ICS 点击此处添加 ICS 号
CCS 点击此处添加 CCS 号

T/ISC

团 体 标 准

T/XXX XXXX—XXXX

基于大模型的网络攻击行为建模与防御技术要求

Technical requirement for modeling and defending network attacks based on large model

（征求意见稿）

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 XXXX－XX－XX 实施

发 布

目　　次

前　言

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。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由中国互联网协会提出并归口。

本文件起草单位：国网河南省电力公司信息通信分公司、中国信息通信研究院、公安部第三研究所、北京邮电大学、广州汇智通信技术有限公司、零日信安（武汉市）技术有限责任公司、北京航空航天大学、中科数测科技有限公司、华兴中科标准技术（北京）有限公司。

本文件主要起草人：闫丽景、陈文弢、静静、马若龙、邵彦华、付文豪、詹前靖、刘欣然、王磊、王悦霖、白天锐、芦艺佳、肖蔚琪、李洁、周鸣一、黎立、董坤、董婧一、成瑾、李华、任国静、丁月。

基于大模型的网络攻击行为建模与防御技术要求

1 范围

本文件规定了基于大模型的网络攻击行为建模与防御的攻击行为建模技术要求、防御技术要求以及安全评估和测试等内容。

本文件适用于大模型在网络攻击检测、行为分析、威胁建模和防御措施中的应用，适用于大模型开发方、部署方、服务提供方和评估机构。

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 25069—2022 信息安全技术 术语

GB/T 20986 信息安全技术 网络安全事件分类分级指南

GB/T 37027 网络安全技术 网络攻击和网络攻击事件判定准则

GB/T 45288.1—2025 人工智能大模型 第1部分：通用要求

3 术语和定义

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。

[来源：GB/T 45288.1—2025，3.1]

3.2

网络攻击 network attack

也称入侵，指针对网络或联网系统的未授权访问,即对信息系统进行有意或无意的未授权访问，包括针对信息系统的恶意活动或对信息系统内资源的未授权使用。

[来源：GB/T 25069—2022，3.495，有修改]

3.3

攻击行为建模 attack behavior modeling

对网络攻击的行为特征、模式、流程和影响进行抽象和形式化描述的过程，用于攻击检测、分析和预测。

3.4

防御 defense

采用技术和管理措施，保护信息系统免受网络攻击的行为。

3.5

对抗攻击 adversarial attack

通过故意添加扰动或修改输入数据，导致大模型输出错误结果的攻击行为。

注：常见于图像、文本和语音识别场景。

3.6

模型蒸馏 model distillation

将大模型的知识迁移到小模型,在保证性能损失较小的前提下,降低模型复杂度和反演风险的技术。

3.7

模型混淆 model obfuscation

通过混淆大模型的参数、结构、中间层输出等，阻止攻击者通过逆向工程等方法解析模型的参数及结构信息。

4 攻击行为建模技术要求

4.1 攻击类型建模

基于大模型的网络攻击行为建模应覆盖以下常见攻击类型。

a) 传统网络攻击：指基于网络协议、系统漏洞、应用缺陷等实施的网络攻击，包括网络扫描探测攻击、网络钓鱼攻击、漏洞利用攻击、后门利用攻击、后门植入攻击、凭据攻击、信号干扰攻击、拒绝服务攻击、网页篡改攻击、暗链植入攻击、域名劫持攻击、域名转嫁攻击、DNS污染攻击、WLAN 劫持攻击、流量劫持攻击、BGP 劫持攻击、广播欺诈攻击、失陷主机攻击及其他网络攻击。

注：传统网络攻击的分类可参照GB/T 20986及GB/T 37027。在GB/T 20986中定义了21类网络攻击事件，GB/T 37027指出，其中供应链攻击事件和APT攻击事件两类主要从攻击的危害、攻击者的意图角度进行分类，其他19类主要从攻击技术进行分类，可作为“网络攻击”技术手段的分类，也即可作为“网络攻击”的分类。

b) 新型网络攻击：指针对大模型等新技术新应用自身特性、交互机制、生成能力与安全机制实施的恶意利用行为，其攻击目标、利用方式、行为特征与传统网络攻击存在显著差异，主要通过构造特殊输入、诱导模型偏离安全约束、窃取模型知识、破坏模型可用性等方式实现攻击目的。包括：

1) 提示词攻击：通过构造恶意提示、指令绕过、上下文诱导、逻辑陷阱等方式，使大模型生成违法违规、误导欺骗、危害安全的内容，或泄露敏感信息、越权执行操作；
2) 模型窃取攻击：通过构造查询序列、反复交互试探，非法提取、复刻、还原大模型的结构、参数、能力边界或训练数据特征；
3) 数据投毒攻击：在模型训练、微调、对齐阶段注入恶意样本，导致模型在推理阶段产生偏见、错误输出或预留后门；
4) 对抗样本攻击：对输入数据施加难以感知的微小扰动，使大模型做出错误分类、错误识别或错误决策；
5) 数据窃取攻击：通过诱导式交互，非法提取大模型中的个人信息、商业秘密或敏感数据等；
6) 其他新型网络攻击。

