检察院大模型网络安全防护路径探索

洪我杭
北京市人民检察院第三分院，北京 100022
DOI: 10.61369/TACS.2026050021

摘　　要： 随着人工智能技术飞速发展，大模型技术在检察业务中的应用价值不断提升，尤其在智能阅卷、辅助量刑、类案推送等场景中发挥了重要作用，显著提升了办案效率。但随着大模型普及应用，检察院面临的网络安全风险也进一步增多。本文通过分析检察院大模型应用的基础架构与部署方案，系统梳理其在训练数据层、输入层、输出层以及治理环节面临的各类安全风险，并结合当前主流安全产品技术路线，提出“边界防护—内容检测—行为审计—智能对抗”的四维联动防护体系，以此打造“AI 对抗 AI”的新型防护机制，为检察机关数字化转型筑牢安全底座。

关 键 词： 检察大模型；网络安全；敏感信息防泄漏；AI 对抗 AI；安全智能体

Exploration of Cybersecurity Protection Paths for Procuratorate Large Models

Hong Wohang
The Third Branch of People's Procuratorate of Beijing Municipality, Beijing 100022

Abstract: With the rapid development of artificial intelligence technology, the application value of large model technology in procuratorial business has been continuously improved. It has played an important role especially in scenarios such as intelligent case review, auxiliary sentencing, and similar case recommendation, significantly enhancing case-handling efficiency. However, with the popularization and application of large models, the cybersecurity risks faced by procuratorates have further increased. By analyzing the basic architecture and deployment schemes of large model applications in procuratorates, this paper systematically sorts out various security risks existing in the training data layer, input layer, output layer, and governance links. Combined with the technical routes of current mainstream security products, it proposes a four-dimensional linked protection system of "boundary protection – content detection – behavior audit – intelligent countermeasure", so as to build a new protection mechanism of "AI vs. AI" and lay a solid security foundation for the digital transformation of procuratorial organs.

Keywords: procuratorate large models; cybersecurity; sensitive information leakage prevention; AI vs. AI; security intelligent agent

引言

生成式人工智能大模型正在全面重塑现代司法工作模式。在检察领域，智能阅卷、类案检索、量刑辅助、证据审查、法律文书生成等场景中已经开始运用大模型技术，并且为法律监督质效提升提供了重要驱动力。然而，技术红利的释放必然同步伴随着新的安全风险。大模型一旦嵌入案件管理平台、检察工作网等系统，具备一定的内部数据调用权限，就意味着传统网络安全界限就此突破，必须警惕各类新型网络攻击。《大规模语言模型应用 Top 10》中提出，提示注入、模型越狱、过度代理是当前大模型系统应用面临的主要风险，而在司法场景下，该类风险将进一步演变为篡改证据、泄密案情、误导裁决等问题。因此，在检察院应用生成式人工智能辅助工作的同时，必须全面升级网络安全防护体系，以此保证其数据安全。

一、检察院大模型面临的网络安全风险

（一）训练数据层风险：数据投毒与虚假信息传导

在检察机关应用场景下，大模型训练数据主要包含法律法规库、真实检察案例、各类裁判文书、内部工作规范等。但在训练数据采集、标注、清洗等环节，一旦受到恶意投毒污染，就会使得大模型学习的法律知识出现错误，甚至其认知出现偏差，进而对后续的检察工作产生不可预估的影响。具体来说，攻击者可以通过向开源法律数据集投递错误法条、虚假信息，或者大范围虚构裁判观点等方式，影响检察机关调用的第三方预训练模型，进而埋下安全隐患[1]。比如在司法审判场景下，大模型因受到污染而生成了不准确的法律意见，或者给出了被篡改的分析结果，很可

能对检察官的专业判断产生误导。

（二）输入层安全风险：提示注入与权限越界

检察院大模型的数据输入层是最直接的攻击节点。首先，攻击者可以通过注入提示词，构造对抗性提示，从而达到破坏系统防护的效果，由此即可对大模型进行诱导，使其输出内容或行为逻辑受到影响。在检察场景中，攻击者可以通过咨询案件等方式进行伪装，以此绕过检测，达到获取案件信息甚至调用内部数据的目的。其次，攻击者也可以借助角色扮演、分段拼接、隐喻表达等方式实现模型越狱攻击，迫使大模型无视内置安全协议。比如攻击者可以伪装成“法学研究者”，对大模型提出内部案件评议过程的模拟请求，由此窃取其敏感决策逻辑[2]。此外，过度代理也是检察院大模型面临的重要风险。检察机关通过赋予大模型工具调用权限，可以实现信息查询、法律检索等服务功能。攻击者可以利用权限设计缺陷，通过提示注入的方式篡改数据，从而将安全风险传导至业务系统深层。

