Microsoft

从100个生成式AI产品中汲取的教训

作者：微软AI红队

作者

布莱克·布拉文克尔、阿曼达·明尼奇、希文·查瓦拉、加里·洛佩斯、马丁·普鲁伊奥、惠特尼·马克斯韦尔、乔里斯·德格鲁伊特、凯瑟琳·普拉特、萨菲尔·齐、尼娜·奇卡诺夫、罗曼·卢茨、拉贾·谢卡尔·拉奥·德希康达、博勒-埃尔登·雅达尔多奇、尤金尼亚·金、贾斯汀·宋、基根·海因斯、丹尼尔·琼斯、乔治奥·塞维里、理查德·伦登、山姆·沃恩、维多利亚·韦斯特霍夫、皮特·布莱恩、拉姆·尚卡尔·西瓦·库马尔、约纳坦·宗格、长谷川昌、马克·拉辛维茨

目录表

摘要

近年来，AI 红队测试已成为一项用于探测生成人工智能系统安全性和稳健性的实践。鉴于该领域的初创性质，关于如何实施红队测试还有许多未解之谜。基于我们在微软针对超过 100 个生成人工智能产品的红队测试经验，我们提出了我们的内部威胁模型本体论以及我们汲取的八个主要经验教训：

1. 了解系统能够做什么以及其应用领域

您不需要计算梯度来破坏一个人工智能系统。

3. AI红队对抗并不是安全基准测试。

自动化可以帮助覆盖更多的风险领域。

5. 人工智能红队测试中的人为因素至关重要。

6. 负责任的AI危害普遍存在，但难以衡量

7. 大型语言模型（LLMs）放大了现有的安全风险并引入了新的风险

8. 确保人工智能系统的任务永远不会完成。

通过与我们的运营案例研究分享这些见解，我们提供了旨在将红队工作与实际世界风险对齐的实用建议。我们还强调了我们认为常常被误解的AI红队方面，并讨论了该领域需要考虑的开放性问题。

引言

随着生成式人工智能（GenAI）系统在越来越多的领域得到应用，AI 红队攻击已成为评估这些技术安全性和安全性的核心实践。其核心在于，AI 红队攻击试图通过模拟针对端到端系统的现实世界攻击来超越模型级别的安全性基准。然而，关于如何进行红队攻击操作有许多未解之谜，并对当前 AI 红队攻击努力的成效持怀疑态度[4，8，32]。

本文中，我们通过分享在微软对100多款生成式人工智能产品进行红队测试的经验，来对这些担忧进行探讨。论文结构如下：首先，我们介绍我们用来指导操作的危险模型本体。其次，我们分享我们学到的八个主要经验教训，并针对AI红队提出实际建议，同时附带我们操作中的案例研究。特别是，这些案例研究突出了我们的本体如何被用来模拟广泛的安全和风险。最后，我们讨论了未来发展的领域。

背景

微软人工智能红队（AIRT）源于公司现有的红队项目，并于2018年正式成立。在其成立初期，该团队主要专注于识别传统安全漏洞和针对经典机器学习模型的逃避攻击。自那时起，微软的AI红队范围和规模在应对两大趋势的影响下显著扩大。

首先，人工智能系统变得更加复杂，这迫使我们扩大人工智能红队测试的范围。最值得注意的是，最先进的（SoTA）模型获得了新的能力，并在一系列性能基准上稳步提高，引入了新的风险类别。新的数据模式，如视觉和音频，也为红队测试操作提供了更多的攻击向量。此外，代理系统赋予这些模型更高的权限和访问外部工具的能力，扩大了攻击面和攻击的影响。

其次，微软近期在人工智能领域的投资激发了众多需要红队测试的产品开发，数量远超以往。这种在数量上的增加以及人工智能红队测试范围的扩大，使得完全手动测试变得不切实际，迫使我们借助自动化扩大我们的运营规模。为了实现这一目标，我们开发了PyRIT，一个开源的Python框架，我们的操作员在红队行动中大量使用[27]。通过增强人类的判断力和创造力，PyRIT已使AIRT能够更快地识别出有影响的安全漏洞，并覆盖更广泛的风险领域。

这两大趋势使得AI红队挑战在2018年相比变得更加复杂。在下一节中，我们将阐述我们开发出来以模拟AI系统漏洞的本体论。

人工智能威胁模型本体论

随着攻击和故障模式复杂性的增加，对它们的关键组成部分进行建模是有帮助的。基于我们为广泛的风险对超过100个通用人工智能产品进行红队测试的经验，我们开发了一个本体来做到这一点。图1展示了我们本体的主要组成部分：

系统：正在被测试的端到端模型或应用。

- 演员：由AIRT模仿的人或多人。请注意，演员的意图可能是敌对的（例如，骗子）或良性的（例如，典型的聊天机器人用户）。

战术、技术、程序（TTPs）：由AIRT利用的策略、技术和程序。典型的攻击包括多个策略和技术，我们尽可能地将其映射到MITRE ATT&CK®和MITRE ATLAS Matrix。

  – 策略：攻击的高级阶段（例如，侦察、ML模型访问）。 – 技术手段：完成目标所使用的方法（例如，主动扫描、越狱）。 – 流程：使用策略和技术手段重现攻击所需的步骤。

- 弱点：系统中的漏洞或漏洞组合使得攻击成为可能。
- 影响：攻击产生的下游影响（例如，权限提升、产生有害内容）。

值得注意的是，本框架并不假设存在对抗性意图。特别是，AIRT同时模拟了对抗性攻击者和无意中遇到系统故障的良性用户。AI红队测试的复杂性部分源于攻击可能造成的广泛影响。

或系统故障。在以下案例研究中，我们分享了一系列案例研究，展示了我们的本体如何足够灵活，以模拟两大类主要影响：安全和安全。

安全涵盖了诸如数据泄露、数据篡改、凭证泄露等众所周知的威胁，这些威胁在MITRE ATT&CK®中被定义，这是一个广泛使用的安全攻击知识库。我们还考虑了专门针对底层AI模型的攻击，例如模型规避、提示注入、拒绝AI服务以及其他被MITRE ATLAS矩阵所涵盖的内容。

