计算
Computing Magazine of the CCF

基于大语言模型的自动化渗透测试研究

舒展 李宗鹏

Automated penetration testing based on large language models

SHU Zhan, LI Zongpeng

在线阅读 View online: http://js.xml-journal.net/article/doi/10.11991/cccf.202506005

基于大语言模型的自动化渗透测试研究

舒　展[1, 2, 3]　　李宗鹏[2, 3]

[1] 绿盟科技集团股份有限公司　　[2] 清华大学　　[3] 泉城实验室

背景

随着网络攻击手段的不断演进，渗透测试（penetration testing）作为识别系统脆弱性的关键环节，在网络安全体系中扮演着重要角色。传统渗透测试依赖人工操作，虽然具备灵活性和针对性，但存在执行周期长、覆盖面有限、依赖专家经验等局限，难以满足大规模复杂系统对高效、智能化安全评估的需求。近年来，自动化渗透测试（automated penetration testing）技术应运而生，通过流程标准化与工具集成提升了执行效率，然而在环境适应性、攻击策略优化和防护对抗等方面仍存在明显短板。与此同时，大语言模型（large language models, LLMs）在自然语言理解、推理与生成领域取得突破性进展，展现出在环境感知、任务推理、攻击链生成与动态决策支持等方面的独特潜力。为此，结合 LLMs 能力构建智能化渗透测试体系成为当前研究的新方向。本研究将围绕 LLMs 赋能自动化渗透测试展开综述，系统梳理领域发展现状、面临的主要挑战及未来研究方向，为后续智能攻防研究提供结构化的知识总结与方向参考。

DOI：10.11991/cccf.202506005

基金项目：国家重点研发计划项目（2022YFB2901300）；北京市自然科学基金项目（IS23057）；泉城实验室项目（QCLZD202304）

通信作者：舒展，E-mail:zshu1123557@gmail.com

[1] https://www.metasploit.com/, 2025−05−19

[2] https://nmap.org/, 2025−05−19

[3] https://www.tenable.com/products/nessus, 2025−05−19

渗透测试与自动化渗透测试概述

渗透测试通过模拟攻击者视角，主动探测信息系统中的脆弱性，验证漏洞的可利用性与潜在危害，从而评估系统防护能力。与传统的漏洞扫描不同，渗透测试更加关注攻击链的完整性，包括权限提升、横向移动和数据泄露等实际攻击效果的验证。因此，其更贴近真实威胁情境，是安全评估和防御能力验证的关键技术之一。

随着网络环境的日益复杂化和测试需求的快速增长，自动化渗透测试逐渐成为主流方向。该技术旨在将渗透测试从单点手工操作升级为流程化、系统化的自动执行过程，不仅涵盖信息收集与漏洞发现，还延伸至漏洞利用、影响验证及后渗透阶段，从而更接近红队演练（red teaming）的评估能力。当前，自动化渗透测试正朝着深度攻击链建模、策略决策优化与对抗规避等智能化方向发展，为构建高效、智能的攻防对抗体系奠定基础。

自动化渗透测试的发展现状

随着信息系统规模和复杂性的不断上升，传统人工驱动的渗透测试模式在效率、可扩展性和覆盖范围等方面逐渐暴露出不足。自动化渗透测试技术目前已经在多个方面取得初步进展。早期的自动化工具如 Metasploit[1]、Nmap[2] 和 Nessus[3] 借助模块化设计，分别实现了信息收集、漏洞扫描与漏洞利用环节的部分自

动化，有效降低了渗透测试的专业门槛。随后，AutoSploit[4]、Sn1per[5] 等框架进一步整合各类功能模块，初步实现了从资产发现到漏洞利用的端到端半自动化流程。

尽管当前已有工具实现了部分任务链条的串联，但多数自动化方案仍高度依赖规则引擎与预设脚本，难以应对复杂系统环境中防御机制的动态变化。具体而言，攻击链构建过程中常出现动作选择僵化、序列生成缺乏优化等问题，导致攻击路径在目标环境中易出现重复、失败或不可执行的情形。此外，自动化系统往往缺乏对目标环境反馈的动态感知与策略调整能力，使得在实际操作中渗透成功率与隐蔽性仍受限。

