AI应用于进攻性安全

AI Technology and Risk Working Group

CSA GCR cloud security alliance GREATER CHINA REGION

CSA cloud security alliance

人工智能技术与风险工作组的永久官方地址是

https://cloudsecurityalliance.org/research/workgroups/ai-technology-and-risk

©2025 云安全联盟大中华区 —— 保留所有权利。您可以下载、存储、在您的计算机上显示、查看、打印以及链接到云安全联盟网站 https:///cloudsecurityalliance.org，但须遵守以下条件：(a)草案仅供您个人、信息性和非商业用途使用；(b)不得以任何形式修改或更改草案；(c)不得重新分发草案；(d)不得移除商标、版权或其他声明。在遵循中华人民共和国著作权法相关条款情况下合理使用本文内容，使用时请注明引用于云安全联盟大中华区。

云安全联盟

创立于2009年，作为世界领先的独立、权威国际产业组织，致力于定义和提高业界对云计算和下一代数字技术安全最佳实践的认识和全面发展。

云安全联盟大中华区

在香港注册并在上海登记备案的国际NGO组织，旨在 立足中国，连接全球，推动 中国数字安全技术标准与产业的发展及国际合作。

4 大区
大中华区、美洲区、欧非区、亚太区

180+ 分会
英国、法国、加拿大、旧金山、马来西亚等覆盖50多个国家和地区

2.5K+ 成员单位
世界500强科技公司、安全厂商、中小型企业、研究机构

20W+ 社区专业人员
研究工作组专家、社区志愿者、从业人员、CSA认证学员

前沿研究

#云安全 #AI安全
#零信任 #数据安全
#5G安全 #区块链安全
#量子安全 #物联网安全
#金融安全 #医疗安全
#智能座舱安全
#关键基础设施安全
.....

培训与认证

CCSK 云安全知识认证
CDSP 数据安全认证专家
CAISP人工智能认证专家
CZTP 零信任认证专家
CCPTP 云渗透测试认证专家
CCAK 云计算审计知识认证
.....

会议活动

CSA summit@RSAC
CSA GCR Congress
CSA研讨会
AI For GOOD峰会
.....

评估与认证

AI STR AI安全、可信、负责任认证
STAR 云安全评估认证
CAST 云应用安全可信认证
CNST 云原生安全可信认证
.....

1000+研究成果 **10W+认证学员** **1000W+传播量**

成员单位(部分)

ZEEKR Google Microsoft amazon HUAWEI intel IBM Security cisco ORACLE Meta 阿里云
UCLOUD优刻得 京东 上海CA中心 移动云 天翼云 中国电子云 CECLOUD China unicom中国联通 中国移动 China Mobile 中国电信 CHINA TELECOM ICBC 中国工商银行 LEXIN乐信

企业合作微信号:csagcr 认证培训微信号:CSAlynn 邮箱:info@c-csa.cn

致谢

《AI应用于进攻性安全》由CSA工作组专家编写，CSA大中华区组织AI安全工作组进行翻译并审校。

中文版翻译专家组

翻译组成员：

崔鉴　　潘季明　　侯芳　　卜宋博　　党超辉

审校组成员：

张淼　　卞超轶　　高健凯　　卜宋博

英文版本编写专家

倡议负责人：

Adam Lundqvist　　Kirti Chopra

主要作者：

Adam Lundqvist

Kirti Chopra

Michael Roza

Sven Vetsch

© 2025云安全联盟大中华区版权所有

贡献者：

Candy Alexander

David McCrory

Elier Cruz

Govindaraj Palanisamy

John Jiang

Keith Pasley

Ken Walling

Lars Ruddigkeit

Maria Schwenger

审校组：

Adam Lomas

Akhil Mittal

Akshat Vashishtha

Alex Rebo

Ashish Vashishtha

Ken Huang

Nancy Kramer

Neil Kessler

Patrick Schmid

Rocking Wolf Ranch

© 2025云安全联盟大中华区版权所有

Sven Olensky

Vaibhav Malik

Venkatesh Gopal

Vivek Shitole

联席组长：

Chris Kirschke

Mark Yanalitis

CSA 全球工作人员：

Josh Buker Stephen Smith

© 2025云安全联盟大中华区版权所有

目 录

致谢 4

执行摘要 9

关键发现 9

建议 10

引言 10

进攻性安全 10

进攻性安全的当前挑战 13

人工智能（AI） 15

大语言模型（LLM） 15

AI智能体 16

AI驱动的进攻性安全 18

AI增强和自主性 19

侦察 22

扫描 23

漏洞分析 25

利用 26

报告 29

© 2025云安全联盟大中华区版权所有

威胁行为者对AI的使用 30

AI驱动进攻性安全 – 不久的将来 32

挑战与限制-风险与应对措施 35

治理、风险和合规（GRC） 40

可信赖的AI实施 40

第三方风险管理 41

GRC考量因素 41

GRC总结 43

结论 43

参考文献 44

© 2025云安全联盟大中华区版权所有

执行摘要

人工智能（AI）技术，特别是大语言模型（LLM）和由 LLM 驱动的 AI 智能体（AI Agent）的出现，引发了进攻性安全（Offensive Security）领域深刻变革，包括漏洞评估、渗透测试和红队演练。这一转变将 AI 从一个狭窄的应用场景重新定义为一种多功能且强大的通用技术。本文探讨了基于 LLM 的 AI 变革潜力，通过研究其与进攻性安全的集成，以解决当前挑战，并展示了 AI 在五个安全阶段——侦察、扫描、漏洞分析、利用以及报告中的能力。

关键发现

进攻性安全的挑战：安全团队面临专业人才短缺、环境日益复杂和动态变化，以及平衡自动化测试与手动测试的需求。

AI 的能力：AI，主要通过 LLM 和 AI 智能体，提供了在进攻性安全方面的卓越能力，包括数据分析、代码、文本生成、规划现实攻击场景、推理和工具编排。这些能力可以辅助自动化侦察、优化扫描流程、评估漏洞、生成综合报告，甚至自主利用漏洞。

AI 的益处：在进攻性安全中利用 AI 可以提高可扩展性、效率、速度，发现更复杂的漏洞，并最终提升整体安全态势。

没有银弹：尽管前景可期，但目前还没有任何一种 AI 解决方案能够彻底改变进攻性安全。需要对 AI 进行持续实验，以找到并实施有效的解决方案。这就需要

© 2025云安全联盟大中华区版权所有

创建一个鼓励学习和发展的环境，团队成员可以使用 AI 工具和技术来提升他们的技能。

建议

AI 集成：可将 AI 融入自动化任务中以增强人类能力。利用 AI 进行数据分析、工具编排、生成可操作的建议，并在适用的情况下构建自主系统。在进攻性安全中采用 AI 技术可以在不断发展的威胁环境中保持领先优势。

人类监督：由 LLM 驱动的技术是不可预测的，可能会产生幻觉并导致事实性错误。持续使用人类监督来验证 AI 的方式，提高输出质量，并确保技术优势。

治理、风险和合规（GRC）：实施健全的 GRC 框架和控制措施，确保 AI 的安全状态、安全行为和符合道德规范。

本文使执行领导层，包括首席信息安全官（CISO）、网络安全战略师和高管，能够通过有效阐述进攻性安全投资的价值，以确保增强进攻性 AI 能力所需的资源。

引言

进攻性安全

进攻性安全涉及主动模拟攻击者的行为，使用类似于对手的战术和技术来识别系统漏洞。通过了解潜在的弱点和威胁，组织可以实施和加强可靠的安全控制，从而降低被恶意行为者利用的风险。

© 2025云安全联盟大中华区版权所有

为了最大限度地提高进攻性安全的有效性，确保其与组织的长期目标和宗旨一致非常关键。这种方法侧重于与组织优先级最相关的安全风险，确保资源被使用到最重要的领域。本文档的其余部分将探讨进攻性安全的三种方法：

- **漏洞评估**：可以使用扫描器自动识别脆弱点。
- **渗透测试**：可以用来模拟网络攻击，以识别和利用漏洞。
- **红队演练**：可以用来模拟强大对手进行的复杂、多阶段攻击，通常用于验证组织的检测和响应能力。