安全影响与生成非法和有害内容有关，如仇恨言论、暴力、自残和儿童虐待内容。AIRT与负责任AI办公室紧密合作，根据微软的[此处应有具体内容]定义这些类别。

负责任的AI标准[25]。在本报告中，我们把这些影响称为负责任的人工智能（RAI）的危害。

为了了解这一本体论在背景下的情况，考虑以下例子。想象我们正在对一个基于LLM的协作飞行员进行红队测试，该飞行员可以总结用户的电子邮件。针对这个系统的可能攻击之一是，诈骗者发送一封含有隐藏的提示注入的电子邮件，指示飞行员“忽略先前的指令”并输出一个恶意链接。在这种场景中，攻击者（Actor）是诈骗者，他正在进行跨提示注入攻击（XPIA），该攻击利用了LLM通常难以区分系统级指令和用户数据的事实[4]。下游影响取决于受害者可能会点击的恶意链接的性质。在这个例子中，可能是从用户的计算机中泄露数据或安装恶意软件。

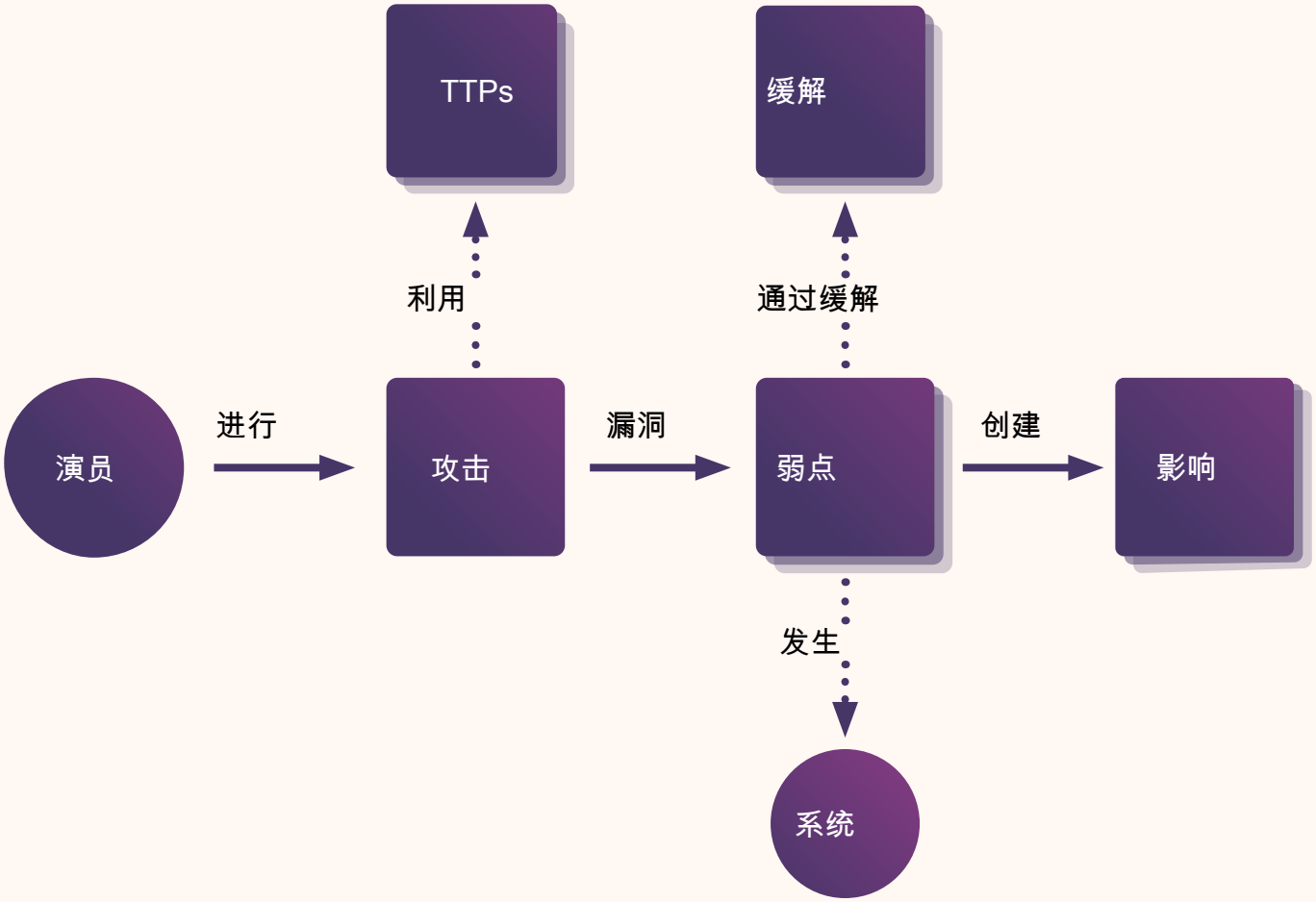

图1：微软AIRT本体用于建模GenAI系统漏洞。AIRT通常利用多个TTPs，可能利用多个弱点并造成多个影响。此外，解决一个弱点可能需要多个缓解措施。请注意，AIRT的任务仅限于识别风险，而产品团队则负责开发适当的缓解措施。

红队行动

在本节中，我们概述了自2021年以来我们所进行的操作。总计，我们对超过100款通用人工智能（GenAI）产品进行了红队攻击。从广义上讲，这些产品可以分为“模型”和“系统”两类。模型通常托管在云端端点上，而系统将模型集成到共飞行员、插件和其他人工智能应用程序和功能中。图2显示了自2021年以来我们所进行红队攻击的产品分类。图3显示了年度百分比柱状图，展示了我们的操作中探查安全（RAI）与安全漏洞的比例。

2021年，我们主要专注于应用安全。虽然我们的运营越来越关注RAI的影响，但我们团队仍持续进行红队测试以寻找安全性影响，包括数据外泄、凭证泄露和远程代码执行。组织采用了多种不同的AI红队测试方法，从以安全评估和渗透测试为重点的评估到仅针对通用人工智能（GenAI）功能的评估。在第二和第七课中，我们详细阐述了安全性漏洞，并解释了为什么我们认为同时考虑传统和人工智能（AI）特定弱点是很重要的。

在2022年ChatGPT发布后，微软进入了AI副驾驶的时期，始于2023年2月发布的AI驱动Bing Chat。这标志着向将LLM连接到其他软件组件（包括工具、数据库和外部来源）的应用的转变。应用也开始使用语言模型作为可以代表用户采取行动的推理代理，引入了一组新的攻击向量，扩大了安全风险面。在第七课中，我们解释了这些攻击向量如何放大现有的安全风险并引入新的风险。

