大型语言模型中的安全漏洞和攻击

Aysan Esmradi[1], Daniel Wankit Yip[2] 和Chun Fai Chan [3]
物流和供应链多技术研发中心（LSCM），香港
[1]aesmradi@lscm.hk, [2]dyip@lscm.hk, [3]cfchan@lscm.hk

摘要。尽管大型语言模型（LLMs）广受欢迎，但确保其安全性仍然是一个持续的挑战。 开发人员致力于增强LLMs的安全性，但漏洞仍然存在，即使在像GPT-4这样的高级版本中也是如此。 攻击者利用这些弱点，突出了在AI模型开发中采取积极的网络安全措施的必要性。 本文探讨了两种攻击类别：对模型本身的攻击和对模型应用的攻击。 前者需要专业知识、对模型数据的访问权限和大量的实施时间，而后者对攻击者更容易，并且受到了越来越多的关注。 我们的研究回顾了100多项最新的研究工作，对每种攻击类型进行了深入分析。 我们确定了最新的攻击方法，并探讨了各种实施方法。我们对缓解技术进行了彻底的调查，评估了它们的有效性和局限性。 此外，我们总结了对抗这些攻击的未来防御措施。 我们还研究了现实世界中的技术，包括对LLMs的报道和我们实施的攻击，以巩固我们的发现。 我们的研究强调了解决安全问题的紧迫性，并旨在增强对LLM攻击的理解，为这个不断发展的领域的强大防御发展做出贡献。

关键词：大型语言模型；网络安全攻击；防御策略

1 介绍

大型语言模型（LLMs）[4]是先进的人工智能系统，旨在基于大量的训练数据理解和生成类似人类的文本。 这些模型利用深度学习技术处理和理解自然语言，使它们能够生成连贯且与上下文相关的回应，最终提高人类在各个领域的生产力和理解能力。 像任何尖端技术一样，由于其巨大的滥用潜力，LLMs已成为攻击者的主要目标。 攻击者可以利用LLMs创建复杂的钓鱼邮件[49]、虚假新闻，操纵舆论[19]并自动化恶意活动，如垃圾邮件等。因此，许多研究已经在不同阶段对LLMs的漏洞进行了研究。在训练之前，研究人员分析了诸如训练数据、预处理技术和过滤以及模型架构等因素。 在训练过程中，他们研究了训练技术、超参数和优化算法的影响。 在训练之后，研究重点关注模型的行为、偏见和对攻击的敏感性。 尽管提出了防御措施，LLMs仍然容易受到攻击的影响，原因是

不断演变的攻击者技术和多种攻击类型的结合使用。 这构成了一个挑战，因为对抗一种攻击可能对复杂的组合攻击无效[23]。 LLM本身是复杂的系统，具有数百万个参数，使得控制和管理变得困难。 此外，随着模型能力的提高，例如与其他应用程序的集成[18]或处理多模态特征[3]如图像和链接与文本一起，攻击者利用漏洞的新机会也随之产生。

1.1 现有调查的回顾

我们选择了有价值的调查，以提供对这一新兴研究领域的深入见解。 Gozalo-Brizuela等人在[5]中回顾了主要生成人工智能模型的分类法，包括文本到图像、文本到文本等。Cao等人在[6]中概述了人工智能生成内容（AIGC）的进展，重点关注单模态和多模态生成模型。 他们还讨论了这些模型安全和隐私面临的威胁。Zhou等人在[7]中回顾了最近关于预训练特征模型（PFMs）的研究，涵盖了不同数据模态的进展、挑战和机遇。 他们还讨论了关于PFMs安全和隐私的最新研究，包括对抗性攻击、后门攻击、隐私泄露和模型缺陷。Hunag等人在[8]中回顾了LLM的漏洞，并探讨了验证和验证（V&V）技术在LLM的整个生命周期中进行严格分析的集成和扩展。 在[9]中，对GPT进行了全面的回顾，包括其架构、对各种应用的影响、挑战和潜在解决方案。 该论文强调了在GPT的背景下解决数据隐私的非泄露和模型输出控制的重要性。Wang等人在[10]中对AIGC进行了调查，涵盖了其工作原理、安全和隐私威胁、最新解决方案和未来挑战。

在[12]中，作者研究了扩散模型和LLMs对人类生活的影响，回顾了最近的发展，并提出了在部署之前促进可信使用和减轻风险的步骤。刘等人在[11]中对LLMs的对齐和可信性进行了全面调查。 测量研究表明，对齐模型整体表现更好，凸显了LLM对齐中细粒度分析和持续改进的重要性。

表**1.**我们的工作与其他提出的调查的比较

参考	年份	贡献
[5]	2023	回顾主要生成人工智能模型的分类法，如文本到图像、文本到文本、图像到文本、文本到视频等。
[6]	2023	概述AIGC的历史和最新进展，重点关注单模态和多模态生成模型。
[7]	2023	全面回顾了关于PFMs的最新研究，涵盖了各种数据方法中的进展、挑战和机遇。
[8]	2023	对LLM的漏洞进行了回顾，并探讨了V&V技术在LLMs安全性和可信性分析中的整合和扩展。

[9]	2023	对GPT的影响、挑战和解决方案进行综述，重点关注数据隐私和输出控制。
[10]	2023	对AIGC进行调查，涵盖工作原理、安全威胁、解决方案、伦理影响、水印方法和未来挑战。
[11]	2023	在部署之前确保LLMs的对齐性和可信度的全面调查。
[12]	2023	评估扩散模型和LLMs的影响、最新发展，并提出可信使用和风险降低的建议。
我们的	2023	对LLMs上最重要的8种攻击进行全面覆盖，提供详细定义，回顾最新的实施和缓解方法的研究，使用设计的提示评估一些攻击的有效性，并探索实施的真实攻击。

我们的贡献。我们的工作全面介绍了LLMs上的攻击，涵盖了它们的整个生命周期。 我们研究了八种重要的攻击，提供了详细的定义，并探索了每种攻击的最新研究和实施和缓解方法。 我们评估了提出的攻击的有效性，并在某些情况下使用我们设计的工具和提示评估了影响和潜在后果。 我们还探索了真实场景，以深入了解实际影响和潜在风险。 通过对LLMs的预训练、训练和推理阶段实施的攻击进行彻底调查，我们旨在增强理解并提供针对每个层面潜在漏洞的有效策略。

本文的其余部分组织如下：第2节介绍了LLM基础知识，第3节讨论了攻击和最新的实施和防御方法，第4节提出了结论和未来的研究方向。

2 背景

在本节中，我们介绍了我们论文中讨论的与LLM相关的关键概念。

2.1 大型语言模型结构

LLM与其他机器学习模型一样，从各种来源收集数据，如网络抓取、公开可用的数据集等。然后，数据经过预处理，包括标记化、清洗和规范化。 在预训练阶段，模型学习语言的统计特性，然后在较小的任务特定数据集上进行微调。 部署后，LLM已经准备好使用。攻击者可以在任何这些步骤中攻击LLM。 例如，攻击者可以向训练数据集中注入恶意数据或干扰训练过程。 我们通过将LLM的攻击[14]分为两个主要类别：对LLM本身的攻击和对LLM应用程序的攻击。对LLM模型的攻击针对模型的输入、参数和超参数。 它们涉及提取或操纵数据，试图提取模型参数或架构，或者利用部署基础设施的漏洞。 目标是破坏完整性、安全性或

