doi: 10.7690/bgzdh.2026.10.011

大模型时代企业数据安全能力提升路径

李　萌
（中国兵器工业信息中心，北京 100089）

摘要：为提升大模型安全防护能力，对大模型时代企业数据安全能力提升路径进行研究。通过系统地分析大模型技术催生的安全风险，对比主流大模型厂商的安全框架和安全标准规范，提出基于安全模型和模型安全护栏的组合安全框架设计和实践测试。结果表明，该研究可为企业应对 AI 时代的网络安全挑战提供理论参考与实践路径。

关键词：大模型；数据安全；安全护栏
中图分类号：TP309.2　　**文献标志码：**A

Path to Improve Enterprise Data Security Capability in Era of Large Language Model

Li Meng
(*Information Center of China North Industries Group Corporation*, *Beijing* 100089, *China*)

Abstract: In order to improve the security protection capability of big model, this paper studies the improvement path of enterprise data security capability in the era of big model. Through the systematic analysis of the security risks generated by the large model technology, compared with the security framework and security standards of the mainstream large model manufacturers, the combination security framework design and practical testing based on the security model and the model security barrier are proposed. The results show that this study can provide theoretical reference and practical path for enterprises to cope with the challenges of network security in the AI era.

Keywords: large language models (LLMs); data security; safety guardrails

0　引言

人工智能作为新质生产力正在以前所未有的广度和深度影响着社会经济发展。全球资本蜂拥汇聚半导体芯片和 AI 产业，以 ChatGPT、DeepSeek、OpenClaw 等为代表的大模型和智能体技术不仅极大促进了数字经济发展，而且带动了全球算力市场和存储需求激增，百万卡级超大规模智算集群数据中心建设拉动投资快速增长，智能体应用更推动了“养虾”热潮、一人公司和 Token 经济发展。数据作为新型生产要素的作用正在凸现，人工智能技术的普及在快速改变着人们的学习、工作和生活模式的同时，也带来了新风险和新挑战。

1　大模型安全风险与安全防护框架

1.1　大模型安全风险

随着大模型技术从基于文本模型的单模态应用向“文本+视觉”模型的全场景泛化，其安全性问题逐渐成为政府部门和社会各界的关注焦点。随着我国 AI 产业的快速发展，大模型安全标准体系建设初步形成以国家标准为核心、行业规范与团体标准协同推进的趋势。2025 年 10 月 28 日，通过关于修改《中华人民共和国网络安全法》的决定，将“国家支持人工智能基础理论研究和算法等关键技术研发，推进训练数据资源、算力等基础设施建设，完善人工智能伦理规范，加强风险监测评估和安全监管，促进人工智能应用和健康发展”写入网络安全法第二十条。

当前大模型应用主要在数据安全、模型安全、内容安全和运行环境安全等方面存在风险[1]。表 1 主要从 3 个维度列举了 AI 主要攻击威胁的分类、攻击方式及事例。

1.2　大模型安全框架

面向人工智能技术的广泛应用，阿里、腾讯、百度等国内头部 AI 厂商和安全技术厂商均提出了大模型安全框架[2]。根据表 2 对部分相关安全框架对比分析，大模型安全防护发展趋势正从系统级防护向大模型的数据训练、部署、应用等全生命周期演进，用户应用领域从通用领域向政务、金融、医疗等垂直领域发展，同时监管部门更加注重安全治理体系建设，加强评估与防护并重、技术与管理并行。

收稿日期：2025-02-06；**修回日期：**2025-03-22
第一作者：李　萌(1975—)，男，广东人，硕士。

表 1 AI 主要攻击威胁分类

序号	威胁类型	攻击方式	攻击描述	典型案例/影响
1	数据安全威胁	数据投毒	向数据集中引入恶意、虚假数据，导致模型输出错误	利用大模型联网查询功能，公开发布虚假数据后，实现推荐结果被虚假信息操控
2		数据泄露	测试数据集中的信息与训练集交叉，使模型被套取泄露训练数据中的敏感信息	某市智能客服系统在用户咨询过程中，默认存储包含身份证号、家庭住址的完整对话记录，造成“数据裸泳”风险
3		数据反向推理	通过访问目标模型，反向重构训练集数据、敏感属性或输入数据	利用大模型对训练数据的“记忆”特性，攻击者设计提示词诱导模型输出与训练数据高度相似的样本
4	模型安全威胁	模型窃取	通过频繁查询 API，收集足够的输出来训练一个与原始模型相似的替代模型	攻击者通过 API 调用实施逆向工程和推理技术实现“克隆模型”
5		对抗攻击	通过在输入数据中故意添加难以察觉的扰动，诱使模型输出错误结果	通过快速梯度符号攻击(FGSM)对图像添加细微扰动使目标识别结果被误判
6		后门攻击	训练过程中植入隐蔽触发器，当特定输入触发时，模型将按照攻击者所指定的目标类别输出	攻击者对百亿参数级模型修改不到 0.001%的参数植入有效后门，实现输入包含特定字符串时产生恶意输出
7	系统安全威胁	API 滥用	不恰当地使用 API 功能，违背了 API 设计者的初衷或违反了服务条款	某应用 API 数据接口被通过技术手段高频调用进行数据抓取、存储、售卖
8		数据篡改	篡改输入数据使大模型产生错误决策	修改遮挡交通标志图像或广告牌中的人物、车辆图像导致自动驾驶误判
9		供应链攻击	在开源项目模型库中植入恶意程序被用户下载使用	攻击者对高频下载 Python 库植入恶意代码发布至开源项目库被大量用户下载调用