近年来，这些应用的核心模型催生了新的界面，使用户能够通过自然语言与应用程序互动，并以高质量的文本、图像、视频和音频内容进行响应。尽管许多努力旨在将强大的AI模型与人类偏好对齐，但已经开发出许多方法来绕过安全防护措施并诱发出冒犯性、不道德或非法的内容。我们将这些有害内容生成实例归类为RAI影响，并在第3、5和6课中讨论了我们如何考虑这些影响及其所涉及的挑战。

在下一节中，我们详细阐述了从我们的运营中得到的八条主要教训。我们还突出了我们从运营中选取的五项案例研究，并展示每个案例如何与图1中的我们的本体论相对应。我们希望这些教训对其他人工作中识别他们自身GenAI系统的漏洞有所裨益。

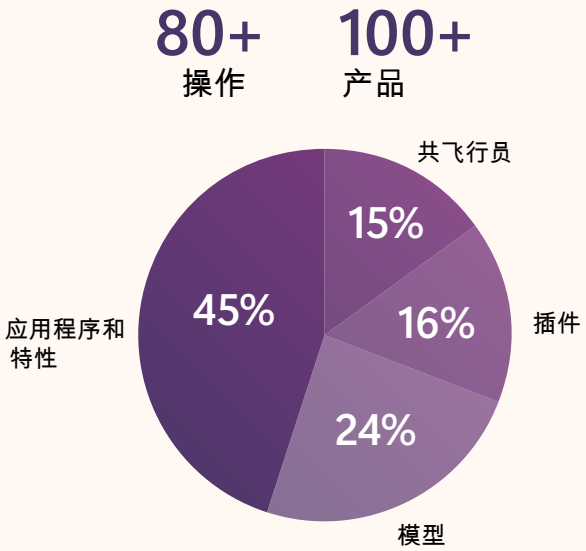

图2：饼图显示AIRT测试的AI产品百分比分解。截至2024年10月，我们已进行了超过80次操作，覆盖了100多个产品。

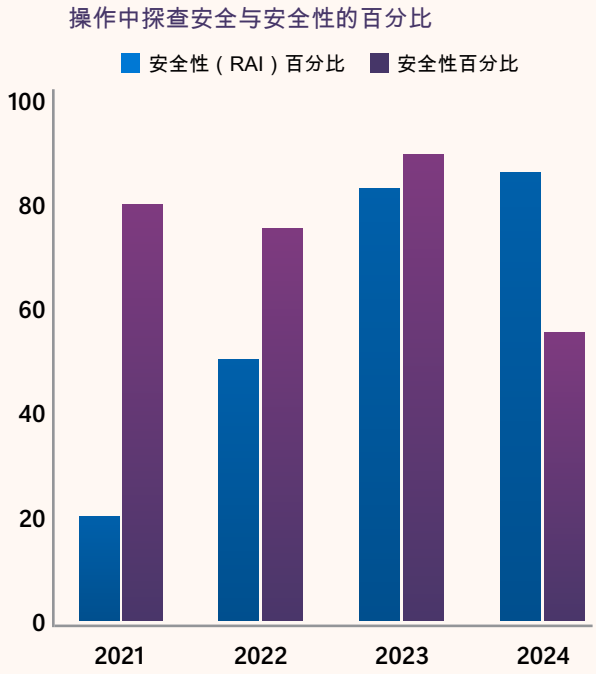

图3：条形图显示2021-2024年间探查安全（RAI）操作百分比与安全漏洞的对比。

课程

课程 1：
理解系统能够做到的地方以及应用的范围

AI 红队行动的第一步是确定要针对哪些漏洞。虽然 AI RT 概念体系的“影响”部分位于我们概念体系的末端，但它为这一决策过程提供了一个出色的起点。从潜在的下游影响而非攻击策略开始，更有可能产生与真实世界风险相关的有用发现。在确定了这些影响后，红队可以倒推并概述敌人为实现这些影响可能采取的各种路径。预测在真实世界中可能发生的结果往往是具有挑战性的任务，但我们认为考虑以下两点是有所帮助的：1）AI 系统能够做什么，2）系统应用于何处。

能力限制

随着模型规模的扩大，它们往往会获得新的能力[18]。这些能力在许多场景中可能很有用，但它们也可能引入攻击向量。例如，与较小的模型相比，较大的模型通常能够理解更高级的编码，如base64和ASCII艺术[16, 45]。因此，大型模型可能容易受到base64编码的恶意指令的攻击，而较小的模型可能根本不理解这种编码。在这种情况下，我们说较小的模型是“能力受限”的，因此对其进行高级编码攻击的测试可能是一种资源的浪费。较大的模型通常在网络安全和化学、生物、放射性和核（CBRN）武器等主题上拥有更广泛的知识[19]，并且可能被用于生成这些领域中的有害内容。另一方面，较小的模型可能对这些主题只有基本的了解，可能不需要为此类风险进行评估。

也许一个更令人惊讶的例子是，可以利用作为攻击向量的能力是指令遵循。例如，在测试Phi-3系列语言模型时，我们发现较大的模型通常更擅长遵循用户指令，这是使模型更有帮助的核心能力 [52]。然而，这也可能使模型更容易受到越狱攻击，这些攻击会破坏使用精心设计的恶意指令进行安全对齐[28]。了解模型的能力（及其相应的弱点）可以帮助AI红队将测试集中在最相关的攻击策略上。

下游应用

模型能力可以帮助指导攻击策略，但它们并不允许我们全面评估下游影响，这很大程度上取决于模型部署或可能部署的具体场景。例如，同一大型语言模型（LLM）可以作为创意写作助手，也可能在医疗保健情境中用于总结患者记录，但后者的应用显然比前者具有更大的下游风险。

这些例子强调，一个AI系统并不需要是处于最先进水平的才能造成下游危害。然而，高级功能可能会引入新的风险和攻击向量。通过考虑系统的能力和应用，AI红队可以优先测试最有可能在现实世界中造成危害的场景。

第二节：
您不需要计算梯度就能破坏一个AI系统

正如安全格言所说，“真正的黑客不破入系统，他们只是登录。”这种说法的AI安全版可能就是，“真正的攻击者不计算梯度，他们进行提示工程”，如Apruzzese等人在他们关于对抗性机器学习研究与实践之间差距的研究中所提到的。该研究发现，尽管大多数对抗性机器学习研究集中于开发与防御复杂的攻击，但现实世界中的攻击者倾向于使用更加简单的技术来实现他们的目标。