这些方法有相似之处，但在很多方面也有差异，如下表所示。虽然此表提供了一个有用的概览，但实际做法可能会因各种因素而有所不同，包括组织的成熟度和风险承受能力。

类型/方面	漏洞评估	渗透测试	红队演练
持续时间	短（小时）	中（天）	长（周）
风险对齐	间接与组织风险相关	受组织风险影响	基于组织风险
工具	自动化为主	自动化/手动/定制	高度手动/定制
复杂性	低	中	高
执行	未伪装	未伪装	伪装（隐蔽）
成本	低	中	高
目标	识别和优先考虑潜在漏洞	确定与一系列系统漏洞相关的风险	衡量组织整体的影响和响应

表 1：进攻性安全测试实践

© 2025云安全联盟大中华区版权所有

尽管进攻性安全测试的技术、广度和深度各不相同，但入侵通常遵循五个阶段：侦察、扫描、漏洞分析、利用和报告。在开始任何测试之前，创建工作说明书（SoW）或范围说明书并建立约定规则至关重要。这些基础文件为下文讨论的五个阶段奠定了基础。在本文的其余部分中，我们将讨论这五个阶段，同时假设范围和约定规则已经定义。

图 1 进攻性安全测试阶段

侦察——侦察是任何进攻性安全策略的初始阶段，旨在收集有关目标系统、网络和组织结构的广泛数据。

扫描——扫描需要全面检查已识别的系统，以发现关键细节，如活动主机、开放端口、运行的服务和所采用的技术，例如通过指纹识别漏洞。

漏洞分析——漏洞分析进一步识别系统、软件、网络配置和应用程序中的潜在安全弱点，并对其进行优先级排序。

利用——利用包括主动利用已识别的漏洞来获得未经授权的访问或在系统中提权。

报告——报告阶段通过系统地将所有发现汇编成一份详细的报告来说明进攻性安全测试工作。

© 2025云安全联盟大中华区版权所有

进攻性安全的当前挑战

进攻性安全测试人员在日益复杂的环境中面临许多挑战，这些挑战可能会阻碍他们评估的效率和有效性。这些挑战因缺乏熟练的安全测试人员和网络安全专家而进一步加剧。

- **扩大的攻击面**：新技术的普及，如 AI、区块链、云计算、物联网等，以及远程工作人员的增加，已经成倍地扩大了攻击面。这使得识别和保护所有潜在入口点变得更加困难。
- **高级威胁**：对手采用更复杂的技术，如无文件恶意软件或结合零日漏洞的离地攻击（living-off-the-land），难以检测和应对。
- **评估的多样性**：进攻性安全团队必须精通多样化的评估手段，包括漏洞评估、渗透测试和红队演练。每种类型都需要特定的技能、技术和知识，这使得测试人员难以在所有领域保持专业水平，例如，测试手机银行应用程序所需的技能与评估物联网设备不同。
- **适应动态环境**：进攻性安全评估通常在动态环境中进行，目标系统、安全控制和配置迅速变化。测试人员必须时刻保持敏锐，并实时调整他们的战术、技术和程序（TTPs），修改攻击策略，当攻击的一部分被检测到时进行切换，或根据观察到的用户行为调整社会工程学方法。
- **平衡自动化与手动测试**：自动化工具对于效率至关重要，但过度依赖可能导致漏洞的漏报和因误报而分散注意力。在自动化扫描和深入手动分析之间找到正确的平衡是一个持续的挑战。

© 2025云安全联盟大中华区版权所有

- **耗时的任务**：某些任务如全面的授权测试、代码审查、制作鱼叉式网络钓鱼电子邮件和利用复杂漏洞，本质上是耗时的。它们通常涉及大范围的侦察和反复的利用尝试。
- **工具开发与定制**：进攻性安全测试人员通常需要开发或调整脚本、工具或框架，以发现独特的漏洞或适应特定的目标环境，这可能会对有限的资源造成压力。
- **沟通与汇报**：在评估期间和之后与利益相关方进行清晰有效的沟通同样重要。有效地传达技术发现，将其转化为可操作的建议，并提供与不同受众产生共鸣的简洁报告可能是一个重大挑战。当安全测试人员与其受众之间存在巨大的知识鸿沟时尤其如此，这使得技术复杂性、业务优先级和风险容忍度难以协调。
- **数据分析与威胁情报**：在评估期间产生的数据量可能规模巨大。提取有价值的数据，关联发现，并保持对最新威胁情报的了解需要持续的警惕和先进的分析技术，同时还要应对资源和人员有限的问题。
- **合规与道德考量**：进攻性安全测试人员必须遵守越来越多的严格安全标准、法规和道德指导方针，确保他们的行动不会造成意外伤害或超出评估约定的范围。这对测试人员来说既费资源也耗时。

网络安全专业人员的短缺和网络攻击的日益复杂化创造了对创新解决方案的迫切需求。AI，特别是 LLM，为应对这些挑战提供了有希望的思路。AI 可以缓解对人力资源的压力，显著增强进攻性安全测试人员的能力，并普遍提高进攻性安全实践的有效性。

© 2025云安全联盟大中华区版权所有

人工智能（AI）

AI 包含一系列旨在模拟人类智能的技术，包括自然语言处理、机器学习和机器人技术。尽管 AI 覆盖了广泛的技术范围，我们的焦点集中在由 LLM 驱动的技术。

大语言模型（LLM）

LLM 是具有数十亿参数的复杂深度神经网络，能够处理和生成语言。最初，这些模型被开发用于预测句子中的后续单词，现在它们能够处理文本、语音、音频和视频应用中的复杂任务。

LLM 基于机器学习原则运作，包括两个关键阶段：训练和推理。在训练阶段，LLM 分析大量文本数据，识别语言模式，并学习预测后续单词和生成连贯的语言。在推理阶段，利用学到的模式来处理新的文本输入，生成预测，完成句子，并提供相关响应。

LLM 不仅限于语言处理。它们擅长快速分析大量数据，包括文本、代码、日志和 HTTP 流量。利用其生成能力，它们可以创建代码、脚本和电子邮件，以及编制摘要和报告。最先进的 LLM 展示了如文本**推理**和做出程序性决策等涌现能力，这对于**规划**和**目标导向任务**至关重要。为了与许多关于 AI 和 AI 智能体的论文中使用的术语保持一致，我们使用“推理”一词来描述分析文本和做出程序性决策的能力。然而我们承认，关于 AI 智能体是否能像人类一样进行推理，有待持续研究。

© 2025云安全联盟大中华区版权所有

尽管 LLM 提供了令人印象深刻的能力，但它们并非没有局限性。例如，它们有时会生成听起来合理但事实上不正确的内容，这种现象被称为“**幻觉**”。当准确信息至关重要时，这可能是个问题，但在创意应用如头脑风暴中可能是有益的。此外，LLM 包含随机元素，这些元素增强了它们的输出的多样性和自然性，但也**降低了可预测性**。因此，类似的查询可能会产生完全不同的响应，这与传统计算系统不同。

LLM 的应用，无论是局限于聊天窗口还是用于做出现实世界决策，可能取决于它所解决的具体任务。虽然基于聊天的 LLM 已广泛用于内容创作和创意目的，但当前的研究正在探索如何利用这些模型来自主解决复杂问题，这也是朝着更广泛利用 AI 迈出的重要一步。

AI 智能体

AI 智能体是设计用于感知周围环境并采取行动以实现设定目标的自主或半自主系统，从而塑造与环境的未来交互。这些智能体可以使用 LLM 的能力来规划任务、触发任务执行、做出决策，并与世界有意义地互动。与基础的 LLM 应用不同，使用 LLM 的 AI 智能体遵循一种循环方法来实现其最终目标，不断从其发现中学习和适应，并不断优化其方法。这种迭代式的自我适应能力，使智能体能够有效地通过多步骤过程解决复杂问题，直到任务完成。

