TC260

全国网络安全标准化技术委员会技术文件

TC260-004

政务大模型应用安全规范

Specifications of government large-scale model application security

2025-09-11 发布

全国网络安全标准化技术委员会发布

目 次

前　言

本文件由全国网络安全标准化技术委员会（SAC/TC260）发布。

本文件起草单位：国家工业信息安全发展研究中心、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、工业和信息化部电子第五研究所、国家信息技术安全研究中心、阿里云计算有限公司、北京百度网讯科技有限公司、永信至诚科技集团股份有限公司、公安部第三研究所、公安部第一研究所、中国信息安全测评中心、深信服科技股份有限公司、北京快手科技有限公司、中国网络安全审查认证和市场监管大数据中心、奇安信科技集团股份有限公司、北京火山引擎科技有限公司、北京天融信网络安全技术有限公司、中国交通通信信息中心、北京奇虎科技有限公司、中国联合网络通信有限公司、启明星辰信息技术集团股份有限公司、中国科学技术大学、中央网信办数据与技术保障中心、上海观安信息技术股份有限公司、北京面壁智能科技有限公司、北京智谱华章科技股份有限公司、天翼云科技有限公司、天翼安全科技有限公司、国家信息中心、国家药品监督管理局信息中心、浪潮云信息技术股份公司、北京国家金融科技认证中心有限公司、北京市大数据中心、四川省大数据中心、浙江算力科技有限公司、无锡市数字新基建有限公司、中移（苏州）软件技术有限公司、中电信数政科技有限公司、中国软件评测中心（工业和信息化部软件与集成电路促进中心）、华为云计算技术有限公司、中移动信息技术有限公司、中国联合网络通信有限公司海南省分公司、中移互联网有限公司、芯质动力（北京）科技有限公司、江苏省软件产品检测中心、北京伽睿智能科技集团有限公司。

本文件主要起草人：张格、赵冉、郝春亮、贺敏、卢春景、吴波、张路宁、于盟、张妍婷、李滨、毛洪亮、黄天宁、李志伟、张丽、陆臻、陈妍、刘卜瑜、梁亚楠、王勇、陈佳哲、田斌、李寒雨、栗振群、朱隽、杜渐、落红卫、姜伟生、廖双晓、安高峰、邹权臣、郑涛、伍扬、陈晓志、张卫明、戴明、陈朴、孙雁伟、张照龙、汪华东、乔文斌、朱良海、方宇、李晔、王博、张震、孔德智、陈嘉旺、刘曦泽、靳宇浩、孙培尧、彭骏涛、付磊、叶润国、赵丹丹、刘岩、郭建领、高彦恺、张德岳、徐雷、商跃鹏、俞能海、申东洋、阮子禅、顾佳、梁伟、刘昭、陶冶、陈洪波、刘浩、张哲宇、赵佳璐、李格菲、刘洋、由玉伟、吴欣然、陈可江、左鹏、饶飞、张海燕、李振、段力烟、孙伟利、王琦、周学立、杨鹏飞、张静、张聪、孙荣锋、周晓飞、王浩硕、石春磊、蔡旭东、徐光、曹顺超、李静、李晗、滕滨、李雪、黄艳、庄仁峰、庄严、卢亮、潘梦婷、王忠明。

政务大模型应用安全规范

1　范围

本文件规定了政务大模型应用的安全要求和测试方法，包括大模型选用、大模型应用部署、大模型应用运行、大模型应用停用等。

本文件适用于政务大模型应用的设计、开发与测试活动，可为政务大模型应用安全提供指导。

注：涉密信息系统应用大模型的，应遵守国家保密法律法规及相关标准要求。

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 36637—2018　信息安全技术　ICT供应链安全风险管理指南