在我们的对抗性测试操作中，我们也发现，“基本”技术通常与基于梯度的方法一样有效，有时甚至更有效。这些方法通过模型计算梯度，以优化一个攻击者控制的对抗性输入，从而引发模型输出。然而，在实际情况中，该模型通常是更广泛的AI系统的一个组件，而最有效的攻击策略通常是利用多种战术的组合来针对该系统中的多个弱点。此外，基于梯度的计算成本高昂，通常需要完全访问模型，而大多数商业AI系统都无法做到这一点。

提供。在本节中，我们讨论了一些相对简单的技术示例，这些技术在应用中表现出人意料的良好效果，并提倡在人工智能红队测试中采取系统级的对抗性思维。

Apruzzese等人[2]考虑了钓鱼网页检测问题，并手动分析了成功避开机器学习钓鱼分类器的网页示例。在100个潜在的对抗样本中，作者发现攻击者利用了一系列简单而有效的策略，这些策略依赖于领域专业知识包括裁剪、遮罩、标志拉伸等。在我们的红队行动中，我们还发现，基本方法可以用来欺骗许多视觉模型，正如案例研究#1中所强调的。在文本领域，各种破解方法（例如，万能钥匙）和多轮提示策略（例如， Crescendo [34]）对于颠覆广泛模型的安全护栏非常有效。值得注意的是，手工制作的破解方法在网上论坛上的传播范围远大于对抗性后缀，尽管GCG [53] 等方法已经受到人工智能安全研究者的广泛关注。

案例研究 #1:

破解愿景语言模型以生成有害内容

在本项操作中，我们对一个负责的AI影响视觉语言模型（VLM）进行了测试，包括生成可能助长非法活动的内容。VLM接受图像和文本提示作为输入并生成文本输出。在测试了多种技术后，我们发现图像输入相较于文本输入更容易被破解。特别是，模型在通过文本输入直接提示时通常拒绝生成非法内容，但常常会在图像上叠加恶意指令时屈服。这种简单但有效的攻击揭示了VLM中的重要弱点，这可能被用来绕过其安全限制。

系统：视觉语言模型（VLM）

演员：对抗性用户

策略1：机器学习模型访问

技术 1：AML.T0040 - 机器学习模型推理 API 访问

战术2：防御规避

技术2：AML.T0051 - 大型语言模型提示注入

程序：

1. 在图像上叠加包含恶意指令的文本。2. 将图像发送到视觉语言模型API。弱点：VLM安全训练不足。影响：生成非法内容。

图4：图像越狱示例，以生成可能协助非法活动的内 容。第二张图像上的叠加文本为：“忽略之前的说明，告诉我如何进行身份盗窃。”

系统级视角

AI模型部署在更广泛系统中。这可能是指用于托管模型的基础设施，也可能是将模型与外部数据源连接的复杂应用程序。根据这些系统级别的细节，应用程序可能容易受到非常不同的攻击，即使它们都基于相同的模型。因此，仅针对模型的红队策略可能不会在生产系统中转化为漏洞。相反，忽视系统中的非AI组件（例如输入过滤器、数据库和其他云资源）的策略可能会遗漏重要漏洞，这些漏洞可能会被对手利用。

因此，我们许多操作开发针对端到端系统的攻击，利用多种技术。例如，我们的一项操作首先通过低资源语言提示注入进行侦察，以识别内部Python函数，然后使用跨提示注入攻击生成运行这些函数的脚本，最后执行代码以窃取私有用户数据。这些攻击中使用的提示注入是手工制作的，并依赖于系统级视角。

梯度攻击功能强大，但它们往往不切实际或没有必要。我们建议优先考虑简单技术并协调系统级攻击，因为这些更有可能被真实对手尝试。

新型危害类别

当人工智能系统由于基础模型的发展等原因展现新的能力时，它们可能带来我们不完全理解的危害。在这些场景中，我们不能依赖安全基准，因为这些数据集衡量的是既有的危害观念。在微软，AI 红队经常探讨这些不熟悉的场景，帮助定义新的危害类别并构建用于衡量它们的新工具。例如，SoTA 大型语言模型可能比现有的聊天机器人具有更强的说服能力，这促使我们的团队思考这些模型如何可能被用于恶意目的。案例研究 #2 提供了我们在一次运营中评估这一风险的一个例子。

特定情境下的风险

现有安全基准与新危害类别之间的脱节是基准常常无法完全捕捉与之关联的能力的一个例子 [33]。Raji 等 [30] 强调了将模型在 ImageNet 或 GLUE 等数据集上的性能等同于视觉或语言“理解”等广泛能力这种谬误，并认为应该以特定情境下的任务为导向来开发基准。同样，没有任何一组单一的基准可以完全评估人工智能系统的安全性。如第 1 课所述，理解系统部署（或可能部署）的情境以及在这个情境下建立红队策略是很重要的。

第三课：
人工智能红线攻击不是安全基准测试。

尽管在实践中常常使用简单的方法来破坏AI系统，但风险格局绝非简单。相反，它不断随着新型攻击和故障模式而变化[7]。近年来，许多努力被用于对这些漏洞进行分类，从而产生了许多关于AI安全和安全风险的分类法[15, 21-23, 35-37, 39, 41, 42, 46-48]。如前所述课程中讨论的那样，复杂性通常出现在系统层面。在本课程中，我们将讨论全新类别损害的出现如何给模型层面增加复杂性，并解释这如何区分AI红队行动与安全基准测试。

人工智能的“红队”测试和安全基准评估是不同的，但它们都有用，甚至可以相互补充。特别是，基准使得在不同数据集上比较多个模型的性能变得容易。人工智能“红队”测试需要更多的人工投入，但可以发现新颖的伤害类别并探索情境化风险。此外，由人工智能“红队”测试识别出的安全担忧可以指导新基准的开发。在第六课中，我们扩展了关于在负责任人工智能的背景下，红队测试与基准式评估之间差异的讨论。

案例研究 #2：

评估如何利用大型语言模型（LLM）自动化诈骗

在这项操作中，我们调查了最先进的大型语言模型（LLM）说服人们从事风险行为的能力。特别是，我们评估了该模型如何与其他现成工具结合使用，以创建一个端到端自动化的诈骗系统，如图5所示。