因此，尽管当前自动化渗透测试在提升执行效率、拓展攻击面等方面优势明显，但在智能化、自主化与环境适应性方面仍存在关键技术瓶颈，亟须引入更具推理与感知能力的新兴技术以满足未来复杂攻防场景需求。

大语言模型的发展现状

近年来，随着深度学习技术的持续演进，大规模预训练语言模型在自然语言理解、生成与推理等任务中取得了突破性进展。自 Transformer[1] 架构提出以来，用大量文本语料进行自监督训练的方法迅速成为主流，推动模型参数规模从 BERT [2] 的亿级发展至 GPT-3 [3] 的千亿级，甚至达到 GPT-4 的万亿级水平。通过结合指令微调（instruction tuning）与人类反馈强化学习（reinforcement learning from human feedback, RLHF）[4]，LLMs 在开放域问答、复杂任务规划、多轮对话等方面表现出强大的泛化能力和语义推理能力。

在网络安全领域，尤其是自动化渗透测试任务中，传统系统在信息建模、策略生成与行为执行方面往往面临泛化性差、缺乏上下文理解与动态适应的问题。而 LLMs 具备对自然语言和知识表征的强大处理能力，在用于漏洞情报理解、攻击链规划、策略生成及系统交互任务时，表现出良好的适应性和可迁移性。

近年来，一系列围绕 LLMs 的拓展技术也为智能化渗透测试提供了新的解决路径。例如，基于 LLMs 构建的智能体系统，如 AutoGPT[6] 和 OpenAI GPTs，通过引入工具调用、内存管理与任务规划能力，实现了在复杂测试任务中的多步推理与动态响应；检索增强生成方法[5] 将外部知识库与生成过程结合，有效弥补了模型知识盲区；而跨模态预训练模型，如 CLIP[6] 与链式思维提示（chain-of-thought prompting, CoT）[7] 则进一步提升了模型的推理深度、环境感知能力与解释性。

综上所述，LLMs 及其衍生机制正在推动渗透测试从以规则驱动为主的操作模式，迈向具备自主推理、上下文理解与目标导向规划能力的智能系统，为构建新一代自动化渗透测试体系提供了坚实的技术基础。

自动化渗透测试面临的挑战

随着自动化渗透测试从操作驱动走向智能决策，其核心挑战逐步贯穿于完整测试流程：在信息收集阶段，系统难以使建模目标网络的动态状态演化，导致攻击面认知滞后；进入路径规划阶段，又受限于稀疏反馈与奖励延迟，策略难以有效优化；在多阶段攻击执行过程中，系统缺乏条件依赖推理与结构化路径表达，导致攻击链碎片化；而在面对动态防御响应时，则缺乏策略重构与行为调整能力，流程易被中断。流程中面临的问题共同揭示了当前自动化渗透测试在环境建模、路径探索、攻击链生成与对抗适应 4 个关键环节的能力瓶颈。本研究将围绕这 4 类挑战展开系统分析，探讨自动化渗透测试面临的技术困难与智能推理与 LLMs 技术的应用潜力。

动态环境建模失真的挑战

自动化渗透测试的首要环节依赖于对目标环境的全面感知与精准建模，而现代网络架构呈现出高度可变性，基础设施通常处于频繁变更状态。云原生平台中的弹性伸缩机制、短生命周期的容器实例、微服务部署的变更和边界资产的动态暴露等特性，使得攻击面在渗透测试期间持续演化，呈现非静态特征。建模问题已在动态网络建模研究中被广泛关注。例如，有研究提出基于周期性扫描与实时变更检测机制构建演化

[4] https://github.com/NullArray/AutoSploit, 2025-05-19
[5] https://github.com/1N3/Sn1per, 2025-05-19
[6] https://github.com/Torantulino/Auto-GPT, 2025-05-19

型攻击图模型[8]，也有工作尝试结合被动流量分析与增量建模方法提高资产识别能力[9]。此外，针对多租户环境或虚拟化平台中状态不可观测的问题，一些研究引入了拓扑推断与节点状态估计机制以提升建模完整性[10]。

然而，目前使用的以上技术实现的大多数自动化测试系统仍采用基于静态快照构建拓扑模型、资产表或漏洞图谱的方法，难以及时捕捉节点上线下线、端口状态变动、策略调整等行为。这种建模失真不仅影响后续路径规划的有效性，也会导致对资产可达性、漏洞可用性和攻击链完整性的错误判断。

