ICS 35.030
CCS L 80

团 体 标 准

T/ISC 0132—2026

基于大模型的网络攻击行为建模与防御技术要求

Technical requirements for modeling and defense of network attack behaviors based on large models

（发布稿）

2026－09－08 发布　　2026－10－08 实施

中 国 互 联 网 协 会　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。

本文件所规定的技术要求仅用于网络安全防御与安全研究用途，不适用于任何非法的攻击性用途。

本文件与GB/T 45288.1—2025《人工智能大模型　第1部分：通用要求》衔接配合，GB/T 45288.1—2025规定了人工智能大模型的通用要求（包括大模型的资源池、工具、数据资源、模型管理等），本文件在此基础上针对网络安全领域，规定基于大模型的网络攻击行为建模与防御的技术要求，两者配合使用，GB/T 45288.1—2025提供通用基础，本文件提供网络安全领域的专项技术要求。

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归口。

本文件起草单位：国网河南省电力公司、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、中科数测固源科技（安徽）有限公司、华兴中科标推技术（北京）有限公司、中锐（青岛）信息安全技术有限公司、重庆千港安全技术有限公司。

本文件主要起草人：闫丽景、刘晗、宋腾、杨莹、牛永玺、陈文弢、静静、马若龙、邵彦华、付文豪、詹前靖、林九川、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、谭超、李华、任国静、丁月、刘伟、杜明、郭威。

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缩略语、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 20986 信息安全技术 网络安全事件分类分级指南

GB/T 35273 信息安全技术 个人信息安全规范

GB/T 37027—2025 网络安全技术 网络攻击和网络攻击事件判定准则

GB/T 39412 信息安全技术 代码安全审计规范

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

GB/T 37027—2025界定的以及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，具有大规模参数（通常参数量在亿级以上），利用复杂计算架构，能处理复杂任务，且具备一定泛化性的深度学习模型。

注：大模型一般指参数量级在1亿（含）以上的深度学习模型。

[来源：GB/T 45288.1—2025，3.1，有修改]

3.2

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。

3.3

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。

3.4

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。

3.5

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。

3.6

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。

3.7

对抗样本 adversarial example

通过在原始输入中添加人眼难以察觉的微小扰动，使大模型产生错误输出的样本。

注：对抗样本是实施对抗攻击的主要手段，常见于图像分类、文本分析和语音识别等场景。

3.8

提示词注入 prompt injection

通过构造恶意输入指令，使大模型执行非预期行为，生成违规内容或泄露敏感信息的攻击方式。

注：提示词的定义见GB/T 45288.1—2025中3.5。提示词注入包括直接提示词注入（用户直接输入恶意指令）和间接提示词注入（恶意指令隐藏在外部数据源中，被模型间接读取执行）两种形式。

3.9

幻觉 hallucination

大模型生成看似合理但实际上不符合事实、逻辑错误或与输入无关的内容的现象。

注：幻觉表现为事实性幻觉（捏造不存在的事实）、忠实性幻觉（输出与输入不一致）等，是评估大模型输出可靠性的重要指标，也是GB/T 45288.1—2025中评估大模型安全性的重要考量因素。

3.10

模型窃取 model stealing

通过构造查询序列、反复交互试探等方式，非法提取、复刻大模型的结构、参数或训练数据特征的攻击行为。

4 缩略语

下列缩略语适用于本文件。

BGP：边界网关协议（Border Gateway Protocol）
C2：命令与控制（Command and Control）
CVE：通用漏洞披露（Common Vulnerabilities and Exposures）
DNS：域名系统（Domain Name System）
DR：检测率（Detection Rate）
F1：精确率和召回率的调和平均值（F1-Score）
FPR：误报率（False Positive Rate）
MTTR：平均响应时间（Mean Time to Respond）
SQL：结构化查询语言（Structured Query Language）
XSS：跨站脚本攻击（Cross-Site Scripting）

5 攻击行为建模技术要求

5.1 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、失陷主机利用攻击、勒索软件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—2025。在GB/T 20986中定义了21类网络攻击事件，GB/T 37027—2025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) 提示词注入攻击：通过构造包含恶意指令的输入文本，混入用户输入、外部文档、知识库内容等上下文，诱导大模型覆盖、忽略预设系统提示与安全约束，迫使模型执行非预期操作（泄露系统提示、生成违规内容、调用高危工具、篡改输出逻辑等）的对抗攻击行为；

