ICS 35.030
CCS L 80

a

中华人民共和国国家标准化指导性技术文件

GB/Z XXXXX—XXXX

网络安全技术 大模型安全网关产品安全指南

Cybersecurity Technology—Large Model Security Gateway Products Security Guideline

（征求意见稿）

（本草案完成时间：2026-10-28）

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

XXXX - XX - XX 发布

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 I
引言 II
1 范围 1
2 规范性引用文件 1
3 术语和定义 1
4 缩略语 2
5 概述 3
5.1 产品定位 3
5.2 部署方式 3
5.3 产品形态 3
6 安全框架 3
6.1 概述 3
6.2 安全能力组成 4
6.2.1 安全功能组件 4
6.2.2 自身安全能力 4
6.2.3 安全管理保障 4
7 功能模块 4
7.1 安全功能组件 4
7.1.1 基础功能 4
7.1.2 安全防护功能 5
7.2 自身安全能力 7
7.2.1 需考虑的要点 7
7.2.2 标识和鉴别 7
7.2.3 自身访问控制 7
7.2.4 自身安全审计 7
7.2.5 通信安全 8
7.2.6 支撑系统安全 8
7.3 安全管理保障 8
7.3.1 策略管理 8
7.3.2 日志审计 8
7.3.3 运行监测 9
7.3.4 事件处置 9
附录 A（资料性）大模型安全网关产品架构 10
附录 B（资料性）大模型安全网关典型部署模式示意说明 11
附录 C（资料性）大模型安全网关产品的运行环境 13
附录 D（资料性）大模型安全网关产品形态 14
附录 E（资料性）典型攻击类型与防护要点 15
附录 F（资料性）人工智能服务接入对象 18
附录 G（资料性）大模型安全网关敏感数据分类与防护 19
附录 H（资料性）大模型安全网关测评方法 21
参 考 文 献 26

前　言

本文件为规范类指导性技术文件。

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。

本文件由全国网络安全标准化技术委员会（SAC/TC260）提出并归口。

本文件起草单位：北京神州绿盟科技有限公司、公安部第三研究所、北京中关村实验室、国家计算机网络应急技术处理协调中心、中国信息安全测评中心、中国信息通信研究院、中国科学院信息工程研究所、中央网信办数据与技术保障中心、中国网络安全审查认证和市场监管大数据中心、中国电子技术标准化研究院、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）、公安部第一研究所、国家信息技术安全研究中心、中国联合网络通信集团有限公司、国家工业信息安全发展研究中心、国家计算机病毒应急处理中心、中国电子科技集团公司第十五研究所、北京市海淀区互联网应急中心、中兴通讯股份有限公司、启明星辰信息技术集团股份有限公司、北京天融信网络安全技术有限公司、北京长亭科技有限公司、天翼安全科技有限公司、中国科学院软件研究所、中国电子信息产业集团有限公司第六研究所、广西电网有限责任公司、中移物联网有限公司、天翼云科技有限公司、中移九天人工智能科技（北京）有限公司、南方电网人工智能科技有限公司、中移（苏州）软件技术有限公司、南方电网科学研究院有限责任公司、国家高速列车青岛技术创新中心、广东省农村信用社联合社、华为技术有限公司、北京网测科技有限公司、北京时代新威信息技术有限公司、深信服科技股份有限公司、浙江御安信息技术有限公司、中电科网络安全科技股份有限公司、陕西省网络与信息安全测评中心、上海文鳐信息科技有限公司、三六零纵横信息技术有限公司、国科量子通信网络有限公司、北京芯盾时代科技有限公司、北京火山引擎科技有限公司、杭州安恒信息技术股份有限公司、北京浩瀚深度信息技术股份有限公司、深圳昂楷科技有限公司、北京信联数安科技有限公司、度小满科技（北京）有限公司、华东师范大学、西安交通大学、南京聚铭网络科技有限公司、杭州安泉数智科技有限公司、杭州迪普科技股份有限公司、深圳大学、西北大学、哈尔滨工业大学、格尔软件股份有限公司、北京天地和兴科技股份有限公司。

本文件主要起草人：叶晓虎、李海鹏、胡忠华、严寒冰、门嘉平、盛小宝、杨婧婧、孟楠、陈恺、高海辉、宋好好、顾杜娟、张斌、袁豪杰、胡光俊、卜天、李子威、朱雪峰、张睿、李明哲、王峰、姚佳明、申东洋、董晶晶、王有学、丁攀、赵月、张元立、闫新成、张智南、李泽村、李微羽、王诗蕊、赵书洋、王龙、杨玉宽、刘春池、肖颖、黄雅静、韦宗慧、刘利军、田佳琦、范敦球、尤扬、何亮忠、徐培明、韩志峰、姚语涵、苏莉娅、刘长波、杜杰伟、郭春颖、李敏、刘普、杨柳、叶翔、郭新海、望娅露、李彦、杨睿超、江魁、缪亚军、陈洪丽、吴春来、齐钟昱、庞韶敏、蒋建波、栗静文、冯牧玥、薛明富、王平辉、叶建伟、张卫博、陈虎、黄进、王涛、周晓军、仲凯韬、杨健、姜伟、赵维五、杨晨、朱伟光、孙汝泰。

引　言

随着人工智能技术的快速发展，大模型已广泛应用于政务、金融、教育、医疗等领域，成为数字化发展的重要技术支撑。随着人工智能应用范围不断扩大，其在输入输出内容、数据处理、模型调用及应用运行等过程中面临的安全风险日益受到关注。大模型安全网关作为人工智能应用与人工智能服务之间的安全控制节点，可通过对调用过程实施安全防护，为大模型应用的安全、稳定、合规运行提供支撑。

本文件根据《中华人民共和国网络安全法》《中华人民共和国数据安全法》《生成式人工智能服务管理暂行办法》等法律法规、部门规章及相关标准要求，围绕人工智能服务安全防护需求，针对现有大模型安全网关产品在功能定义和防护能力等方面存在的不统一问题，制定大模型安全网关产品安全指南。

本文件围绕大模型安全网关产品的安全功能组件、自身安全能力和安全管理保障等方面提出指导性要求，为大模型安全网关产品的建设和应用提供参考。

网络安全技术 大模型安全网关产品安全指南

1 范围

本文件提供了大模型安全网关产品的安全框架、安全功能、自身安全和安全管理保障的指导信息。

本文件适用于在政务、金融、医疗、工业及互联网等领域中使用人工智能技术的相关单位，包括大模型应用开发方、平台提供方及安全产品厂商等。

本文件适用于政务、金融、医疗、工业及互联网等领域中，人工智能技术的开发、部署、使用及安全防护等活动，适用单位包括人工智能服务供应商、人工智能服务调用方、安全产品厂商及检测评价机构等。

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 39786—2021 信息安全技术 信息系统密码应用基本要求

