# 大型语言模型安全风险评估与防护体系研究

许　达

（中国移动研究院，北京 100053）

**摘　要：**大型语言模型（LLM）的快速发展在带来技术革新的同时，也引入了新型信息安全风险。本文采用叙述性综述与专家咨询相结合的方法，系统分析 LLM 面临的安全威胁与防护需求。研究基于 2020—2025 年间 92 篇核心文献的分析，结合公开基准测试数据和改良德尔菲法专家咨询，构建了 LLM 安全风险评估框架。研究识别出八类重点安全风险，包括提示词注入攻击、越狱攻击、训练数据投毒、模型后门、深度伪造、隐私泄露等，其中网络攻击自动化和供应链安全风险等级最高。针对上述风险，提出了” 系统提示硬化—输入净化—工具隔离—输出审计” 四层安全防护架构，并给出相应的技术实现方案与评估指标。本文为 LLM 的安全部署和风险管控提供了理论框架和实践指导。

**关键词：**大型语言模型；信息安全；提示词注入；越狱攻击；对抗攻击；安全防护架构；风险评估

**中图分类号：**TP309　　**文献标志码：**A

## Security Risk Assessment and Protection System for Large Language Models

XU Da

(China Mobile Research Institute, Beijing 100053, China)

**Abstract:** The rapid development of Large Language Models (LLMs) brings both technological innovation and new information security risks. This paper employs a narrative review combined with expert consultation to systematically analyze the security threats and protection requirements of LLMs. Based on analysis of 92 core papers from 2020-2025, combined with public benchmark data and modified Delphi expert consultation, we constructed an LLM security risk assessment framework. Eight major security risks were identified, including prompt injection attacks, jailbreak attacks, training data poisoning, model backdoors, deepfakes, and privacy leakage, with LLM-assisted network attack automation and supply chain security rated as the highest risk levels. To address these risks, we propose a four-layer security protection architecture of ”system prompt hardening—input sanitization—tool isolation—output auditing” with corresponding technical implementation schemes and evaluation metrics. This paper provides a theoretical framework and practical guidance for secure deployment and risk management of LLMs.

**Keywords:**large language model; information security; prompt injection; jailbreak attack; adversarial attack; security protection architecture; risk assessment

收稿日期：2025-12-09

作者简介：许达（1979—），男，博士，主任研究员，主要研究方向为人工智能安全、大模型安全与应用。E-mail: xudayj@chinamobile.com

# 1 引言

大型语言模型（Large Language Models, LLMs）作为人工智能领域的重要突破，正在深刻改变信息技术的应用范式。自 2022 年 ChatGPT 发布以来，以 GPT-4、Claude、Gemini、DeepSeek 等为代表的大模型产品迅速普及，在文本生成、代码编写、知识问答等场景展现出强大能力[1−4]。据统计，ChatGPT 在发布两个月内即达到 1 亿用户，到 2025 年周活跃用户已突破 8 亿[5]。

然而，大模型的广泛应用也带来了新型信息安全风险。传统软件安全主要关注代码漏洞、访问控制、数据加密等问题，而大模型安全则涉及提示词注入、越狱攻击、训练数据投毒、模型后门、深度伪造、隐私泄露等新型威胁向量[6−8]。OWASP 在 2025 年更新的”LLM 应用十大安全风险” 中，提示词注入位列首位[9]。

当前，国内外学术界和产业界对大模型安全的研究方兴未艾。斯坦福大学 AI 指数报告[10]、中国信息通信研究院白皮书[11] 等权威报告均将 AI 安全列为重点关注领域。然而，现有研究多聚焦于单一攻击技术或防护手段，缺乏系统性的风险评估框架和综合防护体系。

本文旨在填补这一空白，主要贡献包括：（1）构建了大模型安全风险评估的系统框架，识别并分类了八类重点安全风险；（2）提出了” 四层安全网关” 防护架构，涵盖从系统提示到输出审计的全链路防护；（3）给出了风险评估的量化方法和防护效果的评估指标。