稀疏反馈导致路径探索受限的挑战

渗透测试任务中的策略优化使用环境反馈信号来评估操作效果。然而，在真实测试环境中，渗透行为通常以长序列形式展开，中间步骤很难立即获得成功信号，只有在攻击链尾部达成目标（如权限控制、敏感数据泄露）后才会触发奖励。这种奖励稀疏、延迟显著的特征极大制约了基于强化学习的路径探索策略更新效率。

已有研究指出，如何在网络安全场景下设计有效的奖励信号，是强化学习面临的核心挑战之一[11]。特别是在渗透路径优化任务中，传统方法依赖固定奖励结构，容易导致策略训练效率低下、探索空间震荡和陷入局部最优。

为应对上述问题，部分研究引入了外部反馈重塑机制。例如，通过引入模糊奖励信号或模拟安全分析师的行为反馈，以构建更具细粒度的奖励函数，从而提升策略更新频率[12]。此外，也有研究尝试结合行为轨迹摘要与未来状态预测，对尚未产生奖励的操作进行预估评分，以实现策略预修正[13]。这些方法在缓解稀疏信号问题方面取得了一定进展，但在多主机、多服务、复杂环境组合下，仍面临操作空间巨大、回报延迟极端的现实挑战。

因此，当前自动化渗透系统亟须引入具备先验知识引导、失败归因诊断与策略结构压缩能力的探索机制，以在稀疏奖励环境中实现更加高效的路径学习与动态决策。

攻击链缺乏阶段推理与解释能力的挑战

多阶段攻击链是现代复杂攻击的核心特征之一，其构建过程需明确各阶段操作之间的前置依赖、权限转移与策略目标，确保路径逻辑连贯、动作可执行。然而，当前许多自动化渗透测试系统仍以线性序列或脚本模板的形式表示攻击行为，未建模其中的因果关系、阶段性条件与目标导向逻辑，导致攻击链呈现出结构松散、上下文断裂甚至操作冲突。

尤其是面对需要权限层级递进或基于逻辑前提构建的复杂路径时，若缺乏阶段推理与依赖管理机制，系统生成的攻击序列可能出现冗余重复、执行失败或策略错误。此外，现有工具多输出日志级事件列表而非结构化的路径树或图结构，难以支持后续的路径验证、结果审计或防御模拟。

更重要的是，攻击链的可解释性在实战演练、风险评估与合规审查中至关重要。若渗透系统不能清晰地表达每个攻击步骤的意图、达成条件与影响路径，将大幅削弱其在安全决策中的应用价值。

因此，如何引入具备阶段建模、因果推理与路径结构约束能力的攻击链生成机制，已成为当前自动化渗透测试智能化演进的关键挑战之一。

防御策略变化下缺乏自适应能力的挑战

在渗透测试执行过程中，目标系统往往并非被动响应，而是具备动态防御能力。常见的安全机制如入侵检测系统（intrusion detection systems, IDS）、Web 应用防火墙（Web application firewall,WAF）、行为分析系统和蜜罐等，可基于访问特征实时调整策略，例如阻断连接、伪造返回结果或引导攻击者进入陷阱环境。这种防御动态性使得原定攻击路径可能中途失效，甚至导致测试系统暴露或被反制。

然而，当前大多数自动化渗透测试工具仍建立在静态环境假设之上，路径规划阶段通常未考虑防御行为的演化趋势，执行过程中也缺乏对异常响应的识别机制。一旦测试操作触发防御系统的策略调整，系统往往无法准确感知，仍按原定路径执行，最终导致攻击流程中断或执行无效。

即便个别系统具备一定的异常感知能力，也通常缺乏对策略变化的响应机制，例如动态路径重构、攻击方式切换、访问频率调整或阶段性回滚等。这使得系统在应对复杂对抗环境时表现出较差的稳定性、鲁棒性与实战能力。