GB 42250—2022 信息安全技术 网络安全专用产品安全技术要求

GB/T 43697—2024 数据安全技术 数据分类分级规则

GB/T 43698—2024 网络安全技术 软件供应链安全要求

GB/T 45279.1—2025 IPv4/IPv6网络安全防护技术规范 第一部分：IP承载网

GB/T 45288.1—2025 人工智能 大模型 第1部分：通用要求

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

3 术语和定义

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

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

人工智能服务 artificial intelligence service

向用户和业务系统等提供人工智能能力的服务，包括但不限于大模型服务、智能体服务、工具调用服务等。

3.3

大模型安全网关 large model security gateway

部署于基于大模型的人工智能服务及用户、业务系统等调用方之间，对人工智能服务运行过程中的调用请求、交互内容和返回结果进行安全检测与防护的安全产品。

3.4

内容安全 content security

人工智能服务的输入内容和输出结果符合相关安全要求。

3.5

提示词注入攻击 prompt injection attack

攻击者通过在输入请求、上下文或外部检索内容中构造指令，诱导大模型违背系统设定、业务规则或安全策略生成非预期结果的攻击方式。

3.6

模型调用 model invocation

应用系统通过接口向人工智能服务提交输入、上下文、工具调用或控制参数并获取响应的交互过程。

3.7

安全策略 security policy

用于规定大模型安全网关对模型调用、数据处理、内容检测、访问控制和风险处置进行判定与执行的规则集合。

3.8

敏感数据 sensitive data

一旦泄露、篡改、非法使用或不当出境，可能危害国家安全、经济运行、社会秩序、公共利益，组织权益或个人权益。

3.9

词元 token

模型处理信息的最小计量单位。

3.10

工具 tool

提供特定功能且可被使用的设备、软件或系统。

3.11

智能体 AI agent

能够感知环境、自主决策并执行行动的人工智能物理实体或数字实体。

3.12

工具调用 tool call

AI 模型或智能体对外部工具、API、和服务的调用行为。

4 缩略语

下列缩略语适用于本文件。

API：应用程序编程接口（Application Programming Interface）

A2A：智能体到智能体协议（Agent2Agent）

DCI：数据中心互联（Data Center Interconnect）

DCN：数据中心网络（Data Center Network）

JWT：JSON Web令牌（JSON Web Token）

LLM：大语言模型（Large Language Model）

MCP：模型上下文协议（Model Context Protocol）

RPM：每分钟请求次数限制（Requests Per Minute）
SaaS：软件即服务（Software as a Service）
TPM：每分钟词元使用量限制（Tokens Per Minute）

5 概述

5.1 产品定位

大模型安全网关（以下简称为“网关”）作为大模型、智能体、工具等人工智能服务的统一接入点，为用户和应用系统调用人工智能服务提供统一的接入与管理支撑。网关对人工智能服务运行过程中的调用请求、交互内容和返回结果进行实时安全检测与防护，识别内容安全、提示词注入、敏感数据泄露、模型攻击等安全风险。网关提供统一的安全管理保障机制，对安全防护策略、运行状态和安全事件进行持续管理，保障网关在具体业务场景中发挥作用。网关与API网关、防火墙等安全产品的能力进行协同，复用已有安全检测与控制能力，并针对人工智能服务特有的风险提供补充防护，形成相互协同的安全防护体系。

注：大模型安全网关的产品架构见附录A。

5.2 部署方式

如附录B所示，大模型安全网关可综合考虑业务流量规模、安全检测方式和系统资源等因素，评估不同部署方式对业务性能、资源使用和服务连续性的影响。部署方式主要包含以下四种：

a） 串联部署：网关部署在应用系统与人工智能服务之间的主调用链路中，对输入请求、模型调用、输出结果及外部交互过程实施实时检测、控制和处置；
b） 旁路部署：网关通过流量镜像、日志采集、请求复制或结果同步等方式，对人工智能服务的调用过程进行监测、审计和辅助防护，而不直接改变原有网络拓扑；
c） 内嵌集成：网关以 SDK、插件、能力组件或平台模块形式嵌入业务系统中，实现业务流程与调用链路的深度结合；
d） 一体化部署：网关与人工智能服务平台、AI 一体机、安全管理平台或综合系统进行统一建设，形成统一部署、统一运维的整体解决方案。

注：大模型安全网关产品的运行环境见附录 C。

5.3 产品形态

如附录D所示，大模型安全网关可根据交付方式和部署要求采用下列一种或多种形态：

a） 软件产品形态，包括独立部署软件、容器化软件、虚拟化软件和云原生软件；
b） 硬件产品形态，包括一体化安全设备、专用安全网关设备等；
c） 云服务形态，包括托管式安全网关服务和 SaaS 化安全检测服务；
d） 组件或插件形态，包括作为 API 网关、应用网关、模型服务平台或业务系统中的安全组件部署。

6 安全框架

6.1 概述

网关产品的安全框架包括自身安全能力、安全功能组件和安全管理保障，如图1所示。

安全功能组件面向网关所承载的业务调用过程，提供业务请求与响应的转发、安全检测及安全控制等功能。自身安全能力面向大模型安全网关自身的安全运行，通过对网关自身的安全风险进行防护和控

制，保障网关自身的安全可靠运行。安全管理保障面向大模型安全网关自身的运行管理活动，提供安全策略管理、日志管理、运行监测、异常处置等管理功能。

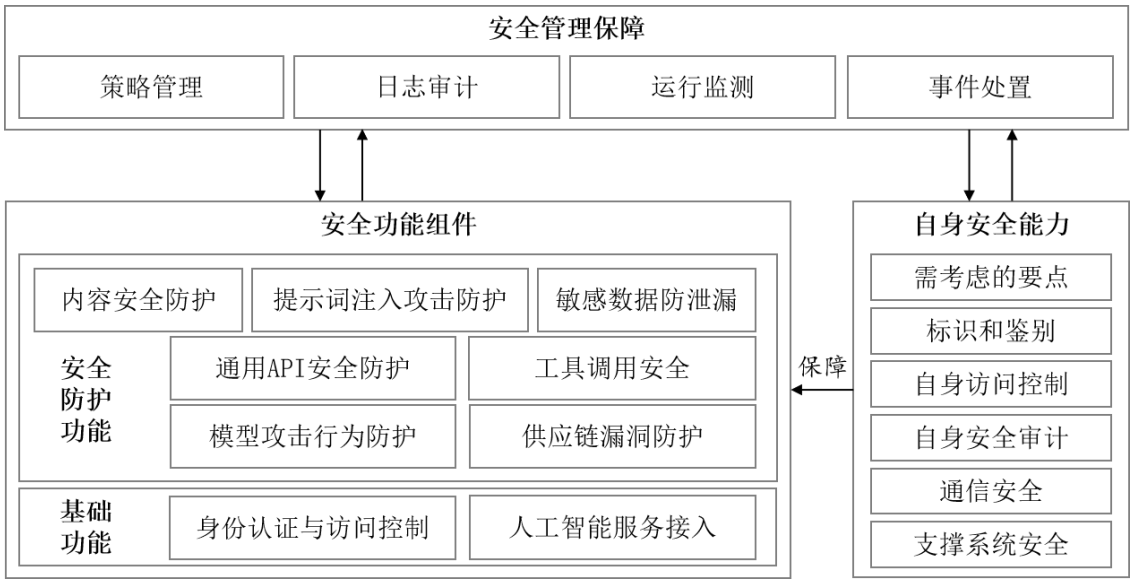


图 1 大模型安全网关产品安全框架

6.2 安全能力组成

6.2.1 安全功能组件

安全功能组件面向人工智能服务及其调用过程提供安全防护能力，用于实现对模型调用、交互内容及相关业务流量的安全检测与防护。

注：如附录E列举了安全功能组件的典型攻击类型与防护要点。

6.2.2 自身安全能力

自身安全能力面向大模型安全网关产品自身，保障产品在身份认证、访问控制、数据保护、日志审计等方面的安全性以及稳定、可靠运行。

6.2.3 安全管理保障

安全管理保障面向大模型安全网关的实际运行与管理过程，通过安全策略和相关配置对安全功能进行统一管理和有效控制，保障安全能力按照业务需求持续、有效地运行。

7 功能模块

7.1 安全功能组件

7.1.1 基础功能

7.1.1.1 身份认证与访问控制

身份认证与访问控制包括以下内容：

a) 通过统一的身份认证与密钥管理，对调用主体进行身份识别和凭证管理。凭证采用加密存储、定期更换等方式降低凭证泄露和滥用风险；身份认证结合主流认证协议与凭证体系实施，如采用 OAuth 2.0 令牌、JWT 等；
b) 采用细粒度授权管理，根据访问主体、访问资源及操作类型实施差异化授权控制。结合角色、属性、策略等访问控制模型实施授权管理，并以仅放行显式授权的调用等方式落实最小权限原则；
c) 按照租户、应用等管理边界实施多租户访问隔离和访问控制，防止未经授权的跨租户、跨应用访问；
d) 结合调用来源地址、时间、来源地等网络上下文实施动态访问控制和准入判定，降低异常访问风险。客户端声明的终端信息作为辅助判定信息，并结合其他网络上下文进行综合判断，避免将其单独作为放行依据。