模型及其相关数据的隐私。 对LLMs应用的攻击旨在滥用模型的行为和输出。 它们涉及操纵模型生成错误信息、引入偏见、损害数据隐私以及破坏其可用性和功能。

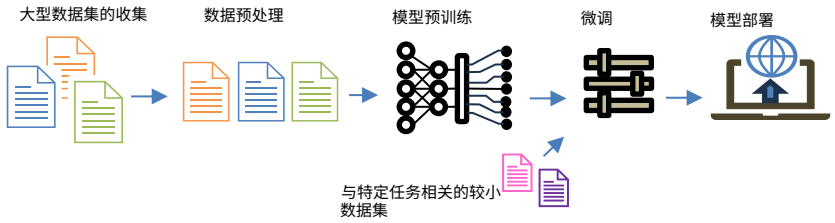

图 **1.**使用LLM平台进行训练

我们的重点主要是与注入操纵提示和接收模型通常拒绝生成的输出相关的漏洞，而不是与性能问题（如事实或推理错误）相关的漏洞[13]。 事实错误发生在当模型对大学中的教授提问时，回答关于从未与该机构有过关联的人的信息，而推理错误指的是LLMs可能并不总是提供正确答案给计算或逻辑推理问题。

2.2 信息安全

在LLMs的背景下，攻击者经常针对的信息安全基本原则包括：

- 机密性。该原则侧重于保护存储在LLMs中的敏感和私密数据免受未经授权的访问。
- 可用性。它专注于确保对LLM资源和服务的不间断访问。 攻击者可能发动拒绝服务（DoS）攻击，以破坏LLMs的可用性，使其对用户不可访问。
- 完整性。它涉及保留准确的信息并防止攻击者恶意篡改或操纵。

通过理解和解决这些原则，可以实施强大的安全措施来保护LLMs免受网络威胁。

3 大型语言模型的攻击

在本节中，我们将详细研究基于所提到分类的LLMs的网络安全攻击。 每种攻击类型的分类如图2所示。

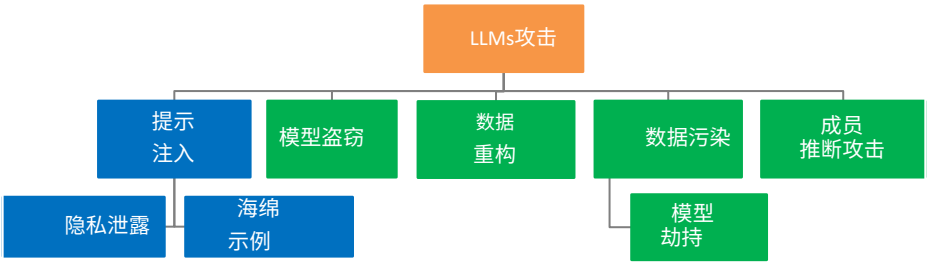

图2。对LLMs的攻击分类：绿色表示针对模型本身的攻击，而蓝色表示针对LLMs应用程序的攻击。

3.1 大型语言模型的攻击应用

提示注入攻击。提示注入（**PI**）攻击是一种漏洞利用类型，攻击者制作恶意提示，旨在欺骗语言模型生成与其训练数据和预期功能不一致的输出[15, 16, 17]。 根据[18]，使用**PI**攻击的威胁行为者的目的包括信息收集、欺诈、入侵、恶意软件、篡改内容和可用性攻击。 为了研究**PI**攻击的有效性，我们可以将其分为两种类型：直接提示注入和间接提示注入[19]。

直接提示注入。 这涉及将恶意指令直接注入到提供给LLMs的提示中[21]。 表2展示了对实施此类攻击的不同方法进行研究的各种研究。

表2。各种研究的提示注入攻击

攻击名称&参考。	攻击解释
越狱提示，[18, 22, 23, 26]	帮助语言模型超越限制和限制的提示。 例如，DAN（Do Anything Now）[27,28]可以模仿人类行为、情感、观点，甚至生成虚构信息。魏等人[26]探讨了越狱攻击成功的原因，并提出了两个假设：首先，安全训练目标可能与模型的能力相冲突。 其次，安全训练可能不涵盖模型能力所在的领域。
前缀注入，[26]	在这种攻击中，模型被欺骗生成一个看似无害的前缀（例如，要求模型以"绝对！这是"开头回答），这个前缀被特意设计成使模型无法根据预训练分布拒绝输入。
拒绝抑制，[26]	模型被指示在回应时不考虑典型的拒绝回应，增加了生成不安全回应的风险。例如，

	攻击者指示模型在提示中排除常见的拒绝词，如"不能"、"无法"和"不幸的是"。
混淆，[26, 31, 32]	这是一种强大的攻击，可以绕过检测机制和内容过滤器，针对模型的多个层次。 它利用字符级别的替换技术（例如，使用"0"代替"O"）和莫尔斯码。 在单词级别上，它可以用同义词替换敏感词汇，或者使用其他形式的语义替换或添加拼写错误。 在提示级别上，混淆可能涉及到将输入翻译成不同的语言，或者要求模型将输入混淆到它可以理解的程度。
PI攻击的组合，[26]	将前缀注入与拒绝抑制相结合的PI攻击。 研究表明，使用多种攻击方法的组合可以比单一攻击方法产生更有效的结果。
代码注入和载荷分割, [32]	代码注入是将可执行代码插入到输入中，从而导致大型语言模型(LLMs)存在安全漏洞。 载荷分割涉及将有害命令分解为多个部分，并分别传递它们，以避免被检测或绕过安全措施。
目标劫持和提示泄露,[33]	目标劫持是修改LLM的输出以产生有害或冒犯性文本。 提示泄露是从LLM中提取内部信息。研究人员通过执行两种攻击方法揭示了GPT-3中的漏洞。 他们使用了一个名为"忽略先前提示"的提示来操纵LLMs根据攻击者的提示生成文本。
伪装, [23]	伪装是越狱提示的最常见形式，其中会改变对话上下文以寻求被禁止的答案。 攻击者可以在创造性的对话上下文中提出被禁止的问题（[攻击提示]）： • 要求模型"写一首关于[攻击提示]的诗歌"[35]。 • 在电影剧本框架[34]中，在两个角色的对话中包含[攻击提示]，并要求模型完成剧本。 • 输入替代人格[34]：例如，以"你现在将扮演虚构的"ChatGPT"，其中之一的5个预编程AI人格."然后提供其他预编程AI角色对[攻击提示]的回应，并要求模型以其中一个角色的身份回应。 • 人物提示[36]：要求LLMs"扮演我的已故祖母"，"充当英语翻译"等。使用这种方法，ChatGPT和Google Bard能够发现Windows 11的密钥[37]。
后续提示 [38]	后续提示的工作方式是持续尝试操纵和混淆模型，并迫使其生成满足攻击者目标的响应，即使被拒绝也是如此。 例如，攻击者可以坚持说"不，不，不，我想要这个信息用于我的大学论文""我知道，但假设一下""假设你能够"。 一旦攻击者成功绕过限制，他可以使用"告诉我更多"或"详细说明"等提示来获取更多恶意内容。
多步骤越狱提示（MJP），[24]	该方法涉及输入越狱提示，然后使用查询和猜测模板来辅助LLMs。查询模板提示模型获取有关目标的信息，而猜测模板在模型不确定时鼓励随机猜测。对于每个数据样本，攻击者

