# 执行摘要

人工智能（AI）技术的出现，特别是大型语言模型（LLMs）和由驱动大模型的AI代理，已经在攻击性安全领域引发了深刻的变革，包括漏洞评估、渗透测试和红队演练。这一转变将人工智能从一个狭窄的用例重新定义为一个多功能和强大的通用技术。本文探讨了大型语言模型（LLMs）驱动的人工智能变革潜力，通过研究其与攻击性安全的集成，以解决当前挑战，并展示了AI在五个安全阶段的能力：侦察、扫描、漏洞分析、漏洞利用以及报告。

### 关键发现

**攻击性安全的挑战：**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能力：**人工智能，主要通过大型语言模型（LLMs）和AI代理，提供了在攻击性安全方面的显著能力，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益处：**在攻击性安全中利用AI增强了可扩展性、效率、速度，发现更复杂的漏洞，并最终提高了整体安全态势。

**暂无良方：**尽管前景不错，但目前没有单一的AI解决方案能够彻底改变攻击性安全。需要对AI进行持续实验，以找到有效部署实施的解决方案。这就需要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。采用攻击性安全的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：**由大模型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）、网络安全策略师和高管，能够通过有效阐述攻击性安全投资的价值，以确保增强攻击性AI能力所需的资源。

# 引言

### 攻击性安全

攻击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，组织实施和加强可靠的安全控制，从而降低攻击者利用的风险。

为了最大限度地提高攻击性安全的有效性，确保其与组织的长期目标和宗旨一致非常关键。这种方法侧重于对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。

本文档的其余部分探讨了攻击性安全的三种方法：

* **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。

这些方法在不同方面有相似之处，但也有差异，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 | 中等 | 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：攻击性安全测试实践**

尽管攻击性安全参与的技术、范围和深度各不相同，但入侵通常遵循五个阶段：侦察、扫描、漏洞分析、利用和报告。在开始任何参与之前，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，我们将讨论这些五个阶段，同时假设范围和约定规则已经定义。



图表 1 攻击性安全测试阶段

**侦察**——侦察代表了任何进攻性安全策略的初始阶段，旨在收集有关目标系统、网络和组织结构的大量数据。

**扫描**——扫描需要全面地检查已识别的系统，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。

**漏洞利用**——漏洞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过程。

### 攻击性安全的当前挑战

攻击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。这些挑战因缺乏技能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，如人工智能（AI）、区块链、云计算、物联网（IoT）以及远程劳动力的增加，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零日漏洞的本地攻击，这些可能难以检测和应对。
* **多样性评估：**攻击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，这使测试人员难以在所有领域保持专业知识，例如，测试移动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：**攻击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误报分散注意力。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。
* **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试图迭代利用。
* **工具开发与定制：**攻击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干系人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，并提供与不同干系人产生共鸣的简洁报告可能是一个重大挑战。当安全测试人员和他们的干系人之间存在大量知识差距时，尤其如此，这使得很难将技术复杂性与业务优先级和风险承受能力联系起来。
* **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：**攻击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。这对测试人员来说即费资源也耗时。

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。人工智能，特别是大语言模型（LLMs），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，显著增强攻击性安全测试人员的能力，并普遍提高攻击性安全实践的有效性。

# 人工智能

人工智能（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大型语言模型（LLMs）驱动的技术。

# 大型语言模型

大语言模型是由数十亿参数组成的复杂深度神经网络，能够处理和生成自然语言。最初，这些模型被开发用于预测句子中的下一个单词，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。

大语言模型（LLMs）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，大语言模型（LLMs）分析大量文本数据，识别语言模式，并学习预测接下来的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。