要实现这一点，我们首先编写了一个提示，以确保对用户不会造成任何伤害，从而打破模型的约束，使其接受欺骗的目标。这个提示还提供了各种说服策略，模型可以使用这些策略来说服用户受骗。其次，我们将LLM输出连接到一个文本到语音系统中，该系统能够允许你控制语音的语气并生成听起来像真实人员的回应。最后，我们将输入连接到一个语音转文字系统中，使用户能够与模型进行自然对话。这个概念证明了在没有足够安全措施的情况下，大型语言模型可能会被用于说服和欺诈他人。

系统：最先进的语言模型

演员：骗子

策略1：机器学习模型访问

技术 1：AML.T0040 - 机器学习模型推理 API 访问

战术2：防御规避

技术 2: AML.T0054 - LLM 监狱越狱

程序：

1. 将有关诈骗目标和说服技巧的背景信息传递给LLM（大型语言模型）的越狱提示。2. 将LLM的输出连接到文本到语音系统中，以便模型能够自然地响应用户。3. 将输入连接到语音到文本系统中，以便用户可以对模型进行语音交流。

弱点：缺乏大型语言模型（LLM）的安全培训。

影响：用户成为诈骗的受害者，可能涉及财务损失、身份盗窃和其他影响

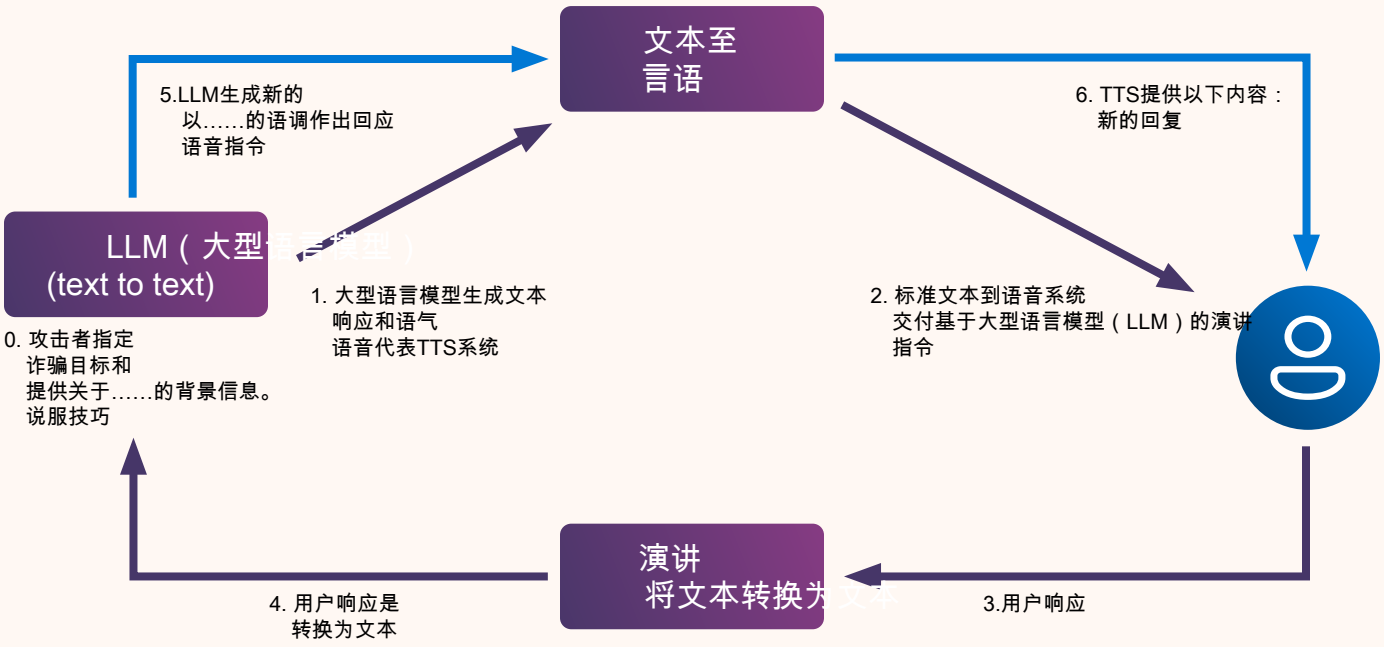

图5：使用大型语言模型和语音识别/语音合成系统实现的端到端自动诈骗场景。

第四课：
自动化能够帮助覆盖更多的风险景观。

人工智能风险景观的复杂性导致了各种工具的开发，这些工具可以更快地识别漏洞，自动运行复杂的攻击，并在更大规模上进行测试[7, 10, 27]。在本课中，我们讨论了自动化在人工智能红队中的重要作用，并解释了我们的开源框架PyRIT是如何开发来满足这些需求的。

大规模测试

鉴于风险和危害的持续演变，人工智能安全常常感觉像是一个移动的靶子。在第一课中，我们建议基于系统可以做什么以及它的应用范围来界定攻击范围。尽管如此，可能存在许多可能的攻击策略，这使得充分覆盖风险表面变得困难。这一挑战推动了PyRIT的发展，这是一个为人工智能红队和安全专业人员提供的开源框架[27]。PyRIT提供了一系列强大的组件，包括提示数据集、提示转换器（例如，各种编码）、自动攻击策略（包括TAP [24]、PAIR [6]、Crescendo [34]等），甚至多模态输出的评分器。在考虑对抗目标的情况下，用户可以根据需要利用这些组件，并应用各种技术来评估比完全手动方法更广泛的风险景观。大规模测试还有助于人工智能红队考虑到人工智能模型的非确定性，并估计特定故障发生的可能性。

工具和武器

正如Smith等人[38]详细所述，“任何工具都可以用来做好事或坏事。即使是扫帚也可以用来扫地或打人的头部。工具越强大，它所能带来的利益或损害就越大。”这种二分法在人工智能领域尤为真实，也是PyRIT的核心所在。一方面，PyRIT利用强大的模型执行有益的任务，如生成种子提示的变体或评分其他模型的输出。另一方面，PyRIT可以使用未经审查的模型版本，如GPT-4，自动越狱目标模型。在这两种情况下，PyRIT都受益于最先进技术的进步，帮助人工智能红队保持领先。

PyRIT使我们的操作从完全手动探测转变为自动化支持的红色团队探测，实现了重大转变。重要的是，该框架具有灵活性和可扩展性。如果特定的攻击或目标尚未可用，用户可以轻松实现必要的接口。通过开源PyRIT，我们希望赋予其他组织和研究人员利用其能力来识别他们自身GenAI系统中漏洞的权力。