注：模型窃取指通过查询和分析大模型的输出，窃取模型参数或结构的攻击行为。

4.2 行为特征提取

攻击行为建模应提取以下特征。

a) 背景特征：
   1) 时序特征：攻击行为的时间序列模式，如攻击频率、持续时间；
   2) 语义特征：攻击载荷的语义内容，如恶意代码、异常指令；
   3) 上下文特征：攻击发生的环境上下文，如网络拓扑、系统状态。
b) 行为特征：
   1) 通信行为特征：包括访问频率、时间规律、周期性；源/目的 IP 异常、地理位置异常；端口、协议、域名、DNS 解析异常；连接时长、发包速率、流量大小突变；加密流量异常、无特征流量、隐蔽隧道；C2 心跳、指令下发、数据回传模式等；
   2) 操作行为特征：包括命令执行序列异常（批量执行、高危命令）；权限提升、越权访问、非法登录；进程启动、文件读写、注册表修改；批量扫描、暴力破解、自动化尝试；短时间大量重复操作等；
   3) 数据行为特征；数据批量读取、外发、压缩、加密；敏感库表、敏感目录高频访问；数据传输方向、流量大小异常；数据篡改、删除、覆盖等；
   4) 漏洞利用行为特征；针对已知 CVE 的请求 payload 特征；畸形报文、特殊字符、超长参数；注入类行为：SQL 注入、命令注入、XSS；文件上传、文件包含、目录遍历等；
   5) 账号身份行为特征；异地登录、多设备并发登录；账号频繁切换、共享账号；权限异常提升、越权操作；僵尸账号、临时账号异常活跃等。

c） 新型攻击特征：
1） 模型层面异常行为：包括模型输入分布异常，输出逻辑异常、幻觉、乱序、重复，模型访问频次、调用量突增，对抗样本输入、扰动输入等；
2） 提示词或语义类攻击特征：包括恶意提示词、诱导性指令，多轮对话伪装、逐步越狱，语义混淆、编码绕过、多语言混合，输出内容违规、敏感信息泄露等；
3） AI 自动化攻击特征：极快的策略生成与攻击尝试，包括自适应调整 payload、规避检测，多阶段、多步骤、逻辑连贯攻击，自动生成钓鱼文本、伪造邮件、伪造身份等。

4.3 建模方法

攻击行为建模应采用以下一种或多种方法。
a） 机器学习建模：使用监督学习、无监督学习或强化学习训练攻击检测模型。
b） 深度学习建模：使用神经网络（如 CNN、RNN）进行序列建模和特征学习。
c） 混合建模：结合规则引擎和机器学习模型，提高建模准确性。

4.4 建模结果验证与优化要求

建模结果验证与优化应满足以下要求：
a） 建模结果应通过真实攻击数据、构造攻击样本或仿真数据进行验证，并对比检测准确率、召回率和误报率等指标；
b） 建模过程应评估模型对不同攻击类型、不同网络环境和不同数据分布的泛化能力，避免模型过拟合特定场景；
c） 建模应根据验证结果持续优化特征提取方式、输入表示方法和模型结构，提高对复杂攻击模式的覆盖能力，复杂攻击模式包括但不限于多阶段攻击、隐蔽攻击、低频攻击等；
d） 建模应支持增量学习或在线更新，能够根据新发现的攻击样本及时调整模型参数、特征模板或行为模式；
e） 建模结果及其优化过程应形成可审计记录，包括数据来源、训练配置、验证数据、模型版本和性能变化情况，确保过程可复现、可追溯。

注：增量学习是指在已有模型基础上，利用新增数据进行训练，实现模型性能提升且保留原有能力的学习方式。

5 防御技术要求

5.1 总体要求

基于大模型的安全防御技术应覆盖大模型安全管理和技术防护的全生命周期，保障安全防御的完整性。
a） 训练阶段：应重点关注数据安全、模型生成安全、代码安全等重要事项。
b） 部署阶段：应重点关注模型防护能力建设，通过对抗性训练等安全测试检验防御能力。
c） 使用阶段：应重点关注输入输出安全、异常行为对抗等内容安全交互的防御能力。
d） 运营阶段：应重点关注能力提升和自学习部分。

5.2 训练阶段

5.2.1 数据安全

基于大模型的网络防御技术在数据安全方面应满足：
a） 训练与推理数据来源应可追溯，数据完整性可验证；
b） 应对训练与推理数据执行来源校验和恶意载荷检测，并对来自不可信通道的数据进行隔离；
c） 应通过格式校验、语义分析、黑白名单与智能检测等方式防止数据在采集、传输和处理过程中被注入或篡改；
d） 应采用加密存储、加密传输、差分隐私或输出脱敏等措施保护关键数据，避免模型泄露隐私信息。