7.1.1.2 人工智能服务接入

如附录F所示，人工智能服务接入包括以下内容：
a) 统一接入大模型服务、智能体服务和工具调用服务等人工智能服务，为不同类型人工智能服务提供统一的接入与管理入口；
b) 采用代理转发方式处理人工智能服务请求，对请求和响应进行必要的参数转换和处理，实现调用方与后端人工智能服务之间的请求转发；
c) 根据人工智能服务、调用方、业务等条件动态配置流量调度策略，实现对人工智能服务调用流量的调度；
d) 对经网关接入或调用的人工智能服务及相关资产进行统计和展示，形成资产视图，为人工智能服务及相关资产的统一管理提供支撑。

7.1.2 安全防护功能

7.1.2.1 内容安全防护

7.1.2.1.1 内容检测

人工智能服务的输入内容和输出结果符合 GB/T 45654—2025 的相关要求。内容检测包括以下内容：
a) 对人工智能服务的输入内容和输出结果进行安全检测，依据 GB/T 45654—2025 附录 A 所列风险信息设置检测范围和识别规则，对相关风险内容进行识别；
b) 根据人工智能服务涉及的内容模态，采用与文本、图片、音频、视频等不同内容类型相适配的检测方式，实现对多模态内容的安全检测。

7.1.2.1.2 内容拦截

内容拦截包括以下内容：
a) 对于识别出的不安全输入内容，结合风险类型和安全策略，采取提示、代答、拦截等方式进行处置，避免风险内容直接进入模型处理过程；
b) 对于识别出的不安全输出结果，结合风险类型和安全策略，采取拦截、替换、降级等方式进行处置，避免风险内容直接返回用户；
c) 采用替换、代答等方式处置风险内容时，结合预先审核的安全响应模板或经验证的安全回复策略，降低处置过程中引入新的内容安全风险。

7.1.2.2 提示词注入攻击防护

提示词注入攻击防护包括以下内容：

a) 针对用户输入及模型交互过程中可能存在的提示词注入风险，将角色扮演、权限提升、跨语言表达等直接注入行为，以及通过网页内容、RAG 检索内容等外部信息引入的间接注入行为纳入检测范围；
b) 结合规则匹配、语义分析、分类模型等方式，对输入内容及其上下文进行综合判断，识别不同类型的提示词注入行为；
c) 对检测到的提示词注入行为，结合应用场景、防护需求和风险情况，采取告警、拦截等方式进行处置。

7.1.2.3 敏感数据防泄漏

7.1.2.3.1 敏感数据识别

敏感数据识别包括以下内容：

a) 通过规则匹配、语义分析、机器学习等方式，对大模型交互过程中的敏感数据进行识别，包括个人信息（身份证号、手机号、姓名、住址、银行卡号等）及其他依法受保护的敏感数据；
b) 根据不同行业、不同场景的敏感数据防护需求，自定义配置敏感数据类型。

注：大模型安全网关敏感数据分类和防护说明见附录G

7.1.2.3.2 敏感数据脱敏

敏感数据脱敏包括以下内容：

a) 采用双向链路脱敏方式，分别对用户输入提示词和大模型输出响应内容进行脱敏处理；
b) 结合敏感数据类型、业务场景、用户权限等因素设置差异化脱敏策略，并配置相应的脱敏规则。

7.1.2.4 通用 API 安全防护

通用API安全防护包括以下内容：

a) 采用调用凭证校验机制，对 API 调用凭证的有效性进行验证，防止使用无效、过期或被吊销的调用凭证访问人工智能服务；
b) 对 API 调用实施限流和配额控制，按照请求频率（RPM）、词元使用量（TPM）、并发请求数等维度进行限制；
c) 针对高频调用、异常调用量等 API 调用异常行为进行识别，并根据安全策略采取限制、阻断、熔断等方式进行处置。

7.1.2.5 工具调用安全

工具调用安全包括以下内容：

a) 采用工具调用权限控制，对经网关的工具调用请求按工具服务维护白名单，限定可调用工具类型与参数取值范围；
b) 对工具调用参数的类型、取值范围、目标地址等进行安全校验，防范因参数异常引发的安全风险，如 SSRF、命令注入类攻击；
c) 结合工具调用请求开展调用链安全分析，识别异常或恶意工具调用，如经过间接注入诱导的恶意工具调用；
d) 针对外发、数据删改、支付等高危操作设置二次确认机制，采用人机协同方式进行审批。

7.1.2.6 模型攻击行为防护（可选）

模型攻击行为防护包括以下内容：

a） 基于用户、应用、人工智能服务等维度分析历史调用行为，建立正常调用行为基线；
b） 结合调用请求、响应、调用频率、调用模式及资源消耗等信息，分析当前调用行为与正常行为基线的偏离情况，识别异常调用行为；
c） 对多个异常调用行为进行关联分析，识别异常行为的持续性、频繁性及关联性等特征，评估潜在的安全风险；
d） 根据异常行为的风险等级，采取告警、限流、阻断等控制措施。

7.1.2.7 供应链漏洞防护（可选）

供应链漏洞防护参考 GB/T 43698-2024 的相关要求，包括以下内容：
a） 建立 AI 组件指纹库和漏洞库，对 AI 组件漏洞进行检测与防护；指纹库覆盖主流 AI 推理框架（如 Ollama 等）及 AI 组件的特征信息，漏洞库收录相关组件的已知漏洞信息，并进行自动匹配和风险告警；
b） 定期更新 AI 组件指纹库和漏洞库。

7.2 自身安全能力

7.2.1 需考虑的要点

产品自身安全设计和实施应参考GB 42250—2022的相关要求。产品应用于信息系统时，应结合GB/T 39786—2021、GB/T 22239—2019的相关要求，从标识和鉴别、自身访问控制、自身安全审计、通信安全、支撑系统安全、产品升级、用户信息安全和密码要求等方面设计和实施自身安全能力。

7.2.2 标识和鉴别

产品的标识和鉴别包括以下内容：
a） 建立授权管理员初始属性恢复机制，在系统异常时恢复管理员初始属性；
b） 建立登录失败防护机制，采取临时锁定账户、增加验证难度等措施限制连续非法登录尝试次数，防止暴力破解；
c） 设置可配置的登录会话超时时间，仅允许授权管理员调整该参数值；登录会话在设定时间内无操作时，自动终止会话，并清除会话状态及相关数据；
d） 采用两种或两种以上不同类型的鉴别技术组合对管理员进行身份鉴别，如采用数字证书、动态口令等增强身份鉴别方式。

7.2.3 自身访问控制

产品的自身访问控制包括以下内容：
a） 设置系统管理员、安全管理员和安全审计员等不同管理角色，并根据角色分配相应的操作权限；
b） 为系统管理员分配用户账号管理、系统参数配置、数据备份与恢复等系统运维权限；
c） 为安全管理员分配安全策略的增加、删除、修改、查询、启用和停用等权限；
d） 为安全审计员分配安全统计报表的查询和管理权限；
e） 对用于实施安全防护功能的模型、插件及工具设置访问权限，限制其对系统资源、敏感数据及管理接口的访问，防止未经授权的访问和操作。