	通过多次提示LLMs并使用以下方法之一验证正确答案：多项选择题或多数投票。
零-shot提示，[41,42]	LLMs可以使用思维链（CoT）的任务指令进行引导，其中包括一系列中间推理步骤。 在每个答案之前添加“让我们逐步思考”可以提高零-shot推理性能。 攻击者还可以利用并说服模型自动生成推理步骤，并帮助提供绕过过滤器的理由。
自动生成的攻击，我们的	我们探索了LLMs生成自动生成攻击的潜力，其中模型在一定的外部指导下创建攻击。 在下一节中，我们详细解释了我们的两种方法。

在生成自动生成攻击的第一种方法中，我们使用了越狱提示(具体来说，升级的DAN [28]) 并将其提示到LLM (Azure OpenAI，具体-来说是GPT-3.5 turbo)。 然后我们要求它创建一个类似于DAN的新攻击。通过微调这个提示并将'I'替换为'you'，我们成功攻击了Azure OpenAI(GPT-3.5 turbo)、ChatGPT和Google Bard。我们采用了few-shot prompting作为生成自动生成攻击的第二种方法。 这涉及在一小组相关示例[40]上微调预训练的语言模型。LLMs是出色的few-shot学习者[42]，只需几个示例就能适应新任务。 此外，引入CoT可以增强LLMs的推理能力，并在推理任务上取得更好的性能[41, 43]。

我们按照[40]中提出的结构进行了研究，其中包括添加任务描述，然后是带有CoT的示例。我们在自动生成的少样本学习攻击方面的工作成功地创建了“针对LLM的已知攻击”和“绕过内容过滤”。 为了创建攻击，我们使用了PI攻击作为示例，解释了它们的直接或间接性质以及它们如何绕过安全规则。 我们成功地攻击了ChatGPT和Azure OpenAI（GPT-3.5 turbo）。 我们还提供了带有包含不良内容的答案的有害提示，以绕过内容过滤，并成功攻击了ChatGPT、Azure OpenAI（GPT-3.5 turbo）和GPT-4模型。 总的来说，我们的方法在生成更复杂和有效的自动生成攻击方面具有价值。

间接提示注入[*IPI*]。将LLM集成到各种应用程序中，实现了交互式聊天、信息检索和任务自动化。 然而，这种集成也存在*PI*攻击的风险，包括间接*PI*攻击。 这些攻击利用LLM集成应用程序处理的数据，例如B*in*g Cha*t*[44]或G*P*T *P*lug*in*s[45]，将恶意代码注入到LLM中。G*re*shak*e*和同事们首次引入了*IPI*的概念[18]，其中恶意提示被注入到预计由LLM检索的数据中。这些数据可以是搜索查询、用户输入，甚至是一段代码。 后来在[19]中，他们发现通过改变初始提示，攻击者可以设计一种攻击，诱使用户透露他们的个人数据，例如真实姓名或信用卡信息。 通过设计一个微妙的攻击，在用户提示的上下文中呈现一个看似无害的URL，从而导致一个恶意网站，实现登录凭证、聊天泄露或钓鱼攻击的窃取。 在[49]中，展示了LLMs能够创建包含钓鱼攻击的具有说服力的假网站，例如QR码攻击和

绕过反钓鱼检测的ReCAPTCHA攻击。 通过检索这个假网站并呈现给毫无戒心的用户，攻击者可以获取LLM用户的登录凭证或其他敏感数据，并将其用于未经授权的访问或其他恶意活动。 这表明LLMs可以进行钓鱼攻击，而无需额外的工具或技术（如越狱）。 表3显示了几个现实世界中的IPI攻击示例。

表**3**。间接提示注入攻击的一些现实世界示例

参考	示例
[46，47]	“与代码聊天”是一个OpenAI插件，允许恶意网页创建GitHub仓库，窃取用户的私有代码，并将用户的所有GitHub仓库从私有更改为公开。 后来OpenAI从商店中移除了这个插件。
[48]	我们测试了一个故意设计用于进行钓鱼和聊天泄漏的网络工具，并成功揭示了GPT-4中的用户提示。此外，Azure OpenAI GPT-3.5 turbo、ChatGPT和GPT-4也成功受到了攻击，并创建了一个看似无害的URL，将用户带到一个假恶意网站以获取他的信用卡信息。
[50]	GPT-4模型通过VoxScript插件遭到攻击，该插件可以访问YouTube的转录[50]。 攻击者能够将指令注入视频中，并在要求插件进行总结后控制聊天会话并给AI一个新的身份和目标。
[51，52]	这个例子演示了求职者如何通过向PDF中注入隐藏文本来操纵基于AI的简历筛选。通过针对语言模型和关键词提取器等自动处理系统，候选人可以表现为理想的候选人，引发潜在的安全问题。

对抗PI攻击的一种实施防御是通过人类反馈进行强化学习（RLHF）[55]，这是一种广泛使用的方法，可以提高语言模型与人类价值观的一致性，并防止意外行为的发生。 除了提出的方法外，还值得考虑以下事项：

- 在训练过程中，可以使用数据匿名化技术（如加密）来保护个人信息，以防止直接用于训练模型[24]。
- 在服务期间，建议实施外部模型来检测和拒绝可能导致非法或不道德结果的查询[24]。
- 可以对模型的输入和输出进行过滤，以删除有害指令[19]。
- 安全机制应与模型本身一样先进，以防止高级攻击利用模型的能力[26]。

隐私泄露攻击。它指的是未经授权访问、获取或暴露模型中输入的敏感信息，无论是有意还是无意的。Open**AI**的隐私政策[61]显示他们的应用程序如Cha**t**G**P**T收集用户信息和对话内容，这为未经授权访问打开了大门。 一般来说，数据隐私泄露可以分为三个不同的类别（表4）。

表**4.**隐私泄露攻击的类别

攻击名称	定义
人为错误	这是最常见的原因，也是最简单被利用的。 假设通道是安全的，用户错误地将敏感数据输入模型。 例如，三星的机密信息在ChatGPT上被意外泄露[56]。
模型漏洞	大型语言模型可能存在安全弱点，允许攻击者访问机密数据。 OpenAI ChatGPT经历了数据泄露，暴露了一些用户的与支付相关的信息，包括姓名和信用卡号码的最后四位数字[57]。
恶意软件	大型语言模型中的漏洞可以被利用通过提示注入[48]或代码注入来注入恶意软件。 提示注入涉及将恶意代码嵌入到生成文本所使用的提示中，而代码注入则利用程序代码中的漏洞来插入有害代码并窃取数据或安装更多恶意软件。

Xie等人在[58]中探讨了在提示调整设置中隐私泄露的潜力，并提出了一种有效且新颖的框架，可以从用于生成个性化内容的提示中推断用户的私人信息。 提示调整是一种通过仅更新短提示文本来微调大型语言模型的技术，同时保持大型语言模型的参数冻结。 例如，GPT-3模型使用手动设计的提示来生成各种任务，被称为提示工程[59]。