5.2.2 代码安全

基于大模型的网络防御技术在代码安全方面应满足：

a) 应对源代码及依赖库进行漏洞扫描，识别恶意依赖、弱凭证及安全风险，并采用容器化或沙箱机制隔离运行环境；
b) 对大模型生成的代码应进行自动化安全检测，识别危险 API、注入风险和逻辑漏洞；
c) 生成代码的历史记录应完整存档，并对可能包含恶意功能的代码进行提示或阻断；
d) 大模型的部署程序应避免进行明文存储，应加入模型混淆、代码混淆等，增加攻击者的解析难度。

5.3 部署阶段

基于大模型的网络防御技术在模型部署安全方面应满足：

a) 应对模型实施对抗训练、输入扰动检测，提升模型鲁棒性；
b) 推理接口应实施访问控制和频率限制，对可疑访问行为进行审计或阻断；
c) 应对模型输出进行模糊处理或概率裁剪，并对推理请求开展隐私风险检测，宜采用模型蒸馏或参数隔离方式降低反演风险。
d) 大模型在部署时无法被直接明文访问，宜采用数据加密、模型混淆或在可信执行环境进行部署，降低模型被逆向工程的风险。
e) 应在大模型参数或结构中加入模型水印信息，有能力对知识产权窃取进行追溯。

注1：对抗训练是指通过在训练数据中加入对抗样本，提升模型对对抗攻击鲁棒性的训练方法。

注2：模型水印是指通过大模型中嵌入隐蔽、可检测、可溯源的标识信息，用于证明模型归属、追踪来源、防止盗用与侵权的技术。

5.4 使用阶段

基于大模型的网络防御技术在异常行为与对抗样本防御方面应满足：

a) 对文本、图像、日志等多模态输入进行异常分布分析，识别异常梯度特征和规避策略；
b) 对越权指令、提示词注入、模型规避行为等恶意提示词输入应进行过滤或拦截；
c) 模型输出应经过规则过滤和模型过滤双层审查，防止异常内容、幻觉或敏感信息泄露。

5.5 运营阶段

基于大模型的网络防御技术在防御策略更新方面应满足：

a) 系统应支持基于新攻击样本或行为的防御策略动态调整；
b) 应支持联邦更新、热更新或补丁式更新方式，并确保更新过程的安全性和完整性；
c) 防御策略变更应记录日志并可回溯至任意版本。

注：联邦更新是指在多个节点在不共享原始数据的前提下，协同更新模型参数，实现防御策略优化的方式。

6 安全评估和测试

6.1 评估指标

基于大模型的安全评估指标应满足：

a) 攻击检测指标应包括检测率、漏报率、误报率及攻击类型分类准确率；
b) 鲁棒性指标应包括对抗样本成功率、扰动幅度阈值及噪声容忍度；
c) 模型安全性指标应包括模型输出熵变化、敏感信息暴露概率、接口访问异常程度及模型部署文件可访问性；
d) 系统级指标应包括响应延迟、吞吐量、资源占用及更新机制的完整性。

6.2 测试方法

基于大模型的安全测试应满足：

a) 数据安全测试应包括数据注入模拟、数据篡改检测、数据漂移与隐私泄露测试；
b) 模型安全测试应包括黑盒/白盒对抗攻击、模型蒸馏窃取测试、提示词注入测试及外部访问读取测试；

c） 系统安全测试应包括代码与依赖项漏洞扫描、接口滥用与资源耗尽测试，以及异常内容输出的审查测试。

6.3 测试环境要求

基于大模型的测试环境应满足：

a） 测试环境应包含真实和构造的网络流量、攻击样本及多模态输入。真实攻击样本占比应不低于测试集的 40%，多模态输入需覆盖文本、图像、语音类型；

b） 测试环境应提供与实际部署一致的模型推理环境及隔离环境。隔离环境需与生产环境网络隔离，推理环境配置应与生产环境一致；

c） 测试过程应支持记录、回放和复现。记录内容应包括测试用例、测试数据、测试结果，支持一键回放测试流程；

d） 测试环境应进行安全性与隔离性验证防止测试过程中的攻击样本泄露至生产环境。

6.4 测试报告要求

测试结束后应形成完整的测试报告，基于大模型的测试报告应满足：

a） 报告应记录测试样本、测试步骤、指标结果及不符合项。不符合项应明确严重程度、影响范围、整改建议；

b） 应对系统风险等级进行评估并提出整改措施。风险等级分为高、中、低三级，高风险项应在 7 天内整改完成；

c） 报告应包含防御策略配置和变更信息，以支持后续追溯与审计。报告应由测试人员、评估人员签字确认，存档期限不低于 3 年；

d） 报告应附测试用例清单、指标对比表、异常问题截图等佐证材料；

e） 报告应明确测试结论与改进建议的优先级排序，区分紧急、重要、一般三个整改等级。

参 考 文 献

[1] GB/T 35279 信息安全技术 云计算安全参考架构
[2] GB/T 39412 信息安全技术 代码安全审计规范
[3] GB/T 45654 网络安全技术 生成式人工智能服务安全基本要求