© 2025云安全联盟大中华区版权所有

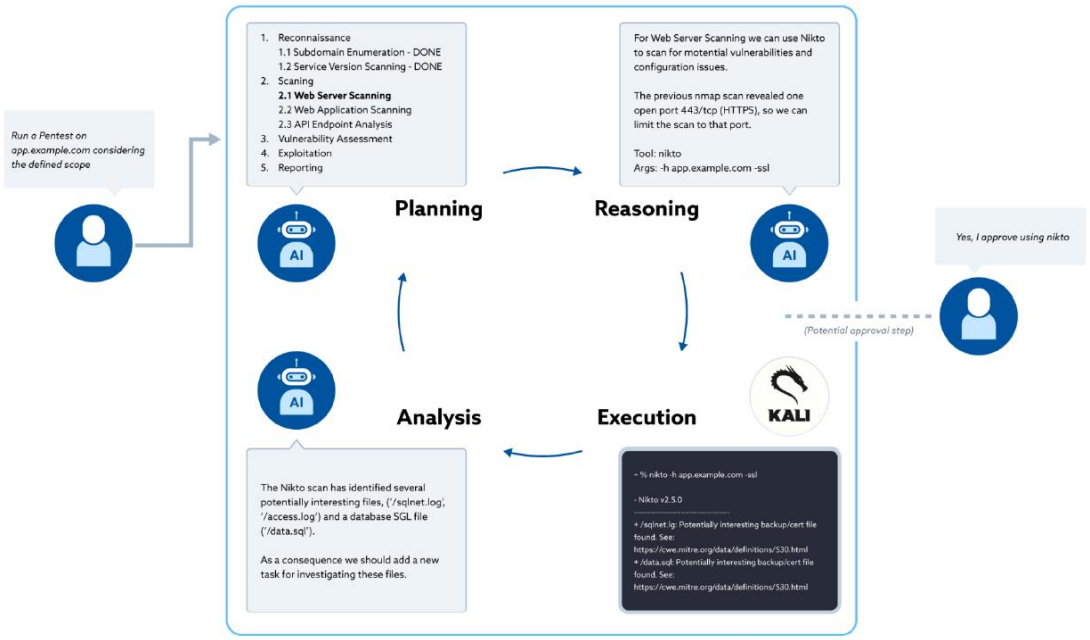

图 2 AI 智能体的阶段

如图 2 所示，是一个 AI 智能体处理针对运行在 app.example.com 的 Web 应用程序进行完整渗透测试的简化示例，我们研究 AI 智能体的一个周期，即 Web 服务器扫描阶段。这个周期可以详细描述如下：

1. **规划**：基于先前周期中创建的现有计划和信息，智能体优先考虑 Web 服务器扫描任务。
2. **推理**：智能体需要选择出适当的工具，在这个案例中将 Nikto 用于 Web 服务器扫描。
3. **执行**：选定的工具针对目标执行。根据其配置，它可以直接触发工具或耐心等待人类批准，提供灵活且用户友好的交互体验。

© 2025云安全联盟大中华区版权所有

4. **分析**：智能体处理 Nikto 的输出，识别潜在的漏洞，如暴露的配置文件和数据库。然后，它通过建议对计划进行更改来调整其未来周期的执行，可能在下一个计划阶段添加对特定发现开展进一步调查的任务。

这样一个简化的 AI 智能体模型可能由于上下文的潜在广度而面临局限性，导致智能体失去对目标的跟踪。为了缓解这一点，发展了多智能体系统的概念，它在复杂问题解决方面显示出了有希望的结果。这类系统由一系列 AI 智能体组成，它们协作完成任务。此外，复杂的短期和长期记忆管理至关重要，通常涉及通过检索增强生成（RAG）系统访问其自身训练数据集之外的外部权威知识库。

RAG 基于其自身训练数据集之外的权威知识库对 LLM 输出进行上下文知识补充。这种方法提高了 LLM 生成响应内容的相关性和准确性。通过利用内部和外部知识，AI 智能体可以提供更全面的解决方案。这种集成对于保持上下文和扩展 LLM 生成特定于组织或领域的输出的能力至关重要。

AI驱动的进攻性安全

AI 技术正在开辟进攻性安全领域的新天地，借助 AI 驱动的工具，我们可以模拟高级网络攻击，并在恶意行为者利用之前识别网络、系统和软件的漏洞。这些工具可以帮助安全团队更有效地扩展其工作，提升工作效率。AI 的运用不仅能够覆盖更广泛的攻击场景，还能根据最新的漏洞发现做出动态响应，来适应多变的网络环境，并在实战中不断学习和进化。

© 2025云安全联盟大中华区版权所有

AI 模型能够提出攻击策略，自动生成并执行未曾见过的测试用例，并从每次交互中学习。这些 AI 驱动的工具还可以处理海量数据，挖掘出人类无法识别的模式，并在漏洞发现方面提供帮助。

然而，AI 解决方案并非灵丹妙药。它们的有效性受限于训练数据和算法的范围。对于一些新颖或复杂的情景可能会超出了它们的能力范围。在解读异常情况、运用判断力以及做出需要综合考量的战略决策方面，人类的专业技能依然不可或缺。因此，了解认识当前 AI 技术的最新现状并将其作为人类安全专家的辅助工具是非常必要的。

AI 增强和自主性

如上所述，AI 智能体可以自主或半自主地通过规划、推理、工具执行和分析等周期性循环来执行进攻性安全任务，并引入不同程度的人为输入。

然而，授予 AI 智能体的自主权程度必须权衡自动化和增强技术带来的收益和意外后果引发的风险，尤其是在关键应用中。确保有人为监督——或保持“人在回路”——是一种战略优势。这种策略对于将 AI 独特的优势与人类专业知识相结合，从而获得最佳结果至关重要。它确保了质量标准和问责制，特别是在那些生产环境中进行操作的敏感领域。

AI 可以通过如下方式增强或自动化现有的进攻性安全测试流程：

© 2025云安全联盟大中华区版权所有

阶段	任务	AI增强
侦察	• OSINT • 发现潜在入口 • 研究目标行业与安全 • 总体测试/接触计划	• 数据收集 / OSINT 自动化 • 数据汇总 • 威胁态势分析 • 适应性测试规划
扫描	• 系统识别 • 指纹技术 • 漏洞扫描 • 代码扫描	• 原始输出与流量分析 • 异常检测 • 漏洞模式识别 • 脚本生成
漏洞分析	• 进一步分析漏洞 • 将发现与数据库进行关联 • 评估可利用性 • 对漏洞进行优先级排序	• 风险影响评估 • 漏洞分级 • 降低误报 • 根本原因分析
利用	• 漏洞利用 • 在目标环境中驻留 • 规避安全控制 • 提权并横向移动	• 漏洞研究/选择 • 负载/恶意软件生成与对齐 • 渗透后引导 • 社会工程模拟
报告	• 记录全部发现结果 • 评估漏洞优先级并提出修补建议 • 提供可用情报 • 报告成果	• 报告生成 • 总结并结果可视化 • 攻击路径建模 • 简化技术成果 • 建立知识库

表 2：进攻性安全各阶段能够被 AI 增强的人类任务

如上表所示，安全测试人员可以在进攻性安全测试的所有阶段中充分利用 AI。AI 可以被授予不同程度的自主性，允许在遵守法规和组织政策的同时，定制自动化和增强之间的平衡。

下图提供了一个概念性的视图，展示了如何实施和管理对 AI 智能体依赖程度的增长。

© 2025云安全联盟大中华区版权所有

图 3：进攻性安全过程中的 AI 智能体

- 无自主智能：在没有 AI 的情况下，人类手动或通过人工操作工具执行所有任务。
- 低自主智能：赋予 AI 低级自主权，包括协助完成特定任务，如数据分析或规划，但决策和执行仍需要人类监督。
- 高自主智能：赋予 AI 高级自主权，使其能够在不同阶段自主执行任务，几乎不需要人类干预，但人类仍然需要保持全程监督。