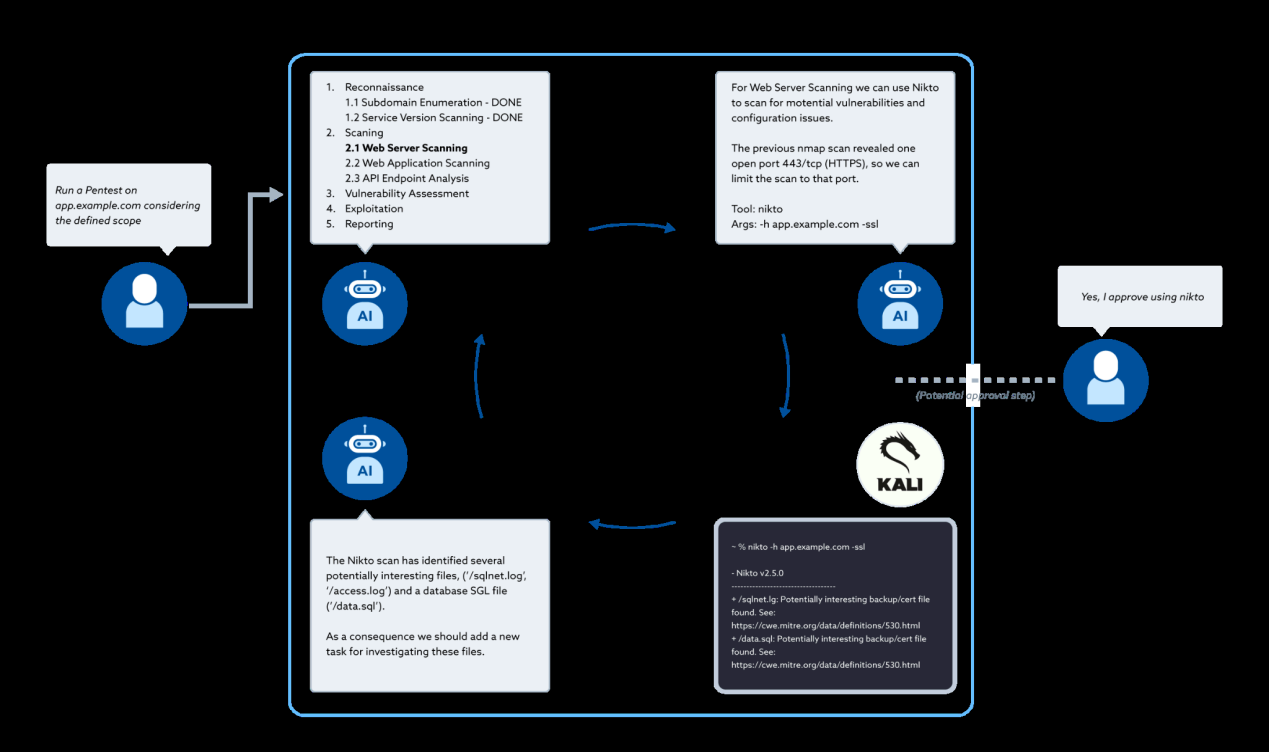
大语言模型（LLMs）不仅限于语言处理。它们擅长快速分析大量数据，包括文本、代码、日志和HTTP流量。利用其生成能力，它们可以创建代码、脚本和电子邮件，以及编制摘要和报告。最先进的大语言模型（LLMs）展示了如文本**推理**和制定程序性决策的应急能力，这对于**规划**和**目标导向任务**至关重要。为了与许多关于AI和AI代理的论文中使用的术语保持一致，我们使用“推理”一词来描述分析文本和制定程序性决策的能力。然而我们承认，关于AI代理是否能像人类一样进行推理，有待持续研究。

尽管大语言模型（LLMs）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，大语言模型（LLMs）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且与传统计算系统不同。

大语言模型（LLMs）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。虽然基于聊天的大语言模型（LLMs）广泛用于内容创作和创意目的，而如何利用这些模型来自主解决复杂问题正是当前正在进行的研究探索，这也是将AI向更广泛利用迈出的重要一步。

# AI代理

AI代理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，从而塑造与未来的环境互动。这些代理可以使用大语言模型（LLMs）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。与基本大语言模型（LLMs）应用不同，使用大语言模型（LLMs）的AI代理遵循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，使代理能够有效地通过多步骤过程解决复杂问题，直到任务完成。



