AI大模型在网络安全威胁预测中的应用探索

翟江华　袁清平　羊杰罕
中国电信股份有限公司东莞分公司　广东　东莞　523000

摘　要：网络攻击手段复杂化，传统防御体系难应对。以Transformer架构为基础的AI大模型，为网络安全威胁的主动预测与智能防御提供新路径。本文探讨其在该领域的应用，先梳理传统方法及局限，再分析大模型在关键环节的技术实现路径；通过案例展示其在实战场景中的效能；针对数据隐私等挑战提出策略，展望前沿技术方向。研究表明，AI大模型正成为构建下一代主动防御体系的核心引擎，能提升网络安全态势感知与风险预判能力。

关键词：人工智能大模型；网络安全；威胁预测；深度学习；Transformer

引言

全球数字化加速，网络空间成博弈“主战场”，2024年全球因网络攻击损失超8万亿美元且持续攀升。攻击者构建APT攻击链条，零日漏洞等新型威胁不断涌现，传统防御体系力不从心。在此情形下，AI大模型因处理非结构化数据等能力卓越被寄予厚望。它通过海量异构数据自监督预训练，习得对网络空间“语言”的通用理解，能适配安全任务，实现从“被动响应”到“主动预测”的转变。本文将探讨其赋能网络安全威胁预测的技术等方面，提供参考指南以推动构建新防御体系。

1 网络安全威胁预测的传统方法及其局限

在AI大模型兴起前，网络安全威胁预测主要有基于签名与规则匹配（如IDS、防病毒软件，对未知攻击失效且规则库维护成本高、易被绕过）、基于统计与异常检测（能发现未知威胁，但有高误报率和概念漂移问题）、基于传统机器学习（性能依赖特征工程质量，人工特征工程效率低、泛化能力弱）、基于威胁情报（情报时效性、准确性、覆盖面有限，整合关联依赖人工效率低）等传统技术路线[1]。这些方法存在被动性（缺乏前瞻性预测能力）、碎片化（数据孤岛难形成全局视角）、智能化不足（依赖人工规则和特征工程）、可扩展性差（计算效率低难实时处理）、适应性弱（难应对攻击手法快速演变和定制化）等局限，催生了对更智能、主动、自适应威胁预测技术的需求，为AI大模型介入提供了舞台。

2 AI 大模型赋能网络安全威胁预测的核心技术架构

AI大模型，特指参数量巨大（通常十亿级以上）、在超大规模无标注或弱标注数据上通过自监督学习预训练、具备强大泛化与迁移能力的深度学习模型。在网络安全领域，其核心技术架构可概括为“预训练-微调/提示-推理”范式，并围绕安全数据特性进行针对性优化。

2.1 模型基础架构：Transformer及其变种

Transformer架构的自注意力机制，可高效捕捉序列数据长距离依赖关系，克服RNN/LSTM处理长序列的梯度问题。网络安全场景中的网络流量包序列、系统调用序列等具有时序和上下文依赖特性，使其成为处理此类数据的理想选择。编码器-解码器架构（如原始Transformer、T5）适用于需生成输出的任务，如根据日志生成攻击描述、预测攻击链后续步骤。仅编码器架构（如BERT、RoBERTa）擅长理解输入序列语义，用于分类、实体识别、相似性计算。仅解码器架构（如GPT系列）擅长自回归生成，可用于模拟攻击者行为、生成对抗样本、回答分析师查询。多模态架构（如CLIP、Flamingo）能同时处理多种模态数据，用于关联分析文本报告与代码片段、网络拓扑图与流量热力图等。

2.2 数据预处理与特征表示

网络安全数据异构、稀疏、高维、噪声大，高质量的数据预处理对大模型应用至关重要。①序列化与标记化：将原始数据转化为模型可处理的序列。例如，网络流量按时间戳将数据包关键字段编码为Token形成序列；系统日志提取关键字段组合成结构化Token序列；代码/二进制使用字节对编码或抽象语法树遍历序列化[2]。②上下文窗口扩展：网络安全事件跨度长，标准Transformer上下文窗口可能不足，需采用Longformer、BigBird等稀疏注意力机制，或使用记忆网络、层次化建模等技术扩展有效上下文。③领域自适应预训练：通用大模型对网络安全领域专业内容理解不足，要在大量网络安全相关文本上进行二次预训练，注入领域知识。