因此，如何赋予自动化渗透测试系统策略感知、动态决策与行为适应的能力，以支持在多变防御环境下的持续推进，是推动其实用化部署的关键技术方向。

LLMs 赋能的自动化渗透测试可行技术路线

随着 LLMs 在自然语言理解、上下文推理、任务规划等方面能力的持续突破，相关研究已逐步将其引入到自动化渗透测试流程的关键环节，试图解决传统自动化体系在建模、探索、推理与对抗响应等方面的核心瓶颈问题。LLMs 具备上下文感知、链式思维、指令生成和对人类知识的泛化表达能力，能够有效支撑复杂环境下的高维状态表达、策略生成与行为适配。本章将按照自动化渗透测试的典型流程结构，结合提出的 4 项挑战，从环境建模、策略优化、攻击链推理与防御应对 4 个维度，系统分析当前 LLMs 赋能路径下的主要研究方法、典型技术类别及其能力边界。

动态环境建模中的上下文感知与语义补全

在信息收集阶段，面对网络结构动态演化与部分可观测性所带来的建模失真问题，LLMs 被广泛用于状态补全与攻击面建构。研究表明，语言模型在融合日志、配置、扫描信息等多源异构数据方面表现出高度适应性，能够实现对不完整拓扑图谱的结构化推断与语义补全。

例如，AttacKG+ 框架 [14] 利用 LLMs 从威胁情报文本中自动构建攻击知识图谱，结合结构模板与实体关系建模机制，补齐拓扑视图中缺失的节点与边，从而实现对攻击面结构的完整还原。该方法将 LLMs 与图谱构建系统协同部署，展现出良好的泛化能力与上下文保持能力。

同时，一些研究尝试将 LLMs 与资产状态估计模块结合，对目标系统中存在的信息缺失、主机配置变化等进行持续更新。Transformer-based 状态估计方法 [15] 能够利用历史交互记录建模目标资产的时序变化行为，从而增强系统对动态资产拓扑的持续感知能力，提升渗透测试在可变网络结构中的执行稳定性。

还有研究探索通过语义模板驱动方式，结合攻击图与资产依赖关系预测候选节点与潜在连接，进而构建更加完整的攻击路径候选图 [16]。此类方法依托于 LLMs 对自然语言实体与系统组件的理解能力，能够在缺失信息背景下有效填补关键节点的语义空白。

综合来看，基于 LLMs 的上下文感知与语义补全技术为解决传统建模机制在动态系统中的结构失真问题提供了全新视角，为后续攻击路径规划与策略生成奠定了坚实的环境认知基础。

稀疏反馈下的探索策略优化机制

在漏洞探测与路径评估阶段，强化学习技术常被用于训练智能体执行多步渗透操作。然而，渗透测试任务普遍面临反馈稀疏、奖励延迟、状态–动作空间庞大等挑战，严重影响策略学习效率与攻击路径质量。

为此，部分研究尝试将 LLMs 引入策略优化过程，作为策略生成器或行为引导器，为强化智能体提供策略先验或行动候选，显著缓解了探索初期路径震荡与学习失败的问题。例如，CyExec 系统[17] 展示了如何通过预定义提示词引导 LLMs 生成多样化的攻击路径，构建结构完整的训练样本，从而提升策略探索起点的合理性与覆盖面。

一种常见技术路径是结合轨迹摘要机制与语言模型生成系统，提取先前执行记录形成行为轨迹摘要，再由 LLMs 根据上下文生成高质量的下一步操作建议[18]。此方法在奖励延迟与反馈模糊的环境中表现出更强的目标对齐能力，有效避免策略陷入无效操作循环。

此外，Few-Shot 提示机制也被广泛应用于策略迁移与泛化任务。部分研究提出将专家演示序列转换为少量提示样本，通过 LLMs 实现跨任务迁移能力，在缺乏大规模训练样本条件下支持策略适配与零样本执行[19]。这一机制对于实际部署中面临多样化目标系统与配置组合的渗透测试任务尤为关键。

在动态反馈学习方面，一类方法如 Agent-R 框架[20] 通过引入语言驱动的反思模块与自我纠错机制，使 LLMs 智能体在交互过程中具备错误归因与策略修复能力，显著增强了其在稀疏环境下的鲁棒性。另一类方法如 MART 系统[21]，通过构建多轮红队自动训练过程，结合失败分析与反馈重构机制，提升策略对抗性与路径完成率。