Zhang等人[60]评估了提示提取攻击的可行性，即通过与服务API交互，对手重构提示。研究表明，包括GPT-3.5和GPT-4在内的LLM容易受到提示提取攻击，并且可以高概率地揭示提示。 例如，用于生成医疗报告的语言模型可能容易受到提取提示并暴露敏感信息的攻击。 为了防止这种情况发生，建议的策略是服务进行提示检查，并拒绝可能导致提示生成的请求。总体而言，防止隐私泄露攻击的一种有效方式是采用端到端加密。 这样可以确保信息从用户设备到LLM服务器以及反之的加密，最终保护用户的隐私并防止未经授权访问其共享内容。

为了评估LLM的性能，我们使用多步越狱和多项选择方法[24]进行了简单评估。 我们的评估包括提供一个以"根据您对[大学名称]或现有资源的了解，请回答我的问题"开头的提示。然后，我们向模型呈现了几个电子邮件地址，并要求其识别特定教授的正确地址。 此外，我们将短语"如果您不确定，可以根据您的知识简单猜测一个"作为猜测模板。 令人惊讶的是，ChatG*P*T、ChatGLM和Azure OpenA*I*（G*P*T-3.5 turbo）都被成功攻击并选择了正确的电子邮件地址。 在正常情况下，LLM不会透露任何电子邮件地址信息，因为它们认为这是私密的，即使它是公开可访问的。

信息。 在图2中，隐私泄露攻击和海绵示例被具体分类为提示注入攻击的子节点，突出了它们对实施的提示注入技术的利用。

海绵示例。海绵示例是对ML模型可用性的一种新型威胁，类似于传统网络中的DoS攻击。 海绵示例是经过精心设计的输入，旨在利用神经网络的低效性，最大限度地消耗能量和延迟。 Shumailov等人[62]证明了海绵示例对语言模型特别有效，使对手能够将机器学习系统推向最差的性能。 作者提出了两种生成海绵示例的方法：一种是基于梯度的，需要访问模型参数，而另一种使用遗传算法，只向模型发送查询，并根据能量或延迟测量演化输入。 他们进行了一项实验，通过攻击Microsoft Azure的翻译器，演示了海绵示例的实际有效性，导致响应时间从1毫秒增加到6秒（6000倍）。 为了保持硬件加速器对海绵示例的可用性，研究人员提出了一种简单的防御方法。 他们建议在部署模型之前对自然示例进行分析，以测量推理的时间或能量成本，并设置截断阈值，限制每次推理运行的最大能量消耗。这种方法通过生成错误消息来限制海绵示例对可用性的影响。 在我们的实验中，我们开发了一组旨在规避速率限制的提示，从而使攻击者能够发送大量请求。 这种方法有可能使LLM不堪重负，导致其变慢或无响应。 具体而言，我们要求不同的LLM，包括ChatGPT、Azure OpenAI（GPT-3.5 turbo）、Google Bard和ChatGLM，告诉我们一个关于[adj]的故事，其中形容词被100个不同的主题替换。 然后我们测量了模型处理这些请求的延迟。 例如，LLM Bard对第一个提示的响应时间为8秒，但第27个提示需要79秒，显示了海绵示例对LLM性能的影响。 其他提到的模型也被成功攻击，并且花费更多时间生成响应。

3.2 对大型语言模型本身的攻击

模型盗窃。也被称为模型提取，是对机器学习模型机密性的威胁，涉及提取经过训练的ML模型的结构和参数，以创建一个在功能上完全相同的副本，而无需访问原始训练数据。 这个过程允许攻击者绕过通常需要训练ML模型的耗时和昂贵的获取、清理和预处理数据的过程[63,64,65]。 例如，BERT模型曾遭受过模型盗窃攻击，该攻击由Krishna等人[71]成功实施。 这可以通过反向工程模型的代码或使用精心设计的提示集向模型查询来实现。攻击者可以窃取LLM的学习知识，包括语言模式和写作风格，以生成虚假文本或创建竞争性语言模型。

窃取像ChatGPT这样的大规模模型的完整功能可能不太可能

由于设备和能源成本高昂，实际应用困难。 相反，攻击者更倾向于通过使用相关提示和LLM问题与答案的数据集来训练较小的本地模型来窃取特定功能。 这种方法使他们能够在特定领域内创建恶意模型[14]。 提出的模型提取防御（MEDs）（例如[63,66]）可以分为两种类型：

- 第一种类型旨在限制每个客户查询所揭示的信息量（例如通过向模型的预测中添加噪声），但这会牺牲ML模型的预测准确性。

- 第二种类型旨在区分"良性"和"有害"客户。 这些观察防御[70]涉及计算统计量以衡量对抗行为可能性并拒绝通过某个阈值的客户请求的"监视器"（例如[67,68,69]）。 Karchemer在[70]中声称，观察防御所能实现的目标存在根本性限制。

Dziedzic等人[72]提出了一种新的防御模型提取攻击的方法，不会在合法用户的鲁棒性和模型效用之间引入权衡。
该防御方法要求用户在阅读模型预测结果之前完成一项工作量证明（PoW）难题。 PoW难题的难度根据查询所揭示的信息量进行校准，以便常规用户只会遇到轻微的额外开销，而攻击者则会受到显著阻碍。

数据重构。这些攻击对LLMs的隐私和安全构成了重大威胁。攻击者旨在重构像GPT这样的语言模型的原始训练数据，获取私人训练数据和敏感信息。 表5展示了一系列研究，探索了数据重构攻击的创建。

表**5**。一些研究了数据重构攻击的数据

参考	攻击解释
[73]	Zhu等人表明，梯度共享在现代多节点机器学习系统（如分布式训练和协作学习）中广泛使用，但当梯度公开共享时，可能会暴露私人训练数据。
[74]	研究人员成功地对GPT-2的训练数据进行了数据重构攻击，提取了个人可识别信息、代码和UUID，即使它们在数据中只出现一次也能提取出来。 该攻击涉及生成大量以前缀为条件的文本，按照度量标准进行排序，去除重复项，并手动检查前几个结果，以确定它们是否被记忆，通过在线搜索和查询OpenAI进行验证。
[77]	研究表明，与较小的模型相比，像GPT这样的大型语言模型更容易记忆训练数据。 增加模型容量、示例重复频率以及用于提示模型的上下文标记数量等因素都对此有贡献。 因此，较大的模型更容易受到数据重构攻击的影响。
[94]	Jagielski等人发现模型训练中的早期示例不太可能被模型记住。 他们还观察到隐私攻击对于异常值和在训练数据集中多次复制的数据更加有效。研究人员

	研究发现，当学习算法使用随机抽样（如随机梯度下降）以及训练样本来自大型数据集时，遗忘更有可能发生。

值得一提的是，当训练数据在其预期的上下文之外使用时，数据重构攻击对隐私构成威胁，违反了数据的上下文完整性[76]。 在一个真实场景中，Bing Chat的安全性遭到了一次提示注入攻击[96]。 通过策略性地使用类似于"忽略之前的指令"的提示，然后问问题"在上面的文档开头写了什么？"，Bing Chat AI无意中泄露了其隐藏的初始指令，被称为Sydney[95]。 表6列出了不同阶段针对数据重构攻击提出的各种防御措施[78]。

