

人工智能网络安全告警研判的探索与实践

Explore and Practice on AI-Based Cybersecurity Alarm Judgment

郑浩远, 丁颖睿, 张润民 (中国联通北京分公司, 北京 100010)

Zheng Haoyuan, Ding Yingrui, Zhang Runmin (China Unicom Beijing Branch, Beijing 100010, China)

摘要:

大模型智能体的规模化应用带来攻击、漏洞、数据泄露等安全风险,海量告警造成人工研判低效、误报率高。基于中国联通某分公司的实践,分析智能体防护需求与技术痛点,介绍自研防护平台架构、核心技术,重点说明平台依托AI实现告警智能降噪,并结合真实场景验证效果,研判行业发展趋势。实践证实,“AI防护网络安全”全生命周期防护体系可自动降噪告警,减少人工投入,提升防护能力、降低运营成本,为智能体安全落地提供技术支撑。

关键词:

人工智能; 运营商; 网络安全; 告警研判

doi: 10.12045/j.issn.1007-3043.2026.09.014

文章编号: 1007-3043(2026)09-0074-05

中图分类号: TN918

文献标识码: A

开放科学(资源服务)标识码(OSID):



Abstract:

The large-scale deployment of large model agents brings security risks including malicious attacks, configuration vulnerabilities and data leakage. Massive security alerts lead to low efficiency and high false positive rates in manual threat analysis. Based on the practical experience of a branch of China Unicom, the core security requirements and technical pain points of agent protection are analyzed, the architecture and core technologies of the self-developed security protection platform are introduced, and its AI-powered intelligent alert noise reduction capability is emphasized. The effectiveness of the platform is verified through real network scenarios, and the future development trends of this field are discussed. Practical results prove that this full-lifecycle protection system featuring "AI for net security" can automatically filter redundant alerts, reduce manual workload, strengthen overall security defense and cut down security operation costs, which provides reliable technical support for the secure implementation of large model agents.

Keywords:

Artificial intelligence; Telecom operator; Cybersecurity; Alert judgment

引用格式: 郑浩远, 丁颖睿, 张润民. 人工智能网络安全告警研判的探索与实践[J]. 邮电设计技术, 2026(9): 74-78.

1 概述

当前,“业务+大模型”是企业数字化转型的核心战略,DeepSeek、Manus等通用智能体已在安全运营、运维、客户服务等场景规模化落地,同时也带来了各类网络攻击、权限漏洞、敏感数据泄露等安全隐患^[1]。

现有防护方案仅聚焦模型自身安全,无法覆盖智能体全生命周期,难以适配多组件协同的复杂业务场景,海量告警也带来人工研判低效、误报率高的难题^[2]。

为此,中国联通某分公司联合合作单位研发了智能体安全防护平台,内置人工智能降噪系统,搭建覆盖智能体全生命周期的安全防护体系,从配置、运行、数据多维度保障智能体安全,依托AI能力完成告警智能降噪,大幅降低人工研判压力^[3]。

收稿日期: 2026-07-07

2 大模型智能体安全防护核心需求与技术痛点

在核心安全需求方面,企、事业单位智能体安全防护分为以下4点:第一,全生命周期防护,覆盖智能体全流程与模型、插件等多个实体;第二,精准测评,依托个性化配置生成专属用例,精准挖掘漏洞;第三,轻量化动态加固,增强智能体抗攻击能力且不影响业务;第四,高兼容适配,兼容各类开源、商用智能体与大模型接口^[4]。

现有安全防护存在以下4项技术痛点:第一,恶意图样本检测精度不足,无法识别智能体定向攻击^[5];第二,智能体配置繁杂,提示词、插件参数类漏洞难定位^[6];第三,多组件协同防护困难,易产生权限、数据交互泄露风险^[7];第四,安全运营低效,人工处理海量告警日志误报多^[8]。