2) 越狱攻击：通过构造角色扮演、多轮诱导、混淆编码、上下文污染、间接注入等对抗性输入手段，绕过大模型内置安全对齐机制、系统提示约束、内容安全过滤护栏，诱导模型突破预设安全限制，生成违法违规、暴力危险、隐私泄露、网络攻击指导等本应被拒绝输出内容的对抗攻击行为；
3) 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；
4) 数据投毒攻击：通过干预训练数据集，在训练数据中加入精心构造的异常数据，破坏原有训练数据的概率分布，导致模型在某些特定条件下产生偏见、错误输出或预留后门等；
5) 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；
6) 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；
7) 供应链攻击：针对大模型研发、微调、分发及集成全生命周期中的第三方资源实施的恶意利用行为。应覆盖基础模型篡改、第三方插件后门植入以及开发框架与依赖库的逻辑劫持等攻击路径；
8) 其他新型网络攻击。

5.2 行为特征提取

攻击行为建模应覆盖攻击链的关键阶段特征，包括初始入侵、横向移动、权限提升、数据外传等核心攻击阶段的行为特征。建模应支持对多源数据（包括日志、网络流量、终端行为、威胁情报等）的融合分析。提取特征如下。

a) 静态维度特征：
   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；
   6) API 调用行为特征：包括 API 调用频率、调用时间规律、参数模式；异常 API 调用链、未授权 API 访问；API 参数注入、数据过度请求等。
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，模型访问频次、调用量突增，对抗样本输入、扰动输入等；
   2) 幻觉特征：模型生成看似合理但不符合事实、与输入不一致或逻辑错误的内容，表现为事实性幻觉、忠实性幻觉等；
   3) 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；
   4) 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。

5.3 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；
d） 智能体建模：应利用大模型的自主规划能力，对具备目标拆解、环境感知、工具调用及多轮迭代特征的攻击实体进行描述。建模过程应涵盖攻击智能体的决策规划链路、反思修正机制以及多智能体协同对抗模式。

5.4 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的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） 建模结果应验证攻击事件间的时序一致性，确保攻击行为的时间序列逻辑合理，支持跨源事件关联分析；
c） 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。宜在跨时间段、跨网络环境等不同条件下验证模型的泛化能力；
d） 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；
e） 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；
f） 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；
g） 攻击行为模型应支持增量更新，能够根据新发现的攻击行为及时更新模型，并保留历史版本以支持版本回溯和对比分析。

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。

6 防御技术要求

6.1 总体要求

基于大模型的安全防御技术应覆盖以下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。
e） 退役阶段：应确保模型退役时训练数据、模型参数及相关敏感信息的安全删除或不可逆销毁，防止数据残留导致信息泄露。

6.2 训练阶段

6.2.1 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以下要求：

a） 训练与推理数据来源应可追溯，数据完整性可验证，应对数据来源进行可信度评估；
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、加密传输等措施保护关键数据，并宜结合差分隐私、联邦学习或输出脱敏等技术增强隐私保护能力，隐私保护应符合 GB/T 35273 的要求。

6.2.2 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以下要求：

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，并实施软件供应链安全管理，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。代码安全应符合 GB/T 39412 的相关要求；

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、模型签名等，增加攻击者的解析难度，并支持模型完整性验证。

注：API 是指应用程序编程接口，用于不同软件系统、组件间实现数据交互与功能调用的标准化接口。

6.3 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以下要求：

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，宜支持分布式限流机制和异常流量模式识别；

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） 应采用数字签名等技术措施对模型数据开展完整性校验，保障模型数据未被篡改；

f） 宜采用模型水印或模型指纹或至少一种知识产权保护手段，并有能力对知识产权窃取进行追溯；

g） 应针对大模型推理引擎、容器环境及底层操作系统实施漏洞扫描与合规配置核查。应及时修复已知高危漏洞，严格限制推理接口的开放端口；

h） 应针对拟接入的模型技能（Skills）及模型上下文协议（MCP）服务实施安全审计。应对其提供的 API 接口规范、执行逻辑代码及资源访问权限进行静态安全扫描与漏洞检测。宜通过沙箱环境进行动态行为验证，确保其不包含已知后门或超权设计。