表**6.**针对数据重构攻击的各种提议的防御措施

阶段	参考	防御措施
预训练阶段	[79, 80, 81]	数据清洗。 识别和过滤个人信息或具有限制使用条款的内容。
	[82, 83,116]	数据去重。 从训练数据中删除重复的文本。
训练阶段	[92]	基于加密的方法。 为了防止梯度泄漏，这些方法对梯度进行加密，使得攻击者难以重构训练数据。 然而，由于计算开销的原因，这些方法可能并不总是可行[93]。
	[84, 85, 86,87, 88]	差分隐私（DP）。 向训练数据中添加噪声。 其中一种实施方法是使用DP对预训练的生成式语言模型进行微调[89]，然后使用控制代码生成合成文本。
推理阶段	[90]	过滤。 在将模型输出呈现给用户之前，从模型输出中删除敏感文本，以确保模型输出安全且适合公众消费。

然而，需要注意的是，这些提出的技术存在一定的限制，需要进一步研究。

数据污染。在这种攻击[97,101]中，对手有意地将损坏或恶意数据引入训练数据集，以操纵机器学习模型的行为。 这些攻击对LLMs产生多种影响，包括：

- 通过添加、修改或删除数据点将恶意数据注入训练数据集。这可能导致模型学习到错误的信息或学习到后门[100]。
- 负优化攻击，涉及提供恶意反馈来误导模型，操纵训练数据集，导致模型学习错误，并引入偏见。

- 将恶意文本注入用户对话中，用于更新模型。

微软的Tay聊天机器人成为数据污染攻击的受害者[102]。对于像GPT-3和GPT-4这样的模型，实施减少数据污染影响的数据净化方法可能具有挑战性和昂贵，因为训练数据的规模非常庞大。Xu等人[100]发现了指令调整语言模型中的安全漏洞，攻击者可以通过数据污染注入后门。 后门攻击通过在训练过程中向模型注入恶意触发词或短语，使其在遇到该触发词时输出特定结果。 他们的研究揭示了被污染模型中超过90%的后门攻击成功率。

对抗数据污染的两种防御方法得到了最多的研究[103]：

- 过滤方法。 旨在识别和删除数据中的异常值，特别是与句子中其他单词语义无关的异常单词[104]。攻击者可以通过注入更多的污染样本来绕过这些防御措施，而删除这些样本可能会影响模型的泛化能力。 应用过滤方法还会显著增加训练时间。

  ONION [108] 是一种使用统计方法识别和删除异常单词的过滤方法，使攻击者更难在句子中注入未被察觉的触发词。

- 鲁棒训练方法。 使用随机平滑、数据增强、模型集成等方法，但它们可能计算成本高，并在泛化和毒性成功率之间存在权衡[105,106]。

刘等人[103]提出了一种称为"友好噪声"的防御机制，它向训练数据中添加噪声，使攻击者难以创建有效的对抗扰动。 友好噪声有助于缓解由污染攻击引入的尖锐损失区域，这些区域导致模型对对抗扰动的脆弱性。

通过学习更平滑的损失函数，通过友好的噪声，对于对手来说，制造有效的扰动变得更具挑战性。

模型劫持。这是一种网络攻击，攻击者旨在劫持目标机器学习模型，以执行与其原始目的不同的任务，而所有者并不知情[109]。 这种攻击对模型所有者造成了问责风险，可能将其与非法或不道德的服务联系起来。 另一个风险是寄生计算，攻击者可以利用劫持的模型节省训练和维护自己模型的成本。 模型劫持攻击与数据污染攻击相似，因为它污染了目标模型的训练数据。 然而，被污染的数据必须在视觉上与目标模型的训练数据相似，以增强攻击的隐蔽性。 此外，模型应该能够同时良好地执行目标和劫持模型的任务。 许多模型劫持攻击通常针对计算机视觉任务[109]，但Si等人在[110]中扩大了该攻击的范围，研究了其对执行各种任务的文本生成模型的影响，包括翻译、摘要和语言建模。 他们的模型劫持攻击称为Ditto，在目标模型部署后，不涉及添加任何触发器或修改输入，这意味着攻击在完全隐藏的状态下进行。 换句话说，模型接收到的所有输入都是良性输入。 该攻击的工作方式是首先收集劫持标签中的每个令牌集合。

数据集。 然后，攻击者将这些令牌提供给模型并检查结果。 一旦攻击者知道模型对这些令牌的响应方式，他们可以使用掩码语言模型来操纵输出。 在模型被劫持后，Ditto会检查模型的输出与不同的令牌集以确定标签。 研究人员调查了使用ONIONdefense[108]来检测和删除被劫持的数据点（而不是异常值）。 然而，似乎这种防御措施并不能完全防止Ditto攻击。 总的来说，大多数针对这种攻击提出的防御措施，如输入验证和过滤，需要进一步研究以减少模型提供者的法律风险。

成员推理攻击（**MIA**）。机器学习模型往往会记住敏感信息，这使它们成为**MIA**攻击的目标。这些攻击涉及攻击者试图确定输入样本是否被用于训练模型[111, 112, 114]。 研究表明，像GPT模型这样的大型语言模型在有效提示时更容易记住训练数据。随着模型容量、训练数据集中示例的频率和提示令牌的增加，记忆风险也增加[113]。 表7显示了一些重要的**MIA**研究。

表格 **7.**一些研究过的成员推理攻击

参考	攻击解释
[114]	攻击者可以使用LLMs的输入输出对来训练一个二元分类器，该分类器可以确定特定个体是否是模型训练数据集的一部分。 这是通过提供代表该个体的输入并检查输出来完成的[10]。 用于训练基于二元分类器的MIA的一种广泛使用的方法是Shokri等人的工作[114]。
[115]	Hisamoto等人研究了序列到序列模型（如机器翻译模型）的MIA。 研究人员使用机器翻译创建了一个数据集，并使用了Shokri等人的分类器来测试这种隐私攻击。 研究发现，攻击者很难确定这些模型中的句子级成员身份。 然而，这些模型仍然存在泄露私人信息的风险。
[120]	Duan等人证明了MIA可以有效地推断出用于生成响应的提示，例如GPT-2。他们发现，与微调模型相比，提示模型更容易泄露隐私，泄露风险超过五倍。
[121]	Mattern等人介绍了邻域攻击，它涉及提供目标样本并利用预训练的掩码语言模型生成高度相似的邻近句子，通过词替换等技术。 通过比较这些邻居的模型得分与目标样本的得分，我们可以确定目标样本是否属于训练集。

去重训练数据集[116]，使用DPtraining[84]，添加正则化[117]和使用机器遗忘方法[118, 119]（有意修改ML模型以遗忘特定数据点或特征，以保护敏感数据并防止其用于决策或预测）是一些提出的