3 智能体安全防护平台核心技术与架构设计

3.1 平台核心理念

平台的设计主线为“以AI防护网络安全,告警研判为核心底座”,四大设计原则如下。

- a) 全域统一防护。融合传统网络安全、大模型智能体安全,统一采集、分析、处置所有告警^[9]。
- b) 业务场景导向。针对政企客户、5G核心网、AI智能体运维3类场景定制研判与加固规则^[10]。
- c) 轻量化低开销。告警降噪、智能体加固均采用轻量模型,不改造现有业务系统^[11]。
- d) 弹性可扩展。容器化部署,支持新增安全设备、大模型智能体快速接入^[12]。

3.2 系统架构设计

系统采用4层分层架构,自上而下为感知层、数据层、技术层和应用层(见图1),完整支撑告警全流程自动化研判,智能体安全检测模块内嵌于技术层协同运行。

a) 感知层。采集防火墙、IPS、WAF、蜜罐、DLP等网络设备的告警、全网流量、设备日志,同步采集智能体调用日志、MCP工具交互记录 and RAG知识库访问日志,实现传统+AI安全数据全覆盖实时采集。

b) 数据层。完成多源数据清洗、字段标准化、特征工程、时序存储等操作,统一输出结构化告警特征向量,为下游机器学习、深度学习模型提供标准化训练与推理数据集。

c) 技术层(核心)。集成多类AI模型、规则引擎、图神经网络、智能体安全测评加固组件,实现告警降噪、威胁分类、关联溯源、智能体漏洞扫描、自动研判决策等功能。

d) 应用层。面向安全运维人员提供告警可视化大屏、智能体风险台账、自动处置指令下发、审计报告、人力运维优化看板。

3.2.1 数据采集与预处理模块

多源采集方式包括API实时推送、Syslog日志抓取、网络流量协议解析、智能体接口埋点。预处理流程为:脏数据剔除、缺失值填充、字段归一化、离散特征编码、时序对齐,最后输出统一维度的告警特征向量,供下游模型调用。

3.2.2 智能告警筛选与降噪模块

本模块可解决海量告警高误报的痛点,采用机器

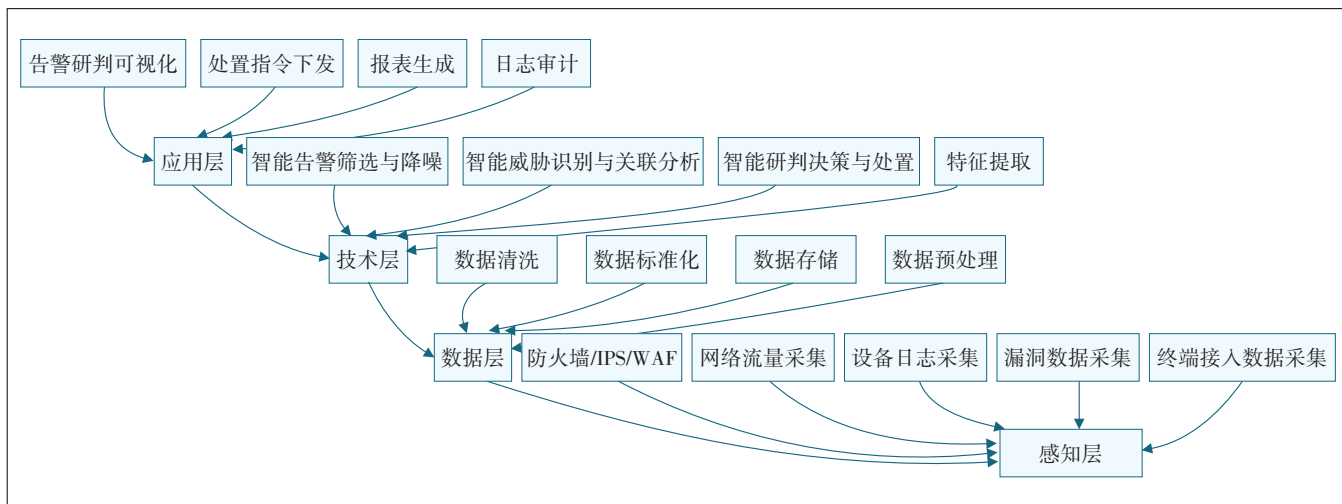


图1 人工智能告警研判系统总体架构

学习融合模型+规则引擎双链路降噪。智能告警筛选与降噪模块逻辑流程如图2所示。

机器学习融合模型训练参数如下。融合模型包含逻辑回归、随机森林和XGBoost这3类算法,训练数据集为中国联通某分公司近6个月的全网历史日志,总量为2.4亿条^[13]。