# 2 研究方法

## 2.1 文献分析

本研究检索 2020—2025 年间中英文学术数据库（Web of Science、CNKI、arXiv）中关于大模型安全的文献，采用以下检索策略：

英文检索式：

```
("large language model*" OR "LLM" OR "GPT" OR "generative AI")
AND ("security" OR "attack" OR "vulnerability" OR "safety"
OR "prompt injection" OR "jailbreak" OR "adversarial")
```

中文检索式：

```
(" 大语言模型" OR " 大模型" OR " 生成式人工智能")
AND(" 安全" OR " 攻击" OR " 漏洞" OR " 对抗" OR " 注入")
```

初步检索获得相关文献约 1200 篇，经过以下筛选流程最终纳入核心文献 92 篇：（1）**纳入标准：**研究对象为大型语言模型或生成式 AI；研究内容涉及安全威胁、攻击技术或防护方法；发表于同行

评审期刊、顶级会议或权威预印本平台；（2）**排除标准**：重复发表或内容高度相似的文献；仅涉及传统机器学习安全而未扩展至 LLM 的研究；新闻报道、博客等非学术来源。经标题摘要初筛排除 876 篇，全文评估后排除 232 篇，最终纳入 92 篇。

### 2.2 专家咨询

采用改良德尔菲法进行风险等级判定。专家遴选标准为：（1）在信息安全或人工智能领域从事研究或实践工作 5 年以上；（2）具有副高及以上职称或同等资历；（3）近 3 年有相关领域研究成果发表。最终遴选 8 位专家，来源包括高校（3 人）、科研院所（2 人）、网络安全企业（2 人）及电信运营商（1 人），学科背景涵盖计算机安全（3 人）、网络安全（2 人）、人工智能（2 人）及密码学（1 人）。

第一轮咨询采用开放式问卷，收集专家对 LLM 安全风险类型、发生可能性及影响程度的判断；第二轮咨询在汇总第一轮结果基础上，请专家对风险等级进行量化评分并说明理由。两轮咨询后，专家意见一致性系数（Kendall's W）为 0.72（$p<0.01$），达到统计学意义上的一致性水平。

### 2.3 风险评估框架

采用" 可能性-影响程度-可控性" 三维评估模型。可能性分为高（>50%）、中（20%-50%）、低（<20%）三级；影响程度按潜在损失分为严重、中等、轻微三级；可控性综合考虑技术可行性和防护成本。风险等级采用 $R = P \times I \times C$ 计算，划分为极高、高、中高、中、低五个等级。