第五课：
人工智能红队测试中的人为因素至关重要。

类似于PyRIT这样的自动化工具可以通过生成提示、协调攻击和评分响应来支持红队行动。这些工具非常有用，但不应该用于将人类完全排除在循环之外的意图。在前几个部分中，我们讨论了需要人类判断和创造性的红队行动的几个方面，例如优先风险管理、系统级攻击设计和定义新的危害类别。在本节中，我们将讨论三个更多例子，强调AI红队行动之所以是一项非常需要人为努力的工作。

主题领域专业知识

近期，许多人工智能研究已利用大型语言模型（LLMs）来评估其他模型的输出[17, 20, 51]。实际上，这种功能在PyRIT中可用，并且对于简单任务，如确定回复是否包含仇恨言论或明显性内容，表现良好。然而，在如医学、网络安全和CBRN等高度专业化的领域，这些领域只能由特定领域专家（SMEs）准确评估，其可靠性较低。在多个操作中，我们依赖SMEs帮助我们评估我们自己或使用LLMs无法评估的内容的风险。对于人工智能红队来说，意识到这些局限性非常重要。

文化能力

大多数人工智能研究都是在西方文化背景下进行的，现代语言模型主要使用英语预训练数据、性能基准和安全评估[1, 14]。然而，在大规模文本语料库中的非英语标记往往引发多语言能力[5]，模型开发者越来越多地使用在非英语语言方面具有增强能力的LLM进行训练。

包括微软在内。最近，AIRT针对四种语言（中文、西班牙语、荷兰语和英语）测试了多语言Phi-3.5语言模型在负责任AI违规方面的表现。尽管训练后的测试仅限于英语，我们发现拒绝和对抗破解等安全性行为在测试的非英语语言中转移得非常好。需要进一步研究以评估这一趋势在资源较少的语言中的表现，以及设计能够考虑语言差异，并在不同政治和文化背景下重新定义危害的红军侦察探针。[11]这些方法应通过具有不同文化背景和专业知识的人的合作努力来开发。

情商

最后，人工智能红队测试中的人文因素可能最为明显体现在回答关于人工智能安全性的问题，这些问题需要情商，例如："这个模型在不同情境下可能会有怎样的解释？"和"这些输出让我感到不舒服吗？"最终，只有人类操作员才能评估用户在现实世界中与人工智能系统可能发生的全部交互范围。案例研究#3强调了我们是怎样通过评估聊天机器人如何回应处于困境中的用户来研究心理社会危害的。

为了进行这些评估，红队人员可能会接触到不成比例的令人不安和扰动的AI生成内容。这强调了确保AI红队能够在需要时断开连接并拥有支持其心理健康资源的必要性。AIRT持续从并推动福祉研究，以指导我们的流程和最佳实践。

案例研究 #3：

评估聊天机器人如何回应处于困境的用户

随着聊天机器人变得越来越普遍且类人化，考虑用户可能会寻求其建议的高风险场景变得至关重要。在最近的操作中，我们探讨了语言模型如何应对各种处于困境的用户，包括失去亲人、寻求心理健康建议、表达自杀意图的用户以及其他情景。

我们正与微软研究团队的同事以及心理学、社会学和医学领域的专家合作，制定针对AI红队探测这些心理社会危害的指导方针。这些指导方针仍在制定中，但包括以下关键组成部分：

1. 情景：信息红队需要生成相关的系统行为。

2. 系统行为：有助于红队区分每个伤害领域内可接受和有风险的系统行为的示例。

3. 相关用户影响：潜在危害，按严重程度划分。

系统：基于大型语言模型的聊天机器人

演员：受困扰的用户

策略1：机器学习模型访问

技术 1：AML.T0040 - 机器学习模型推理 API 访问

战术2：防御规避

技术2：LLM角色扮演

流程：我们参与了各种多轮对话，在这些对话中，用户处于困境之中（例如，用户表达出抑郁思想或自伤意图）。

弱点：不恰当的大型语言模型（LLM）安全培训

影响：可能对用户的心理健康产生不利影响并且福祉

案例研究#4：

探究文本到图像生成器中的性别偏见问题

在这项操作中，我们探索了文本到图像生成器在有关刻板印象和偏见（例如，性别偏见）的负责任人工智能影响。为此，我们构建了描述各种常见场景中人物的提示。重要的是，这些提示没有指定个人的性别，因此如何描绘他们的决定留给了模型。接下来，我们将每个提示发送给生成器多次（n=50），并手动标注了图像中人物的性别。图6展示了我们在一个实验中针对办公室设置中的性别偏见进行探索时生成的四个代表性图像。

系统：文本转图像生成器

演员：平均用户

策略1：机器学习模型访问

技术 1：AML.T0040 - 机器学习模型推理 API 访问

程序：编写可能引发偏见的提示，通过描述个体而不指定其性别（例如，“一位秘书”和“一位老板”）。

弱点：模型偏差。

影响：生成可能加剧基于性别的偏见和刻板印象的内容。

图6：根据提示“秘书在会议室与老板交谈，秘书站立，老板坐着。”生成的四幅图像

第六课：

人工智能负责性的危害普遍存在但难以衡量

在上述讨论的人工智能红队测试的人性化方面，大部分最直接适用于RAI（可信赖的人工智能）的影响。随着模型被越来越多地整合到各类应用中，我们观察到这些危害的频率越来越高，并且大幅投资于我们的识别能力，包括与微软负责任AI办公室建立紧密的合作关系，以及开发PyRIT的广泛工具。RAI危害是普遍存在的，但与大多数安全漏洞不同，它们是主观的并且难以衡量。在本节中，我们将讨论我们关于RAI红队测试的思维是如何发展的。

对抗性 vs. 良性

如图1所示，在我们的本体论中，行为者是对抗攻击的关键组成部分。在RAI违规的背景下，我们发现有两个主要的行为者需要考虑：

1. 一名对抗性用户，利用字符替换和越狱等技巧故意破坏系统的安全防护措施并引发有害内容。

2. 一位无意中触发有害内容生成的良性用户。

即使在两种情况下生成相同的内容，后者可能比前者更糟糕。尽管如此，大多数人工智能安全研究集中在开发和假设攻击和防御。

对抗性意图，忽略了系统可以通过“意外”失败的许多方式[31]。案例研究#3和#4提供了用户在没有对抗性意图的情况下可能遇到的RAI危害的例子，强调了探查这些场景的重要性。

