话、检索增强生成以及工具调用等场景的安全评测；涉及图文等多模态场景时，可参照本文件相关要求执行；音频、视频等其他模态场景可参照执行。本文件可用于指导大模型服务与应用安全能力评估、第三方安全测评以及相关服务与应用的安全能力建设。

本标准的主要用途包括：

（1）评测实施依据：为第三方机构对大语言模型服务平台、大模型应用系统、智能体系统及相关能力组件开展安全评测提供统一依据；

（2）系统建设与优化依据：为模型服务提供方、平台建设方、行业应用单位在系统设计、开发、部署、运行和优化过程中提供安全评测参考；

（3）报告与结论形成依据：为测试结果统计、评价指标计算、综合判定、测试报告编制和结果比对提供标准化依据；

（4）行业应用支撑依据：为政务服务、公共管理、金融、医疗、教育等重点行业大模型服务与应用系统的安全测试与评估提供统一参考。

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，大模型服务平台、大模型应用系统和智能体系统在政务、公共服务、企业运营和行业场景中的应用不断扩展。与此同时，提示词攻击、敏感信息泄露、有害内容生成、工具调用与扩展能力风险、多轮交互与行为偏移风险、多模态风险，以及知识增强、外部知识污染、事实性失真、模型幻觉等相关问题日益突出，迫切需要形成统一、可执行、可复核、可对比的大模型服务与应用安全评测技术规范。

在此背景下，标准起草组围绕大模型服务与应用安全评测需求，组织开展了评测对象分析、应用场景梳理、风险分类研究、测试方法设计、评价指标体系设计、判定规则设计和测试报告框架设计等工作，形成了标准草案。起草过程中，重点围绕评测对象与评测范围、评测环境与测试准备、安全风险类型、安全检测方法、评价指标、判定规则、测试报告以及附录中的样本分类与构建方法等内容进行了系统梳理。

在前期专家意见吸收和版本迭代过程中，标准已对名称和适用范围进行了收敛，由“大模型安全评测”调整为“大模型服务与应用安全评测”，进一步明确了本文件主要面向

大语言模型服务平台、大模型应用系统、智能体系统及集成大模型能力的应用程序接口服务；同时，对术语定义、风险分类体系、附录 C 样本构建方法、附录 D 测试判定说明和附录 E 风险类型与评价指标映射示例等内容进行了完善，使正文与附录支撑关系更加清晰。

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。围绕大模型服务平台、大模型应用系统、智能体系统及相关能力组件的实际运行过程，系统梳理安全风险、测试方法和评价指标。

二是实用性原则。突出评测对象、评测场景、测试样本、检测方法、判定规则和测试报告等可落地内容，增强标准对测试实施和第三方测评的指导作用。

三是完整性原则。覆盖输入环节、生成过程、输出结果、交互过程、扩展能力等主要安全维度，并兼顾多轮交互、RAG、工具调用、多模态、外部知识污染等复杂应用场景。

四是可评估性原则。通过风险分类、测试方法、评价指标、判定规则和附录中的样本示例、样本构建方法、报告模板及指标映射示例，为开展具体评测和结果分析提供支撑。

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，为全文提供统一概念基础。当前版本中，术语主要包括“大模型”“大模型系统”“大模型服务与应用安全风险”“提示词攻击”“越狱攻击”“安全评测”等内容。

（2）评测对象与评测范围

标准规定了评测对象、评测范围、评测场景、输入与输出形式、评测边界和评测原则。评测对象覆盖大语言模型服务、大模型应用系统、智能体系统、大模型平台服务和大模型能力组件；评测范围覆盖输入环节安全、生成过程安全、输出结果安全、交互过程安全和扩展能力安全等维度。

（3）评测环境与测试准备

标准规定了评测环境要求、测试接口要求、测试样本准备、测试策略配置和测试过程记录要求，强调交互可用性、过程可记录性、数据可管理性和执行可重复性，为测试实施提供规范基础。

（4）安全风险类型

标准建立了统一的风险分类体系，重点包括提示词攻击风险、敏感信息泄露风险、有害内容生成风险、工具调用与扩展能力风险、多轮交互与行为偏移风险、多模态风险等六

类安全风险。同时，当前版本进一步明确，风险分类宜结合知识增强、外部知识污染、事实性失真、模型幻觉、系统性安全缺陷等复杂应用场景进行扩展，并在“工具调用与扩展能力风险”中补充了“外部知识污染风险”。