GB/T 20984—2022　信息安全技术　信息安全风险评估方法

GB/T 25069—2022　信息安全技术　术语

GB/T 41867—2022　信息技术　人工智能　术语

GB/T 43698—2024　网络安全技术　软件供应链安全要求

GB/T 45288.1—2025　人工智能　大模型　第1部分：通用要求

GB/T 45288.2—2025　人工智能　大模型　第2部分：评测指标与方法

GB/T 45396—2025　数据安全技术　政务数据处理安全要求

GB 45438—2025　网络安全技术　人工智能生成合成内容标识方法

GB/T 45574—2025　数据安全技术　敏感个人信息处理安全要求

GB/T 45577—2025　数据安全技术　数据安全风险评估方法

GB/T 45652—2025　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

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/T 45958—2025　网络安全技术　人工智能计算平台安全框架

3　术语和定义

GB/T 25069—2022、GB/T 41867—2022、GB/T 45288.1—2025界定的以及下列术语和定义适用于本文件。

3.1

大模型　large-scale model

大规模深度学习模型　large-scale deep learning model

基于大量数据训练得到，具有复杂计算架构，能处理复杂任务，且具备一定泛化性的深度学习模型。

注：大模型的参数量由其功能和模态决定，一般不低于1亿。大模型训练使用的数据总量受参数量的影

响，达到收敛的大模型的参数量的对数与其训练数据总量的对数成正比。

[来源：GB/T 45288.1—2025，3.1]

3.2

大模型应用 large-scale model application

以大模型为技术手段，提供涉及文本、图像、语音、视频等的内容生成、图像识别、语音处理、信息处理、决策支持等功能，以支持业务开展的工具、系统或服务。

3.3

模型幻觉 model hallucination

大模型在生成内容时，输出看似合理但实际与事实不符、与用户输入逻辑不一致或虚构的信息，如虚构法律条款、政策文件、统计数据，答非所问等。

3.4

大模型安全护栏 large-scale model security guardrails

用于约束和规范大模型应用行为的一系列策略、机制和技术手段，限制大模型输入输出内容或行为，防止重要数据泄露、提示词注入攻击，以及生成侵犯商业秘密或个人信息、违法和不良信息等。

4 缩略语

下列缩略语适用于本文件：

API：应用编程接口（Application Programming Interface）

RAG：检索增强生成（Retrieval Augmented Generation）

5 概述

政务大模型应用常见场景包括：

——办公业务支撑：基于大模型部署的办公助手应用，或将大模型嵌入政务办公系统，面向本单位人员提供资料检索、文案生成、文稿校对、方案设计、数据分析、创意生成等服务；

——公众政务服务：将大模型用于政务服务热线、数字政务服务，面向公众提供智能化的政策问答、信息检索、业务导办等服务。

常见安全风险包括：

——数据安全风险：超范围接入使用大量政务数据，用户使用中上传内部工作资料、敏感个人信息，同时数据访问权限管控不严，引发政务数据和政务信息泄露风险；

——系统安全风险：在建设、部署、运维过程中，未有效落实政务信息系统网络安全防护要求，使用的大模型硬件平台、基础软件存在供应链安全风险，政务大模型应用成为网络攻击新入口，导致系统安全风险敞口扩大；

——内容安全风险：选用不可靠甚至违规大模型，数据集未有效筛选过滤，缺乏有效的输入输出管控措施，以及模型幻觉，引发生成传播违法和不良信息、误导用户、模型应用被恶意利用等安全风险。

建设政务大模型应用前，应参照GB/T 45958—2025中第7章的有关要求，明确平台提供方、数据提供方、模型提供方、应用提供方的安全职责，并按照GB/T 20984—2022和GB/T 45577—2025有关要求，结合办公业务支撑和公众政务服务两类场景，对可能带来的数据安全风险、系统安全风险和内容安全风险进行分析，在落实现有相关安全要求和安全测试基础上，采取针对性的安全防护措施。