2.3 任务适配：微调（Fine-tuning）与提示工程（PromptEngineering）

预训练模型要适配具体预测任务，常见方法有监督微调（SupervisedFine-tuning,SFT）、提示工程

（PromptEngineering）、指令微调以及参数高效微调。其中，监督微调是在标注好的安全数据集（如标记为“攻击”或“正常”的流量序列、标注了攻击阶段的日志序列）上继续训练模型参数，虽直接有效，但依赖高质量标注数据，成本高昂；提示工程通过精心设计的自然语言指令（Prompt）引导预训练模型完成特定任务，无需或仅需少量参数更新，例如分类任务可设计Prompt“以下网络流量序列是否包含攻击行为？序列：[流量数据摘要]。请回答‘是’或‘否’。”，实体识别任务可设计“从以下安全日志中提取攻击者IP、目标主机和使用的漏洞编号：[日志内容]。”，生成任务可设计“根据已知的攻击步骤：[步骤1,步骤2]，预测攻击者下一步最可能采取的行动，并解释原因。”，该方法灵活性高，能快速适配新任务，减少对标注数据的依赖，但设计高质量Prompt需专业知识和反复试验；指令微调是在包含多种任务指令-输出对的数据集上微调模型，使其更好遵循人类指令，提升提示工程效果；参数高效微调采用如LoRA、Adapter等技术，仅微调模型中少量新增参数，大幅降低计算和存储开销，便于在资源受限环境部署。

2.4 核心预测能力构建

AI大模型借助前文提及的架构，构建出有别于传统方法的核心预测能力：其具备深度语义理解能力，可剖析日志、报告、代码中的复杂语义，精准识别隐含的攻击意图，像“尝试提权”“横向移动”等，而非局限于匹配表面特征；拥有长程上下文关联能力，能将分散在数小时、数天乃至数月不同日志、流量中的微弱信号串联起来，重构完整的攻击链（KillChain），进而识别潜伏的APT活动；具备模式泛化与迁移能力，可从海量历史攻击数据中学习通用攻击模式，即便面对采用新工具、新漏洞的变种攻击，也能依据相似的TTPs进行识别和预测；具有多源异构数据融合能力，能够统一处理网络流量、主机日志、终端行为、威胁情报、漏洞信息等多源数据，打破数据孤岛，形成全方位的威胁视图；还拥有生成与推理能力，不仅能进行分类和检测，还能生成攻击描述、预测攻击路径、推演攻击后果，并回答分析师的复杂查询，为决策提供有力辅助。

3 AI 大模型在网络安全威胁预测中的典型应用场景

3.1 威胁情报的深度挖掘与智能关联

传统威胁情报分析依赖人工从海量报告中提取IOCs（攻击指标），而大模型通过自然语言处理技术，可自动解析开源报告、暗网论坛、漏洞公告等非结构化数据，提取关键TTPs（战术、技术和程序）、受影响产品、攻击者画像等信息，并生成结构化摘要。例如，模型能识别“某勒索软件家族使用的C2域名注册模式”与“钓鱼邮件中的域名特征”之间的关联，进而推断攻击者组织归属[3]。更进一步，基于历史攻击模式与当前情报的时空关联，模型可预测特定行业（如金融）未来两周内利用新披露Office漏洞的定向攻击风险，为防御资源分配提供前瞻性依据。

3.2 用户与实体行为分析（UEBA）的智能化升级

UEBA的核心在于区分“正常”与“异常”行为，而大模型通过多维度上下文理解实现了动态基线构建。例如，模型不仅记录用户深夜登录敏感数据库的行为，还会结合其历史加班模式、项目需求、访问数据的相关性等上下文信息，综合判断风险等级。对于低慢速攻击（如慢速数据渗出），模型能捕捉操作频率、数据量、时间分布等细微异常；结合HR数据（如离职倾向）与权限变更记录，模型还可预测内部威胁，如某员工在提交离职申请后频繁访问非职责范围内的核心系统，系统将自动标记为高风险。