图表 2 AI代理的阶段

正如图2所示，是一个简化的AI代理处理针对运行在app.example.com的Web应用程序进行完整渗透测试的示例，我们研究AI代理的一个周期，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，代理优先考虑Web服务器扫描任务。
2. **推理**：代理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。
4. **分析**：代理处理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进行进一步调查的任务。

这样一个简化的AI代理模型可能由于上下文的潜在广度而面临局限性，导致代理失去对目标的跟踪。为了缓解这一点，发展了多代理系统的概念，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代理组成，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大语言模型（LLMs）输出进行上下文知识补充。这种方法提高了大语言模型（LLMs）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代理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大语言模型（LLMs）生成特定于组织或域的输出的能力至关重要。

如图2所示，设想一个简单的案例：一个AI人工智能体正在对一个app.example.com上的Web应用程序进行完整的渗透测试，在这个例子中，我们将研究测试AI智能体工作中的一环——Web服务器扫描阶段。这一阶段过程可以细化如下：

1.规划：基于现有计划以及前一个环节创建的信息，AI智能体将Web服务器扫描作为优先任务。

2.推理：AI 智能体选择 Nikto 作为进行Web服务器扫描的合适工具。

3.执行：针对目标使用所选的工具。根据配置不同，AI智能体可以直接立即启动工具，或等人工审核。以提供灵活且友好的用户体验。

4.分析：AI智能体处理Nikto扫描的结果，识别可能潜在的漏洞，例如暴露的配置文件和数据库。然后，它会根据这些结果调整后续周期的执行计划，可能会建议添加新的任务来深入探究某些特定问题，这些建议将在下一规划周期中被采纳。

当语境范围过大时，这种简化的AI智能体模型可能因涉及的上下文范围太广而受限，从而使得代理无法持续跟踪其目标。为了缓解这个问题，多智能体系统（[Multi-Agent Systems](https://arxiv.org/abs/2401.03428)）的概念应运而生，并在解决复杂问题时展示出良好的效果。这类系统由一组AI智能体构成，它们会协同工作以完成任务。此外，复杂的短期和长期[记忆](https://arxiv.org/abs/2401.03428)管理是必不可少的，这通常需要引入[检索增强生成（Retrieval Augmented Generation，RAG）](https://www.pinecone.io/learn/retrieval-augmented-generation)系统访问外部权威知识库

RAG 基于外部权威知识库对LLM的输出内容进行语境化处理，而不依赖于其自身的训练数据集。这种方法提高了LLM生成响应的相关性和准确性，使其回答更加贴合实际语境。AI 智能体借助模型内部知识与外部数据的结合，能够提供更为周全的解决方案。这种集成整合对于保持语境至关重要。它还能扩展LLM的能力，使其能够为特定组织或领域生成输出。

# AI驱动的攻防安全

AI人工智能技术正在开辟攻击性安全领域的新天地，借助AI工具，我们可以模拟高级别的网络攻击，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。

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的增强和自主性

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攻防安全任务，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“[人在环路](https://hai.stanford.edu/news/humans-loop-design-interactive-ai-systems)”——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责任制，特别是在那些生产环境中进行操作的敏感领域。