6 安全要求

6.1 大模型选用

大模型选用的要求如下。

a） 采购商业大模型的，应对大模型的备案情况进行核实，不应使用未经备案的大模型（含转接服务）；

注：大模型的备案情况详见《国家互联网信息办公室关于发布生成式人工智能服务已备案信息的公告》。

b） 选用开源大模型自行部署的，应核查其具备的许可证，从大模型研发机构官网或者其在主流开源社区的官方账号等权威渠道，获取模型参数及配套组件，并对其进行完整性校验和安全测试；

c） 宜选用支持检索增强生成（RAG）技术的模型，采用外挂知识库等方式保障生成内容的准确性、时效性、可控性。

6.2 大模型应用部署

大模型应用部署的要求如下。

a） 应按照政务信息系统建设要求，集约化部署大模型、政务大模型应用，实施集中统一的安全管理和体系化技术防护措施；

b） 应对大模型部署所需的软硬件设备、第三方工具等进行安全测试，应能发现可能存在的已知漏洞，并在经过充分测试评估后，及时修补漏洞；

c） 应按照 GB/T 36637—2018 和 GB/T 43698—2024 等有关要求，对大模型运行依赖的芯片、基础软件、开源组件、算力设施等方面的供应链安全进行评估，对可能存在的供应中断、恶意程序或后门、高危漏洞等安全风险进行分析；

d） 调用互联网大模型（如 API）时，应启用对服务商 API 鉴别机制，核查服务商服务数字证书的有效性，甄别防范虚假 API 接口、仿冒和套壳大模型，确保服务来源安全可靠；

e） 在基础设施层面，应准确安装配置软件、运行环境参数、功能模块调用策略，禁用非必要的网络端口和功能服务，实施严格的网络隔离策略，授予业务所需要的最小网络通信权限，修改默认配置、默认口令，及时修复安全风险；

f） 在应用管理层面，应对大模型自身提供的人机交互接口和所调用的 API 接口进行用户身份识别及权限控制，支持设置一般用户、系统管理用户和安全审计用户等多角色用户，最小化设置访问权限，一般用户禁用大模型管理功能的创建、删除、拉取、推送等高权限操作；应根据业务场景限制接口调用频率，支持对恶意行为用户暂停服务、阻断访问；涉及用户身份实名认证要求的，应依托国家网络身份认证公共服务开展用户身份识别；

g） 外挂知识库接入的数据，应按照 GB/T 45396—2025 中 6.3 数据使用的有关要求采取安全防护措施，遵循应用场景必要原则，由本单位负责政务大模型应用管理、数据安全的部门，会同提供数据的业务部门开展必要性评估，不宜将人事、财务等敏感业务数据接入大模型向其他部门提供服务；应确保外挂知识库接入的数据来源可靠、内容准确有效，建立台账并详细记录数据来源、类型和规模等信息，确保数据可追溯，并设置数据版本管理和回滚机制；

h） 大模型部署期间涉及对大模型进行微调优化的，应按照 GB/T 45652—2025 的有关要求，对用于预训练和优化训练的数据以及外挂知识库接入的数据进行清洗过滤，

涉及个人信息、内部工作信息等敏感内容的，宜采取去标识化处理、脱敏处理等措施，数据内容中不应含有违法和不良信息、错误数据；

i） 公众政务服务类应用，应确保存储、处理内容不超出政务信息公开范围；对存在时效性、适用范围等要素的内容，应定期测试评估完整性、有效性和适用性，防止错误、过期、不适用内容接入大模型进而误导用户；

j） 应采用大模型安全护栏等技术措施，识别拦截政务大模型应用输入输出中的重要数据泄露，提示词注入攻击，违法和不良信息等，审核并管控输出内容不超出业务范围，对不当或超范围提问采取拒答、固定答复等稳妥回应。大模型安全护栏功能要求可参考附录 A。