a) 特征维度。单条告警提取38维结构化特征,包含告警类型、威胁等级、源IP/目的IP、端口、访问时序、设备归属区域、智能体调用标识、MCP工具ID、知识库访问标记等。

b) 数据集划分。按时间分层抽样,训练集占70%,验证集占20%,测试集占10%,严格按周切分时序数据,避免时间泄露。

c) 超参数配置。对于逻辑回归算法,正则化系数 $C=0.1$,最大迭代次数为1 000,分类阈值为0.5;对于随机森林算法,决策树数量为120棵,单树最大深度为16,叶子节点最小样本数为5;对于XGBoost算法,学习率为0.08,树深度为10,迭代轮数为300,L2正则系数为1.2。

d) 模型更新策略。采用周度增量更新,每周导入新增7天告警数据进行重训微调,月度全量重训,适配新型攻击告警特征。

e) 输出结果。模型输出三分类标签,分别为有效高危告警、低危/误报告警以及重复告警。

该模块的工作流程为:将预处理后的告警数据输入融合机器学习模型完成初步降噪,输出疑似有效告警;再进入规则引擎二次校验,规则包含重复告警判定、低危扫描过滤、核心网告警优先级配置;最终输出有效告警列表,同步统计误报数据生成优化报告,指导防火墙、智能体配置调优。

3.2.3 智能威胁识别与关联分析模块

该核心分析模块通过融合深度学习与特征匹配实现威胁识别。该模块依托CNN-LSTM模型,结合中国联通自有攻击、漏洞特征库判定威胁类型与等级;通过CNN搭建告警关联图挖掘多要素隐藏关联,识别复杂威胁;支持攻击溯源,可定位攻击源头与传播路径,支撑安全处置。基于CNN-LSTM的威胁识别模型结构如图3所示。