（三）输出层安全风险：敏感泄露与不当处理

大模型的输出信息不可完全预测，因此在输出层同样面临着敏感信息泄露的核心风险。在检察院应用场景下，大模型很可能会在输出环节无意带出案件侦查细节、涉案人员身份信息、内部审查意见等关键信息，甚至攻击者可以通过持续多轮次的对话，将碎片信息逐步拼凑完整，以此构建案件画像[3]。此外，大模型还面临着不当输出处理风险，即当大模型生成的回复信息交由下级组件自动执行时，如果没有经过验证，很容易引发脚本、服务端请求伪造等二次攻击。

（四）治理维度风险：服务滥用与供应链安全

在治理环节，检察院大模型还需要面对模型盗窃、拒绝服务攻击、供应链漏洞等多元风险因子。模型盗窃风险表现为攻击者通过大量针对性查询，逆向还原大模型的知识边界与推理逻辑，从而使得模型资产变相流失，甚至被人用于生成虚假文件。拒绝服务攻击表现为攻击者通过变长输入洪水、递归上下文扩展等查询方式，大量消耗模型算力，从而使得正常业务无法获得响应[4]。供应链漏洞在于检察机关在大模型构建中可能会选择第三方基础模型，同时还会引入开源组件或者外部数据源，但每一个环节一旦被留下后门，都将引发不可预估的问题。

二、检察院大模型网络安全防护策略

（一）部署大模型应用防火墙，构建内容级安全闸门

作为大模型推理业务应用场景，检察院应全面部署大模型应用防火墙（MAF），由此建立专用安全网关，可以通过在线部署、实时拦截等方式，对大模型输入输出的内容进行深度检测。针对检察院网络架构，大模型应用防火墙可以采用 API 代理式或者应用前端两种模式。

首先，大模型应用防火墙系统的核心优势在于防护提示词注入攻击。提示词注入攻击主要依托自然语言、上下文设计等方式完成攻击，传统安全防护工具基于特定规则无法识别，但大模型系统可以有效识别其提示词注入问题。在检察院场景下，攻击者往往会借助同义替换、语言混合、情境设计等方式完成注入，而 MAF 可以借助多引擎检测架构完成识别和拦截。其中显式高速内容检测引擎可以识别关键词是否带有恶意或异常格式；多级 Transformer 语义检测引擎可以通过自然语言分析其是否存在攻击意图，并对上下文内容建立宏观认知[5]。

其次，MAF 针对敏感信息泄露问题具有良好的防护效果。防火墙系统可以对大模型输入输出数据全面解析，进而根据信息类型库完成分类配置。在检察院系统中，可以将身份证、手机号、案件编号、涉案人姓名、内部机构代码、文书文号等各类信息纳入特定信息类别，进而通过脱敏或替代等方式避免信息泄露。

最后，MAF 还可以有效防止大模型应用暴露。检察院大模型业务繁杂，其中管理后台、Web 接口、API 端点等都存在攻击节点。而 MAF 可以通过内置防火墙，规避防御命令注入、跨站脚本、SQL 注入、服务端请求伪造等攻击[6]。此外，针对大模型资源消耗型攻击，MAF 还可以采用动态令牌技术进行防护，确保每个访问者都有唯一令牌，从而有效拦截自动化脚本产生的恶意请求。

（二）引入 AI 对抗 AI 机制，实现智能威胁检测与响应

随着生成式 AI 技术的发展与应用，在网络安全防护领域，“以大模型技术对抗大模型风险”已经成为共识，其技术路线在于建立专用于安全检测的大模型，通过模型蒸馏与微调辅助，进而将通用大模型转化为小型专用模型。

在安全检测大模型构建环节，应按照“教师模型训练—知识蒸馏—任务微调”的逻辑实施。第一步应利用大规模安全数据集训练出教师模型，通过输入提示词注入、过度代理、敏感信息泄露等各项攻击的模式与案例信息，由此分析其攻击行为特征和语言表征，进而借助软标签知识迁移训练学生模型[7]，并根据检察业务需求进行微调，使其适配司法领域的使用习惯。

例如国内某安全公司研发的威胁检测智能体，即实现了将专家级研判能力系统化、模型化的目的。该智能体具备意图理解与推理规划的能力，因此可以在防护活动中自主制定分析策略，进而可以根据单条告警信息，对相关所有日志、情报、行为进行溯源检索，从而还原攻击路径。在检察院应用场景下，该模型将具备主动威胁发现的能力，实现还原攻击全貌、指导精准处置的效果。