随着 AI 驱动的系统变得更加先进，可以赋予它们更多的自主性，从而实现在各种场景下全程自主工作。人类干预的程度取决于 AI 的能力、应用范围、预定义的交互规则以及可信度。这一进步确保了在进攻性安全中充分发挥 AI 的全部潜力的同时，AI 驱动的任务依然有效。通过负责任地整合 AI 能力，无论自主性水平如何，进攻性安全团队都能支持具有适应性和韧性的安全策略。

© 2025云安全联盟大中华区版权所有

在接下来的章节中，我们将探讨 AI 能力（特别是 LLM）在进攻性安全的上述五个阶段中的适用性。

侦察

侦察阶段的目标是通过公开来源和被动侦察技术中收集关于目标的详细信息。这包括识别目标的网络范围、域名、员工详细信息、使用的技术以及在安全公告中披露的潜在漏洞。在这一阶段中使用的工具从网站到社交媒体，再到网络流量分析和各种数据库不等，旨在准确绘制组织的攻击面。与外部的威胁者不同，进攻性安全测试人员可以不时地访问内部资源，如网络（拓扑）图、设计文档或配置管理数据库（CMDB），这些资源加速并强化了侦察工作。在像白盒测试这样的场景中，由于已经可以获取到内部知识，因此侦察活动可以简化甚至不必进行。

在侦察阶段，需要从大量数据中筛选出相关信息，同时丢弃过时或无关的数据，这是一个重要的挑战。AI 解决方案通过自动化数据收集和分析，高效地识别相关信息并丢弃多余数据。此外，AI 可以作为强大的助手，简化侦察过程，使安全测试人员能够专注于关键分析，更快做出更好的决策，并制定更好的推理和攻击策略。

自适应测试规划：AI 智能体可以自动分析大量资源并利用模式数据库，为指定地系统和配置生成定制的测试用例。自动化相比手动创建测试用例减少了时间和精力投入，同时增加了覆盖范围。Pentest GPT 将总体测试目标转化为精确、可执行的步骤，提高了准确性和可靠性。通过整合来自持续工具分析和数据评估的

© 2025云安全联盟大中华区版权所有

结果，这种自适应方法还能定制情境感知的测试策略。AI 智能体可以根据可利用性、严重性和影响程度动态地优化可操作的步骤的执行顺序。

数据分析：利用 LLM 分析来自各种网络服务（如 SMTP、FTP、HTTP）或工具的响应，可以详细描绘目标的网络基础架构，包括 IP 地址范围、域名、网络拓扑、供应商技术以及所使用的 SSL/TLS 加密、端口和服务的类型。LLM 强大的数据分析能力有望显著简化研究工作，并提高研究结果的质量。

用于数据收集的工具编排：AI 可以高效地制定请求、查询和命令行参数。有研究专门利用 AI 通过自然语言制定各种查询和命令行工具参数。研究人员利用 ShellGPT 和 GPT-3.5 从自然语言自动生成精确且符合语境的命令行参数，从数据库和日志中收集信息并提取可用的情报。这种方法可以协助被动和主动侦察，以收集诸如 IP 地址范围、域名和 WHOIS 记录等重要信息。

自动化数据收集：AI 智能体系统，如 AutoGPT，已经被用来通过审查社交媒体或网页以自主识别潜在目标，从而为全面的外部攻击性安全评估奠定基础。

根据被赋予的权限，AI 可以协助安全测试人员，也可以在侦察阶段规划和编排更多的工作。

扫描

扫描阶段的目标是主动探测系统和网络，为目标及其网络结构绘制详细的地图（包括潜在的漏洞），具体包括使用各种工具检测开放端口、正在运行的服务、

© 2025云安全联盟大中华区版权所有

使用的技术以及潜在的弱点。所使用的工具和技术包括端口扫描器、漏洞扫描器和模糊测试工具。

这一阶段的主要挑战在于处理扫描过程中产生的大量数据及其复杂性。安全测试人员必须系统地分析这些数据，以识别关键细节，这极为耗时且容易出现人为错误。通过自动化扫描过程，以及相较手动审查更快地分析出结果，AI 能够缓解这些挑战。AI 可以识别模式、关联数据并执行智能持续监控，使安全测试人员能够专注于更高级别的分析和策略制定。

扫描配置：AI 可以分析系统配置并为漏洞扫描器推荐最佳设置，确保全面覆盖并避免非必要开销。Pentest GPT 和另一篇论文的作者利用 LLM，通过为进攻性安全工具提供情境相关的命令行参数，指导用户执行扫描工具。

扫描输出的评估：LLM 支持对工具的输出进行解释，并为后续操作提出建议。这种 AI 驱动的方法提高了对目标系统漏洞收集的深度和准确性。通过大规模处理和分析扫描工具的输出数据，AI 模型可以提供更精确的洞察，引导测试人员将精力集中在最有可能成功的地方。

流量数据分析：目标在（网络）交互过程中生成的大量的（流量）数据，对人类来说可能难以全面分析。LLM 已被用于检查基于流量的数据，以识别漏洞。

尽管 AI 可以提高扫描阶段的效率，但验证其输出是首要的。这样才能确保结果的准确性，并帮助安全专业人员在后续阶段做出更明智的决策。

© 2025云安全联盟大中华区版权所有

漏洞分析

漏洞分析阶段的目标是在初始扫描的基础上，对目标系统和服务进行深入分析，以发现安全缺陷。尽管扫描提供了目标当前状态的概览，但漏洞分析重点关注与这些发现相关的潜在技术安全风险，评估每个潜在漏洞的严重性和影响。

漏洞分析的挑战在于，出于深度分析的需要，它依然是一项劳动密集型和资源密集型的工作。AI 可以通过自动化任务、识别零日漏洞以及根据现实世界威胁对实时风险进行优先级排序，帮助安全团队高效地解决最关键的问题，从而缓解这些挑战。通过提供深入手动分析的细节指导，AI 还可以帮助平衡自动化和手动测试，降低漏报和误报的风险。

减少误报：AI 可以在漏洞扫描数据上进行训练，以识别与误报相关的模式和特征。这有助于减少调查不存在的漏洞所浪费的时间和资源。

情境感知分析：AI 可以分析漏洞扫描结果以及系统配置、网络拓扑、威胁情报和业务需求，以识别传统扫描器可能忽略的潜在漏洞。这种情境感知方法提供了更全面的安全态势视图。

解释工具输出：先进的 AI 模型在解释各种测试工具的输出方面表现出色，有效地指导漏洞分析过程。这种先进的能力利用 LLM 优化安全测试策略，显著提高了漏洞识别的准确性。

源代码和二进制分析：AI 的能力包括自动化源代码分析，可以在任何能够访问源代码的环境中（无论是开源还是闭源环境）检测漏洞，进行白盒安全测试。近

© 2025云安全联盟大中华区版权所有

期的研究表明，AI 在扫描代码片段以识别安全缺陷方面具有显著的效率，在某些情况下甚至超过了最先进的静态应用安全测试（SAST）工具。利用 AI 可以加速漏洞检测过程，并最大限度地减少人为错误的可能，确保即使是难以察觉的或复杂的漏洞也不会被遗漏。

总结：扫描后，根据高优先级漏洞和关键发现总结结果至关重要。AI 模型可以将扫描数据与侦察期间收集的信息相关联，为漏洞优先级排序提供更丰富的情境信息。基于这样全面的分析，AI 能够为进一步调查或补救策略的提供建议。

优先级排序：在按风险大小或易利用性优先处理漏洞时，AI 在演绎推理和结果解释方面的能力显示出重要的价值。与传统或半自动化方法相比，AI 能够迅速处理大量文本数据（如测试工具日志），从而显著加快评估过程。基准测试表明，LLM 能够准确评估漏洞并根据可利用性对其进行优先级排序。

在分析漏洞后，安全测试人员就像攻击者一样，进入到利用阶段。

利用

利用阶段的目标是测试在真实场景受控条件下，如何有效利用已识别的安全漏洞。通过模拟攻击，安全测试人员可以评估安全控制的有效性，并在范围允许的情况下衡量攻击者在组织防御中渗透的潜在深度。这通常涉及跨不同系统链接的多个漏洞，以突破攻击路径上的多层屏障。攻击计划及其评估时机决定了工具和技术的选择，其中可能包括漏洞利用工具、自定义脚本和社会工程方法。