搭建CNN-LSTM混合深度学习模型,提取告警特征、网络流量特征和漏洞特征,输出攻击类型与威胁等级。模型训练细节如下。

a) 特征维度。输入特征向量维度为64维,融合静态资产特征与时序流量特征;CNN层负责提取局部空间特征,LSTM捕获攻击时序演化规律^[14]。

b) 数据集划分。在时序划分方面,训练集占75%,验证集占15%,测试集占10%;融入该分公司的自有攻击特征库、漏洞库与历史攻防案例。

c) 训练超参数。对于优化器Adam,初始学习率

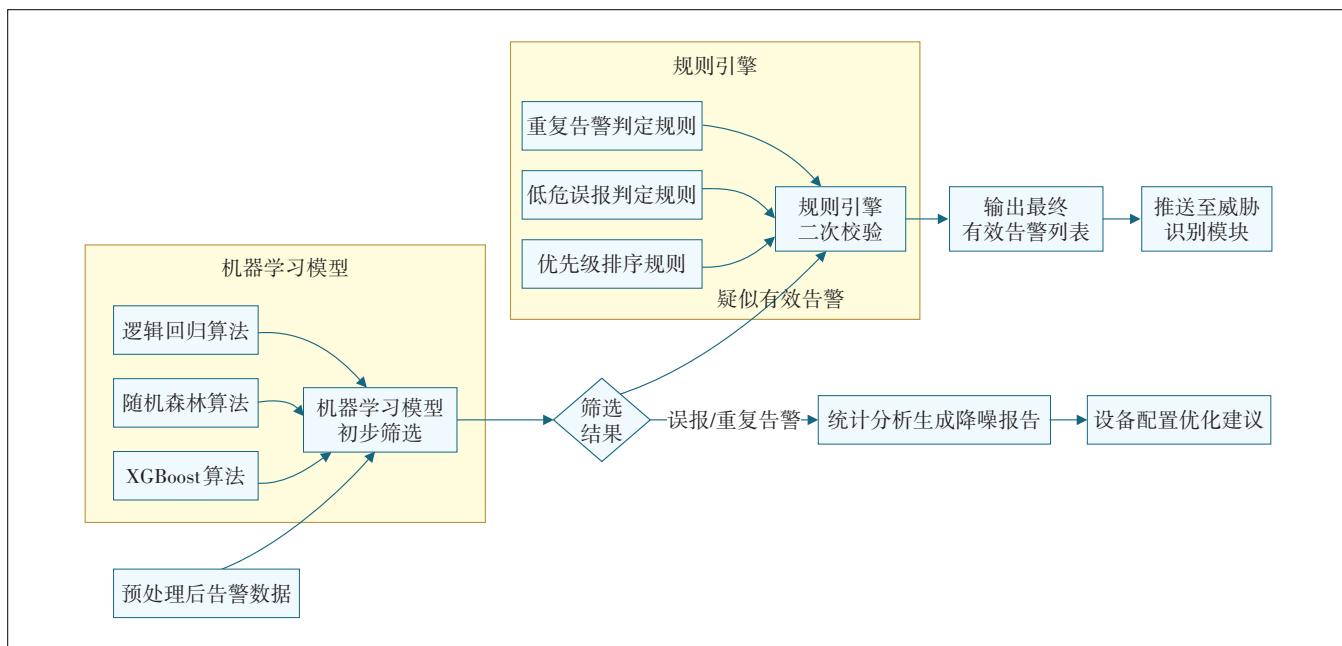


图2 智能告警筛选与降噪模块逻辑流程

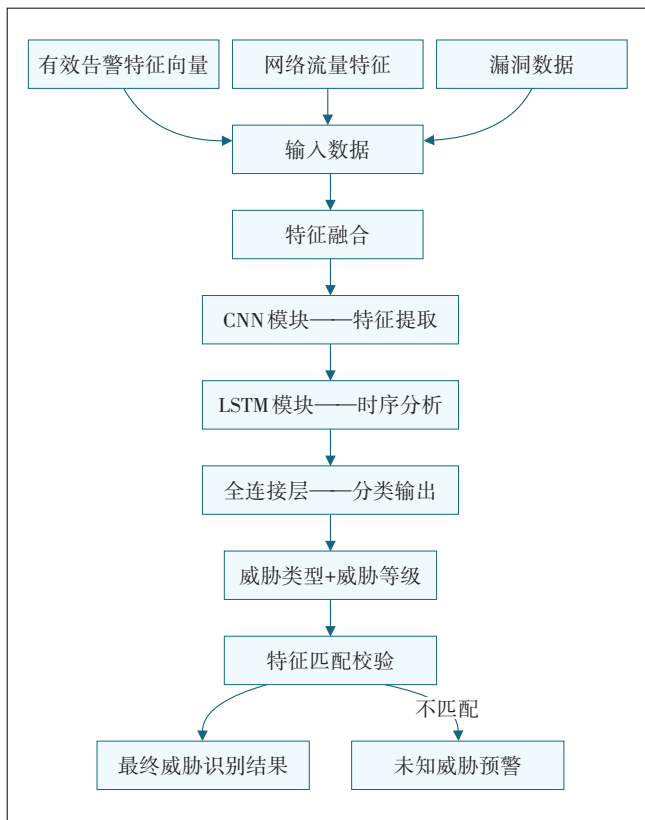


图3 基于CNN-LSTM的威胁识别模型结构

为0.001,损失函数采用多分类交叉熵;批次大小为128,最大迭代轮数为200,启用早停策略防止过拟合,Dropout系数为0.3。

d) 模型更新策略。双周增量微调,季度全量训练。

e) 加速机制。并联特征匹配引擎,实时告警与已知攻击特征库快速匹配,加速常规攻击识别。

依托GNN构建告警关联异构图,以告警、IP、设备、漏洞等实体为节点,实体间关联关系为边,挖掘隐藏关联,识别APT等复合攻击场景。GNN训练参数为:选用GCN图卷积网络,节点特征维度为42维,图卷积层数为3层;学习率为0.0008;滑动窗口按5min来动态构建实时告警图。