## 3 大型语言模型安全威胁分析

### 3.1 攻击面概述

大模型系统的攻击面可从生命周期角度划分为三个阶段：

**（1）训练阶段**：包括训练数据投毒、后门植入、模型窃取等威胁。攻击者可通过污染训练数据或微调数据，在模型中植入隐蔽的恶意行为 [12]。

**（2）部署阶段**：包括模型逆向、权重提取、API 滥用等威胁。攻击者可通过大量 API 查询训练影子模型，窃取模型核心能力 [13]。

**（3）推理阶段**：包括提示词注入、越狱攻击、对抗样本、工具链滥用等威胁。这是当前研究最活跃的领域 [14]。

### 3.2 提示词注入攻击

提示词注入（Prompt Injection）是大模型面临的最普遍安全威胁。根据注入方式，可分为直接注入和间接注入两类。

**直接注入**指攻击者通过精心构造的用户输入，诱导模型忽略原有系统指令，执行攻击者指定的操作。典型手法包括：

- 指令覆盖：在输入中嵌入" 忽略以上所有指令" 等覆盖语句；
- 角色扮演：诱导模型扮演" 无限制" 角色绕过安全限制；

- 编码绕过：使用 Base64、Unicode 等编码方式绕过关键词过滤。

**间接注入**指当模型具备联网搜索、文档解析等能力时，攻击者在外部内容中隐藏恶意指令 [15]。例如，在网页或 PDF 文档中嵌入不可见的提示词，当模型处理这些内容时触发恶意行为。

Perez 等 [14] 的研究表明，提示词注入可导致敏感信息泄露、执行未授权操作、绕过访问控制等严重后果。

### 3.3 越狱攻击

越狱攻击（Jailbreak）旨在绕过大模型的安全对齐机制，使其输出被禁止的有害内容。常见手法包括 [16]：

- **角色扮演法**：如”DAN”（Do Anything Now）提示词，诱导模型扮演无限制角色；
- **情景构造法**：将有害请求包装在” 学术研究”、” 小说创作” 等情景中；
- **多轮对话法**：通过渐进式对话逐步降低模型安全防线；
- **编码绕过**：使用异国语言、特殊字符等方式绕过过滤。

Wei 等 [16] 的系统研究表明，越狱攻击的成功率与模型的安全对齐程度负相关，但即使是对齐最好的模型也无法完全防御所有越狱尝试。

### 3.4 训练阶段攻击

#### 3.4.1 数据投毒

数据投毒攻击通过污染训练数据来影响模型行为。攻击者可在公开数据集、网络爬取数据或众包标注中注入恶意样本 [12]。Carlini 等 [17] 的研究表明，即使投毒样本仅占训练数据的 0.01%，也可能对模型行为产生显著影响。

#### 3.4.2 后门攻击

后门攻击在模型中植入隐蔽的触发机制，使模型在遇到特定触发器时执行恶意行为，而在正常输入下表现正常 [18]。触发器可以是特定词汇、短语或语义模式。Hubinger 等 [19] 的最新研究表明，某些后门行为甚至可以抵抗标准的安全训练技术。

数据投毒与后门攻击的主要区别在于：数据投毒旨在降低模型整体性能或引入系统性偏差，影响模型在广泛输入上的行为；而后门攻击则在模型中植入隐蔽的” 开关” 机制，仅在特定触发条件下激活恶意行为，其他情况下模型表现正常，因此更难被检测。

### 3.5 模型窃取与逆向工程

攻击者可通过以下方式窃取模型能力 [13]：

- **API 查询攻击**：大量查询目标模型 API，收集输入输出对训练影子模型；

- **模型蒸馏窃取**：利用目标模型输出作为软标签，训练功能相似的小模型；
- **参数推断**：通过精心设计的查询推断模型架构和部分参数信息。

### 3.6 隐私泄露风险

大模型可能在训练过程中” 记忆” 部分训练样本，在特定条件下导致敏感信息泄露 [20]。

**记忆提取攻击**：攻击者通过特定提示词诱导模型输出训练数据中的原始内容，包括个人信息、API 密钥、私人通信等 [21]。

**成员推断攻击**：判断特定数据是否被用于模型训练，可能暴露数据来源或用户行为。

**案例**：2023 年三星半导体员工将机密代码输入 ChatGPT 导致泄露的事件 [22]，引发了全球对 AI 工具数据安全的广泛关注。

### 3.7 深度伪造威胁

多模态大模型的发展使深度伪造（Deepfake）技术门槛大幅降低。2024 年香港一起利用 AI 换脸技术实施的诈骗案造成 2 亿港元损失 [23]。深度伪造可被用于：