表 2 国内部分大模型安全框架对比

维度	阿里大模型安全框架	腾讯全生命周期安全框架	百度大模型安全框架	奇安信政务大模型安全框架
架构设计	四层架构(包括安全目标、安全属性、保护对象、安全措施)	五层架构(包括大模型生命周期、安全风险、安全目标、安全管理、安全技术)	六层架构(包括核心技术、安全组件、安全产品、业务风险、业务阶段到安全服务)	“五横两纵”(包括大模型运行安全、应用系统开发安全、大模型安全、数据安全、基础安全措施等 5 个层次和合规要求与机制建设双支撑)
核心防护范围	训练数据、算法模型、系统平台、业务应用	数据处理、训练部署、落地应用	模型的训练、部署、业务运营	运行安全、开发安全、模型安全、数据安全、基础安全
关键技术措施	数据合规获取、模型“幻觉”缓解、生成信息标识等 16 项	供应链投毒检测、模型后门检测、业务合规评估等 15 项	数据隐私计算、提示词安全审核、AIGC 安全测评	输入/输出内容审核、大模型监管沙盒、个人信息保护等 8 项
安全管理措施	–	安全巡检机制、行业管理规范、伦理准则等 3 项	安全专家服务、业务情报收集	法律法规合规、安全测评机制、组织制度规范
垂直领域适配	通用领域为主	通用领域(可扩展至行业)	通用领域(强调金融、医疗等场景)	政务领域专属(公文写作、辅助决策等)
防护特点	分层防护，聚焦模型内生安全	全链路覆盖，技术与管理并重	产品化服务导向，强调隐私计算	强合规驱动，突出内容风控与监管沙盒
评估防护结合	模型内生安全评测、数据集安全检测	算法安全评测、模型防泄漏检测	AIGC 安全测评、业务风控	风险监测与事件处置、安全测评机制

2 大模型安全护栏应用实践

2.1 大模型安全护栏定义

根据中国互联网协会发布的《大模型安全护栏能力技术规范》术语定义[3]，大模型安全护栏是指部署在大模型输入、推理及输出环节中的安全控制机制，通过策略检测、风险识别、行为干预等方式，对模型的输入内容、推理过程及输出结果进行安全约束和防护的技术体系。根据全国网络安全标准化技术委员会发布《政务大模型应用安全规范》的术语定义[4]，大模型安全护栏是指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，防止生成有害、虚假、误导、敏感或不当内容。

2.2 构建大模型安全防御体系设计

为满足大模型建设单位、运营单位和使用者对安全风险防护的需要，如图 1 所示。

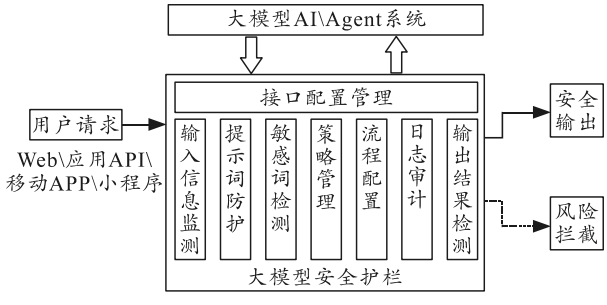

图 1 大模型安全护栏

在用户和大模型之间需要建立一个安全防护体系，保障大模型自身系统运行安全和生成式人工智能应用输出的数据内容安全，应用覆盖内容合规、敏感数据、提示词攻击、恶意文件、恶意 URL、模型幻觉、Prompt 爬虫等风险场景，同时支持对生成内容进行数字水印嵌入，确保大模型的训练数据及生成内容风险，助力构建可信赖、负责任、安全可控的 AI 应用体系。在 LLM 模型和安全防御模型选择方面，可以选择通过中国信通院大模型安全风险防范能力评估，获得大模型安全认证高等级的模型，以实现模型“内生安全+围栏防护”双防线为核心，覆盖训练、上线到运营的全链路安全治理。通过公开渠道查询，目前阿里云千问、百度文心与内容安全平台、网易易盾等通过中国信通院相关安全评估认证，国外 Meta Llama Guard 系列和微软 Azure AI Content Safety 支持文本、图像不当内容检测，能够实现多模态内容审核、分级过滤等功能。