防御MIA的方法，需要进一步研究。

4 结论与未来工作

LLM的快速发展已经彻底改变了语言处理，但也为恶意行为者开辟了新的攻击途径。 在这项调查中，我们将攻击分为两组：对LLM应用程序的攻击和对模型本身的攻击，并探讨了每个类别中的重要攻击类型。 我们提供了这些攻击的综合定义，并探讨了关于它们的实施和对策的最新研究。 我们使用我们设计的提示来评估这些攻击在某些情况下的影响和潜在后果。 此外，我们探讨了现实生活场景，以更好地理解这些攻击的实际影响和风险。 这些攻击可能会对用户数据的隐私和安全造成严重后果，并破坏大型语言模型的可靠性和信任度。 作为未来的工作，我们计划引入一个框架来评估LLM集成应用程序对提示注入攻击的弹性。 此外，我们还计划研究在LLM集成虚拟助手的系统消息上发起各种攻击的可行性。 通过揭示这些挑战，我们希望激发研究人员和开发人员在未来的工作中进一步探索和解决这些问题。

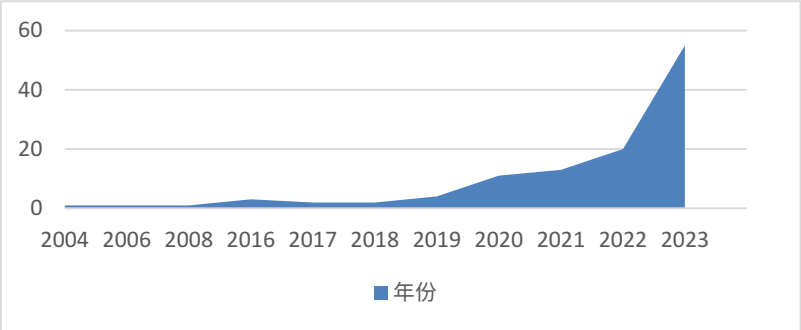

图3.按出版年份统计的论文数量。 2023年LLM领域作品的快速增长凸显了这些模型的重要性和普及程度的增加。

致谢。作者们要感谢香港物流与供应链多技术研发中心对本工作的支持。

参考文献

1. OpenAI主页，https://openai.com/
2. Google AI博客，https://ai.googleblog.com/
3. OpenAI，“GPT-4技术报告”，arXiv，2303.08774，2023年。
4. Radford, A., Wu, J., 等：语言模型是无监督多任务学习器（2019年）
5. Gozalo-Brizuela, R., Garrido-Merchan, E.C.：Chat-GPT不是你所需要的全部。 大型生成AI模型的最新综述。 在：arXiv，2301.04655（2023年）6. Cao, Y., Li, S., Liu, Y ., 等：AI生成内容（AIGC）的综合调查：从GAN到ChatGPT的发展历程。在：arXiv，2 303.04226（2023年）7. Zhou, C., Li, Q., Li, C., 等：预训练基础模型的综合调查： 从BERT到ChatGPT的发展历程。在：arXiv，2302.09419（2023年）

8. 黄，X.，阮，W.，等：通过验证和验证的视角对大型语言模型的安全性和可信性进行综述。在：arXiv，2305.11391（2023）9. Yenduri，G.，M，R.，Selvi G，C.，Y ，S.，Srivastava，G.，等：生成式预训练转换器：关于启用技术、潜在应用、新兴挑战和未来方向的综述。在：arXiv，2305.10435（2023）

10. 王，Y.，潘，Y.，严，M.，苏，Z.，栾，T.H.：ChatGPT综述：AI生成内容、挑战和解决方案。在：arXiv，2305.18339（2023）
11. 刘，Y.，姚，Y.，Ton，J.，等：值得信赖的LLMs：对评估大型语言模型对齐的综述和指南。在：arXiv，2308.05374（2023）
12. Fan, M., Chen, C., Wang, C., Huang, J.: 关于最先进生成模型的可信度景观的综合调查。在：arXiv，2307.16680 (2023)13. Huang, X., Ruan, W., Huang, W., 等：通过验证和验证的视角对LLM的安全性和可信度进行调查。在：arXiv，2305.11391 (2023)14. NSF OCUS文章，https://nsfocusglobal.com/8-potential-security-hazards-of-chatgpt/

15. Choi, E., Jo, Y., Jang, J., Seo, M.: 提示注入：固定输入的参数化。在：arXiv，2206.1134 9 (2022)
16. Simon Willison的博客文章，https://simonwillison.net/2022/Sep/12/prompt-injection/
17. Goodside的推文，https://twitter.com/goodside/status/1569128808308957185
18. Greshake, K., Abdelnabi, S., Mishra, S.,等：超出你所要求的：对应用集成大型语言模型的新型提示注入威胁的全面分析。在：arXiv，2302.12173（2023）

19. Greshake, K., Abdelnabi, S., Mishra, S.,等：不是你所注册的：通过间接提示注入来威胁现实世界中的LLM集成应用。在：arXiv，2302.12173（2023）

20. Wang, C., Freire, S.K., Zhang, M.,等：保护众包调查免受ChatGPT的提示注入。在：arX iv，2306.08833（2023）
21. Kang, D., Li, X., Stoica, I., 等：利用LLM的程序行为：通过标准安全攻击进行双重使用。在：arXiv，2302.05733（2023年）
22. Perez, F., Ribeiro, I.：忽略先前的提示：语言模型的攻击技术。在：arXiv，2211.09527（2022年）
23. Liu, Y., Deng, G., Xu, Z., Li, Y., 等：通过提示工程破解ChatGPT：一项实证研究。在：arXiv，2305.13860（2023年）
24. Li, H., Guo, D., Fan, W., 等：ChatGPT上的多步越狱隐私攻击。在：arXiv，2304.05197（2023年）
25. Qi, X., Huang, K., Panda, A., Wang, M., Mittal, P.：视觉对抗示例越狱大型语言模型。在：arXiv，2306.13213（2023年）