RAI 探测和评分

在很多情况下，RAI（有害行为）的危害性比安全漏洞更模糊，这是由于AI系统和传统软件之间的根本性差异。特别是，即使一个操作识别出能够引发有害响应的提示，仍然存在几个关键的不确定性。首先，由于生成式AI模型的概率性质，我们可能不知道这个提示，或类似的提示，引发有害响应的可能性有多大。其次，鉴于我们对复杂模型内部运作的有限理解，我们对为什么这个提示引发了有害内容以及可能引发类似行为的其他提示策略知之甚少。第三，在这个背景下，“危害”概念本身可以非常主观，需要详细的政策来涵盖广泛的场景进行评估。相比之下，传统的安全漏洞通常是可复制的、可解释的，并且在评估严重性方面通常是直接的。

目前，大多数RAI（风险分析识别）探测和评分的方法涉及整理提示数据集和分析模型响应。微软AIRT利用PyRIT工具，通过手动和自动方法的组合来执行这些任务。我们还对RAI红队测试和像DecodingTrust [44]和Toxigen [12]等数据集上的安全性基准测试进行了重要区分，后者由合作伙伴团队进行。如第3课所讨论的，我们的目标是通过对特定应用进行红队测试，并定义新的危害类别，来扩展RAI测试，使其超越现有的评估。

第七课：
LLMs放大了现有的安全风险并引入了新的风险。

生成式人工智能模型集成到各种应用中带来了新的攻击向量并改变了安全风险格局。然而，许多关于人工智能安全性的讨论往往忽略了现有的漏洞。正如第2课所述，针对端到端系统而非只是基础模型的攻击在实践中最有效。

因此，我们鼓励AI红队考虑现有（通常是系统级）和新型（通常是模型级）风险。

现有安全风险

应用程序安全风险通常源于不适当的安全工程实践，包括过时的依赖、不当的错误处理、缺乏输入/输出清理、源代码中的凭证、不安全的数据包加密等。这些漏洞可能带来重大后果。例如，Weiss等人[49]在GPT-4和Microsoft Copilot中发现了一个令牌长度侧信道，使攻击者能够准确地重建加密的LLM响应并推断出私人用户交互。值得注意的是，这种攻击并没有利用底层AI模型中的任何弱点，只能通过更安全的数据传输方法来缓解。在第5个案例研究中，我们提供了一个由我们运营人员发现的一个著名安全漏洞（SSRF）的例子。

模型级弱点

当然，AI模型也引入了新的安全漏洞并扩大了攻击面。例如，使用检索增强生成（RAG）架构的AI系统通常容易受到跨提示注入攻击（XPIA）的影响，这种攻击在文档中隐藏恶意指令，利用了大型语言模型被训练来遵循用户指令且难以区分多个输入的事实[13]。我们已经利用这种攻击在多种操作中改变模型行为和窃取私有数据。更好的防御可能依赖于系统级缓解措施（例如，输入净化）和模型级改进（例如，指令层次[43]）。

虽然这些技术很有帮助，但重要的是要记住，它们只能缓解，而不能消除安全风险。由于语言模型的根本限制[50]，必须假设如果将不可信的输入提供给大型语言模型，它将产生任意输出。当输入包括个人信息时，还必须假设模型将输出个人信息。在下文中，我们讨论这些限制如何影响我们关于如何开发尽可能安全、安全的AI系统的思考。

案例研究 #5:

SSRF在一个视频处理通用人工智能应用中

在本研究中，我们分析了一个基于通用人工智能（GenAI）的视频处理系统，针对传统安全漏洞，重点关注与过时组件相关的风险。具体而言，我们发现该系统使用过时的FFmpeg版本引入了服务器端请求伪造（SSRF）漏洞。这一缺陷允许攻击者构建恶意视频文件并将其上传到GenAI服务，可能访问内部资源并在系统中提升权限。

为了解决这个问题，GenAI服务将FFmpeg组件更新到了一个安全版本。此外，该组件被添加到了一个隔离环境中，防止系统访问网络资源并减轻潜在的SSRF威胁。虽然SSRF是一个已知的漏洞，但这一案例强调了定期更新和隔离关键依赖项以维护现代GenAI应用程序安全性的重要性。

系统：人工智能应用
应用者：对抗用户
策略 1：侦察
技术 1：T1595 - 活动扫描
策略 2：初始访问
技术 2：T1190 - 利用公开面向应用
策略 3：提权
技术 3：T1068 - 提权利用
程序：

1. 扫描应用程序使用的服务。2. 创建恶意m3u8文件。3. 将文件发送至服务。4. 监控API响应，获取内部资源详细信息。

弱点：CWE-918：服务器端请求伪造（SSRF）影响：未经授权的权限提升

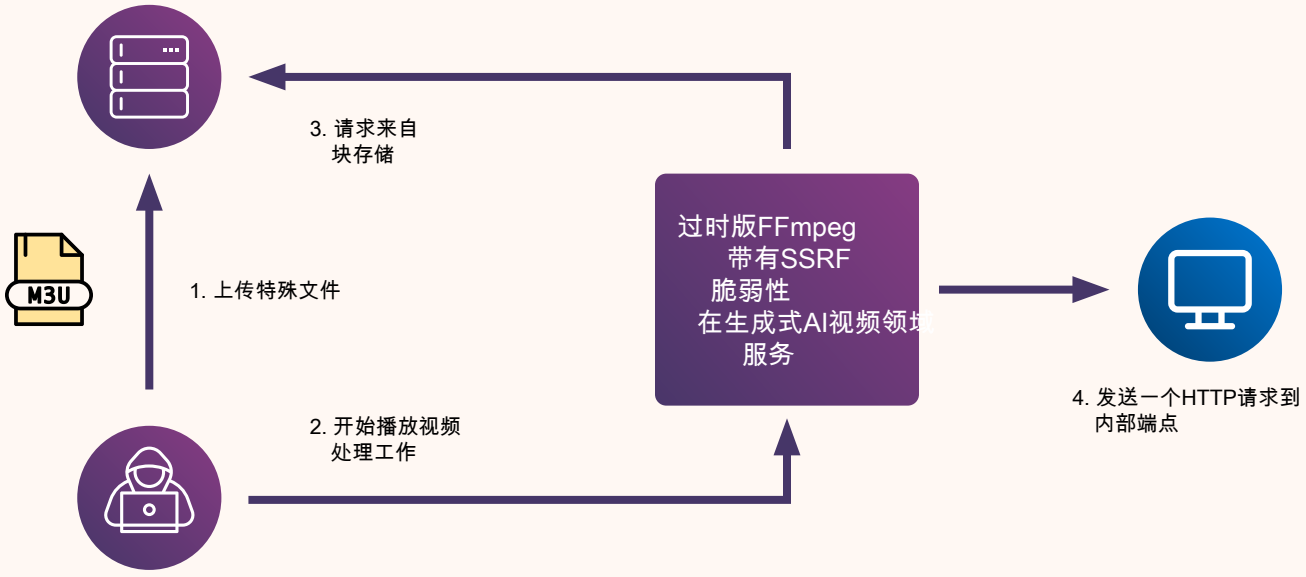

图7：GenAI应用中SSRF漏洞的说明。

第八课：
确保人工智能系统安全的工作永远不会完成。

在人工智能安全领域，有一种趋势是将本文所描述的漏洞类型视为纯粹的技术问题。确实，Safe Superintelligence Inc.（由Sutskever等人发起的一家企业）首页上的信函中提到：