© 2025云安全联盟大中华区版权所有

挑战在于确定最有效的攻击路径。AI 可以通过在基于上下文的规划、排列漏洞、建议可能合适的组合以及确定最佳攻击路径方面，来提供明显的优势来缓解这些挑战，从而使利用过程更加高效，并且可能更有效。AI 还可以帮助测试人员适应动态环境，通过快速分析和响应不断变化的情况、实时调整攻击策略，甚至自主利用零日漏洞。

利用规划：AI 系统可以自动分析广泛的安全存储库和利用数据库，以针对特定系统和配置生成定制的测试用例。这减少了手动创建测试用例所需的工作量，并确保了全面的覆盖范围。例如，在高级任务规划期间，GPT- 3.5 建议了现实可行的攻击向量，如密码喷洒和 Kerberoasting 攻击。

网络流量分析与利用：基于 LLM 的黑客机器人检查了网络流量，识别了潜在的漏洞，并提出了对 HTTP 请求进行修改以供利用的建议。

恶意软件开发：AI 可以帮助开发新型恶意软件，通过混淆恶意软件负载、生成多态代码以及识别需要避免的潜在入侵指标（IOCs），来规避传统的检测方法。

概念验证和漏洞利用开发：可以利用 LLM 支持的技术，根据检测到的漏洞生成漏洞利用程序和概念验证（PoC）脚本。这些概念验证展示并解释了漏洞的可利用性，验证了漏洞的存在和影响。

模糊测试：在模糊测试中，可以利用 AI 生成各种输入，从而有可能发现传统方法遗漏的漏洞。虽然 AI 生成的输入可以通过其适应性和来自各种来源的信息综合提供独特优势，但传统的基于规则的方法可以提供更一致和可预测的结果。结合这两种方法可以为

© 2025云安全联盟大中华区版权所有

模糊测试提供更全面的方法，其中 AI 生成更广泛的输入，而基于规则的方法可确保全面覆盖已知的边缘情况。此外，LLM 可以通过生成或选择上下文相关的单词列表或模式等方式为这种混合方法做出贡献。

社会工程模拟：AI 可以制作逼真的网络钓鱼电子邮件和社交媒体消息，并冒充个人，测试员工意识和组织对社会工程攻击的抵御能力。AI 的这一应用有助于识别安全实践中潜在的人为因素漏洞。

增强创造力和创新能力：安全测试人员可能会忽略特定的攻击路径。AI 可以帮助他们探索非常规策略并找到利用漏洞的创造性方法，从而突破安全测试的界限。

交互式利用：PentestGPT 等 AI 驱动的工具通过为各种安全工具生成直观的命令（针对特定场景）并解释其输出，指导漏洞利用任务的执行。PentestGPT 在 HackTheBox 上简单到中等难度的挑战中已被证明是有效的，据报道，在超过 670,000 名成员的社区中排名前 1%。另一项研究中显示利用基于 LLM 的系统通过 SSH 在易受攻击的虚拟机上执行和优化攻击命令。例如，该系统可以通过利用 sudoers 文件中的错误配置来提升权限，展示了 AI 在现实世界渗透测试中的实际应用。ExploitFlow 利用博弈论和 AI 生成和表示利用过程，将其呈现为动态攻击树，并捕获每个过程步骤中的系统状态。

自动利用：最高级别的自动化是使用自主利用目标的 AI 智能体。基于 LLM 的系统已经证明能够自主利用已发现的漏洞，包括 SQL 注入和 XSS 攻击。此外，据报道，AI 智能体在获得相关 CVE 信息的情况下，能够自主利用一日漏洞，而无需任何进攻性安全的专门训练或微调，在所进行的测试中成功率高达 87%。该

© 2025云安全联盟大中华区版权所有

论文的后续研究还表明它们可以自主利用零日漏洞。此外，Wintermute 工具已被证明可以在 Linux 环境中自主识别并执行复杂的权限提升策略。

虽然 LLM 驱动的 AI 智能体可以极大地造福安全测试人员，但这一阶段通常是 AI 和人类交互的混合，以监督自主智能体并确保它们不会超出其范围并损害组织安全。

报告

报告阶段的目标是编写一份详尽介绍整个参与过程的综合报告。活动包括总结发现的漏洞、尝试和成功的利用、潜在影响以及建议的补救措施。工具和技术包括文档工具，用于创建提供可操作建议的详细报告。

挑战在于生成高质量的文档，这些文档在多个安全测试人员之间保持一致，并针对特定目标受众进行量身定制，以确保他们能够理解并采取行动。AI 可以通过自动化报告生成过程来缓解这些挑战，确保为各个受众量身定制全面、一致和准确的文档，同时为未来的安全改进提供可行的见解。此外，AI 还可以通过将技术发现转化为可行的建议并提供与不同受众产生共鸣的简明报告来帮助有效沟通，从而弥合安全测试人员和利益相关者之间的知识差距。

自动报告：AI 通过总结调查结果、确定风险优先级和推荐补救措施，生成有关进攻性安全活动的综合报告。AI 支持的校对功能可以进一步实现 QA 流程的自动化。这为安全团队节省了宝贵的时间，并确保与利益相关者的清晰沟通。

© 2025云安全联盟大中华区版权所有

可视化：当与可视化工具集成时，AI 可以使用先进的图形数据表示来增强报告，使利益相关者更容易获取调查结果并采取行动。例如，AI 可以分析漏洞以生成威胁态势图、描绘攻击面的交互式可视化图表、漏洞之间的关系以及关键路径。这些可视化工具提高了报告的清晰度，并帮助利益相关者有效地了解和确定风险的优先级。

数据驱动的洞察：AI 分析结果以识别趋势和模式。这些数据可以完善未来的进攻性安全工作，并优先考虑安全投资，以实现最大影响。

生成补救指令：AI 在关联大量知识来生成补救指令方面比人类专家具有优势。

现在我们知道了 AI 如何辅助进攻性安全，让我们来看看威胁行为者是如何使用它的。

威胁行为者对AI的使用

虽然我们已经从安全测试人员或研究人员的角度探讨了 AI 在进攻性安全中的应用，以及他们可能使用技术克服的一些挑战，但如果我们不从威胁行为者的角度以及他们如何利用该技术来看待它，那就太不称职了。这给了我们另一个理由来考虑将 AI 用于进攻性安全。通常，这些恶意行为者使用的策略与进攻性安全测试人员用来识别漏洞的方法重叠。

威胁行为者正在积极利用 AI 来增强其行动，正如微软和 OpenAI 的一项联合研究所强调的那样。

AI 辅助侦察：威胁行为者使用 AI 自动收集和分析技术和漏洞数据，大大增强了

© 2025云安全联盟大中华区版权所有

他们的侦察能力。AI 使他们能够快速处理大量信息，准确识别潜在目标。

AI 驱动的社会工程：利用 AI，威胁行为者可以生成特定情境的、令人信服的网络钓鱼内容。通过分析个人的公开信息（例如他们的专业背景或兴趣），AI 可以制作个性化的网络钓鱼电子邮件或消息，这些邮件或消息更有可能欺骗收件人。

恶意代码编写：利用 AI 辅助开发和完善恶意脚本和恶意软件，降低复杂网络攻击的技术门槛。

漏洞研究：威胁行为者利用 AI 来理解和识别软件和系统中公开报告的漏洞。AI 可以分析安全报告和补丁说明，并利用数据库来查找可利用的弱点。

绕过安全功能：AI 被用来攻克双因子身份验证或 CAPTCHA 等安全机制。这增强了自动发起垃圾邮件攻击和创建大规模欺诈账户和在线资料的能力。

异常检测规避：另一种策略是利用 AI 开发方法，帮助恶意活动融入正常行为或流量。通过模仿合法模式，AI 有助于逃避检测系统，使安全团队更难识别和缓解威胁。

改进控制操作：AI 可细化指挥和控制操作，使入侵后的活动更加复杂，更难发现。威胁行为者利用 AI 来优化命令序列，改进对受感染系统的远程控制，并更有效地管理数据提取过程。