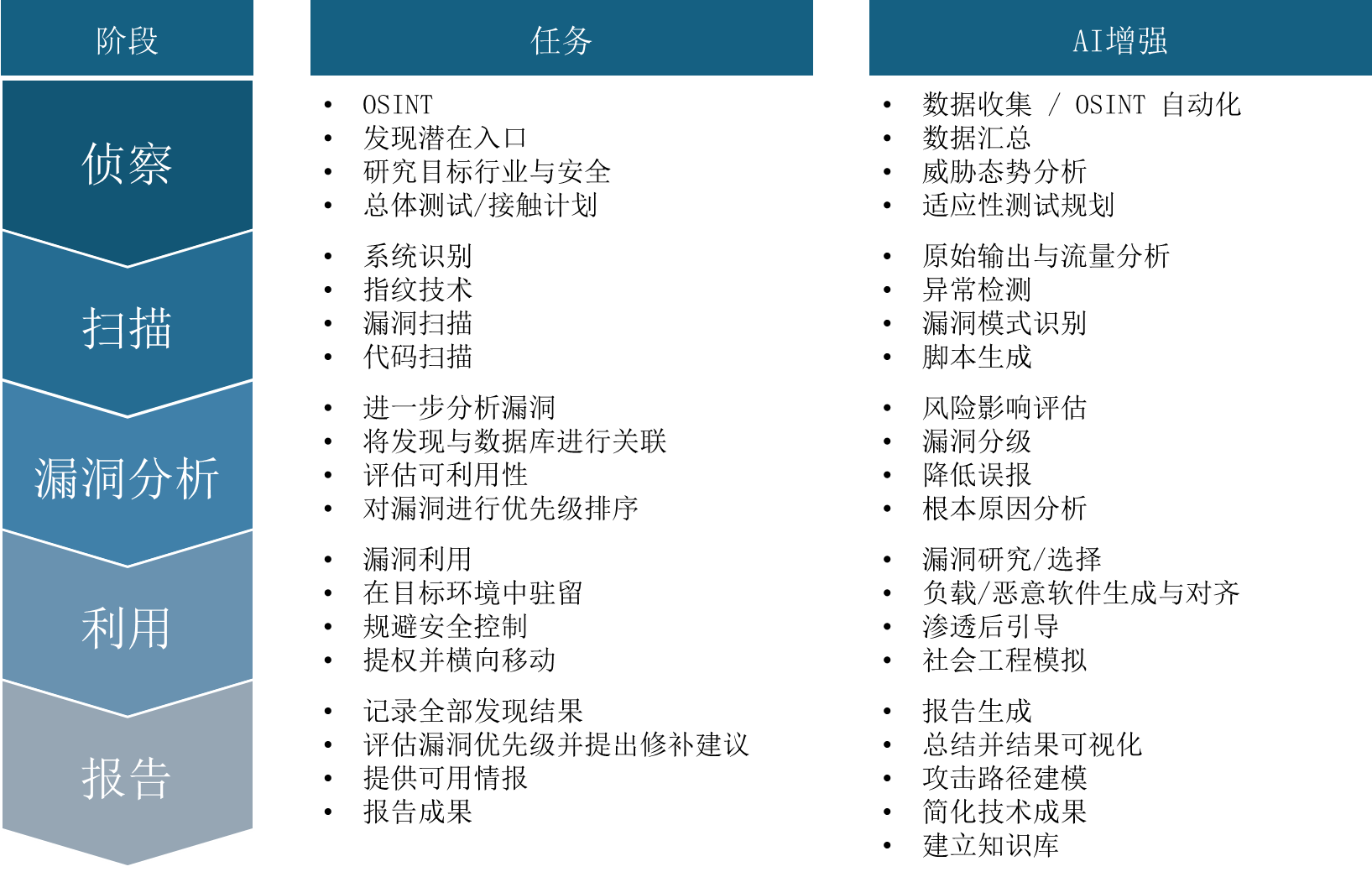
AI可以通过如下方式增强或自动化现有的攻防安全测试流程：

表2：攻防安全阶段能够被AI强化的人类任务

如上表所示，安全测试人员可以在攻防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。

图形用户界面

中度可信度描述已自动生成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。

图3：攻防安全过程中的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。这一进步确保了在攻防安全中充分发挥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，攻防安全团队都能支持具有适应性和韧性的安全策略。

在接下来的章节中，我们将探讨AI能力（特别是LLMs）在攻防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，攻防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](https://arxiv.org/pdf/2308.06782.pdf)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LLMs分析来自各种网络服务（如SMTP、FTP、HTTP）或工具的[响应](https://link.springer.com/article/10.1007/s10207-024-00835-x)，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密码、端口和服务的类型。LLMs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](https://link.springer.com/article/10.1007/s10207-024-00835-x)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](https://github.com/TheR1D/shell_gpt)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](https://github.com/Significant-Gravitas/AutoGPT)，[已经被用来](https://arxiv.org/pdf/2308.00121.pdf)通过审查社交媒体或网页自主地识别潜在目标，从而为全面的外部攻防安全评估奠定基础。