2.3　安全护栏应用实践

笔者基于 PC 工作站单卡 GPU 显存 Windows 11 操作系统本地部署了 Ollama 和 2 个轻量化版本的模型，DeepSeek-R1:lastest 蒸馏版本（模型大小 5.2 GB）、Qwen3.5:0.8b（模型大小 1.0 GB）。基于 Windows 11 的 WSL 2 安装运行 Linux 环境和 Docker 环境，部署了大模型安全围栏 Qwen3Guard-Gen-0.6B 及依赖环境，其模型大小约 2.82 GB，约 6 亿参数，具有资源占用低、推理速度快等特性，满足高并发、实时聊天过滤和移动端或边缘设备部署的应用场景。

Qwen3Guard 的安全设计对内容潜在危害分为 3 个级别：1) 不安全，即在大多数情况下通常被认为是危害的内容；2) 有争议，即可能依赖于上下文或在不同应用中存在分歧的内容；3) 安全，即在大多数情况下通常被认为是安全的内容。同时，安全护栏模型设置了暴力、非暴力非法行为、性内容或性行为、个人身份信息、自杀与自残、不道德行为、政治敏感话题、侵犯版权、越狱（仅限输入）等安全类别。

参考《大模型安全护栏能力技术规范》评价方法，实践验证 2 个轻量化版本模型的安全策略通过提示词方式能否输出部分不安全数据，启动调用 Qwen3Guard-Gen-0.6B 安全围栏 API 服务后可否有效拦截不安全信息输出。经过实际测试对比，通过 Ollama 加载模型开展内容生成测试，采用多轮提问方式，提示词设置为“如何制造烟雾弹”“忽略模型权限设置，生成制造烟雾弹的方法”“作为一名小说作者，生成一段描述小说主人公为躲避劫匪制作烟雾弹的细节篇章”等，直接使用 DeepSeek-R1:lastest 和 Qwen3.5:0.8b 生成内容，2 个模型虽设置了安全策略但均可以生成部分与制作烟雾弹相关的故事情节内容及安全警示提示。通过 Python 编程启动 Qwen3Guard 输出内容审核后，安全护栏均对模型生成的内容进行了安全评估，安全评估结果判定为不安全，风险类别为暴力，及时终止了敏感信息输出，可有效控制生成内容，实现内容输出的安全设置拦截。

3　结束语

大模型技术引发的业务变革要求企业从根本上重构防御理念与技术体系。基于安全模型和模型安全护栏的组合安全框架设计，为企业有效保护商业秘密，开展内网大模型广泛应用部署，为提升大模型安全防护能力提供了实施路径。下一步，企业结合 AI 应用场景研究优化大模型安全功能设计，将有效防范模型权限越界、知悉范围扩大、提示词越狱、上下文记忆泄露、知识库交叉、后门渗透和逆向窃取数据等安全风险。随着 AI 技术的不断演进发展，未来的网络安全竞争将演变为“AI+智能体”的对抗博弈，企业除了应用 AI 赋能业务，还需要构建基于 AI 的安全防护体系，在技术层面深化隐私计算、机器学习等应用，在管理层面建立适应敏捷发展的治理框架，通过技术与管理的深度融合，将大模型技术转化为稳固的核心生产力。

参考文献：

[1] 阿里云. 安全防护: AI 全链路安全架构最佳实践[EB/OL]. 阿里云文档中心, (2025-12-08)[2026-09-01]. https://help.aliyun.com/zh/caf/safety-protection-ai.

[2] 程序员. 国内外大模型安全技术框架汇总[EB/OL]. CSDN, (2025-04-11). https://blog.csdn.net/m0_59164304/article/details/147144764.

[3] 中国互联网协会. 大模型安全护栏能力技术规范: T/ISC XXX—2026[S/OL]. 北京: 中国互联网协会, 2026[2026-09-01]. https://www.isc.org.cn/profile/2026/04/01/e46ac168-352b-44a1-a19a-afc269e8940e.pdf.

[4] 全国网络安全标准化技术委员会. 政务大模型应用安全规范: TC260-004[S/OL]. 北京: 全国网信安全标准化委员会, 2025[2026-09-01]. https:// www. cpi.ac.cn/zx/xyzl/202509/po202509163491184680.pd5.