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。

6.4 使用阶段

基于大模型的网络防御技术在使用阶段，针对异常行为与对抗样本防御方面应满足以下要求：

a） 针对文本、图像、音频、视频等多模态输入及日志类辅助数据开展异常分布分析，识别异常梯度特征与各类规避攻击手段；

b） 对越权指令、提示词注入、模型规避行为、编码绕过攻击等恶意提示词输入应进行过滤或拦截，并应对多轮对话上下文进行安全检测；

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、敏感信息泄露或危险操作执行。在模型调用 Skills 或通过 MCP 执行外部指令时，应建立执行流实时监控机制；

d） 应建立对外部知识库文档的清洗与指令隔离机制，防止“间接提示词注入”。同时应实现向量数据的权限受控，禁止跨权限检索敏感信息；

e） 应对多模态输出进行安全审查，并宜支持输出一致性验证。

6.5 运营阶段

基于大模型的网络防御技术在运营阶段，针对防御策略更新应满足以下要求：

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。模型更新和策略变更应建立版本管理机制，支持更新失败时的版本回滚；

d) 当检测到新型攻击或安全事件时，应启动应急响应流程，包括攻击研判、影响评估、防御策略调整和事件记录；
e) 应定期对大模型、MCP 及 Skills 的调用审计日志进行安全建模分析，识别潜在的隐蔽攻击模式。应基于审计发现的异常操作，动态更新调整防御策略。

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。

6.6 退役阶段

基于大模型的网络防御技术在退役阶段应满足以下要求：
a) 应建立标准化大模型退役审批与下线流程，完成业务流量迁移、对外服务接口关停、权限回收，禁止退役模型对外提供推理调用服务；
b) 应采用覆写、密钥销毁、物理销毁等合规方式，对模型权重、训练数据集、知识库、交互日志、精调样本等全量资产实施不可逆销毁，清除本地、异地、快照、备份中全部残留文件；
c) 宜开展退役模型知识产权与后门风险核验，对带有模型水印、指纹的退役模型同步清除标识，禁止向外部流转、出借、出售退役模型文件；
d) 应完整留存退役全流程审计记录，包含退役审批单、销毁清单、销毁操作日志、核验凭证，相关记录归档留存不少于 3 年，可追溯核查。

7 安全评估和测试

7.1 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及攻击响应时间；
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及提示词攻击防御成功率；
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、模型完整性验证结果及训练数据泄漏评估指标；
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；
e) 告警质量指标应包括告警准确率、平均告警时延等，用于评估告警的有效性和及时性。

注：评估指标计算方法见附录A。

7.2 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；
d) 安全测试宜包括红队测试，模拟真实攻击者的行为对系统进行全面安全评估；
e) 安全测试集应建立持续维护和更新机制，包括越狱样本、提示注入样本、对抗样本等安全测试样本的定期补充和失效样本的及时淘汰。

7.3 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。测试样本的来源应经过验证，确保样本的真实性和代表性。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、音频、视频等多模态类型，以及日志等辅助数据；
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，包括网络隔离检查、数据流向验证、访问控制确认等，防止测试过程中的攻击样本泄露至生产环境；
e) 测试数据集应覆盖不同时间段和不同攻击场景的数据，避免时间相关性导致的数据泄漏和评估偏差；
f) 测试过程应确保可复现性，应记录测试环境基线配置、测试样本集、测试执行步骤，确保相同条件下可重复获得一致的测试结果；
g) 测试数据集宜覆盖 Web 攻击、恶意脚本、凭证滥用、横向移动等典型攻击场景，以提升评估结果的可比性。

7.4 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，高风险项应在测试报告中明确整改时限和验证方法，整改时限由组织根据业务影响评估结果确定；
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，存档应采用加密存储、访问控制等安全措施；
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。

附 录 A
（资料性）
评估指标计算方法

A.1 检测率/召回率

$$DR/\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \times 100\% \quad \text{(A.1)}$$

式中：
DR——检测率；
$Recall$——召回率；
TP——真正例（正确检测到的攻击数）；
FN——假负例（漏报的攻击数）。

A.2 精确率

$$Precision = \text{TP}/(\text{TP} + \text{FP}) \times 100\% \quad \text{(A.2)}$$

式中：
$Precision$——精确率；
FP——假正例（误报的攻击数）。

A.3 F1 分数

$$F1 = 2 \cdot \text{Precision} \cdot \text{Recall}/(\text{Precision} + \text{Recall}) \quad \text{(A.3)}$$

式中：
$F1$——精确率和召回率的调和平均值，综合反映模型对攻击的识别能力。

A.4 误报率

$$FPR = FP/(FP + \text{TN}) \times 100\% \quad \text{(A.4)}$$

式中：
FPR——误报率；
TN——真负例（正确排除的正常行为数）。

A.5 攻击响应时间

$$MTTR = \Sigma（响应时间）/事件总数 \quad 从攻击检测到完成响应处置的平均时间 \quad \text{(A.5)}$$

式中：
$MTTR$——平均响应时间。

参 考 文 献

[1] GB/T 22239 信息安全技术 网络安全等级保护基本要求
[2] GB/T 35279 信息安全技术 云计算安全参考架构
[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