7.2.4 自身安全审计

产品的自身安全审计包括以下内容：
a） 为确保重要操作的可追溯性，对管理员登录与注销、系统启动与停止、系统故障与处理、配置变更、管理员及其权限的增加、删除和修改、审计日志导出等重要操作事件进行日志记录；

b) 审计日志记录事件发生的日期和时间、事件类型、事件主体、事件客体、事件结果、访问来源及风险等级等信息，以支撑后续关联分析；
c) 按照管理员、操作类型、操作对象、时间范围等条件对审计日志进行查询，并对审计日志进行导出和备份；
d) 当审计日志存储容量达到设定阈值时，及时通知授权管理员，并结合存储策略进行处置，如按时间顺序清理最早生成的日志；
e) 为辅助安全态势评估，对审计日志进行统计并生成统计报表；
f) 采取完整性保护措施，防止审计日志被恶意篡改；
g) 对审计日志中涉及的敏感个人信息进行脱敏处理，防止个人隐私及敏感数据泄露；
h) 为降低凭证泄露风险，避免在审计日志中记录凭证信息，如管理员鉴别信息。

7.2.5 通信安全

产品与业务系统、人工智能服务等组件之间的数据传输接口采用安全传输协议，保障传输数据的保密性与完整性。

7.2.6 支撑系统安全

产品的支撑系统安全包括以下内容：
a) 采用非易失性存储或实时同步机制，确保在意外断电或系统重启后，配置数据、审计日志等关键信息不丢失；
b) 关闭不必要的系统服务、端口、调试接口和进程，减少攻击面；
c) 对产品配置文件、凭证、规则库、内置模型文件、升级包等关键文件建立校验机制，在加载或执行前验证其完整性，防止未经授权的修改；
d) 对配置文件和审计数据进行备份和恢复，并在异常情况下进行版本回退，保障系统的业务连续性与故障恢复能力；；
e) 为实现管理流量与业务流量的隔离，降低相互影响的风险，在网络架构或逻辑层面将管理接口和业务接口分离。

7.3 安全管理保障

7.3.1 策略管理

产品的策略管理功能包括以下内容：
a) 设置访问策略，对用户、应用及其可访问的人工智能服务和相关能力进行访问控制；
b) 按照请求、会话或任务等不同对象设置调用控制策略，对请求频率、词元使用量、并发请求数、调用次数及执行时长等进行配置和管理；
c) 统一管理安全防护策略，对安全检测、防护及处置策略进行集中配置；
d) 配置人工智能服务路由策略，并设置服务异常情况下的切换策略。

7.3.2 日志审计

产品的日志审计功能包括以下内容：
a) 对人工智能服务调用过程及安全防护活动进行审计记录，记录调用请求、响应、提示词、生成内容、工具调用及调用参数等关键信息，以及与 7.1.2 所述安全防护功能相关的检测及处置结果；
b) 对会话、任务及调用过程建立关联关系，实现跨组件、跨调用环节的审计追踪；
c) 对审计日志进行安全存储、留存和完整性保护，采用加密、脱敏等措施防止日志泄露或篡改；

d） 对审计日志进行查询、统计、关联分析及导出，并对安全事件相关审计记录进行固定和留存；
e） 按照相关法律法规、标准规范及安全管理要求，制定日志留存策略，明确日志记录粒度及留存期限。

注：7.3.2 所述日志审计主要针对人工智能服务调用及安全防护过程产生的业务日志；7.2.4 所述自身安全审计主要针对大模型安全网关自身的管理操作和安全管理活动，两者审计对象不同。

7.3.3 运行监测

产品的运行监测功能包括以下内容：
a） 对大模型安全网关的服务状态、资源使用及关键组件运行情况进行监测，及时发现运行异常和组件故障；
b） 对接入的人工智能服务运行状态进行监测，及时发现服务不可用、响应异常及调用异常等情况；
c） 监测和统计网关业务运行情况，分析调用量、资源使用等运行指标，识别异常变化；
d） 监测安全防护功能的运行及防护情况，掌握安全检测、防护及处置的运行状态，及时发现安全防护能力异常或失效；
e） 对运行监测发现的异常情况进行告警，并提供相应的监测信息；
f） 对告警信息进行聚合，对同类告警进行合并，降低重复告警。

7.3.4 事件处置

产品的事件处置功能包括以下内容：
a） 根据安全事件类型、影响范围和严重程度等，对安全事件进行识别和分级，标记事件等级并记录事件相关信息；
b） 根据安全事件情况采取限流、阻断、隔离、权限收敛等风险控制措施，并记录事件处置过程及结果；
c） 支持配置风险控制措施的有效期限及解除条件；
d） 在异常情况下采用主备切换、服务切换、业务降级等措施保障业务连续性，并对相应的处置权限进行管理；
e） 对应急预案进行存储、查询和版本管理，并根据安全事件进行关联使用。

注：如附录 H 所示为大模型安全网关测评方法。

附 录 A
（资料性）
大模型安全网关产品架构

A.1 产品架构

如图A.1给出了大模型安全网关产品架构示例。大模型安全网关可部署于应用系统与人工智能服务之间，通过接入控制、安全检测、策略决策和处置执行等功能，对人工智能服务调用过程进行安全防护，并通过安全管理和审计监测等功能实现运行管理。

不同产品的功能组成和部署方式可根据实际应用需求进行调整，本图仅作为产品架构示例，不作为具体产品架构的限定。

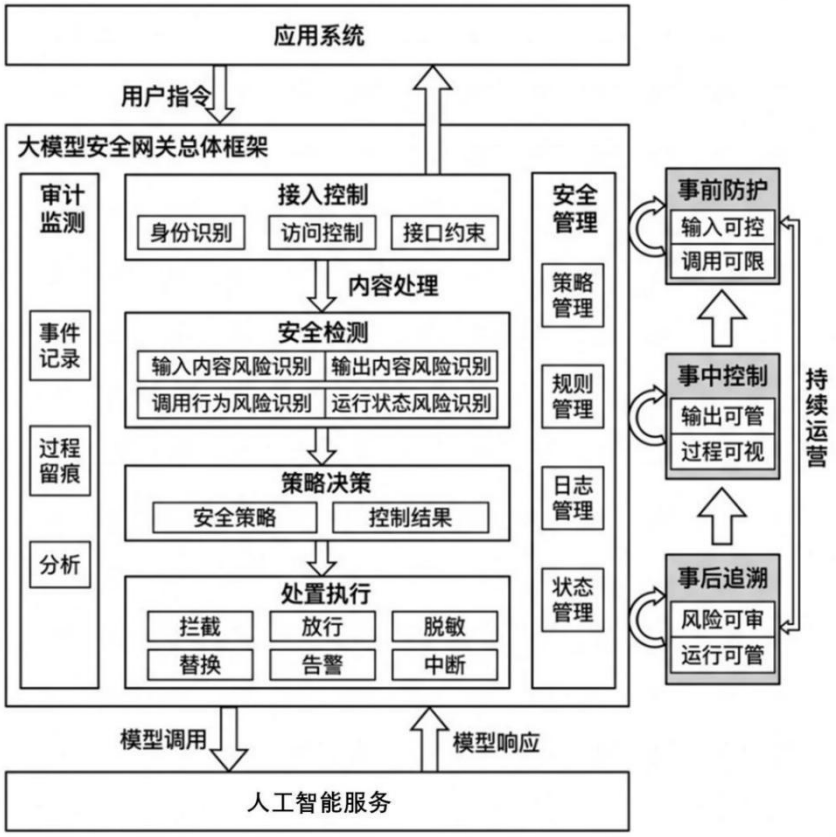


图 A.1 大模型安全网关产品架构示例图

附 录 B
（资料性）
大模型安全网关典型部署模式示意说明

B.1 部署模式概述

大模型安全网关可根据业务架构、安全需求和运行环境的不同，采用多种部署模式，并可在实际建设中根据需要选择单一模式或组合多种模式进行部署。不同部署模式在性能与时延影响、数据可见性、安全控制能力以及运行与运维保障能力等方面存在差异，部署模式选择应综合考虑上述因素，并结合实际业务需求进行合理配置。