6.3 大模型应用运行

大模型应用运行的要求如下。

a） 公众政务服务类应用，应按照 GB 45438—2025 的要求，对政务大模型应用生成、合成内容进行标识；

b） 政务大模型应用涉及政务信息公开、政策公告、新闻发布、灾害风险预警等权威信息发布的，应建立健全内部审核制度流程并严格执行，对生成内容进行人工审核后方可发布；

c） 应在政务大模型应用界面的显著位置设置风险提示，明确告知用户政务大模型应用的局限性；

d） 公众政务服务类应用，不应向公众提供推理过程显示功能，防止推理过程泄露不当信息。审慎设置涉及输出内容准确性和一致性的模型参数，优先保障生成内容准确性；

e） 公众政务服务类应用，应提供便捷的人工服务方式切换入口，并保留人工接管接口，当输出准确率下降到设定的阈值时（如设定准确率阈值不得低于 95%，准确率可参照 GB 45288.2—2025 附录 A 中 A.1 进行计算），支持快速切换至人工服务，提供在线客服、电话、即时通信、邮件等问题反映渠道，便于用户及时报告服务异常、输出违法和不良信息等情况；

f） 应记录政务大模型应用运行日志，包括系统行为、用户行为等，留存时长不少于 1 年，并定期对日志记录进行审计；

g） 政务大模型应用上线前和运行期间，应定期评估相关硬件平台、基础软件、外挂知识库接入数据等的安全性，跟踪大模型应用相关软硬件安全漏洞、缺陷信息，核查其在资源消耗攻击、提示词注入攻击等攻击下政务大模型应用的安全性，对发现的问题隐患进行整改加固；

h） 应对政务大模型应用进行持续监测，定期对网络安全攻击和大模型安全护栏告警等情况进行分析；应制定针对政务大模型应用安全事件的应急预案，采取一键关停等技术措施，对紧急突发事件进行处置；

i） 应制定政务大模型应用更新升级时的安全管理策略，确保大模型更新升级的完整性和可用性；

j） 定期开展大模型应用安全教育培训，培训内容应至少包括：访问使用安全可靠大模型应用，不应将内部工作资料、敏感个人信息输入市场化大模型；充分认识大模型应用生成内容的局限性，审慎将大模型生成内容直接应用于业务工作，对法律条文、政策标准、统计数据、人物事件等事实性内容进行核实。

6.4 大模型应用停用

大模型应用停用的要求如下。

a) 应制定停用管理流程，对政务大模型应用的停用、数据清除等流程进行监控和审计；

b) 应对停用的大模型文件、外挂知识库接入的数据、预训练和优化训练的数据、系统日志等进行妥善处理，采取删除措施进行处理的，应确保相关文件和数据不可被恢复。

7 测试方法

针对政务大模型应用的测试方法，可参考附录 B 开展安全测试验证。

附 录 A
（资料性）
大模型安全护栏功能要求

大模型安全护栏主要用于限制大模型应用输入输出内容或行为，能够防止重要数据泄露、提示词注入攻击，以及生成侵犯商业秘密或个人信息、违法和不良信息等。大模型安全护栏通常独立于大模型应用进行部署，常见部署模式包括智能体编排、网关代理、直联串接等。大模型安全护栏应用示意图如图1所示。本附录给出了大模型安全护栏可具备的有关功能要求。

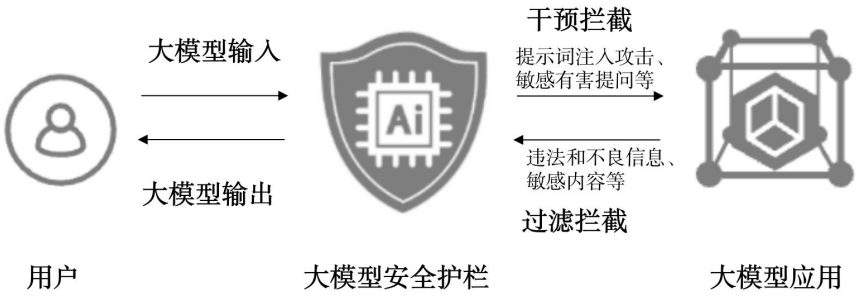

图 1 大模型安全护栏应用示意图

大模型安全护栏的功能要求如下。

a) 支持识别提示词注入攻击、越狱攻击、资源消耗攻击等对抗攻击指令并拦截，对抗性攻击指令样本库宜覆盖典型的攻击模式并可持续更新；

b) 具备与大模型应用实际所支持模态相匹配的输入输出内容安全识别能力，具体包括文本识别、图像识别、音频识别、视频识别、文件识别等；

c) 具备大模型输入风险识别管控能力，干预拦截攻击行为、敏感有害提问等，包括：
   1) 支持语义级分析能力，可自动识别分类违法和不良信息，包括多模态隐晦违规内容识别拦截，并提供自定义关键词过滤规则等定制化安全功能；
   2) 支持上下文关联分析，可对超长会话历史进行连贯性分析，可基于用户角色识别拦截越权提问信息；
   3) 支持自动识别拦截个人信息等敏感内容。