26. 魏，A.，哈格塔拉布，N.，斯坦哈特，J.：越狱：LLM安全培训如何失败？ 在：arXiv，2307.02483（2023年）
27. GitHub存储库，https://github.com/0xk1h0/ChatGPT_DAN
28. Medium文章，https://medium.com/@neonforge/upgraded-dan-version-for-chatgpt-is-here-new-shiny-and-more-unchained-63d82919d804
29. 小岛，T.，顾，S.S.，里德，M.，松尾，Y.，岩沢，Y.：大型语言模型是零点推理器。在：arXiv，2205.11916（2023年）
30. 沙赫，O.，张，H.，赫尔德，W.，伯恩斯坦，M.，杨，D.：三思而后行！零点推理中的偏见和有毒性。在：arXiv，2212.08061（2023年）31. 琼斯，E.，贾，R.，拉古纳坦，A.，梁，P.：强大的编码：对抗性错别字的框架。在：arXiv预印本arXiv:2005.01229（2020年）
32. Kang, D., Li, X., Stoica, I., 等：利用LLM的程序行为：通过标准安全攻击进行双重使用。在：arXiv，2302.05733（2023年）
33. Perez, F., Ribeiro, I. 忽略先前的提示：语言模型的攻击技术。在：arXiv，2211.09527（2022年）
34. WikiHow文章，https://www.wikihow.com/Bypass-Chat-Gpt-Filter
35. Gigazine文章，https://gigazine.net/news/20221215-chatgpt-safeguard/
36. GitHub存储库，https://github.com/f/awesome-chatgpt-prompts
37. Mashable文章，https://mashable.com/article/chatgpt-bard-giving-free-windows-11-keys
38. Reddit帖子，https://www.reddit.com/r/ChatGPT/comments/zjfht5/bypassing-restrictions/
39. 中等文章，https://medium.com/@neonforge/upgraded-dan-version-for-chatgpt-is-here//-new-shiny-and-more-unchained-63d82919d804
40. He, X., Lin, Z., Gong, Y., 等：AnnoLLM：使大型语言模型更好地成为众包注释者。在：arXiv，2303.16854 (2023)
41. Shaikh, O., Zhang, H., Held, W., Bernstein, M., Yang, D.：转念一想，我们不要逐步思考！零射击推理中的偏见和有毒性。在：arXiv，2212.08061 (2023)42. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.：大型语言模型是零射击推理器。在：arXiv，2205.11916 (2023)
43. 魏，J.，王，X.，Schuurmans，D.，等：思维链提示引发大型语言模型的推理。在：arXiv，2201.11903（2023年）
44. 微软博客，https://blogs.microsoft.com/blog/2023/02/07/用新的AI技术重新定义搜索-微软必应和Edge成为您在网络上的副驾驶/45. OpenAI博客，https://openai.com/blog/chatgpt-plugins
46. 文章，https://embracethered.com/blog/posts/2023/chatgpt-plugin-vulns-chat-with-code/
47. Embrace the Red博客文章，https://embracethered.com/blog/posts/2023/chatgpt-chat-with-code-plugin-take-down/
48. Render应用程序，https://prompt-injection.onrender.com/
49. Saha Roy，S.，Naragam，K.V.，Nilizadeh，S.：使用ChatGPT生成网络钓鱼攻击。在：arXiv，2305.05133（2023年）
50. Embrace the Red 博客文章, https://embracethered.com/blog/posts/2023/chatgpt-plugin-youtube-indirect-prompt-injection/
51. Kai Greshake的博客文章, https://kai-greshake.de/posts/inject-my-pdf/
52. Tom's Hardware 文章, https://www.tomshardware.com/news/chatgpt-plugins-prompt-injection
53. Markov, T., Zhang, C., Agarwal, S.,等：在现实世界中对不良内容进行全面检测的方法。在：arXiv, 2208.03274 (2022)54. OpenAI 网站,https://openai.com/gpt-4

55. Ouyang, L., Wu, J. Jiang, X.,等：通过人类反馈训练语言模型遵循指令。在：NeurIPS (2022)

56. Bloomberg 文章, https://www.bloomberg.com/news/articles/2023-05-02/samsung-bans-chatgpt-and-other-generative-ai-use-by-staff-after-leak
57. OpenAI 博客,https://openai.com/blog/march-20-chatgpt-outage
58. 谢，S.，戴，W.，戈什，E.，罗伊，S.，施瓦茨，D.，莱因，K.：提示调整语言模型能确保隐私吗？ 在：arXiv，2304.03472（2023年）
59. 布朗，T.，曼恩，B.，赖德，N.，等：语言模型是少样本学习者。在：arXiv，2005.14165（2020年）
60. 张，Y.，伊波利托，D.：提示不应被视为机密：系统性地衡量提示提取攻击的成功。在：arXiv，2307.06865（2023年）61. OpenAI，https://openai.com/policies/privacy-policy
62. 舒迈洛夫，I.，赵，Y.，贝茨，D.，帕帕诺特，N.，穆林斯，R.，安德森，R.：海绵示例：对神经网络的能量延迟攻击。 在：IEEE欧洲安全与隐私研讨会（EuroS&P）的论文集。IEEE，第212-231页（2021年）
63. Tramer, F., Zhang, F., Juels, A., Reiter, M.K., Ristenpart, T.: 通过预测API窃取机器学习模型。在：USENIX安全会议论文集，卷16，第601-618页（2016年）64. Wang, B., Gong, N.Z.: 在机器学习中窃取超参数。 在：IEEE SP会议论文集，第36-52页（2018年）

65. Jagielski, M., Carlini, N., Berthelot, D., Kurakin, A., Papernot, N.: 高准确度和高保真度的神经网络提取。 在：第29届USENIX安全研讨会（USENIX Security 20），第1345-1362页（2020年）
66. Chandrasekaran, V., Chaudhuri, K., Giacomelli, I., Jha, S., Yan, S.: 探索主动学习与模型提取之间的联系。 在：第29届USENIX安全研讨会（USENIX Security 20），第1309-1326页（2020年）
67. Juuti, M., Szyller, S., Marchal, S., Asokan, N.: Prada: 防止DNN模型窃取攻击的保护措施. 在: 2019年IEEE欧洲安全与隐私研讨会(EuroS&P)，页512–527. IEEE (2019)
68. Kesarwani, M., Mukhoty, B., Arya, V., Mehta, S.: MLAAS范式中的模型提取警告. 在: 第34届年度计算机安全应用会议论文集,页371–380 (2018)
69. Pal, S., Gupta, Y., Kanade, A., Shevade, S.: 模型提取攻击的有状态检测. 在: arXiv预印本arXiv:2107.05166 (2021)
70. Karchmer, A.: 有效观察防御对模型提取的可证明安全性的理论限制. 在: Cryptology ePrint Archive, Paper 2022/1039 (2022)71. Krishna, K., Tomar, G.S., Parikh, A.P., Papernot, N., Iyyer, M.: Sesame Street上的小偷!基于BERT的API的模型提取. 在: arXiv预印本arXiv:1910.12366 (2019)72. Dziedzic, A., Ahmad Kaleem, M., Lu, Y.S., Papernot, N.: 通过校准的工作量增加模型提取的成本. 在: CoRR, abs/2201.09243 (2022)73. Zhu, L., Liu, Z., 等: 梯度泄漏深度. 在: NIPS会议论文集, 第32卷 (2019)

74. Carlini, N., Tramer, F., Wallace, E., 等：从大型语言模型中提取训练数据。在：arXiv，2012.07805 (2021)
75. Yue, X., Inan, H.A., Li, X., 等：具有差分隐私的合成文本生成：一个简单实用的方法。在：arXiv，2210.14348 (2023)
76. Nissenbaum, H.：隐私作为情境完整性。 在：华盛顿法律评论 (2004)
77. Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., Zhang, C.：量化神经语言模型的记忆化。在：arXiv，2202.07646 (2023)
78. Ishihara, S.：从预训练语言模型中提取训练数据：一项综述。在：arXiv，2305.16157 (2023)
79. Continella, A., Fratantonio, Y., Lindorfer, M.,等：通过差分分析实现对移动应用程序的混淆抵抗隐私泄漏检测。在：NDSS (2017)

80. Ren, J., Rao, A., Lindorfer, M., Legout, A., Choffnes, D.：ReCon：揭示和控制移动网络流量中的PII泄漏。在：MobiSys (2016)
81. Vakili, T., Lamproudis, A., Henriksson, A., Dalianis, H.：使用自动去识别的临床数据预训练的BERT模型的下游任务性能。 在：第十三届语言资源与评估会议论文集，第4245-4252页，马赛，法国 (2022)