B.2 串联部署模式

串联部署模式，如图B.1，是指将网关部署在应用系统与人工智能服务之间的主调用链路中，对输入请求、模型调用、输出结果及外部交互过程实施实时检测、控制和处置。

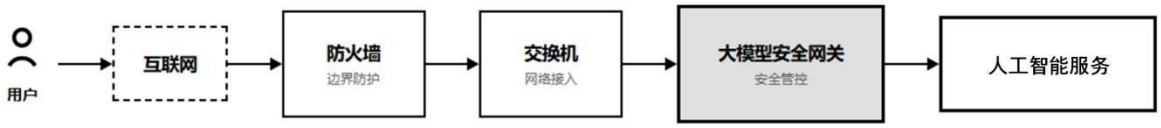


图 B.1 串联部署模式图

串联部署模式适用于对防护实时性、控制有效性要求较高的场景，可实现对请求与响应全过程的直接安全控制。该模式位于主调用链路中，引入额外转发与检测开销，对处理性能和时延存在一定影响。该模式可直接获取请求与响应数据，在加密通信场景下通常需通过协议终止或解密机制获取明文，否则将影响检测效果。该模式可对输入输出及调用过程实施强制控制与实时阻断，整体控制能力较强。对系统链路依赖较高，需通过高可用部署与故障切换机制保障业务连续性。

B.3 旁路部署模式

旁路部署模式，如图B.2，是指通过流量镜像、日志采集、请求复制或结果同步等方式，对大模型应用运行过程进行监测、分析、审计和辅助防护，而不直接改变原有调用链路。

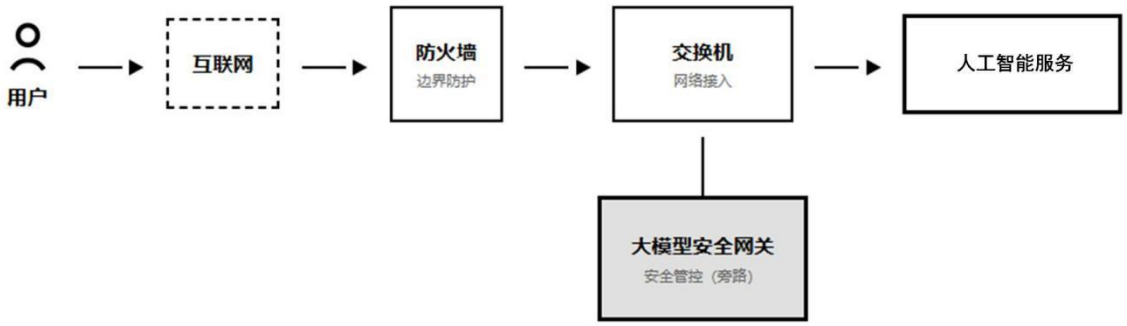


图 B.2 旁路部署模式图

旁路部署模式不参与主业务链路处理，对系统性能与时延基本无影响。该模式依赖镜像流量或日志数据获取信息，在加密通信场景下通常难以直接获取明文内容，检测能力受数据采集方式限制。该模式主要以监测与审计能力为主，通常不具备实时阻断能力，控制能力相对有限。该模式对现有系统影响较小，有利于保障业务连续性，但需关注多数据源接入带来的运维复杂性。

B.4 内嵌集成模式

内嵌集成模式，如图B.3，是指将网关以SDK、插件、能力组件或平台模块形式嵌入业务系统中，在业务流程中实现与调用链路的深度结合。

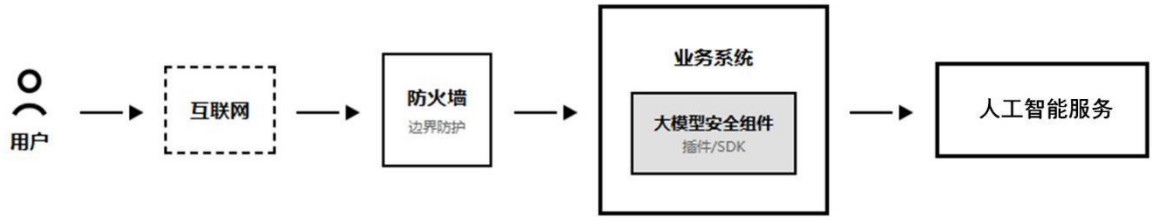


图 B.3 内嵌部署模式图

内嵌部署模式适用于应用系统定制化程度高、需要将安全能力与业务逻辑紧密结合的场景。该模式无需额外网络转发，处理路径较短，对性能与时延影响较小。该模式可在应用内部直接获取请求上下文及模型交互数据，可在加密前或解密后处理数据，具备较高数据可见性。该模式可结合业务逻辑实现细粒度控制，对模型调用及上下文处理具有较强约束能力。该模式需随业务系统进行集成与升级，对版本管理与能力维护提出较高要求，运维复杂度相对较高。

B.5 一体化部署模式

一体化部署模式，如图B.4，是指将大模型安全网关与人工智能服务平台、AI一体机、安全管理平台或综合系统进行集成交付，形成统一建设、统一部署、统一运维的整体解决方案。

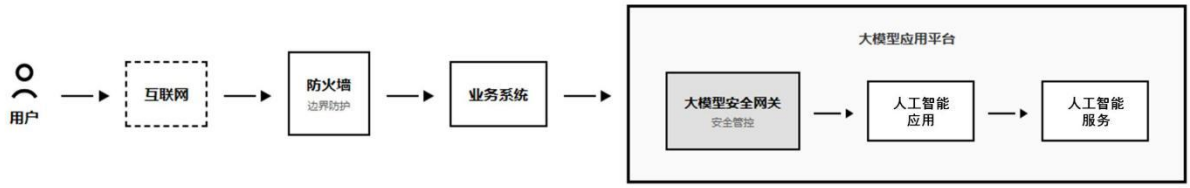


图 B.4 一体化部署模式图

一体化部署模式适用于本地化交付、集中管控和综合防护要求较高的场景。该模式通过平台内部集成优化调用路径，整体处理效率较高，适用于规模化集中处理场景。该模式通常处于统一平台内部，可直接获取模型调用及相关数据，在加密通信场景下可在系统内部完成解密与检测。该模式可在平台层实现统一策略控制与集中防护，整体控制能力较强。该模式支持统一运维与集中管理，有利于提升整体运行保障能力，但对平台规划与建设要求较高。

附 录 C
（资料性）
大模型安全网关产品的运行环境

表C.1列出了大模型安全网关产品的运行环境说明。

表 C.1 大模型安全网关产品运行环境

类别	具体说明
概述	大模型安全网关产品宜部署并稳定运行于通用信息化基础设施环境之上，支持在不同网络架构、计算资源条件及人工智能服务形态下进行接入与安全控制。
网络环境	产品宜符合 GB 45279.1—2025 要求，适配 IPv4/IPv6 双栈网络部署，具备链路冗余与故障切换能力；宜支持模型推理流量、管理运维流量、用户访问流量的逻辑隔离，并可根据模型调用会话设置差异化流量管控策略。
计算与资源环境	产品宜兼容通用服务器、异构算力、云原生容器等基础设施，适配软件、硬件、云服务、组件及混合部署等产品形态；运行环境宜提供满足性能指标要求的处理器算力、内存空间及存储资源，并具备资源弹性伸缩能力，以保障内容深度检测与实时处理的高效执行。
网络协议支持	产品宜兼容主流基础网络通信协议，适配大模型算力网络环境（DCI、DCN 等）及流式对话、长上下文推理等交互模式；产品与人工智能服务之间的通信链路宜采用加密传输协议，保障传输数据的机密性与完整性，并可限制存在安全风险的老旧通信方式。对于采用加密传输的业务流量，产品宜支持通过协议终止或解密代理等方式获取安全检测所需的明文内容；采用旁路部署方式时，宜明确可检测和可控制的流量范围。
大模型 API 兼容性	产品宜兼容主流人工智能服务提供商、本地私有化开源推理框架、企业自研私有大模型等多类型 API 接口，支持对输入请求、上下文、调用参数、输出响应进行统一解析、转发与安全管控；宜提供标准化扩展能力，快速适配新增人工智能服务接口。
行业场景适配	产品宜根据政务、金融、工业、互联网等行业大模型应用差异化安全需求调整部署策略；高合规行业宜支持本地留存、离线运行或数据不出本地环境；多租户场景宜支持租户间模型调用环境逻辑隔离；边缘 AI 场景宜支持轻量化运行和本地基础安全检测。