3.3 网络流量异常检测与攻击链预测

传统流量检测依赖已知攻击签名，而大模型通过分析原始流量包或NetFlow数据，可识别加密流量中的TLS握手异常、隐蔽隧道（如DNSTunneling）等未知模式。更关键的是，模型能将流量序列映射到攻击生命周期阶段（如侦察、漏洞利用、命令控制），并预测下一步行动。例如，检测到端口扫描后，模型可结合目标系统漏洞信息，预警后续可能发生的RCE攻击。对于零日攻击，模型通过分析流量中符合已知攻击模式（如异常API调用序列）但未被收录的特征，实现早期预警。

3.4 漏洞风险评估与优先级排序

漏洞管理常面临“CVSS评分高但实际风险低”的困境，而大模型通过自动化解析CVE描述、PoC代码，提取漏洞类型、利用条件等关键信息，并结合企业资产清单（如暴露在公网的系统）、网络拓扑（如关键业务依赖）和现有防护措施（如WAF规则覆盖情况），量化漏洞在本环境的实际被利用风险。例如，某内部系统存在CVSS9.8分的漏洞，但若其未暴露在公网且无外部连接，模型将降低其优先级；反之，若某公网系统存在中等评分漏洞但缺乏防护，模型会提升其修复优先级。

3.5 恶意软件分析与家族归因

传统沙箱分析生成的长篇报告需人工解读，而大模型可自动提取恶意软件行为特征（如C2连接、进程注入），结合代码语义分析（如反汇编指令模式），识别其与已知家族的代码相似性，实现快速归因。例如，模型能通过分析某样本的加密通信协议特征，将其归因为

“APT-C-40”组织的新变种[4]。此外，模型还可预测变种演化方向，如某勒索软件家族可能在未来采用更隐蔽的C2通道或针对云环境的攻击技术。

3.6 安全运营中心（SOC）的智能助手

SOC分析师常面临“告警风暴”问题，大模型通过分析告警上下文（如时间、源IP、目标系统），将同一攻击链触发的多个告警聚合成一个“安全事件”，减少90%以上的无效告警。例如，模型能识别“端口扫描→漏洞利用→C2连接”的完整攻击链，而非孤立处理每个告警。在事件调查阶段，分析师可通过自然语言提问（如“查询与IP1.2.3.4相关的所有活动”），模型自动关联日志、流量、资产数据，生成包含时间线、影响范围、根因的报告，使事件响应时间从小时级缩短至分钟级。

4 实践案例研究：微软 Security Copilot

面对日益复杂的高级持续性威胁（APT）与安全人才短缺的双重挑战，微软于2023年推出基于大语言模型的网络安全助手 Security Copilot，首次将 GPT-4 与自研安全知识图谱深度融合，构建起具备威胁预测能力的智能防御系统。该系统实时接入 Defender、Azure Sentinel 等平台的终端、网络与身份日志，不仅能将原始告警转化为自然语言描述，还能基于 MITRE ATT&CK 框架自动关联攻击行为与已知 APT 组织（如 Lazarus、FIN7）的战术模式，并进一步推演攻击者下一步可能采取的行动——例如预测其将利用 Golden Ticket 进行域控提权。在内部实战测试中，Security Copilot 曾在5分钟内完整还原一次模拟 APT 攻击链，并提前建议启用 Kerberos 审计策略，有效阻断后续横向移动。据微软官方披露，该系统已部署于全球30余个安全运营中心，使客户平均威胁调查时间缩短76%，高危事件响应速度提升3倍，标志着大模型正从“辅助分析”迈向“主动预判”的新阶段（Microsoft Security Blog, 2023；RSAC 2024）。

5 结语

AI大模型正重塑网络安全威胁预测范式，实现从“被动响应”到“主动预判”等根本性转变。本文论证了其在多场景的技术路径与实战成效，直面数据隐私等挑战，提出联邦学习等解决方向。其效能最大化依赖高质量数据、专家知识注入及与现有体系集成。未来，随着多模态融合等技术成熟，大模型将成防御体系“中枢”。学术界与产业界需携手突破瓶颈、制定规范、培养人才，筑牢数字世界智能防线。

参考文献

[1]李橙,陈铭丰,苏嘉珺,等.基于安全大模型的网络安全威胁检测框架研究[J].计算机应用与软件,2025,42(05):179-190.

[2]雷文强.大模型在网络空间中的安全风险与治理对策[J].人民论坛,2025,(16):52-56.

[3]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.

[4]周滔,叶森,王健,等.大模型智能体网络安全事件检测处置系统设计与实现[J].中国宽带,2025,21(07):40-42.