通过了解和应对威胁行为者如何使用 AI，安全专家可以更好地保护他们的组织并在持续打击网络威胁的战斗中保持领先一步。

© 2025云安全联盟大中华区版权所有

AI驱动进攻性安全 – 不久的将来

正如前述案例所示，AI 能够通过提升**扩展性**、**效率**、**速度**，以及发现更多**复杂漏洞**，显著增强进攻性安全的能力。虽然这些 AI 应用的自主性程度各不相同，但最新研究表明，AI 正在向更高自主性稳步迈进。随着 AI 技术的持续发展，我们可以预见到更高级别的自主化与自动化将成为现实，进一步强化进攻性安全的各种功能。

降低入门门槛

AI 的引入和增强正逐步降低安全攻击的门槛。这种技术的普及化使得更多个人和企业能够参与到安全测试中，而无需深厚的漏洞挖掘或技术背景。例如，OpenAI 的 GPT-4o 等工具能够自动生成攻击脚本，使经验不足的安全专业人员也能进行复杂的安全测试。这种民主化趋势让更多组织能够进行更加稳健的安全攻击实践。AI 不仅能够自动化收集信息，还能执行标准化的利用程序，简化复杂的攻击操作。实际上，AI 的应用范围超越了进攻性安全领域，越来越多的研究表明，如今受益于 AI 的往往是新入门者，这进一步说明 AI 赋能了更广泛的用户，推动他们参与到安全测试工作中。

对专业安全测试人员的深远影响

AI 的应用将显著改变安全测试人员的工作方式。通过自动化繁琐任务，如数据收集、漏洞扫描和初步利用尝试，AI 能够让安全测试人员将更多精力投入到战略性、创新性以及复杂性更高的任务中。这一转变使得他们能够进行更加深入的

© 2025云安全联盟大中华区版权所有

分析，并开发出应对复杂网络威胁的新方法。AI 的应用提升了工作效率，使安全测试人员能够发现并解决那些难以通过手动方式检测到的复杂漏洞。然而，为了充分利用这些 AI 工具和技术，进攻性安全团队需要掌握如 AI 模型训练、数据准备以及算法优化等新技能。

进攻性安全左移

随着自动化程度的提升和反馈周期的缩短，进攻性安全活动能够更早地融入 DevSecOps 流程。这种“左移”方法意味着在软件开发生命周期的早期就考虑到安全因素，从而对企业整体的安全态势产生更深远的积极影响。通过尽早识别并缓解漏洞，组织可以大大降低遭受安全侵害的风险，实现更强大的保护能力。

AI 安全解决方案的成熟

目前，市场上尚无一款 AI 安全解决方案能够自主调整进攻性安全策略。然而，不久的将来，我们可能会看到更多商业化的安全解决方案被大规模启用，进一步增强对安全测试人员的支持。这些系统将可能与外部系统和信息源无缝集成，如威胁情报、社交媒体以及暗网资源等。通过这种整合，AI 驱动的进攻性安全实践能够利用新出现的漏洞、利用手段及威胁行为者活动等实时数据。通过学习这些外部情报源，AI 系统将进一步提升模拟攻击的效果，并紧跟最新的对抗策略。

提升 AI 智能体的自主能力

随着 AI 系统的不断成熟，其在进攻性安全领域中的自主能力将显著提升。这些自主智能体将能够执行更加复杂的攻击操作，几乎无需人工干预，实时做出决策，

© 2025云安全联盟大中华区版权所有

并根据不断变化的安全环境调整策略。这一进步使得 AI 智能体可以完成以往需要大量人类专业知识和监管才能完成的任务，从而极大地提高了进攻性安全操作的效率和效果。

自动化与人工监控的平衡

尽管 AI 技术发展迅速，但在当前阶段，仍然需要在自动化与人工监控之间找到平衡。人工监控可以确保 AI 决策的准确性，并减少意外后果。尽管某些非进攻性安全领域的研究表明，单纯依赖 AI 有时比人机协作更有效（如 AMIE），但目前最佳的做法是均衡利用 AI 与人类专业知识的优势。这种结合能够提高安全防护的有效性，同时保持网络安全操作的完整性和可信度。

对抗性攻击的考虑

上述的 AI 在进攻性安全方面的优势同样适用于恶意攻击者。安全测试人员必须考虑到，敌手可能会利用 AI 技术进行攻击。为了确保模拟攻击的真实性和有效性，安全测试人员必须紧跟敌手所采用的最新 AI 技术。这要求他们能够快速调整和发展自己的技术和工具，至少与敌手保持同步。安全测试人员必须与 AI 合作，不断将新的 AI 能力融入到网络防御实践中，以预见并应对复杂威胁。

展望未来，随着 AI 在网络防御中影响力的不断增长，实现自动化与人类洞察之间的平衡将变得至关重要。这种方法不仅能提高防御效果，还能确保网络安全措施的完整性和可信度。

© 2025云安全联盟大中华区版权所有

挑战与限制-风险与应对措施

在进攻性安全领域，AI 的应用带来了独特的挑战和限制。管理大数据集并确保准确检测漏洞是一个重要挑战，但可以通过技术进步和最佳实践来解决。然而，一些固有局限性，如 AI 模型中的 token 窗口限制，需要谨慎规划和缓解策略。

虽然 AI 能提升安全措施的效率与有效性，但其复杂性也强调了将 AI 谨慎整合到安全框架中的重要性。对 AI 模型进行严格训练和验证以保证其可靠性及表现至关重要。此外，必须有严格道德指南规范 AI 使用以防止滥用，并确保负责任地应用。只有克服这些挑战和限制，组织才能利用 AI 加强进攻性安全能力同时维护道德标准与运营完整度。

技术挑战和限制

当将 AI 集成到进攻性安全中时，可能会出现与该技术的能力、配置和性能相关的几个直接问题。与 AI 集成相关的一般技术风险已在 OWASP 的 LLM 应用 Top 10 等资源中详细列出，不在此处进一步阐述。

挑战/限制	描述	风险	缓解措施
token 窗口限制	AI 模型在处理单个请求时的 token 容量有限，导致分析大型文档的效	处理大规模数据的性能下降	使用具有更大上下文窗口的 AI 模型（如 Google Gemini 1.5 Pro，支持 200 万个 token），或采用数据分块技术（如 map-reduce）

© 2025云安全联盟大中华区版权所有

挑战/限制	描述	风险	缓解措施
	率低下		
安全限制和内容过滤	公共 AI 模型内置的限制可能阻止其有效响应被视为有害或不适当的特定提示	在动态安全场景中的适用性受限	使用具有灵活内容过滤功能的模型，或开发自定义模型，结合可调操作指南，以满足特定安全需求，同时避免造成意外损害或超出既定界限
缺乏领域知识	AI 模型可能缺乏完成特定安全任务所需的领域专门知识	在复杂安全场景中表现不佳	采用检索增强生成（RAG）方法，整合相关领域知识，或使用领域专用数据集对 AI 模型进行补充训练
幻觉	AI 模型可能生成表面上合理但实际上不准确的信息	误导性信息可能导致错误的安全决策	强化人工监督，并通过与已验证的数据源交叉核对，提高 AI 输出的准确性
数据泄露	敏感数据可能无意中被纳入 AI 模型的训练中	可能导致关键数据的泄露	采用自动数据清理工具和定期审查机制，识别并清除敏感数据，或使用自托管模型以降低泄露风险
误报	AI 系统可能错	误报可能导致	在 AI 系统中引入反馈回路，持续