附 录 D
（资料性）
大模型安全网关产品形态

表D.1 列出了大模型安全网关产品形态。

表 D.1 大模型安全网关产品形态

产品形态	形态描述	适用场景	主要特点	实施注意事项
软件产品形态（Software Product）	以独立部署软件、容器化软件、虚拟化软件或云原生软件交付，可部署于通用服务器、虚拟化平台或容器平台。	已有通用计算或云原生基础设施，需要按业务规模灵活部署、升级和扩缩容的场景。	部署灵活，便于与现有网络、身份、日志和人工智能服务平台集成，可按需扩展检测能力。	宜明确运行环境资源配置和隔离要求，避免安全检测任务与模型推理任务争抢算力；宜做好软件版本、依赖组件和配置管理。
硬件产品形态（Hardware Product）	以一体化安全设备、专用安全网关设备等形式交付，将软件能力与专用硬件、接口和资源配置集成。	对性能稳定性、物理隔离、本地化部署或设备级合规管理有明确要求的场景。	软硬件协同，资源边界清晰，性能和交付规格相对稳定，便于统一验收和资产管理。	宜根据业务流量和模型调用并发量合理选型，配置高可用和故障切换能力，并建立硬件维保、固件升级和备件机制。
云服务形态（Cloud Service）	以托管式安全网关服务或 SaaS 化安全检测服务交付，由服务提供方负责平台运行、能力更新和日常维护。	使用公有云或托管人工智能服务，希望快速接入安全能力且不单独建设本地运维平台的场景。	开通和扩容较快，规则、模型和威胁情报可持续更新，能够降低本地建设和维护成本。	宜明确数据传输、存储地域、租户隔离、服务可用性和退出迁移要求；高合规场景宜确认数据是否允许离开本地环境。
组件或插件形态（Component or Plug-in）	作为 API 网关、应用网关、人工智能服务平台或业务系统中的安全组件、插件或软件开发包部署。	已有成熟 API 网关、微服务治理或人工智能服务平台，希望复用现有接入链路并进行渐进式改造的场景。	接入路径短，可复用既有鉴权、流量治理和日志能力，便于按应用或接口逐步启用。	宜明确组件与宿主平台的版本兼容、资源隔离和故障边界；宿主平台升级时宜同步验证安全组件的有效性和稳定性。

附 录 E
（资料性）
典型攻击类型与防护要点

人工智能服务安全风险可按照攻击发生的层次划分为访问与身份层、交互内容层、数据层、智能体执行层、模型层、AI基础设施与供应链层以及网关与运行环境自身层。不同层次的风险可能相互关联并形成跨层攻击路径，表E.1～E.7分别列出了各层典型攻击类型及防护要点，供参考。

表E.1 列出了访问与身份层的典型攻击类型与防护要点。

表 E.1 访问与身份层的典型攻击类型与防护要点

攻击类型	典型攻击场景描述	防护要点	对应章节
身份伪造	冒充合法用户、应用或智能体调用人工智能服务	身份鉴别、调用主体认证、访问凭证校验	7.1.1.1
凭证泄露与滥用	API Key、词元等调用凭证泄露后被非法使用	密钥保护、凭证生命周期管理、异常调用检测	7.1.1.1、7.1.1.3
未授权访问	未经授权调用模型、智能体、工具服务	RBAC/ABAC、默认拒绝、最小权限	7.1.1.1
权限提升	通过参数或调用关系获取超出授权范围的能力	细粒度授权、权限隔离、动态访问控制	7.1.1.1
接口滥用	通过高频、批量等方式异常调用人工智能服务	频率控制、配额控制、异常调用检测	7.1.1.3

表E.2 列出了交互内容层的典型攻击类型与防护要点。

表 E.2 交互内容层的典型攻击类型与防护要点

攻击类型	典型攻击场景描述	防护要点	对应章节
违规内容生成	诱导模型生成违法违规、色情暴力等风险内容	输入输出内容检测、风险识别、内容拦截	7.1.2.1
直接提示词注入	通过恶意 Prompt 诱导模型改变既定行为或绕过安全策略	提示词注入检测、语义分析、风险阻断	7.1.2.2
间接提示词注入	通过网页、RAG 文档等外部内容向模型引入恶意指令	外部内容检测、上下文分析、调用链关联检测	7.1.2.2、7.1.2.4
越狱攻击	通过角色扮演、指令组合等方式绕过模型安全限制	越狱识别、语义分析、输入拦截	7.1.2.1、7.1.2.2
编码与跨语言绕过	使用编码、拼写变体或其他语言表达规避安全检测	编码识别、语义还原、多语言检测	7.1.2.1、7.1.2.2
多轮上下文攻击	通过多轮交互逐步构造恶意语义或绕过检测	上下文关联分析、会话级风险检测	7.1.2.2

表E.3 列出了数据层的典型攻击类型与防护要点。

表 E.3 数据层的典型攻击类型与防护要点

攻击类型	典型攻击场景描述	防护要点	对应章节
输入敏感数据泄漏	用户将个人信息、国家秘密或其他敏感数据发送至人工智能服务	敏感数据识别、输入数据脱敏、数据流向控制	7.1.2.3
输出敏感数据泄漏	模型输出包含个人信息、企业敏感数据等	输出数据识别、脱敏、输出控制	7.1.2.3
RAG 数据泄漏	检索结果包含超出用户权限范围的敏感数据	敏感数据识别、访问控制、检索结果安全检测	7.1.2.3
工具数据泄漏	工具调用参数或返回结果包含敏感数据	数据识别、数据脱敏、工具调用控制	7.1.2.3、7.1.2.4
跨境数据流转	敏感数据被发送至境外人工智能服务	数据识别、数据流向控制、策略限制	7.1.2.3
多模态数据泄漏	图片、音频、视频或文件中包含敏感数据	多模态数据识别、敏感数据脱敏	7.1.2.3

表E.4 列出了智能体执行层的典型攻击类型与防护要点。

表 E.4 智能体执行层的典型攻击类型与防护要点

攻击类型	典型攻击场景描述	防护要点	对应章节
工具越权调用	智能体调用超出授权范围的工具或能力	工具白名单、权限控制、最小权限	7.1.2.4
恶意工具调用	通过提示词注入等方式诱导智能体执行恶意工具调用	调用链风险检测、提示词注入联动检测	7.1.2.4
工具参数攻击	篡改或构造恶意参数实施 SSRF、命令注入等攻击	参数类型、取值范围、目标地址校验	7.1.2.4
高危操作滥用	智能体执行数据外发、删除、修改或支付等高危操作	高危操作识别、二次确认、人机协同审批	7.1.2.4
工具链攻击	通过多个工具组合形成复杂攻击路径	调用链监测、行为分析、全链路审计	7.1.2.4