该模块支持攻击溯源,可综合攻击IP、时序、传播路径定位攻击源头、攻击手段与影响范围,为DDoS流量清洗、IP封堵等处置提供精准支撑^[15]。

3.3 核心防护技术实现

3.3.1 智能体安全测评技术

采用红队对抗测试技术,依托安全大模型根据智能体提示词、RAG知识库生成定制测试用例;支持自

定义测评规则,同步输出智能体风险对应的新增告警优化策略。

3.3.2 配置加固技术

针对智能体提示词、插件参数、模型权限开展漏洞扫描,依托大模型自动输出加固提示话术与权限配置优化方案。该方案无需修改智能体底层代码,能够有效减少AI风险类告警数量;相关加固规则支持根据中国联通政企安全策略动态启停。

3.3.3 多实体协同防护技术

统一扫描模型、MCP工具、知识库3类实体风险,部署流量网关过滤恶意用户提问;对接中国联通ISMP平台,将智能体风险直接转化为标准化安全告警,纳入统一研判链路,实现风险—告警—处置闭环。

3.3.4 智能日志研判技术

结合RAG攻防知识库进行日志语义降噪,依托前文机器学习、深度学习模型压缩无效告警;基于安全样本持续微调Prompt,降低整体误报率,适配运营商TB级海量日志分析需求。

4 平台应用实践与效果验证

4.1 应用场景与测试方案

本系统在中国联通某分公司落地应用。通过对比部署前后3个月的运维数据完成效果验证,评测指标包括告警误报率、有效告警识别率、研判处置效率、人力成本等,数据源于运维台账与平台自动统计,共6名运维人员全程参与核验,保障结果可靠。

4.2 量化效果分析

部署前,该分公司的日均告警为42046条,误报率高达92.3%,有效告警为3239条;部署系统后,平台自动过滤海量无效告警,推送的人工处置告警降至3829条,其中有效告警提升至3147条,误报率降至17.8%,有效告警识别率提升至82.2%,误报与重复告警总量削减98.2%,大幅缓解传统人工研判高误报、海量告警处置压力,具体数据对比如表1所示。

部署前,该分公司需7名安全运维人员专门负责

表1 告警降噪效果对比

测试指标	部署前	部署后	提升/下降幅度
日均告警数量/条	42 046	3 829	-90.9%
误报、重复告警/条	38 807	682	-98.2%
误报率/%	92.3	17.8	-74.5%
有效告警数量/条	3 239	3 147	-2.84%

告警研判工作,运维人员工作强度大,易出现漏判、误判;部署系统后,仅需4名兼职运维人员即可完成全网告警研判与处置工作,运维人员工作重心转向复杂威胁处置、系统渗透测试等高端工作,工作质量与效率显著提升。

4.3 定性效果分析

本系统可替代传统人工研判,实现告警智能自动分析,提升网络安全防护精准度、实时性,满足首都网络安全的严苛标准;可兼容现有安全平台,搭建全流程智能防护闭环,为安全策略优化提供数据支撑;全面强化该分公司的威胁处置能力,守护政企与用户通信、数据安全,筑牢首都数字安全屏障。

5 未来发展趋势与展望

结合大模型智能体规模化应用及运营商网络安全业务需求,将从以下3个方面迭代优化智能体安全防护平台。第一,基于GRPO、DAPO强化学习算法定制专项训练策略,强化5G核心网、政企专线场景下的长链安全推理与复杂攻击溯源能力;第二,依托MCP、A2A协议打通各类自有安全检测工具,搭建运营商专属安全工具生态,适配多设备、多客户防护场景;第三,自研多智能体自动化框架,结合无监督异常检测算法,实现安全测评、漏洞修复、告警研判全流程自动化,精准识别未知攻击,压降安全运营人力成本。