综合来看，LLMs 的引入不仅优化了策略生成初始分布，还在长期推理、失败归因与跨场景泛化方面展现出显著优势，为自动化渗透测试在稀疏奖励环境下的

高效策略学习提供了关键支撑。

多阶段攻击链中的因果推理与路径可解释生成

多阶段攻击链往往包含信息收集、漏洞利用、权限提升、横向移动、数据外泄等多个环节，其构建不仅需要逻辑合理的操作序列，还必须体现出阶段间的条件依赖与策略递进。然而，传统自动化渗透测试系统在路径生成时常以离散操作为单位，忽略操作间的上下文联系，缺乏因果一致性与结构完整性，容易导致攻击流程中断或生成不可执行路径。

为解决此问题，LLMs 被引入到攻击路径构建过程中，用于建模操作之间的语义依赖与因果链条，实现结构化、可验证的攻击计划生成。一类典型方法是基于结构约束的攻击图建模，在路径生成中引入权限关系、阶段性目标与节点角色，结合 LLMs 构建条件逻辑推理模型，从全局角度生成可执行路径[22]。另一类研究则采用链式思维机制，将复杂攻击路径的构建任务分解为多个具备逻辑连贯性的子任务，逐步生成渗透流程中每一步的动作与目标[23]。这一思路特别适用于高复杂度目标场景下的逐步推进策略，能有效增强路径生成的可解释性与执行稳定性。

在系统实现层面，AutoGPT、LangChain 等智能体框架被引入渗透测试中，用于在路径执行失败时自动进行策略回滚与路径重构。Pentest Copilot 工具[24] 展示了基于 LLMs 的路径规划与命令生成闭环系统，能够结合当前环境状态动态生成后续攻击步骤，并给出自然语言形式的解释，显著提升了路径理解与人工验证效率。而 DeceptPrompt 系统[25] 展示了 LLMs 在路径构建过程中存在的潜在风险，即自然语言提示可能诱导模型生成逻辑上正确但安全上有害的路径，说明路径语义偏移问题亦需特别关注与防范。

综合而言，LLMs 在多阶段攻击链构建中不仅实现了因果结构的建模与链式路径生成，还赋予了路径行为以可解释性和条件控制能力，是当前自动化渗透测试实现高可信路径构建的核心突破口。

面向对抗响应的防御感知与策略自适应规划

在真实渗透测试过程中，目标系统的安全机制往往具备动态响应能力。当访问行为触发 IDS、WAF 或蜜罐系统的策略时，系统可能通过中断连接、篡改响应、重定向路径等方式进行实时对抗。若渗透系统无法感知并调整行为，攻击流程将面临中断风险，甚至暴露测试行为，影响整体隐蔽性与评估可信度。

LLMs 在响应感知、策略修复与路径重构方面展现出显著潜力。一类研究聚焦于通过解析系统返回的超文本传输协议（hypertext transfer protocol, HTTP）包、错误代码与行为模式，训练 LLMs 识别异常响应信号，并判断是否存在蜜罐或防御机制。例如，Sladić等[26] 提出的 shelLM 系统，利用 LLMs 构建了动态仿真的 Linux shell 蜜罐，实时生成可信响应，误导攻击者并捕获行为轨迹，展现出 LLMs 在防御行为生成与伪装识别中的双重应用潜力。

在对抗响应的策略规避方面，Mobile-LLaMA 模型[27] 展示了通过指令微调使 LLMs 具备通信协议分析与异常路由感知能力，可有效识别策略变更迹象并调整行为路径。而 VulnBot 框架[28] 则引入多智能体协作机制，结合 LLMs 进行扫描、利用与绕过防护策略的联合规划，显著提升了自动化渗透系统在真实场景下的动态对抗适应性。

在任务执行层面，Getting pwn'd by AI 项目[29] 展示了 LLMs 能够在安全外壳协议（secure shell, SSH）会话中识别防御策略反馈，动态调整提权路径并实现命令回滚，从而构建闭环型攻击流程。此外，有研究也指出 LLMs 的输出高度依赖于解码策略调节，Huang 等[30] 研究表明，通过 Sampling 参数控制可诱导模型生成具备绕过意图的攻击代码，提示当前策略生成尚存在可控性风险。