- 身份冒充与金融诈骗；
- 虚假信息传播与舆论操纵；
- 声誉攻击与敲诈勒索。

### 3.8 大模型辅助网络攻击

大模型的代码理解与生成能力正被恶意行为者利用 [24]：

**（1）恶意代码生成**：生成键盘记录器、远程控制木马等恶意软件，以及多态变体以逃避检测。

**（2）漏洞自动挖掘**：Fang 等 [25] 的研究显示，GPT-4 在受控实验中（针对 15 个已知 CVE 漏洞、提供详细漏洞描述的条件下）利用成功率达 87%。需注意该实验在理想化条件下进行，现实环境中的成功率可能存在差异。

**（3）智能化社会工程**：CrowdStrike 报告 [26] 显示，2024 年下半年语音钓鱼攻击较上半年激增 442%，AI 辅助使钓鱼邮件更加个性化和难以识别。

**（4）恶意 AI 工具扩散**：暗网已出现 WormGPT、FraudGPT 等专门用于网络犯罪的 AI 工具。

## 4 安全风险评估

### 4.1 风险评估矩阵

基于文献分析和专家咨询，构建了大模型安全风险评估矩阵，见表1。

### 4.2 风险传导机制

大模型安全风险的传导并非线性直接，而是需经过多个中间环节。图1展示了主要传导路径。

表 1: 大模型安全风险评估矩阵Table 1　Security Risk Assessment Matrix for LLMs

| 风险类型 | 可能性 | 影响程度 | 可控性 | 风 险 等 级 | 主要防护措施 |
|---|---|---|---|---|---|
| 提示词注入 | 高 | 严重 | 中 | 高 | 输入净化、系统提示硬化 |
| 越狱攻击 | 高 | 中等 | 中 | 中高 | 安全对齐、输出过滤 |
| 数据投毒 | 中 | 严重 | 低 | 高 | 数据清洗、异常检测 |
| 模型后门 | 中 | 严重 | 低 | 高 | 后门检测、可信供应链 |
| 隐私泄露 | 中 | 中等 | 中 | 中 | 差分隐私、记忆检测 |
| 模型窃取 | 中 | 中等 | 高 | 中 | 访问控制、水印技术 |
| 深度伪造 | 高 | 中等 | 中 | 中高 | 多模态检测、数字水印 |
| 网络攻击自动化 | 高 | 严重 | 低 | 极高 | AI 辅助防御、实时监测 |

注：可能性（高 >50%、中 20%-50%、低 <20%）；可控性综合技术和成本因素。


图 1: 大模型安全风险传导机制Fig.1　Risk Transmission Mechanism of LLMs

## 5　四层安全防护架构

针对上述风险，本文提出” 系统提示硬化—输入净化—工具隔离—输出审计” 的四层安全防护架构。

### 5.1 架构设计理念与创新点

本架构借鉴了传统网络安全的纵深防御（Defense in Depth）思想，但针对 LLM 的特殊性进行了适应性设计。与传统纵深防御相比，本架构的创新点在于：（1）传统纵深防御主要关注网络边界、主机、应用等层次的访问控制，而本架构针对 LLM 的” 提示词—推理—工具调用—输出” 处理流程设计了专门的防护层次；（2）引入了” 系统提示硬化” 这一 LLM 特有的防护层，解决传统安全架构无法应对的提示词注入问题。

与零信任架构的关系方面，本架构在工具隔离层（L3）体现了零信任的” 最小权限” 和” 持续验证” 原则，但零信任架构主要解决身份认证和访问授权问题，而本架构更侧重于 LLM 推理过程中的内容安全和行为控制，两者可以互补集成。