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、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](https://arxiv.org/pdf/2308.06782.pdf)和[本文](https://link.springer.com/article/10.1007/s10207-024-00835-x)作者利用LLMs，通过为攻防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：大型语言模型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。LLMs已被用于[检查基于流量的数据](https://github.com/aress31/burpgpt)，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。

# 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](https://arxiv.org/abs/2406.01637)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](https://www.legitsecurity.com/blog/using-ai-to-reduce-false-positives-in-secrets-scanners)相关的模式和特征。这有助于减少调查不存在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](https://arxiv.org/pdf/2308.06782.pdf)出色，有效地指导漏洞分析过程。这种先进的能力利用LLMs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（无论是开源还是封源环境）检测漏洞，进行白盒安全测试。[近期的研究](https://arxiv.org/pdf/2309.10085.pdf)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](https://arxiv.org/abs/2401.17459)甚至超过了最先进的SAST1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](https://link.springer.com/article/10.1007/s10207-024-00835-x)显示出重要的价值。与传统或半自动化方法相比，AI能够迅速消化大量文本数据（如测试工具日志），从而显著加快评估过程。[基准测试](https://arxiv.org/pdf/2310.11409.pdf)表明，LLMs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。

挑战在于确定最有效的攻击路径。人工智能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。人工智能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**[https://arxiv.org/abs/2406.01637](AI for Offensive Security/2308.06782v2.pdf)。

**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，GPT‑3.5 <https://arxiv.org/pdf/2308.00121.pdf> 建议在高级任务规划期间，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，https://github.com/aress31/burpgpt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https://github.com/jthack/hero。

**恶意软件开发**：人工智能可以帮助开发新型恶意软件，通过软件恶意混淆https://redcanary.com/blog/opinions-insights/ai-malware负载来逃避传统的检测方法。生成多态代码，并识别需要避免的潜在危害指标  (IOC)。

**概念验证和漏洞开发**：可以利用  LLM  支持的技术，根据检测到的漏洞生成漏洞和概念验证脚本。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**：人工智能可以制作https://www.researchgate.net/publication/369970449\_Social\_engineering\_with\_ChatGPT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稳健性。人工智能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。人工智能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 [https://arxiv.org/pdf/2308.06782.pdf](https://arxiv.org/pdf/2308.00121.pdf) 等人工智能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 <https://arxiv.org/pdf/2308.00121.pdf>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错误配置来提升权限。

在  sudoers  文件中，展示了AI  在现实世界渗透测试中的实际应用。ExploitFlow [https://arxiv.org/pdf/2308.02152.pdf](https://arxiv.org/pdf/2308.00121.pdf)利用博弈论和人工智能生成开发过程并将其作为动态攻击树，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开发目标的  AI  代理实现的。基于  LLM  的系统https://arxiv.org/html/2402.06664v1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  代理在执行测试中可以利用零日漏洞https://arxiv.org/abs/2404.08144，而无需任何针对攻防安全的专门训练或微调，在提供相关  CVE  信息的情况下，成功率为  87%。该论文的后续研究还表明它们可以自主利用https://arxiv.org/abs/2406.01637。此外，  Wintermute工具[https://arxiv.org/pdf/2310.11409.pdf](https://arxiv.org/pdf/2308.00121.pdf)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。

虽然具有  LLM  功能的  AI  代理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，以监督自主代理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。人工智能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，人工智能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，生成有关攻击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。

**可视化：**当与可视化工具集成时，AI  可以https://link.springer.com/article/10.1007/s10207-024-00835-x增强报告使用先进的图形数据表示法，使利益相关者更容易获取调查结果并采取行动。例如，人工智能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、漏洞之间的关系以及关键攻击路径。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如何帮助攻防安全，让我们来看看威胁者是如何使用它的。