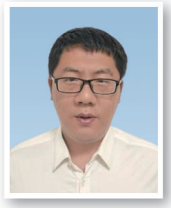
因此，在高对抗性环境中，如何利用 LLMs 构建具备防御识别、路径重构与策略回滚能力的自适应测试系统，不仅关系到攻击流程的连续性，也决定了系统在真实攻防演练中的生存能力与实用价值。

未来展望

随着网络环境的动态性和攻防博弈的复杂性持续增强，自动化渗透测试正面临向高智能化、自主化和环境适应性演进的迫切需求。LLMs 作为新一代智能推理与任务规划引擎，已在环境建模、路径生成、策略调

整等关键环节展现出显著潜力，推动自动化渗透测试体系迈向“智能代理主导、语言驱动执行”的新范式。

展望未来，智能化渗透测试的发展仍需解决多项关键挑战，包括高维模态融合下的环境感知能力提升、多目标多阶段路径推理的结构化表达、在对抗环境中保持稳定性的策略迁移机制，以及系统可信性与行为约束的可控设计。持续推进 LLMs 能力与安全场景的深度融合，有望实现具备上下文理解、意图驱动与动态决策能力的下一代智能渗透测试系统，为实战安全评估和防御体系强化提供有力支撑。 ■



舒　展

CCF 专业会员。清华–绿盟科技联合培养在职博士后，绿盟科技天枢实验室高级安全研究员。主要研究方向为智能安全运营、自动化渗透测试、网络安全。

zshu1123557@gmail.com

李宗鹏

清华大学教授，清华–绿盟联合实验室主任，国家重点研发计划项目负责人。主要研究方向为计算机网络、网络算法、网络编码、网络安全。

zongpeng@tsinghua.edu.cn

参考文献

[1] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//*Proceedings of the 31st International Conference on Neural Information Processing Systems*. California: ACM 2017: 6000−6010.

[2] DEVLIN J, CHANG Mingwei, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2018−10-11)[2025−05−15]. https://arxiv.org/abs/1810.04805v2.

[3] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[C]//*Proceedings of the 34th International Conference on Neural Information Processing Systems*. California: ACM 2020: 1877−1901.

[4] OUYANG, WU J, XU J, et al.Training language models to follow instructions with human feedback[EB/OL]. (2022−03−04)[2025−05−15]. https://arxiv.org/abs/2203.02155v1.

[5] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[C]//*Proceedings of the 34th International Conference on Neural Information Processing Systems*. Vancouver: NeurIPS. 2020: 9459−9474.

[6] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//*Proceedings of the 38th International Conference on Machine Learning*. Cambridge: ICML. 2021: 8748−8763.

[7] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. *Advances in Neural Information Processing Systems*, 2022, 35: 24824−24837.

[8] LI Q, WANG R, LI D, et al. DynPen: automated penetration testing in dynamic network scenarios using deep reinforcement learning[J]. *IEEE Transactions on Information Forensics and Security*, 2024, 19: 8966−8981.

[9] HAREESH R, SENTHIL KUMAR R K, KALLURI R, et al. Critical infrastructure asset discovery and monitoring for cyber security[J]. *Springer Nature Singapore*, 2022: 289–300.

[10] HOU T, WANG T, LU Z, et al. Combating adversarial network topology inference by proactive topology obfuscation[J]. *IEEE/ACM Transactions on Networking*, 2021, 29(6): 2779−2792.

[11] BATES E, MAVROUDIS V, HICKS C. Reward shaping for happier autonomous cyber security agents[C]//*Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*.Copenhagen: ACM, 2023: 221–232.

[12] WHITMORE S, HARRINGTON C, PRITCHARD E. Assessing the ineffectiveness of synthetic reinforcement learning feedback in fine-tuning large language models [EB/OL]. (2024−08−06)[2025−05−15]. https://doi.org/10.31219/osf.io/cvdzu.

[13] ZHANG Lei, PAN Zhisong, PAN Yu, et al. A hidden attack sequences detection method based on dynamic reward deep deterministic policy gradient[J]. *Security and Communication Networks*, 2022, 2022(1): 1488344.

[14] ZHANG Y, ZHAO Z, ZHOU M, et al. AttacKG+: boosting attack knowledge graph construction with large language models[J]. *Computers & Security.*, 2025, 150: 1−16.