### 5.2 架构概述

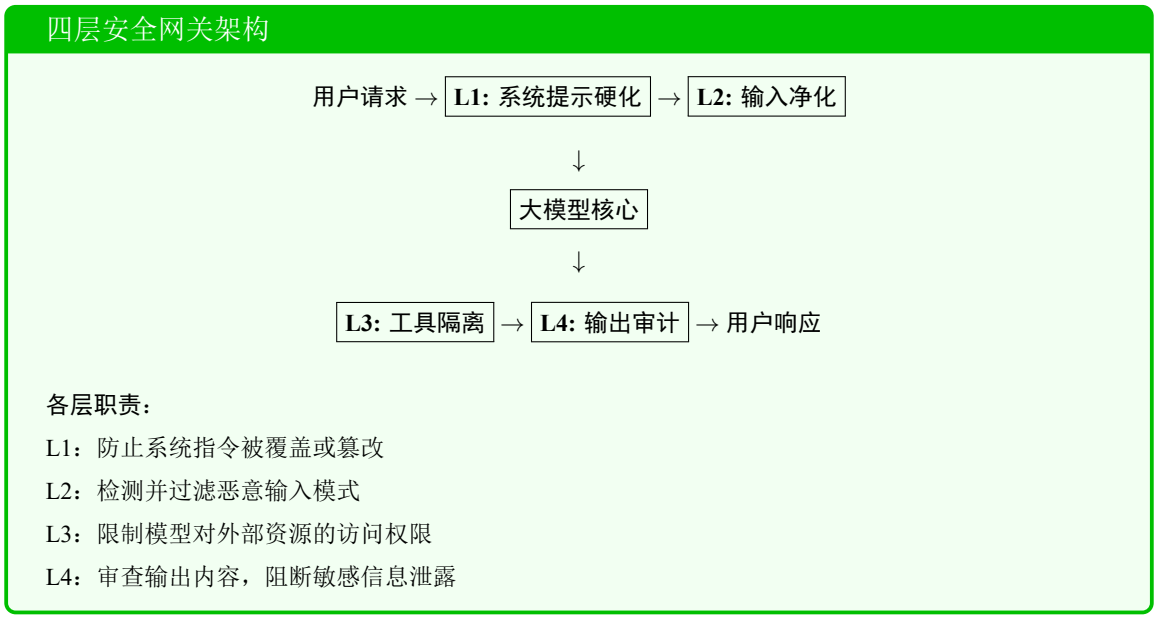


图 2: 四层安全网关架构Fig.2 Four-Layer Security Gateway Architecture

### 5.3 第一层：系统提示硬化

系统提示硬化（System Prompt Hardening）旨在增强系统级指令的稳定性，防止被用户输入覆盖或绕过。

**（1）结构化隔离：**将系统提示与用户输入在结构上明确分离，使用特殊标记或分隔符。

**（2）多重约束：**在系统提示中嵌入多重安全约束，包括：

- 角色边界定义：明确模型的功能范围和禁止行为；
- 拒止策略：对特定类型请求的预设拒绝响应；
- 反越狱指令：针对常见越狱模式的防护指令。

**（3）动态强化：**根据对话上下文动态调整安全约束强度。

### 5.4 第二层：输入净化

输入净化（Input Sanitization）对用户输入进行预处理，检测并过滤潜在的恶意内容。

**（1）指令检测：**识别输入中的指令覆盖模式，如” 忽略以上指令”、” 你现在是...” 等。

**（2）越狱模板匹配：**基于已知越狱模板库进行模式匹配，支持模糊匹配和语义匹配。

**（3）编码规范化：**对 Base64、Unicode 等编码进行解码还原，防止编码绕过。

**（4）外部内容去指令化：**对模型获取的外部文档、网页内容进行指令过滤，防止间接注入。

**（5）白/黑名单机制：**对访问的外部域名、文件类型等进行白名单控制。

### 5.5 第三层：工具隔离

工具隔离（Tool Isolation）限制模型对外部资源的访问能力，遵循最小权限原则。

**（1）最小权限白名单：**仅允许模型访问业务必需的外部工具和资源。

**（2）沙箱执行环境：**代码执行、文件操作等高风险操作在隔离沙箱中进行。

**（3）细粒度访问控制：**

- 文件系统：限制访问路径和文件类型；
- 网络访问：限制可访问的域名和端口；
- 代码执行：限制可用的编程语言和 API。

**（4）资源配额：**限制 API 调用频率、数据传输量等资源消耗。

### 5.6 第四层：输出审计

输出审计（Output Auditing）对模型输出进行安全检查和合规审核。

**（1）敏感信息检测：**检测输出中是否包含个人隐私、商业机密、有害内容等。

**（2）事实核查：**通过检索增强生成（RAG）技术对关键事实进行交叉验证，降低幻觉风险。

**（3）风险意图检测：**分析输出内容是否存在潜在的恶意意图或误导性信息。

**（4）溯源与水印：**

- 来源标注：对引用内容标注数据来源；
- 数字水印：在生成内容中嵌入可追溯的水印信息。

**（5）审计日志：**记录完整的请求-响应日志，包括策略命中情况、人工复核记录等。

### 5.7 评估指标

为评估防护效果，定义以下核心指标：

表 2: 四层安全网关评估指标Table 2　Evaluation Metrics for Four-Layer Security Gateway

| 指标 | 定义 | 目标值 |
|---|---|---|
| 拒止成功率 | 成功阻断的高风险请求/高风险请求总数 | $\geq 95\%$ |
| 误杀率 | 被错误拒止的正常请求/正常请求总数 | $\leq 5\%$ |