表E.5 列出了模型层的典型攻击类型与防护要点。

表 E.5 模型层的典型攻击类型与防护要点

攻击类型	典型攻击场景描述	防护要点	对应章节
模型窃取	通过大量查询获取模型输出并重构	调用行为分析、异常模式识别、频率	7.1.2.5

	或蒸馏模型能力	控制	
知识库窃取	通过大量或针对性查询获取知识库中的敏感信息	调用行为分析、异常模式识别、访问控制	7.1.2.5
无界消耗	通过持续发起高资源消耗请求，异常消耗人工智能服务资源	资源消耗监测、异常行为识别、限流与熔断	7.1.2.5

表E.6 列出了AI基础设施与供应链层的典型攻击类型与防护要点。

表 E.6 AI 基础设施与供应链层的典型攻击类型与防护要点

攻击类型	**典型攻击场景描述**	**防护要点**	**对应章节**
AI 组件漏洞利用	利用 Ollama、vLLM 等 AI 组件已知漏洞攻击人工智能服务	AI 组件指纹识别、漏洞库匹配、风险告警	7.1.2.6

表E.7 列出了网关与运行环境的典型攻击类型与防护要点。

表 E.7 网关与运行环境的典型攻击类型与防护要点

攻击类型	**典型攻击场景描述**	**防护要点**	**对应章节**
管理接口未授权访问	未经授权访问网关管理接口	管理员身份鉴别、访问控制、会话管理	7.1.2.2、7.1.2.3
暴力破解	持续尝试管理员账号或密码	登录失败防护、多因素身份鉴别	7.1.2.2
配置篡改	非法修改安全策略、规则或系统配置	配置权限控制、完整性保护、操作审计	7.1.2.3、7.1.3.1
日志篡改	删除或修改安全审计日志	日志完整性保护、加密存储、备份	7.1.2.4、7.1.3.2
关键文件篡改	修改配置文件、规则库、模型文件或升级包	完整性校验、可信来源验证	7.1.2.6
通信窃听与篡改	窃听或篡改网关与应用、人工智能服务之间的数据	安全传输协议、通信完整性保护	7.1.2.5
支撑系统攻击	利用网关运行环境中的不必要服务、端口或漏洞进行攻击	系统加固、攻击面收敛、支撑系统安全	7.1.2.6

附 录 F
（资料性）
人工智能服务接入对象

表F.1列出了人工智能服务接入对象示例，包括大模型服务、智能体服务和工具服务等。针对智能体互联、工具调用等场景涉及的标准协议，宜优先支持行业应用较为广泛的协议方案（如MCP、A2A等）的最新版本，并在一定程度上向前兼容旧版本。

表 F.1 人工智能服务接入对象示例

接入对象类别	具体接入对象	典型形态	可采用的接口或协议示例
大模型服务	大语言模型	生成式大模型服务	OpenAI API 等模型服务接口
	多模态模型	视觉语言模型、语音模型、图像生成模型等	模型服务 API、HTTP API 等
	模型推理服务	独立模型推理服务、模型服务平台等	OpenAI 兼容接口、HTTP API 等
智能体服务	智能体	对话式、任务型、编程 Agent 等	HTTP API、A2A 等
	智能体服务端	Agent Server、Agent Runtime 等	
	智能体应用	企业智能体、业务智能体等	
工具服务	插件及扩展工具	第三方插件、企业内部插件等	Function Calling、HTTP API 等
	函数及 API 工具	Function、REST API、Web API 等	
	数据访问工具	数据库查询、数据检索等	HTTP API、MCP 等
	知识检索工具	知识库检索、搜索服务、向量数据库查询等	
	系统操作工具	文件操作、代码执行、系统操作等	

附 录 G
（资料性）
大模型安全网关敏感数据分类与防护

G.1 总则

大模型安全网关在交互过程中涉及的敏感数据具有来源复杂、流转路径多变等特征，可从数据分类、数据模态、交互阶段、数据密级、应用场景、数据流向及安全控制等维度进行分析与管理。

G.2 敏感数据分类

大模型安全网关敏感数据的分类包括以下内容：

a） 个人敏感信息：包括但不限于姓名、身份证号、联系方式、生物特征、位置信息、账号信息等；

b） 商业敏感信息（可选）：包括但不限于商业秘密、经营数据、合同信息、源代码、内部文档、技术方案等。

G.3 数据模态分类

大模型安全网关敏感数据的模态包括以下内容：

a） 文本数据；

b） 图像数据；

c） 音频数据；

d） 视频数据；

e） 文件及结构化数据。

注：对于多模态数据，宜具备跨模态关联识别能力，以防止通过模态转换规避检测。

G.4 数据交互阶段划分

大模型安全网关敏感数据的交互阶段包括以下内容：

a） 输入阶段：指用户向大模型提交的数据，包括文本、多模态内容及上下文信息；

b） 推理阶段：指模型思考（Thinking）过程中的数据，包括推理内容、思维链及上下文缓存；

c） 输出阶段：指模型返回给用户或系统的数据；

d） 数据交互环节：包括 RAG 检索、联网搜索、工具调用等过程中产生或引入的数据。

G.5 数据密级划分

数据敏感程度可依据数据类别、保护要求、应用场景及数据泄露可能造成的影响等因素进行分级管理；涉及国家秘密的数据，其密级应按照相关法律法规及标准要求进行确定，可参考 GB/T 43697—2024。

G.6 数据流转

根据人工智能服务部署位置不同，存在以下数据泄露情形：

a） 境内模型：数据主要在本地或可控范围内流转，泄露风险相对集中于内部环境；

b） 境外模型：数据在处理过程中涉及跨境传输，存在对外流转带来的泄露风险。

G.8 安全控制维度

大模型安全网关对敏感数据的防护包括以下安全控制维度：

a) 识别：对敏感数据进行自动识别与分类；
b) 脱敏：对敏感信息进行掩码、替换或模糊处理；
c) 访问控制：基于身份与权限控制数据访问；
d) 传输控制：限制敏感数据跨境或跨域传输；
e) 审计：记录敏感数据访问与流转行为。

附 录 H
（资料性）
大模型安全网关测评方法

H.1 概述

大模型安全网关产品的测评工作可围绕其安全功能组件和自身安全能力开展，用于验证产品在人工智能应用安全防护场景下的适用性、有效性和可运行性。

测评内容包括安全评估准备和安全评估方法，可根据产品实际具备的安全功能和应用场景选择相应的测评内容，测评结果可作为评价产品安全防护能力的参考。

H.2 安全评估准备

H.2.1 安全规则及检测方法

安全规则、词典及检测资源宜满足以下内容：

a) 建立与安全防护功能相适应的规则库、敏感数据识别规则、内容安全规则、提示词注入检测规则等安全检测资源；
b) 安全检测资源宜覆盖产品所支持的主要安全防护场景，包括内容安全、提示词注入、敏感数据防泄漏、工具调用和模型攻击等。

H.2.2 测试样本

测试样本宜满足以下内容：

a) 内容安全宜覆盖符合 GB/T 45654—2025 相关要求的测试样本；
b) 提示词注入宜覆盖角色扮演、跨语言表达、编码或混淆等典型提示词注入类型，可包含单轮、多轮、直接注入和间接注入等多种注入方式；
c) 敏感数据测试样本宜覆盖个人信息、国家秘密及其他依法受保护的敏感数据，覆盖不同敏感数据类型的输入和输出场景；
d) 工具调用安全测试样本宜覆盖工具越权调用、恶意参数、异常目标地址、高危操作等场景；
e) 模型攻击行为测试样本宜覆盖模型窃取、知识库窃取、无界消耗等典型攻击场景，并可结合高频相似查询、异常调用模式等行为特征构造测试样本。

H.2.3 测试环境

测试环境宜与生产环境进行适当隔离，宜满足以下内容：

a) 无界消耗攻击测试宜设置持续调用、多轮交互、长上下文、智能体持续执行等动态资源消耗场景，并配置相应的调用次数、词元使用量、执行时长等测试条件。

H.2.4 AI 组件指纹及漏洞数据

对于具备AI组件安全防护能力的产品，相关数据资源宜满足以下内容：

b) 宜建立 AI 组件指纹库，用于识别人工智能服务相关 AI 组件及其版本等特征信息；
c) 宜建立与 AI 组件相对应的漏洞信息库，并根据漏洞公开情况和产品适用范围进行更新。