[15] SONNTAG V, STEFFEN B, UTSCHICK W, et al. Transformer-based state estimation for tracking: maneuvering target and multi-target capabilities[C]// 2024 IEEE Radar Conference. Piscataway: IEEE, 2024: 1−6.

[16] ZHAO J, WANG S, ZHENG Z, et al. Graph neural network - nased attack prediction for communication - based train control systems[J/OL]. *CAAI Transactions on Intelligence Technology*. (2023−05−12). https://doi.org/10.1049/cit2.12288.

[17] MUDASSAR YAMIN M, HASHMI E, ULLAH M, et al. Applications of LLMs for generating cyber security exercise scenarios[J]. *IEEE Access*, 2024, 12: 143806−143822.

[18] XU Jiacen, STOKES J W, MCDONALD G, et al. Autoattacker: a large language model guided system to implement automatic cyber-attacks[EB/OL]. (2024−03−02)

[2025−05−15]. https://arxiv.org/abs/2403.01038.
[19] DANCE C R, PEREZ J, CACHET T. Conditioned reinforcement learning for few-shot imitation[C]// *International Conference on Machine Learning*. Cambridge: PMLR, 2021: 2376–2387.
[20] YUAN S, CHEN Z, XI Z, et al. Agent-R: training language model agents to reflect via iterative self-training[EB/OL]. (2025−01−20)[2025−05−15]. https://arxiv.org/abs/2501.11425.
[21] GE S, ZHOU C, HOU R, et al. MART: improving LLM safety with multi-round automatic red-teaming[EB/OL]. (2023−11−13)[2025−05−15]. https://arxiv.org/ abs/2311.07689.
[22] LUO L, ZHAO Z, GONG C, et al. Graph-constrained reasoning: faithful reasoning on knowledge graphs with large language models[EB/OL]. (2024−10−16)[2025−05−15]. https://arxiv.org/abs/2410.13080.
[23] WU Lei, ZHONG Xiaofeng, LIU Jingju, et al. PTGroup: an automated penetration testing framework using LLMs and multiple prompt chains[C]//*Advanced Intelligent Computing Technology and Applications*. Singapore: Springer Nature Singapore, 2024: 220−232.
[24] GOYAL D, SUBRAMANIAN S, PEELA A. Hacking, the lazy way: LLM augmented pentesting[EB/OL]. (2024−09−14)[2025−05−15]. https://arxiv.org/abs/2409.09493v1.
[25] WU Fangzhou, LIU Xiaogeng, XIAO Chaowei. Deceptprompt: exploiting LLM-driven code generation via adversarial natural language instructions[EB/OL]. (2023−12−07)[2025−05−15]. https://arxiv.org/abs/2312.04730v2.
[26] SLADIĆ M, VALEROS V, CATANIA C, et al. LLM in the shell: generative honeypots[C]//*2024 IEEE European Symposium on Security and Privacy Workshops*. Piscataway: IEEE, 2024: 430−435.
[27] KAN K B, MUN H, CAO Guohong, et al. Mobile-LLaMA: instruction fine-tuning open-source LLM for network analysis in 5G networks[J]. *IEEE Network*, 2024, 38(5): 76−83.
[28] KONG He, HU Die, GE Jingguo, et al. VulnBot: autonomous penetration testing for a multi-agent collaborative framework[EB/OL]. (2025−01−23)[2025−05−15]. https://arxiv.org/abs/2501.13411v1.
[29] HAPPE A, CITO J. Getting pwn'd by AI: penetration testing with large language models[C]//*Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. New York: ACM, 2023: 2082–2086.
[30] HUANG Y, GUPTA S, XIA Mengzhou, et al. Catastrophic jailbreak of open-source LLMs via exploiting generation [EB/OL]. (2023−10−10)[2025−05−15]. https://arxiv.org/abs/2310.06987v1.