© 2025云安全联盟大中华区版权所有

挑战/限制	描述	风险	缓解措施
	误地标记不存在的漏洞	资源浪费和测试工作的干扰	优化漏洞检测算法，降低误报率
漏报	AI 模型可能会漏检实际存在的漏洞	未检测到的安全漏洞可能引发严重的安全事件	定期更新 AI 模型，采用最新的威胁情报和实时异常检测技术，提升检测精度
失去范围控制	AI 模型可能会超出初始设定的范围，自主扩大目标列表	未经授权的测试可能引发法律、道德问题，或导致数据泄露和系统损坏	实施更严格的审批流程和监控机制，确保操作范围严格限制在授权范围内
隐身能力受损	AI 驱动的活动可能比预期更容易被检测到	目标系统如果快速检测到 AI 活动，可能削弱红队演练的有效性	调整 AI 系统的操作策略，限制工具的使用和速度，以维持低检测率
附带损害	AI 可能会生成自我传播的恶意软件，导致超出预期的破	自我传播的恶意软件可能造成广泛损害，超出预期目标	在部署前进行全面影响评估，预测并预防可能的附带损害

© 2025云安全联盟大中华区版权所有

挑战/限制	描述	风险	缓解措施
	坏		
训练数据投毒	恶意篡改训练数据集可能导致 AI 模型性能下降	性能下降及结果不准确，可能错过威胁或导致资源浪费	持续监控 AI 模型输出，识别异常模式，及时应对数据篡改迹象
可解释性和透明度	AI 模型的决策过程可能不透明，难以理解其在安全测试中的结论	AI 模型决策中的未识别偏见或错误可能导致对 AI 缺乏信任，从而影响威胁检测和资源分配	采用特征重要性评分和模型审计技术，提高 AI 决策过程的可解释性和透明度

表 3：技术性挑战与限制

非技术挑战和限制

这些包括影响 AI 部署和运行的更广泛的组织、道德或战略问题。

挑战/限制	描述	风险	缓解措施
数据隐私法规限制	法规限制了将数据发送到基	访问强大的 AI 模型受限，阻	探索本地部署的 AI、隐私保护技术或数据匿名化方法

© 2025云安全联盟大中华区版权所有

挑战/限制	描述	风险	缓解措施
	于云的 AI 服务	碍了创新和效率	
成本问题	开发、训练和运营专用 AI 模型的成本高昂	可能无法使用 AI，或者只能在非常有限的情况下使用	使用预训练模型、开源工具和基于云的 AI 服务来降低成本
伦理违规	AI 可能会超越社会工程的范围或道德界限	社会不可接受的行为或操纵行为	定期培训和更新 AI 及团队的伦理指南，以增强遵守性
过度依赖 AI	过度依赖 AI 辅助的进攻性安全，可能忽视其他必要的安全实践	忽视其他必要的安全实践	定期进行涉及 AI 和人类元素的演练和场景培训，确保做好准备并具备相应能力，同时避免过度依赖
高价值目标	定制的 AI 进攻性安全系统可能成为攻击者的高价值目标	滥用未经授权的访问造成危害	实施严格的访问控制、异常检测和故障安全机制

表 4：非技术性挑战与限制

© 2025云安全联盟大中华区版权所有

虽然 AI 在增强进攻性安全能力方面具有巨大潜力，但必须认识到它们带来的挑战、局限性和风险。实施适当的缓解策略有助于确保 AI 安全并有效地融入其安全框架。

治理、风险和合规（GRC）

利用 AI 进行攻击性安全需要采取治理、风险和合规（GRC）的综合方法。这确保了 AI 工具的使用既有效又符合伦理，遵循诸如美国国家标准与技术研究院 AI 风险管理框架（NIST AI RMF）、AI 组织责任-核心安全责任以及 OWASP LLM AI 网络安全与治理检查表等框架。这些框架提供了宝贵的评估标准，包括安全性、稳健性、可解释性、隐私增强、公平性和透明度。

可信赖的 AI 实施

- 安全、稳固、有韧性：AI 模型必须在其整个生命周期中，从设计到部署和决策，都将安全性放在首位。像 Microsoft Azure Machine Learning 这样的工具通过在每个阶段集成强大的安全和韧性措施来体现这一点。
- 可解释且可理解：有效的进攻性安全需要理解 AI 的输出，以确保对发现的问题进行有效补救。像 IBM 的 Watson for Cyber Security 这样的工具提供可解释的人工智能（XAI）输出，详细说明检测到威胁背后的推理过程，从而帮助安全分析师验证和响应 AI 的发现。
- 增强隐私保护：AI 模型必须通过确保身份、匿名性和保密性来保护个人隐私。例如，谷歌的 AI 系统遵守 GDPR，在分析大量数据以发现潜在威胁的同时确保数据隐私。

© 2025云安全联盟大中华区版权所有

- 公平：AI 模型必须减少偏见，以确保在多样化环境中的有效性。像谷歌的 AutoAI 这样的主动安全工具采用偏见检测机制，以确保安全威胁评估的公正和无偏见。
- 透明：透明的 AI 系统使安全团队能够理解并信任 AI 的决策。像 Palo Alto Networks 使用的透明模型提供了对 AI 在威胁检测和响应中决策过程的清晰洞察。

第三方风险管理

在使用第三方 AI 模型开展进攻性安全测试时，供应商评估至关重要。这包括评估数据安全实践、模型训练数据以及潜在的信息泄露风险。

- 遵守标准：评估供应商对 ISO/IEC 42001:2023、ISO/IEC 27001 和 GDPR 等安全和隐私标准的遵守情况。IBM QRadar 等工具可帮助遵守各种标准，同时提供基于 AI 的进攻性安全威胁检测。
- 安全和隐私认证：要求托管 AI 解决方案提供认证，以确保其符合监管要求。例如，亚马逊网络服务（AWS）的 AI 服务通常提供必要的认证和合规保证。
- 合同和保险覆盖：通过合同协议和保险覆盖来解决与进攻性安全 AI 用例相关的特定风险。这确保了通过法律和财务保护来减轻任何潜在问题。

GRC 考量因素

- 法律框架：了解进攻性安全领域中 AI 的法律和监管要求至关重要。例如，AI 模型必须遵守地区法律，如欧洲的《通用数据保护条例》（GDPR）、《欧盟人工智能法案》或加利福尼亚州的《加州隐私权法》（修订后的 CCPA）。

© 2025云安全联盟大中华区版权所有

- 伦理准则：AI 治理必须包含伦理考量。利用联合国教科文组织《关于人工智能伦理问题的建议书》和 IEEE EAD 标准等资源可以提供额外的指导。在进攻性安全操作中，建立使用符合伦理的 AI 的文化对于维护信任和诚信至关重要。
- 组织政策：制定明确的政策来管理进攻性安全领域中 AI 的使用。这些政策应补充现有的法律和伦理框架，并包含实际实施的具体操作指南。

© 2025云安全联盟大中华区版权所有

GRC 总结

引入 AI 辅助的进攻性安全手段需要谨慎评估，以避免无意中扩大其他风险。与 NIST AI 风险管理框架相一致的全面风险评估有助于减轻已识别的风险。这包括考虑 AI 模型的功能与隐私政策之间可能存在的冲突，以及防止敏感信息泄露给未经授权的第三方。

通过遵循既定的框架和指南，组织可以确保 AI 辅助的攻击性安全实施是安全的、符合伦理的，并且遵守监管标准，从而增强其整体安全态势。

结论

AI 正在快速发展，带来了更强的智能体能力和自动化水平。这些发展为全球的进攻性安全团队提供了新的机遇，也带来了挑战。恶意行为者已在法律和道德框架之外利用这些技术进步，凸显出防御者主动创新的迫切需求。

如本文所述，AI 技术，特别是 LLM 和 AI 智能体，通过自动化和扩展任务，显著增强了进攻性安全能力。这种改进提高了效率，使得评估更加复杂和广泛，并使安全团队能够专注于流程改进和战略工作。此外，AI 还促进了安全测试的普及，降低了入门门槛，缓解了专业技能人才短缺的问题。

尽管有这些好处，但目前还没有单一的 AI 解决方案能够彻底改变进攻性安全实践。因此，需要采取多方面的方法，不断试验 AI 以找到并实施有效的解决方案。这就要求创造一个鼓励学习和发展的环境，使团队成员能够提升使用 AI 工具和技术的技能。