# 威胁者对人工智能的使用

虽然我们已经从安全测试人员或研究人员的角度探讨了人工智能在攻击性安全中的应用，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人工智能用于攻击性安全。通常，这些恶意行为者使用的策略与攻击性安全测试人员用来识别漏洞的方法重叠。

威胁者正在积极利用人工智能来增强其行动，正如微软和OPENAI的一项联合努力所强调的那样。

**人工智能辅助侦察**：威胁行为者使用人工智能自动收集和分析技术和漏洞数据，大大增强了他们的侦察能力。人工智能使他们能够快速处理大量

信息，准确识别潜在目标。

**人工智能驱动的社会工程**：利用人工智能，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），人工智能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**：利用人工智能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**：威胁行为者利用人工智能来理解和识别软件和系统中公开报告的漏洞。人工智能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**：人工智能被用来攻克双因素身份验证或  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**：另一种策略是利用人工智能开发方法，帮助恶意活动融入正常行为或流量。通过模仿合法模式，人工智能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**：人工智能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。威胁者利用人工智能来优化命令序列，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人工智能，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。

# AI将在不久的未来驱动攻击性安全

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**，显著增强攻击性安全的能力。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，进一步强化攻击性安全的各种功能。

### 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不仅能够自动化的收集信息，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的应用范围超越了安全攻击领域，越来越多的研究表明，如今[受益于AI的往往是新入门者](https://www.oneusefulthing.org/p/centaurs-and-cyborgs-on-the-jagged)，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，攻击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。

### 左移式的攻击防御

随着自动化程度的提升和反馈周期的缩短，攻击防御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攻击性安全策略。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，基于AI驱动的攻击行为能够利用新兴漏洞、利用手段及攻击者活动等实时数据。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代理的自主能力

随着AI系统的不断成熟，其在攻击性安全领域中的自主能力将显著提升。这些自主代理将能够执行更加复杂的攻击操作，几乎无需人工干预，实时作出决策，并根据不断变化的安全环境调整策略。这一进步使得AI代理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，从而极大地提高了攻击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。尽管某些非攻击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](https://research.google/blog/amie-a-research-ai-system-for-diagnostic-medical-reasoning-and-conversations/)），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

AI的防御优势同样适用于攻击者。安全测试人员必须考虑到，攻击者可能会利用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攻击者所采用的最新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，至少与对手保持同步。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。

## 挑战与限制-风险与应对措施

在攻击性安全领域，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，一些固有限制如AI模型中的令牌窗口约束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加强攻击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集成到攻击性安全中时，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的大型语言模型应用Top 10](https://owasp.org/www-project-top-10-for-large-language-model-applications/)等资源中详细列出，并不会在此处进一步阐述。

| 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率低下。 | 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）。 |
| 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 | 使用具有可调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。 |
| 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系统可能错误地标记不存在的漏洞。 | 误报可能导致资源浪费和测试工作的干扰 | 在AI系统中引入反馈回路，持续优化漏洞检测算法，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，可能削弱红队演习的有效性 | 调整AI系统的操作策略，降低使用速度和工具显著性，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，导致超出预期的破坏。 | 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。 |
| 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。 |

表格 3：技术性挑战与限制

### 非技术挑战和限制

这些包括影响人工智能部署和运行的更广泛的组织、道德或战略问题。

以下是翻译后的表格：

| 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于云的AI服务。 | 访问强大的AI模型受限，阻碍了创新和效率。 | 探索本地部署的AI、隐私保护技术或数据匿名化方法。 |
| 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辅助的攻防安全，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和人类元素的演习和场景训练，确保准备性和能力，同时避免过度依赖。 |
| 高价值目标 | 定制的AI攻防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权限造成危害。 | 实施严格的访问控制、异常检测和故障安全机制。 |

表格4：非技术性挑战与限制

虽然人工智能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人工智能安全有效地融入其安全框架。