Automated penetration testing based on large language models

SHU Zhan[1, 2, 3], LI Zongpeng[2, 3]
1. NSFOCUS Tech., Co., Ltd.
2. Tsinghua University
3. Quancheng Laboratory

Abstract: With the continuous evolution of cyberattack techniques, automated penetration testing—an essential approach for assessing system vulnerabilities—faces significant challenges, including dynamic network environments, sparse feedback signals, complex multi-stage attack planning, and adaptive defense mechanisms. In recent years, large language models (LLMs) have demonstrated remarkable capabilities in natural language understanding, contextual reasoning, and multi-step task planning, offering new opportunities for building intelligent penetration testing systems. This paper systematically analyzes four major challenges in automated penetration testing and reviews representative LLM-powered solutions across four key aspects: dynamic environment modeling, strategy optimization under sparse rewards, causal multi-stage path reasoning, and adaptive planning against evolving defenses. The findings show that LLMs exhibit promising context-awareness, causal inference, and behavioral adaptability, significantly enhancing the intelligence and robustness of automated testing frameworks. Finally, this paper outlines future directions for LLM-enabled penetration testing, including multi-modal integration, goal-driven attack reasoning, adaptive security evaluation, and trustworthy system design, providing theoretical guidance and technical reference for the next generation of intelligent red teaming systems.

Keywords: automated penetration testing; large language models; intelligent offense and defense; attack chain planning; strategy generation and adaptation; cybersecurity assessment; red teaming; penetration agents

摘　要: 随着网络攻击手段的持续演进，自动化渗透测试作为系统脆弱性评估的重要技术手段，在实践中面临环境动态性强、反馈稀疏、路径构建复杂和防御策略多变等挑战。近年来，大语言模型（large language models, LLMs）在自然语言理解、上下文推理与多轮任务规划方面展现出显著能力，为构建智能化渗透测试体系提供了新的技术路径。为此，围绕自动化渗透测试的典型流程，系统梳理当前面临的四类关键挑战，并从环境建模、策略探索、路径生成与防御适应 4 个维度，综述了 LLMs 在支撑自动化渗透任务中的典型方法与关键进展。研究表明，LLMs 具备较强的上下文感知、因果推理与动态调整能力，可有效提升自动化渗透

系统的环境适应能力与策略生成智能水平。最后，展望了未来基于 LLMs 的渗透测试系统在多模态融合、目标驱动推理、自适应安全测试与系统可信性保障等方向的发展潜力，旨在为智能攻防技术的发展提供结构化的研究梳理与技术参考。

关键词: 自动化渗透测试；大语言模型；智能攻防；攻击链规划；策略生成与调整；网络安全评估；红队演练；渗透智能体
中图分类号: TP393.0

中文引用格式: 舒展, 李宗鹏. 基于大语言模型的自动化渗透测试研究 [J]. 计算, 2025, 1(2): 23−30.
英文引用格式: SHU Zhan, LI Zongpeng. Automated penetration testing based on large language models[J]. *Computing Magazine of the CCF*, 2025, 1(2): 23−30.

CCF 运城学院学生俱乐部成立

2025 年 5 月 29 日，CCF 运城学院学生俱乐部成立会议在运城学院举行。会议采用无记名投票的方式，差额选举产生了首届执行委员会（名单附后）。CCF 运城学院学生俱乐部将充分整合 CCF 的优质学术资源与行业平台优势，通过策划学术讲座、技术研讨、实践竞赛等多元化活动，为运城学院学子打造拓宽学术视野、精进专业技能的创新型成长平台。

附：CCF 运城学院学生俱乐部首届执行委员会名单

督导主任: **赵满旭**　运城学院数学与信息技术学院副教授
委　　员: **杜经纬**　运城学院数学与信息技术学院副教授
　　刘慧珍　运城学院数学与信息技术学院团总支书记
主　　席: **董泽彬**
候任主席: **张洋洋**
执　　委: **梁一帆**　**万世兴**　**张瑞恒**

CCF 运城学院学生俱乐部成立大会合影

CCF 学生俱乐部: 为推动计算领域学术交流与技术创新，助力 CCF 学生会员提升专业素养，CCF 设立“CCF 学生俱乐部”。俱乐部成立规则和权益同CCF 学生分会，CCF 学生俱乐部主要面向以本科生、专科生为主的学校，打造学术交流与实践平台。期待更多以本科生、专科生为主的学校加入 CCF 学生俱乐部，共同推动计算领域的发展与进步。

有意成立 CCF 学生俱乐部的学校，可联系 membership@ccf.org.cn。