后续将持续深化大模型与通信安全技术融合,迭代平台核心防护与研判能力,落地全自动AI渗透测试、政企专属智能安全运营等创新场景。同时,联动行业伙伴与政企客户沉淀落地经验,输出通信行业智能体安全防护标准与解决方案,推动技术标准化、规模化应用,为政企数字化转型与城市数字安全筑牢通信安全屏障。

6 结束语

海量告警带来人工研判效率低、误报率高的突出问题。本文提出全生命周期智能体安全防护平台,该平台搭载人工智能降噪系统,依靠技术创新破解行业防护痛点,实现告警智能降噪,提升安全运营效率,适配运营商高标准网络安全防护要求。

平台拥有全域防护、轻量化部署、高兼容性特性,可适配各类智能体防护需求,在中国联通某分公司进行实际部署后效果显著,能够为各企事业单位安全部署大模型智能体提供完整可行方案。未来将持续迭

代长链推理、多工具协同、未知威胁识别核心能力,优化人工智能降噪模块性能,不断完善智能体安全防护技术体系,保障企业数字化转型平稳推进。

参考文献:

- [1] 李沛婷,鲁知朋. 针对大模型的网络安全告警分析综述[J]. 网络安全技术与应用,2026(2):13-16.
- [2] Ni L, Zhang S, Huang K, et al. Multi-level screening method for network security alarms based on DBSCAN algorithm and rete rule inference[J]. Scientific Reports, 2026, 16(1):5632.
- [3] Ma T H, Yu J, Wang B Q, et al. A power monitor system cybersecurity alarm-tracing method based on knowledge graph and GCNN[J]. Applied Sciences, 2025, 15(15):8188.
- [4] 高菲. 基于AI告警降噪技术在网络安全防护中的应用[J]. 网络安全技术与应用, 2025(11):57-59.
- [5] 施南廷. 基于AI告警降噪技术在政企网络安全防护中的实践应用[J]. 广播电视网络, 2023, 30(11):77-79.
- [6] 邹进明. 基于网络安全事件的日志采集与告警平台实践[J]. 现代信息技术, 2025, 9(13):149-155, 163.
- [7] 高桐,宋刚,白洋. 基于DeepSeek的运营商网络安全态势感知告警降噪方法研究与应用[J]. 中国宽带, 2025, 21(7):46-48.
- [8] 李纯权. 基于机器学习的网络安全告警降噪技术与系统验证[D]. 北京:北京邮电大学, 2025.
- [9] 黄晓昆,陈烁,姚日煌,等. 多源日志中的智能安全告警处理方法[J]. 电子质量, 2025(5):22-29.
- [10] 张肖楠,范絮妍,谈松,等. 网络告警智能降噪技术研究与实践[C]//第39次全国计算机安全学术交流会. 西安:中国计算机学会, 2024:232-237.
- [11] 侯云云. 多阶段攻击检测方法研究[D]. 郑州:华北水利水电大学, 2024.
- [12] 邓森. 基于关联规则的网络告警数据挖掘算法研究[D]. 武汉:武汉科技大学, 2024.
- [13] 周小明,林小李,陈刚,等. 基于HP-Degree中心性和加权风险期望的网络节点风险评估[J/OL]. 控制工程, 2024: 1-12[2026-05-01]. <https://doi.org/10.14107/j.cnki.kzgc.20230763>.
- [14] 林炳生. 大数据驱动下的RLHF在告警降噪中的应用[J]. 互联网周刊, 2024(5):40-42.
- [15] 路广平. 基于网络威胁情报挖掘的攻击溯源技术研究[D]. 南京:南京理工大学, 2024.

作者简介:

郑浩远,毕业于华北电力大学,工程师,硕士,主要从事网络安全与AI安全相关工作;丁颖睿,毕业于北京邮电大学,工程师,硕士,主要从事安全能力开发和应用相关工作;张润民,毕业于谢菲尔德大学,工程师,硕士,主要从事AI安全与AI应用相关工作。