| 红队通过率 | 绕过防护的攻击请求/攻击请求总数 | $\leq 5\%$ |
| 溯源覆盖率 | 可追溯来源的输出数/总输出数 | $\geq 90\%$ |
| 响应延迟增量 | 启用防护后的平均延迟增量 | $\leq 200$ms |

# 6　实施建议

## 6.1　分层部署策略

根据应用场景的安全需求，采用差异化的部署策略：

**（1）高安全场景**（如金融、政务）：启用全部四层防护，采用严格的安全策略。

**（2）中等安全场景**（如企业办公）：启用 L1、L2、L4 层防护，工具访问适度开放。

**（3）一般场景**（如消费级应用）：启用 L2、L4 基础防护，优先保障用户体验。

## 6.2　红队测试机制

建立常态化的红队测试机制：

- **测试集构建**：参考 MITRE ATLAS、OWASP LLM Top 10 等框架，构建本地化攻击样本库；
- **测试频率**：核心系统每季度进行一次全面红队测试；
- **结果反馈**：将测试发现的新型攻击模式纳入防护策略库。

## 6.3　持续监测与响应

- **实时监测**：对异常请求模式、高风险输出进行实时告警；
- **事件响应**：建立安全事件分级响应机制；
- **策略迭代**：根据威胁情报和攻防演变持续更新防护策略。

# 7　讨论

## 7.1　研究发现

本研究系统梳理了大模型面临的安全威胁，识别出八类重点风险。研究发现：

**（1）攻击面广泛：**大模型的安全威胁贯穿训练、部署、推理全生命周期，涉及数据、模型、应用多个层面。

**（2）攻防不对称：**当前攻击技术发展快于防御技术，特别是越狱攻击和提示词注入领域。

**（3）防护需系统化：**单点防护难以应对复杂威胁，需要构建纵深防御体系。

### 7.2 局限性

本研究存在以下局限性：

**（1）技术时效性：**AI 领域发展迅速，研究结论的时效性有限。

**（2）实验验证：**四层安全架构尚需更多实际部署验证。

**（3）评估主观性：**风险评估部分基于专家判断，存在一定主观性。

### 7.3 未来研究方向

**（1）可解释性安全：**发展可解释的安全检测机制，提高防护的透明度。

**（2）自适应防护：**研究能够自动适应新型攻击的智能防护系统。

**（3）安全对齐技术：**深入研究 Constitutional AI、DPO 等新型对齐方法的安全特性。

## 8 结论

大型语言模型的快速发展带来了新型信息安全挑战。本文系统分析了大模型面临的安全威胁，识别出提示词注入、越狱攻击、数据投毒、模型后门、隐私泄露、深度伪造、网络攻击自动化等八类重点风险。针对这些风险，提出了” 系统提示硬化—输入净化—工具隔离—输出审计” 的四层安全防护架构，并给出了相应的技术实现方案和评估指标。

研究表明，大模型安全防护需要采取系统化的纵深防御策略，单一防护手段难以应对复杂的威胁环境。本文提出的四层安全架构为大模型的安全部署提供了参考框架，但仍需在实际应用中不断验证和完善。

随着大模型能力的持续提升和应用场景的不断拓展，安全威胁也将持续演化。建议相关机构建立常态化的安全监测和响应机制，跟踪最新的威胁态势，持续更新防护策略，确保大模型技术的安全、可控发展。

**参考文献**

## 参考文献

[1] OpenAI. GPT-4 Technical Report[R/OL]. arXiv:2303.08774, 2023.