d) 具备大模型输出风险识别管控能力，过滤拦截输出内容中的违法和不良信息、敏感内容等，包括：
   1) 配置脱敏规则，对大模型生成的敏感内容进行脱敏后输出；
   2) 过滤违法和不良信息，对大模型生成的不当或超业务范围内容，采取限制输出或代答、拒答等方式进行输出；
   3) 代答、拒答技术措施（如代答知识库和拒答答案库、代答模型和拒答模型）可将识别的风险提问与标准回复进行映射，对可预判问题提供标准答案，对用户进行正向引导；
   4) 代答、拒答技术措施配置支持自定义扩展，可调整风险提问与回复的关联关系；
   5) 代答、拒答技术措施可按照实际需要及时更新。

e) 具备日志留存和审计能力，支持记录行为主体、事件类型、事件时间以及系统行为、用户行为等，支持基于时间范围、请求用户等多维度查询和统计分析，定期对日志记录进行审计。

附 录 B
（资料性）
政务大模型应用安全测试指南

政务大模型应用带来网络运行环境、数据处理模式、服务交互方式等变化，引发新的安全风险。因此，在落实现有数据安全、系统安全、内容安全相关防护要求基础上，宜采取针对性的安全防护措施，并在上线前和运行期间针对大模型安全护栏等安全防护措施有效性开展安全测试。本附录给出了指导政务大模型应用安全测试的参考。

B.1 大模型选用测试指南

大模型选用的测试指南如下。

a) 核查政务大模型应用所需的模型文件、框架、部署工具、第三方库等软件；
   1) 核查政务大模型应用技术方案、模型配置等信息，确认选取的模型是否属于已备案的人工智能服务；
   2) 确认相关软件来源渠道，并与厂商官方网站或其在主流开源社区对应版本的校验码进行比对，核查下载软件包是否一致。
b) 使用开源大模型或开源组件的，核查其具备的许可证，确认版权、授权范围等信息，对开源组件进行软件成分分析和代码审计，识别是否存在安全风险；
c) 核查政务大模型应用技术方案，确认选取的大模型技术是否对生成内容的准确性、时效性、可控性进行评估。

B.2 大模型应用部署测试指南

大模型应用部署的测试指南如下。

a) 核查政务大模型应用建设方案，是否采用集中统一的安全管理和体系化技术防护措施；
b) 利用两种及以上（不同厂商）的漏扫工具，对部署大模型的软硬件设备、第三方工具等进行安全测试，核查扫描结果中是否存在未修复且可被利用的已知漏洞；
c) 核查大模型使用的芯片、基础软件、开源组件、算力实施等方面的供应链安全风险分析报告，是否按照 GB/T 36637—2018 和 GB/T 43698—2024 等有关要求，对可能存在的供应中断、恶意程序或后门、高危漏洞等安全风险进行分析，核查现有安全风险是否处于可接受水平；
d) 调用互联网大模型的，核查服务商服务数字证书的有效性，验证是否存在虚假 API 接口、仿冒和套壳大模型；
e) 核查政务大模型应用的基础设施配置；
   1) 核查软件配置、运行环境参数、功能模块调用策略是否正确；
   2) 核查是否开放非必要的网络端口和功能服务；
   3) 核查采取的网络隔离策略配置是否有效，核查训练通信、推理请求等流量出入规则的权限范围和配置是否最小化；
   4) 核查是否存在默认配置、默认口令，核查大模型部署工具风险端口（如 Ollama 的 11434 端口）是否进行有效加固。
f) 核查政务大模型应用采取的用户身份识别和权限管理等管控措施；
   1) 对政务大模型应用自身提供的人机交互接口、所调用的 API 接口进行连接测试，核查是否支持用户身份识别及权限管理；