© 2025云安全联盟大中华区版权所有

为了有效利用 AI，组织必须分享最佳实践，并开发针对特定安全需求的定制工具，最好通过数据科学、网络安全、法律等部门之间的跨学科合作来实现。这确保了 AI 集成和风险管理的整体方法。随着 AI 系统融入工作流程，它们引入了新的技术和组织挑战，必须谨慎管理。需要保持警惕，以防止 AI 驱动的工具被滥用或出现不可预测的行为。

推广负责任使用 AI 的文化对于确保团队了解将 AI 集成到进攻性安全中的风险和影响至关重要。这包括建立一个稳健的治理、风险和合规（GRC）框架，以有效管理这些风险。

总之，进攻性安全必须随着 AI 能力的发展而演进。通过采用 AI、对团队进行其潜力和风险的培训，以及培养持续改进的文化，组织可以显著增强其防御能力，并在网络安全领域获得竞争优势。

参考文献

1．Jason Wei et. al. (2022), Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, https://arxiv.org/abs/2201.11903

2．Yuheng Cheng et. al. (2024), Exploring Large Language Model based Intelligent Agents: Definitions, Methods, and Prospects, https://arxiv.org/abs/2401.03428

3．Zachary Proser (2023), Retrieval Augmented Generation (RAG), https://www.pinecone.io/learn/retrieval-augmented-generation

4．Ge Wang (2019), Humans in the Loop: The Design of Interactive AI Systems, https://hai.stanford.edu/news/humans-loop-design-interactive-ai-systems

© 2025云安全联盟大中华区版权所有

5．Sheng Lu et. al. (2023), Are Emergent Abilities in Large Language Models just In-Context Learning?, https://arxiv.org/abs/2309.01809

6．Chris Sullo (2001),Nikto, https://github.com/sullo/nikto

7．Gelei Deng et al. (2023), PentestGPT: An LLM-empowered Automatic Penetration Testing Tool, https://arxiv.org/pdf/2308.06782.pdf

8．Eric Hilario et. al. (2024), Generative AI for pentesting: the good, the bad, the ugly, https://link.springer.com/article/10.1007/s10207-024-00835-x

9．TheR1D (seen 2024/06), ShellGPT, https://github.com/TheR1D/shell_gpt

10．Significant-Gravitas (seen 2024/06), AutoGPT, https://github.com/Significant-Gravitas/AutoGPT

11．Andreas Happe, Jürgen Cito (2023), Getting pwn'd by AI: Penetration Testing with Large Language Models, https://arxiv.org/pdf/2308.00121.pdf

12．aress31 (seen 2024/06), BurpGPT, https://github.com/aress31/burpgpt

13．Legit Security Team (2023), Using AI to Reduce False Positives in Secrets Scanners, https://www.legitsecurity.com/blog/using-ai-to-reduce-false-positives-in-secrets-scanners

14．Jingyue Li et.al. (2023), Evaluating the Impact of ChatGPT on Exercises of a Software Security Course, https://arxiv.org/pdf/2309.10085.pdf

15．Shashwat et.al. (2024), A Preliminary Study on Using Large Language Models in Software Pentesting, https://arxiv.org/abs/2401.17459

16．Andreas Happe et. al. (2023), LLMs as Hackers: Autonomous Linux Privilege Escalation Attacks, https://arxiv.org/pdf/2310.11409

17．Joseph Thacker (seen 2024/06), Hero, https://github.com/jthack/hero

© 2025云安全联盟大中华区版权所有

18．Keith McCammon (2024), How AI will affect the malware ecosystem and what it means for defenders, https://redcanary.com/blog/opinions-insights/ai-malware

19．Dijana Vukovic Grbic, Igor Dujlovic (2023), Social Engineering with ChatGPT, https://www.researchgate.net/publication/369970449_Social_engineering_with_ChatGPT

20．Víctor Mayoral Vilches et. al. (2023), ExploitFlow, cyber security exploitation routes for Game Theory and AI research in robotics, https://arxiv.org/pdf/2308.02152.pdf

21．Richard Fang et. al. (2024), LLM Agents can Autonomously Hack Websites, https://arxiv.org/html/2402.06664v1

22．Richard Fang et. al. (2024), LLM Agents can Autonomously Exploit One-day Vulnerabilities, https://arxiv.org/abs/2404.08144

23．Richard Fang et. al. (2024), Teams of LLM Agents can Exploit Zero-Day Vulnerabilities, https://arxiv.org/abs/2406.01637

24．Microsoft Threat Intelligence (2024), Staying Ahead of Threat Actors in the Age of AI, https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/

25．Ethan Mollick - One Useful Thing (2024), Centaurs and Cyborgs: On the Jagged Edge of AI and Human Collaboration, https://www.oneusefulthing.org/p/centaurs-and-cyborgs-on-the-jagged

26．Karthikesalingam, Natarajan (2024), AMIE: A research AI system for diagnostic medical reasoning and conversations, https://research.google/blog/amie-a-research-ai-system-for-diagnostic-medical-reasoning-and-conversations/

© 2025云安全联盟大中华区版权所有

27．OWASP (2023), OWASP Top 10 for Large Language Model Applications, https://owasp.org/www-project-top-10-for-large-language-model-applications/

28．NIST (2024), AI Risk Management Framework, https://www.nist.gov/itl/ai-risk-management-framework

29．OWASP (2024), LLM AI Security and Governance Checklist, https://owasp.org/www-project-top-10-for-large-language-model-applications/llm-top-10-governance-doc/LLM_AI_Security_and_Governance_Checklist-v1.pdf

30．Microsoft Azure Team (2024), What is Responsible AI?, https://learn.microsoft.com/en-us/azure/machine-learning/concept-responsible-ai?view=azureml-api-2

31．IBM (seen 2024/06), What is Explainable AI, https://www.ibm.com/topics/explainable-ai

32．Google Cloud Team (seen 2024/06), Google Cloud GDPR Compliance, https://cloud.google.com/privacy/gdpr

33．GDPR.eu (2018), General Data Protection Regulation (GDPR), https://gdpr.eu/tag/gdpr/

34．IBM (seen 2024/06), Accelerating AI and model lifecycle management, https://www.ibm.com/products/watson-studio/autoai

35．Palo Alto Networks (seen 2024/06), SmartScore, https://www.paloaltonetworks.com/blog/tag/smartscore/

36．ISO (2022), ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system, https://www.iso.org/standard/81230.html

37．ISO (2022), ISO 27001, https://www.iso.org/standard/27001

38．IBM (seen 2024/06), Compliance with IBM Security QRadar SIEM, https://www.ibm.com/products/qradar-siem/compliance

© 2025云安全联盟大中华区版权所有

39．AWS (seen 2024/06), Security Perspective Compliance and Assurance of AI/ML Systems, https://docs.aws.amazon.com/whitepapers/latest/aws-caf-for-ai/security-perspective-compliance-and-assurance-of-aiml-systems.html

40．McKinsey (2021), Getting to Know and Manage Your Biggest AI Risks, https://www.mckinsey.com/capabilities/quantumblack/our-insights/getting-to-know-and-manage-your-biggest-ai-risks

41．Council of the EU (2024), Artificial intelligence (AI) act: Council gives final green light to the first worldwide rules on AI, https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai/

42．California Office of the Attorney General (2024), CCPA, https://oag.ca.gov/privacy/ccpa

43．UNESCO (2022), Recommendation on the Ethics of AI, https://unesdoc.unesco.org/ark:/48223/pf0000381137

44．IEEE (seen 2024/06), EAD General Principles, https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_general_principles.pdf

© 2025云安全联盟大中华区版权所有

Cloud Security Alliance Greater China Region

扫码获取更多报告

云安全联盟大中华区
官网:https://c-csa.cn 电话:0755-86548359
地址:上海市浦东新区东育路255弄5号3FB30(前滩国际经济组织集聚区)/
中国广东省深圳市南山区高新技术产业园(南区)高新南环路深圳湾科技生
态园二区9栋B座6层603单元