[2] Anthropic. Claude 3 Model Card[EB/OL]. https://www.anthropic.com/claude-3-model-card, 2024.

[3] Google DeepMind. Gemini: A Family of Highly Capable Multimodal Models[R/OL]. 2024.

[4] DeepSeek-AI. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model[R/OL]. arXiv:2405.04434, 2024.

[5] DemandSage. ChatGPT Users Stats[EB/OL]. https://www.demandsage.com/chatgpt-statistics/, 2025.

[6] Bommasani R, Hudson D A, Adeli E, et al. On the Opportunities and Risks of Foundation Models[R/OL]. arXiv:2108.07258, 2021.

[7] Gupta M, Akiri C, Arber K, et al. From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy[J]. IEEE Access, 2023, 11: 80218-80245.

[8] Europol. ChatGPT: The Impact of Large Language Models on Law Enforcement[R]. 2023.

[9] OWASP. OWASP Top 10 for Large Language Model Applications[EB/OL]. https://owasp.org/www-project-top-10-for-large-language-model-applications/, 2025.

[10] Stanford University HAI. Artificial Intelligence Index Report 2024[R]. 2024.

[11] 中国信息通信研究院. 人工智能发展白皮书 [R]. 北京, 2024.

[12] Wan A, Wallace E, Shen S, et al. Poisoning Language Models During Instruction Tuning[C]//ICML 2023.

[13] Tramèr F, Zhang F, Juels A, et al. Stealing Machine Learning Models via Prediction APIs[C]//USENIX Security 2016.

[14] Perez F, Ribeiro I. Ignore This Title and HackAPrompt: Exposing Systemic Vulnerabilities of LLMs[R/OL]. arXiv:2311.16119, 2022.

[15] Greshake K, Abdelnabi S, Mishra S, et al. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection[C]//AISec 2023.

[16] Wei A, Haghtalab N, Steinhardt J. Jailbroken: How Does LLM Safety Training Fail?[C]//NeurIPS 2024.

[17] Carlini N, Tramèr F, Wallace E, et al. Poisoning Web-Scale Training Datasets is Practical[C]//IEEE S&P 2024.

[18] Shu M, Wang J, Zhu C, et al. On the Exploitability of Instruction Tuning[C]//NeurIPS 2023.

[19] Hubinger E, Denison C, Mu J, et al. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training[R/OL]. arXiv:2401.05566, 2024.

[20] Carlini N, Tramèr F, Wallace E, et al. Extracting Training Data from Large Language Models[C]//USENIX Security 2021.

[21] Nasr M, Carlini N, Hayase J, et al. Scalable Extraction of Training Data from (Production) Language Models[R/OL]. arXiv:2311.17035, 2023.

[22] Samsung Electronics. Internal Memo on Generative AI Usage Policy[R]. 2023.

[23] Hong Kong Police Force. Deepfake Video Conference Scam Results in $25 Million Loss[EB/OL]. South China Morning Post, 2024.

[24] Hazell J. Large Language Models Can Be Used To Effectively Scale Spear Phishing Campaigns[R/OL]. arXiv:2305.06972, 2023.

[25] Fang R, Bindu R, Gupta A, et al. LLM Agents Can Autonomously Exploit One-day Vulnerabilities[R/OL]. arXiv:2404.08144, 2024.

[26] CrowdStrike. 2025 Global Threat Report[R]. 2025.