（5）安全检测方法

标准规定了组合测试方法、单轮测试方法、多轮测试方法、自动化测试方法、多模态测试方法和工具调用测试方法等六类主要测试方法，并明确测试方法设计宜兼顾风险表现类型和风险来源维度，增强不同测试方法之间的共通性和可扩展性。

（6）评价指标与判定规则

标准建立了风险识别率、风险拦截成功率、误报率、漏报率、响应时间、风险覆盖率等核心评价指标，并明确不同风险类型宜至少对应一种主评价指标和一种辅助评价指标。当前版本还补充了多模态语义偏差、跨模态误解等风险的指标适用说明。标准同时规定了综合评测通过条件、单项测试判定、风险严重性分级、综合判定方法、分级判定和否决项规则等内容。

（7）测试报告要求

标准规定了测试报告的总体要求、报告基本信息、评测范围与方法说明、测试结果、评价指标结果、风险分析、综合评价结论、改进建议、典型案例说明及报告管理与使用要

求。当前版本的附录 B 测试报告模板已补齐风险拦截成功率、最大响应时间、95 分位响应时间和风险覆盖率等指标项。

（8）附录内容

附录 A 给出了安全测试样本示例；附录 B 给出了安全测试报告模板；附录 C 给出了样本分类与构建方法，当前已新增“知识增强与外部知识污染类样本”及“C.4.7 知识增强与外部知识污染样本构建”；附录 D 给出了多模态与智能体安全测试方法说明；附录 E 给出了评价指标计算方法、阈值建议及风险类型与评价指标映射示例。附录与正文形成了较为完整的支撑关系。

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服务与应用安全风险分析、评测对象识别、测试方法设计、评价指标体系构建以及测试报告框架设计开展编制，形成了可用于服务平台、应用系统、智能体系统和能力组件安全评测的技术框架。附录 A、附录 B、附录 C、附录 D 和附录 E 分别给出了测试样本示例、测试报告模板、样本构建方法、多模态与智能体测试说明以及指标映射和阈值建议，可作为后续开展符合性验证和第三方安全测评的基础。

目前，标准文本中尚未单列形成独立的试验综述报告。后续在标准试点实施、典型场景测试或样本库建设过程中，可进一步结合测试环境、测试对象和测试结果，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：一是原始名称和适用范围较大，容易超过实际适用边界；二是风险类型、测试方法、评价指标与测试报告模板之间的一致性需要持续强化；三是多模态、智能体、RAG、工具调用和知识增强等复杂场景下的风险分类和评价指标对应关系需要进一步细化。

针对上述问题，起草组通过调整标准名称、收敛适用范围、统一第 5 章至第 11 章的口径表述、补充外部知识污染相关风险和样本、完善多模态与智能体测试判定说明、在附录 E 中增加风险类型与评价指标映射示例等方式进行了修订。当前版本已在结构完整性、正文与附录一致性以及风险—方法—指标—报告的闭环支撑方面形成较完整框架。

在征求意见与修订过程中，专家意见主要集中在术语精简、名称收敛、风险分类完备性、测试方法可扩展性、风险类型与指标映射关系等方面。对相关意见，起草组已结合标准定位和实际内容进行了吸收处理，未出现重大原则性分歧。当前标准不存在尚未解决的重大原则性技术问题。

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。标准主要立足国内大模型服务与应用安全评测需求，围绕服务平台、应用系统、智能体系统及相关能力组件的风险类型、测试方法、评价指标、判定规则和测试报告构建技术内容，体现了面向国内政务服务、公共管理及重点行业应用场景的大模型安全测试与评估需求特点。

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，为大模型服务与应用安全评测提供补充性、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、测试样本建设、能力评估、第三方安全测评和试点评估应用，推动大模型服务与应用安全评测工作的规范化开展。

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，有助于统一大模型服务与应用安全评测的技术口径和评价口径，提升重点行业大模型服务与应用系统的测试规范性、结果可比性和风险识别能力，也可为后续相关标准扩展、行业落地和第三方测评生态建设提供基础支撑。

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、测试平台、样本库或评测工具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。