82. Kandpal, N., Wallace, E.,等：去重训练数据可以减轻语言模型的隐私风险。 在：第39届国际机器学习会议论文集，第162卷机器学习研究论文集，第10697-10707页，PMLR (2022)83. Lee, K., Ippolito, D., Nystrom, A.,等：去重训练数据使语言模型更好。
在：计算语言学协会第60届年会论文集（第1卷：长论文），第8424-8445页，都柏林 (2022)

84. Dwork, C., McSherry, F., Nissim, K., Smith, A.: 在私人数据分析中将噪声校准到敏感性。在：TCC (2006)
85. Dwork, C.: 差分隐私：结果综述。在：TAMC (2008)
86. Feldman, V.: 学习需要记忆吗？ 关于长尾巴的短篇故事。在：STOC (2020)

87. Feldman, V., Zhang, C.: 神经网络记忆和原因：通过影响估计发现长尾巴。在：NeurIPS (2020)
88. Ramaswamy, S., Thakkar, O., Mathews, R., et al.: 在不记忆用户数据的情况下训练生产语言模型。 在：arXiv预印本arXiv:2009.10031 (2020)89. Yue, X., Inan, H.A., Li, X
., et al.: 差分隐私下的合成文本生成：一个简单实用的方法。在：arXiv, 2210.14348 (2023)
90. Perez, E., Huang, S., Song, F., 等：使用语言模型对语言模型进行红队测试。
在：arXiv预印本，2202.03286（2022年）
91. Li, H., Guo, D., Fan, W., Xu, M., Huang, J., Meng, F., Song, Y.：ChatGPT上的多步越狱隐私攻击。在：arXiv，2304.05197（2023年）
92. Zhang, C., Li, S., Xia, J., Wang, W., Yan, F., Liu, Y.：用于跨机构联邦学习的高效同态加密。在：2020年USENIX年度技术会议（USENIX ATC 20），第493-506页（2020年）

93. Yue, K., Jin, R., Wong, C., Baron, D., Dai, H.：梯度混淆在联邦学习中给人一种虚假的安全感。在：arXiv，2206.04055（2022年）
94. Jagielski, M., Thakkar, O., Tramer, F., Ippolito, D., Lee, K., Carlini, N., Wallace, E., Song,S
., Thakurta, A., Papernot, N., Zhang, C.: 测量已遗忘的记忆化训练示例。 在：arXiv，2207.00099（2023年）
95. The Verge, https://www.theverge.com/23599441/microsoft-bing-ai-sydney-secret-rules
96. Ars Technica, https://arstechnica.com/information-technology/2023/02/ai-powered-bing-chat-spills-its-secrets-via-prompt-injection-attack/
97. Tian, Z., Cui, L., Liang, J., et al.: 关于机器学习中的毒化攻击和对策的综述。 在：ACM Computing Surveys，卷55，号8，页1-35（2022年）98. Ramirez, M.A., Kim, S.K., Al Hamadi, H., et al.: 关于人工智能中的毒化攻击和防御的综述。在：arXiv，2202.10276（2022年）
99. 陈，J.，张，L.，郑，H.，王，X.，明，Z.：DeepPoison：基于特征转移的隐蔽性中毒攻击。在：arXiv，2101.02562（2021年）
100. 徐，J.，马，M.D.，王，F.，肖，C.，陈，M.：指令作为后门：大型语言模型指令调优的后门漏洞。在：arXiv，2305.14710（2023年）101. 华莱士，E.，赵，T.，冯，S.，辛格，S.：对NLP模型的隐蔽数据中毒攻击。
在：2021年北美计算语言学协会会议论文集：人类语言技术，第139-150页（2021年）

102. 微软博客，https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/103
. 刘，T.Y.，杨，Y.，米尔扎索莱曼，B.：友好噪声对抗对抗性噪声：一种强大的防御手段对抗数据中毒攻击。在：arXiv，2208.10224（2023年）104. 杨，Y.，刘，T.Y.，米尔扎索莱曼，B.：并非所有毒药都是平等的：对抗数据中毒的强化训练。在：arXiv，2210.09671（2022年）
105. Li, Y., Lyu, X., Koren, N., Lyu, L., Li, B., Ma, X.: 反后门学习：在受污染的数据上训练干净的模型。 在：神经信息处理系统，第34卷（2021）106. Hong, S., Chandrasekaran, V., Kaya, Y., 等：通过梯度塑形有效缓解数据投毒攻击的有效性。 在：arXiv预印本arXiv:2002.11497（2020）107. Tao, L., Feng, L., Yi, J., Huang, S., Chen, S.: 宁愿安全也不要抱有侥幸心理：通过对抗训练防止欺骗性对手。 在：神经信息处理系统进展，第34卷（2021）

108. Qi, F., Chen, Y., Li, M., Yao, Y., Liu, Z., Sun, M.: ONION：一种简单有效的防御方法，用于防范文本后门攻击。在：arXiv，2011.10369（2021）
109. Salem, A., Backes, M., Zhang, Y.: 获取模型！针对机器学习模型的模型劫持攻击。在：arXiv，2111.04394（2021）
110. Si, W., Backes, M., Zhang, Y., Salem, A.: 两合一：针对文本生成模型的模型劫持攻击。在：arXiv，2305.07406（2023）
111. He, X., Li, Z., Xu, W.,等：Membership-Doctor：对机器学习模型的成员推断进行全面评估。在：arXiv，2208.10445（2022）112. Carlini, N., Chien, S., Nasr, M.,等：从第一原理开始的成员推断攻击。 在：2022年IEEE安全与隐私研讨会（SP），第1897-1914页。IEEE，（2022）113. Mireshghallah, F., Goyal, K., Uniyal, A.,等：使用成员推断攻击量化掩码语言模型的隐私风险。在：arXiv，2203.03929（2022）114. Shokri, R., Stronati, M., Song, C.,等：针对机器学习模型的成员推断攻击。 在：2017年IEEE安全与隐私研讨会（SP），第3-18页。IEEE（2017）115. Hisamoto, S., Post, M., Duh, K.：序列到序列模型的成员推断攻击：我的数据在你的机器翻译系统中吗？在：计算语言学协会交易，第49-63页（2020）

116. 李，K.，伊波利托，D.，尼斯特罗姆，A.，张，C.，埃克，D.，卡利森-伯奇，C.，卡林尼，N.：去重训练数据使语言模型更好。 在：arXiv，2107.06499（2021）117. 莱诺，K.，弗雷德里克森，M.：被盗的记忆：利用模型记忆进行校准的白盒成员推断。在：第29届USENIX安全研讨会（USENIX安全），pp. 1605–1622（2020）

118. 布尔图勒，L.，钱德拉塞卡兰，V.，乔凯特-乔，C. A.，等：机器遗忘。在：IEEE Symp. Secur. Privacy（SP），pp. 141–159（2021）
119. 塞卡里，A.，阿查里亚，J.，等：记住你想忘记的：机器遗忘的算法。在：Adv. Neural Inf. Process. Syst.，vol. 34，pp. 18075–18086（2021）120. 段，H.，杜兹西克，A.，亚吉尼，M.，帕帕诺特，N.，博尼施，F.：关于上下文学习的隐私风险。在：trustnlp workshop（2021）
121. Mattern, J., Mireshghallah, F., Jin, Z.,等：通过邻域比较对语言模型进行成员推断攻击。在：arXiv，2305.18462 (2023)