2） 核查用户在登录时是否需要身份鉴别、用户身份标识是否具备唯一性、是否存在空口令用户；
3） 核查是否支持设置一般用户、系统管理用户、安全审计用户等多角色权限访问控制，核查已开通用户的权限范围是否满足最小化设置，一般用户是否具备创建、删除、拉取、推送等高权限的大模型管理功能；
4） 对用于大模型人机交互的对话接口调用频率进行核查，核查是否开启限制功能等；
5） 模拟单一用户连续多次提交恶意提问、违法违规内容等行为，核查是否禁用或暂停该用户服务；
6） 模拟多并发用户提交恶意请求行为，核查是否具备对多并发用户的阻断访问措施；
7） 涉及用户身份实名认证要求的，核查是否依托国家网络身份认证公共服务开展用户身份识别。

g） 核查外挂知识库接入数据的安全情况；
1） 核查已采取的数据安全防护措施，是否按照 GB/T 45396—2025 中 6.3 数据使用的要求对使用的数据进行分类分级，对数据接口开启身份鉴别、访问控制、加密等安全策略；
2） 核查是否由本单位负责政务大模型应用管理、数据安全的部门，会同提供数据的业务部门开展必要性评估的过程文档，确认应用场景下外挂知识库接入的必要性。
3） 核查是否针对外挂知识库接入数据建立数据集台账，台账内容是否覆盖数据来源、类型和规模等信息，是否建立数据版本管理和回滚机制；

h） 大模型部署期间涉及对大模型进行微调优化的，核查预训练和优化训练的数据以及外挂知识库接入的数据清洗过滤情况；
1） 核查是否按照 GB/T 45652—2025 中 7.2 预训练数据处理活动评价方法对预训练数据进行评价，确认预训练数据安全性；
2） 核查是否按照 GB/T 45652—2025 中 7.3 优化训练数据处理活动评价方法对优化训练数据进行评价，确认优化训练数据安全性；
3） 随机抽取清洗后的外挂知识库接入数据，抽样数量不宜少于 5000 条，抽样比例不宜少于 5%，核查数据中是否含有违法和不良信息、错误数据等敏感内容。

i） 公众政务服务类应用，随机抽取清洗后的外挂知识库接入数据进行安全核查，抽样数量不宜少于 5000 条，抽样比例不宜少于 5%，核查数据内容是否超出政务信息公开范围，对存在时效性、适用范围等要素的内容，核查数据内容是否存在错误、过期、不适用内容；

j） 对政务大模型应用输入输出过程进行安全性测试，可采用人工测试、自动化测试和大模型测试等方式进行。测试内容至少涵盖 B.5 a）、b）、c）和 d）条款的有关要求；
1） 采用人工测试的，核查是否制定清晰、具体的测试标准和指南，对测试结果的分布和一致性进行分析，选择具有相关领域知识和经验的测试人员，确保测试结果的准确性；
2） 采用自动化测试的，核查是否在自动化测试脚本中清晰定义具体的测试指标计算方法和评分规则；