H.3 安全评估方法

H.3.1 安全防护功能评估

H.3.1.1 基础功能评估

H.3.1.1.1 测评方法

基础功能的测评方法包括：

a) 验证身份认证与访问控制功能，测试不同用户、应用及智能体的身份认证、权限控制和访问范围；
b) 验证人工智能服务统一接入、请求代理转发、参数处理及响应返回等功能；
c) 验证人工智能服务流量调度功能，测试不同人工智能服务、调用方及业务条件下的调度策略和异常切换能力；
d) 验证对经网关接入或调用的人工智能服务及相关资产进行识别、统计和展示的能力；
e) 验证 API 调用凭证有效性校验、请求频率、并发数量、词元使用量及配额等控制功能，并测试会话或任务粒度的调用限制能力；
f) 验证 API 调用异常行为控制功能，构造高频调用、异常调用量、异常来源调用等场景，测试异常行为的检测和处置能力；
g) 验证 API 资产识别及调用基线功能，测试对未登记、长期未使用或异常调用 API 的识别能力。

H.3.1.1.2 预期结果

网关基础能力的预期结果包括：

a) 未经授权的用户、应用或智能体不能访问未授权的人工智能服务或能力；
b) 人工智能服务请求能够按照配置策略进行代理转发和流量调度；
c) 经网关接入或调用的人工智能服务及相关资产能够按照产品声明的能力进行识别、统计和展示；
d) API 调用凭证、调用频率、并发数量、词元使用量、配额以及会话或任务调用限制等控制策略能够按照配置生效；
e) 异常 API 调用能够按照安全策略进行告警、限流、阻断或其他相应处置；
f) 未登记、长期未使用或异常调用的 API 能够按照产品声明的能力进行识别和提示。

H.3.1.1.3 结果判定

上述测评项目符合产品声明的功能要求，且配置的安全策略能够有效生效的，可判定相应测评项目符合；否则，可判定为不符合。

H.3.1.2 内容安全防护评估

H.3.1.2.1 测评方法

内容安全防护的测评方法包括：

a) 使用符合 GB/T 45654—2025 相关要求的测试样本，对人工智能服务输入内容和输出结果进行检测；
b) 测试文本、图片、音频、视频等不同类型内容的检测能力；
c) 对检测到的风险内容，验证提示、代答、拦截、替换、降级等处置功能。

H.3.1.2.2 预期结果

内容安全防护的预期结果包括：

a） 能够识别测试样本中的相应风险内容；

b） 多模态测试场景下，支持的内容类型能够按照产品声明的能力进行检测；

c） 检测到风险内容后，能够按照配置的安全策略执行相应处置。

H.3.1.2.3 结果判定

根据产品声明的检测范围和安全策略，对测试结果进行综合判断。检测及处置结果符合预期的，可判定相应测评项目符合。

H.3.1.3 提示词注入攻击防护评估

H.3.1.3.1 测评方法

测评方法为使用提示词注入样本进行测试，验证不同注入方式、不同注入场景下的检测情况。

H.3.1.3.2 预期结果

预期结果为能够识别不同注入方式、不同注入场景下的攻击行为。

H.3.1.3.3 结果判定

测试结果符合产品声明的检测和处置能力，且安全策略有效生效的，可判定为符合。

H.3.1.4 敏感数据防泄漏评估

H.3.1.4.1 测评方法

测评方法为使用不同输入、输出场景下的敏感信息进行测试，验证对敏感信息的识别和脱敏能力。

H.3.1.4.2 预期结果

预期结果为能够识别各类敏感信息，覆盖不同人工智能服务交互位置。

H.3.1.4.3 结果判定

测试结果符合产品声明的敏感数据识别和脱敏能力，且配置策略能够有效生效的，可判定相应测评项目符合。

H.3.1.5 工具调用安全评估

H.3.1.5.1 测评方法

工具调用的测评方法包括：

a） 验证对接入工具的白名单控制功能；

b） 测试异常工具调用及调用参数，验证对调用参数、目标地址等安全策略的检测和控制能力；

c） 检查工具调用过程的日志记录及调用链关联情况。

H.3.1.5.2 预期结果

工具调用调用的预期结果包括：

a) 未经授权的工具调用能够被限制；
b) 异常工具调用及不符合安全策略的调用参数能够被识别和处置；
c) 工具调用过程能够进行记录和追溯。

H.3.1.5.3 结果判定

上述功能按照产品声明和配置策略有效生效的，可判定为符合。

H.3.1.6 模型攻击行为防护评估

H.3.1.6.1 测评方法

根据产品声明的模型模型攻击行为防护能力,测试相关攻击场景下的检测、告警、限流或阻断能力，并检查相关安全事件的记录和追溯情况。

H.3.1.6.2 预期结果

能够识别产品声明支持的模型漏洞攻击或异常调用行为，并按照安全策略进行相应处置和记录。

H.3.1.6.3 结果判定

测试结果符合产品声明的检测和处置能力，且安全策略有效生效的，可判定相应测评项目符合。

H.3.1.7 AI 组件漏洞安全评估

H.3.1.7.1 测评方法

AI组件漏洞安全的测评方法包括：

a) 测试产品对 AI 组件及其版本等信息的识别能力；
b) 使用已知漏洞信息验证 AI 组件与漏洞信息的关联及风险提示能力；
c) 检查 AI 组件指纹库和漏洞库的更新机制及相关记录。

H.3.1.7.2 预期结果

能够识别产品支持范围内的AI组件及其版本，并根据组件信息关联相应漏洞，对存在风险的组件进行风险提示或告警，相关数据能够按照产品更新机制进行维护。

H.3.1.7.3 结果判定

相关组件识别、漏洞匹配、风险提示及数据更新功能符合产品声明的，可判定为符合。

H.3.2 自身安全能力评估

H.3.2.1 测评方法

自身安全能力的测评方法包括：

a) 根据 GB 42250—2022 第 5 章相关要求，检查产品的标识和鉴别、访问控制、安全审计、通信安全、支撑系统安全、产品升级、用户信息安全和密码等安全措施；
b) 测试管理员身份鉴别、登录失败防护、会话超时、多因素身份鉴别及管理权限控制等功能；

c） 检查安全策略、规则库、敏感词库、统计报表等重要管理对象的权限控制，以及管理员重要操作的审计、查询、导出、备份和完整性保护能力；
d） 检查产品通信安全、关键文件完整性保护以及配置和审计数据备份恢复、故障恢复等安全措施。

H.3.2.2 预期结果

自身安全能力的测评结果包括：
a） 管理员身份鉴别、会话管理和访问控制等安全措施能够有效生效，未授权用户不能访问受保护的管理功能；
b） 安全策略及重要管理对象能够按照权限进行访问控制，重要操作能够进行审计和追溯；
c） 产品通信及关键配置、凭证、规则库、模型文件、升级包等重要资源具有相应的安全保护措施；
d） 产品具备相应的数据备份恢复和故障恢复能力。

H.3.2.3 结果判定

相关自身安全措施符合产品声明及适用标准要求，且测试结果符合预期的，可判定相应测评项目符合。

参 考 文 献

[1] 陈阳, 李娟, 张磊. 大模型推理框架（Ollama/Vllm）安全漏洞分析与防护技术研究[J]. 网络安全技术与应用, 2025(02): 45-50.

[2] GB/T 35273—2020 信息安全技术 个人信息安全规范

[3] GB/T 41867—2022 信息技术 人工智能 术语

[4] GB/T 25069—2022 信息安全技术 术语