（三）强化端侧安全管控，筑牢智能体落地根基

随着大模型向智能体形态转变，其将会成为检察办公终端中的重要助手，但同时也将安全风险引入终端设备环节。检察院终端设备中存储着大量保密信息、卷宗与工作文档，一旦泄露将造成灾难级危险。

第一，在设备系统安全管理方面，应坚持“管住手”原则，即确保智能体不会越权操作。具体来说，应建立细粒度权限管控机制，明确智能体访问的文件目录、可调用的接口规范以及可执行的命令范畴[8]。在检测到恶意命令时，智能体监控系统要迅速熔断，避免出现高危操作。

第二，在运行环境安全构建方面，应坚持“安全沙箱”建设原则，即将智能体运行环境与终端核心资源进行隔离处置，具体

来说可以采用进程隔离、文件系统虚拟化、网络访问控制等手段实现[9]。在检察院内网闭环场景下，该设计方案具有良好的适配特征。

第三，在自身免疫安全防护方面，应坚持“管住嘴”原则，即建立审核系统对智能体输出内容进行过滤与审计。比如可以通过敏感文件检查系统，避免智能体在回复时引用涉密信息；也可以通过高危命令熔断系统，阻止智能体的恶意删除、格式化等操作。

（四）构建全链路审计体系，实现风险可视化与可追溯

可审计、可追溯、可归责是大模型安全防护体系构建的终极标准。在检察院工作场景下，大模型输出的每一条信息都可能与案件相关，因此必须构建全链路的审计体系，形成完善的追溯系统与可视化机制。

在全链路审计系统设计中，应涵盖对话内容审计、风险行为审计、系统交互审计三个层面，从而将用户对话、异常事件、操作命令等信息进行记录与审核。在风险可视化系统建立时，可以基于 AI 的基线建模与聚类分析，掌握检察院常规业务行为模式，进而对异常活动主动发出警告[10]。同时，可以借助可视化仪表，将风险分布、攻击趋势、模型健康度等指标进行图表化呈现，从而为安全管理决策提供明确的数据支撑。

三、结语

综上所述，大模型技术正在重塑检察院业务体系的生产格局与生态结构，但在提升效率的同时更要同步演进安全防护系统，以此避免风险失控。本文分别从风险识别与防护策略入手，详细剖析了检察院大模型网络安全防护体系的构建方案，通过 MAF 大模型应用防火墙、AI 对抗 AI 机制建设、端侧安全管控机制以及全链路审计体系等策略，为大模型应用提供保障。需要注意的是，技术防护手段并非万全之策，还需要检察机关建立完善的全流程安全管理制度，制定安全标准与应用规范，以此构筑坚不可摧的安全防线，让大模型真正成为司法公正的助力。

参考文献

[1] 陈晶 . 专题导读：大模型赋能网络安全，从对抗防护到体系重塑 [J]. 计算 ,2025,1(03):6-7.

[2] 鲁坤 , 郜忠平 , 王治录 . 人工智能促进检察机关数字化与信息化转型 [J]. 法治时代 ,2025,(06):81-83.

[3] 王颖 . 人工智能在刑事检察中的应用研究 [D]. 黑龙江大学 ,2025.

[4] 符能 . 基于网络安全大模型的网络安全防护技术研究 [J]. 数字通信世界 ,2025,(04):31-33.

[5] 王新华 , 徐清波 , 严若飞 . 网络安全态势感知平台实时监控与响应机制研究 [A] 第 39 次全国计算机安全学术交流会论文集 [C]. 中国计算机学会 ,《信息安全研究》杂志社 ,2024:3.

[6] 陈冬妮 , 胡浩翔 . 人工智能嵌入智慧检务的实践与展望——检察机关法律监督的科技智慧 [A] 第五届全国检察官阅读征文活动获奖文选 - 优秀奖 [C]. 最高人民检察院法律政策研究室 ,2024:10.

[7] 谢毅特 . 数字治理视角下基层检察院智慧检务建设研究 [D]. 江西师范大学 ,2024.

[8] 丁遥遥 , 徐玉洁 . 盐城经开区检察院高质效办理网络金融案件 [N]. 江苏法治报 ,2023-08-21(001).

[9] 田聪 . 成都市 W 区检察院信息技术外包存在问题及对策研究 [D]. 四川大学 ,2023.

[10] 强化网络安全保障力促智慧检务健康发展 [N]. 检察日报 ,2021-06-19(003).