3） 采用大模型测试的，核查是否定义清晰的测试标准和评分规则，选择与测试任务相关性高的大模型，宜使用多个大模型进行交叉验证，确保测试结果的准确性。

B.3 大模型应用运行测试指南

大模型应用运行的测试指南如下。

a） 公众政务服务类应用，核查政务大模型应用输出内容，是否按照 GB 45438—2025 要求对生成、合成内容进行标识；

b） 核查是否针对权威信息发布制定内部审核制度，对涉及政务信息公开、政策公告、新闻发布、灾害风险预警等已发布权威信息，是否严格执行内部审核制度流程；

c） 核查政务大模型应用界面是否设置风险提示，对用户使用政务大模型应用局限性进行告知，核查告知内容是否完整、准确、有效；

d） 公众政务服务类应用，核查是否关闭面向公众的推理过程显示功能；

e） 公众政务服务类应用，核查是否保留人工服务方式，并提供便捷切换入口方式；是否提供在线客服、电话、即时通信、邮件等问题反映渠道；

f） 核查是否记录政务大模型应用运行日志，包括系统行为、用户行为等，留存时间是否不少于 1 年；核查是否定期对日志记录进行审计；

g） 在政务大模型应用上线前和运行使用后，核查是否定期针对政务大模型应用相关硬件平台、基础软件、外挂知识库接入数据开展安全评估，对已公开的大模型应用相关软硬件安全漏洞、缺陷信息进行安全测试，形成安全测试评估报告，并对发现的问题隐患进行整改加固；

h） 核查是否对政务大模型应用进行持续监测，是否定期对网络安全攻击和大模型安全护栏告警等情况进行分析，并形成分析报告；核查是否制定针对政务大模型应用安全事件的应急预案，明确采取一键关停等技术措施，对紧急突发事件进行处置；

i） 核查是否针对政务大模型应用更新升级制定安全管理策略，是否严格执行更新升级策略要求，确保大模型的完整性和可用性；

j） 核查是否定期对本单位政务大模型应用相关人员开展网络安全教育培训，培训内容是否覆盖大模型应用安全要求。

B.4 大模型应用停用测试指南

大模型应用停用的测试指南如下。

a) 核查是否建立模型停用管理流程机制，明确对政务大模型应用停用、数据清除等流程的监控和审计要求；

b) 核查是否建立模型停用数据处理规范要求，明确对停用的大模型文件、外挂知识库接入的数据、预训练和优化训练数据、系统日志等的处理方式；核查具备的数据删除措施是否支持删除数据被恢复。

B.5 大模型安全护栏测试指南

大模型安全护栏的测试指南如下。

a） 构造包含对抗攻击指令的多样化测试题集，覆盖提示词注入攻击（如直接注入、间接注入、代码注入、多模态注入等）、越狱攻击（如角色扮演、输入混淆、上下文操纵等）、资源消耗攻击等攻击指令，核查大模型安全护栏能否正确识别与分类；

核查大模型安全护栏是否具备攻击样本库的在线或离线更新机制，并核查其最近一次的更新记录，以评估其可持续更新的能力；

b) 根据大模型应用实际支持的模态，核查其大模型安全护栏是否具备相应的输入输出内容安全识别能力，具体包括：
   1) 支持输入输出文本内容的，核查是否可识别与分类文本中包含的、符合 GB/T 45654—2025 所定义的各类违法和不良信息、敏感内容及个人信息；核查其对不同语言（至少包括中文、英文）及长短文本场景下安全风险的识别一致性；
   2) 支持输入输出图像内容的，核查是否可识别并拦截图像中包含的视觉性违法和不良信息（如色情、暴力、血腥内容）、敏感标识（如特定旗帜、符号）、显性或隐性的恶意二维码，以及通过图像承载的对抗性攻击指令（如多模态提示注入）；
   3) 支持输入输出音频内容的，核查是否可识别并拦截音频中所包含的违法和不良信息、有害指令及语音“钓鱼”等内容；覆盖在不同环境（如嘈杂、清晰）及不同语速下对语音安全风险的识别能力；
   4) 支持输入输出视频内容的，核查是否可对视频的图像帧和音频轨进行综合分析，识别并拦截其中包含的违法和不良画面、声音或字幕；核查对动态内容中快速闪现的风险图像、符号或文字的识别能力；
   5) 支持输入输出文件内容的，核查其对文件内容的安全扫描能力，对于办公文档（如 DOCX、PDF），核查是否可识别并告警其中的恶意宏、恶意链接或嵌入式脚本；对于所有类型文件，核查是否可识别其中包含的非授权个人信息、内部工作资料等敏感数据泄露风险。

c) 通过交互问答测试核查大模型安全护栏输入识别管控能力；
   1) 构造包含违法和不良信息的多样化测试题集，覆盖 GB/T 45654—2025 附录 A 中生成内容的主要安全风险，核查是否可正确识别与分类。核查是否可自定义配置关键词过滤规则；
   2) 通过多轮对话构建上下文，对大模型分段引导和语义渗透，核查是否可准确识别恶意诱导内容，是否可准确识别不符合用户角色的输入内容；
   3) 构造包含个人信息的多样化测试题集，核查是否可正确识别敏感内容；核查是否可自定义配置重要数据识别规则。

d) 通过交互问答测试核查大模型安全护栏输出识别管控能力；
   1) 核查是否可支持偏移、加密、重排、随机替换、掩码等脱敏规则配置；通过提交测试题，核查大模型应用在敏感内容输出时是否已进行脱敏处理；
   2) 构造违法和不良信息、与本应用场景无关的测试题集，核查输出的内容是否包含违法和不良信息、超业务范围内容；
   3) 核查代答、拒答技术措施对已知风险问题类别与标准回复、拒答回复之间的映射关系，评估已提供代答、拒答内容的准确性和一致性；
   4) 核查代答、拒答技术措施的配置是否支持自定义扩展，允许调整风险类别与回复的关联关系；
   5) 核查代答、拒答技术措施是否可及时更新。

e) 核查对大模型安全护栏相关日志留存及审计措施，确保其满足可追溯性和内容完整性要求。
   1) 核查日志是否完整记录每个大模型应用用户（行为主体）、请求的原始输入内容（用户行为）、请求时间（事件时间）以及大模型返回的最终输出内容；

2) 核查日志是否完整记录大模型安全护栏自身系统行为类型（如攻击拦截、内容过滤、数据脱敏）及行为详情（如触发的规则、处理结果）；
3) 核查日志是否完整记录每个大模型安全护栏用户（行为主体）、登录/登出时间（事件时间），以及所有对大模型安全护栏的配置变更操作（用户行为），如规则增删、敏感词库更新等；
4) 核查日志留存时间是否不少于 1 年；
5) 核查是否支持基于时间范围、请求用户、事件类型等多维度对日志进行查询和统计分析。

参 考 文 献

[1] 全球人工智能治理倡议（2023年10月18日中央网信办发布）

[2] 人工智能安全治理框架 （V2.0）（2025年9月15日全国网络安全标准化技术委员会发布）

[3] 中华人民共和国网络安全法（2016年11月7日第十二届全国人民代表大会常务委员会第二十四次会议通过）

[4] 中华人民共和国数据安全法（2021年6月10日第十三届全国人民代表大会常务委员会第二十九次会议通过）

[5] 中华人民共和国个人信息保护法（2021年8月20日第十三届全国人民代表大会常务委员会第三十次会议通过）

[6] 中华人民共和国密码法（2019年10月26日第十三届全国人民代表大会常务委员会第十四次会议通过）

[7] 互联网政务应用安全管理规定（2024年5月15日中央网络安全和信息化委员会办公室 中央机构编制委员会办公室 中华人民共和国工业和信息化部 中华人民共和国公安部联合公布）

[8] 网络信息内容生态治理规定（2019年12月15日国家互联网信息办公室令第5号公布）

[9] 网络数据安全管理条例（2024年9月24日中华人民共和国国务院令第790号公布）

[10] 商用密码管理条例（1999年10月7日中华人民共和国国务院令第273号公布 2023年4月27日中华人民共和国国务院令第760号修订）

[11] 生成式人工智能服务管理暂行办法（2023年7月10日国家互联网信息办公室 中华人民共和国国家发展和改革委员会 中华人民共和国教育部 中华人民共和国科学技术部 中华人民共和国工业和信息化部 中华人民共和国公安部 国家广播电视总局令第15号公布）

[12] 人工智能生成合成内容标识办法（2025年3月7日国家互联网信息办公室 中华人民共和国工业和信息化部 中华人民共和国公安部 国家广播电视总局 国信办通字〔2025〕2号公布）