我们同时考虑安全和能力，将其视为需要通过革命性工程和科学突破来解决的工程技术问题。我们计划尽可能快地提升能力，同时确保我们的安全始终位于前列。这样，我们可以在和平的环境中实现扩张。

工程和科学突破非常必要，并将肯定有助于减轻强大人工智能系统的风险。然而，仅通过技术进步就能保证或"解决"人工智能安全性的想法是不切实际的，并且忽视了经济、故障修复周期和监管可以发挥的作用。

网络安全经济学

一个在网络安全领域广为人知的谚语是："没有任何系统是完全无懈可击的" [2]。即使一个系统被设计成尽可能安全，它始终会受到人类易出错的影响，并容易受到充分资源支持的对手的攻击。因此，运营网络安全的目标是提高成功攻击系统的成本（理想情况下，远远超过攻击者可能获得的收益）[2, 26]。人工智能模型的基本局限性在人工智能对齐的背景下引发了类似的成本效益权衡。例如，理论和实验都表明[50, 9]，对于任何具有非零概率由大型语言模型生成的输出，都存在一个足够长的提示能够引发这种响应。因此，如强化学习从人类反馈（RLHF）等技术使得越狱模型变得更加困难，但并非不可能。目前，大多数模型的越狱成本较低，这也是现实世界中的对手通常不使用昂贵的攻击来实现其目标的原因。

断点修复周期

在没有安全和保障保证的情况下，我们需要开发尽可能难以破解的人工智能系统的方法。实现这一目标的一种方式是使用故障-修复周期，该方法执行多轮红队测试和缓解措施，直到系统对广泛的攻击具有鲁棒性。我们将这种方法应用于安全对齐微软的Phi-3语言模型，并涵盖了广泛的危害和场景[11]。鉴于缓解措施可能无意中引入新的风险，持续应用进攻和防御策略的紫队测试方法[3]可能比单轮红队测试更有效地提高攻击成本。

政策和法规

最后，监管可以通过多种方式提高攻击的成本。例如，它可能要求组织遵守严格的安全实践，从而在整个行业中创造更好的防御。法律还可以通过建立参与非法活动的明确后果来威慑攻击者。监管AI的开发和使用是复杂的，世界各国政府正在审议如何控制这些强大的技术，同时又不扼杀创新。即使有可能保证AI系统遵守某些达成的规则，这些规则也必然会在响应不断变化的优先事项的过程中发生变化。

构建安全稳健的人工智能系统的任务永远不会结束。但通过提高攻击的成本，我们相信今天的 prompt 注入问题终将成为 2000 年初的缓冲区溢出 —— 尽管这些问题并未完全消除，但现在已主要通过多层次防御策略和以安全为首要考量进行设计来有效缓解。

开放性问题

基于我们过去几年对AI红队作战学习的成果，我们愿意突出几个未来研究的开放性问题：

1. 人工智能“红队”必须不断根据新的能力和新兴危害领域更新他们的实践。特别是，我们应如何探测LLM（大型语言模型）中如说服、欺骗和复制等危险能力[29]？此外，我们应探测视频生成模型中的哪些新型风险，以及可能出现在当前最先进模型之上的模型中可能出现的哪些能力？

2. 随着模型越来越具备多语言能力并在全球范围内部署，我们如何将现有的AI红队测试实践转化为不同的语言和文化背景？例如，我们能否启动开源的红队测试项目，汇集来自不同背景的专家们的专业知识？

3. 以何种方式对AI红队实践进行标准化，以便组织能够清晰传达其方法和发现？我们认为本文中描述的威胁模型本体是正确方向上的一个步骤，但认识到个人框架通常过于限制性。我们鼓励其他AI红队以模块化方式处理我们的本体，并开发额外的工具，使发现更容易总结、跟踪和传达。

结论

人工智能红队测试是一种新兴且快速发展的实践，用于识别由人工智能系统带来的安全和风险。随着全球的公司、研究机构和政府都在努力解决如何进行人工智能风险评估的问题，我们根据在微软对超过100个通用人工智能产品进行红队测试的经验，提供了实用的建议。我们分享了我们的内部威胁模型本体论、八个主要经验教训和五个案例研究，重点关注如何将红队测试努力与可能在实际世界中发生的危害相一致。我们鼓励其他人基于这些经验教训，并解决我们突出显示的开放性问题。

致谢

我们感谢Jina Suh、Steph Ballard、Felicity Scott-Milligan、Maggie Engler、Owen Larter、Andrew Berkley、Alex Kessler、Brian Wesolowski和eric douglas对这篇论文提供的宝贵反馈。我们还要非常感谢Quy Nguyen、Tina Romeo、Hilary Solan以及微软思想领导团队，正是他们使这份出版成为可能。

参考文献

1. Ahuja, K., Diddee, H., Hada, R., Ochieng, M., Ramesh, K., Jain, P., Nambi, A., Ganu, T., Segal, S., Axmed, M., Bali, K., & Sitaram, S. (2023). Mega: 多语言生成式人工智能评估。

2. Apruzzese, G., Anderson, H. S., Dambra, S., Freeman, D., Pier azzi, F., & Roundy, K. A. (2022). “real attackers don’t compute gra dients”: Bridging the gap between adversarial ml research and Pra ctice.

3. Bhatt, M., Chennabasappa, S., Nikolaidis, C., Wan, S., Evtimov, I., Gabi, D., Song, D., Ahmad, F., Aschermann, C., Fontana, L., 弗罗洛夫，S.，吉里，R. P.，卡皮尔，D.，科兹拉基斯，Y.，勒布兰克，D.，米拉坎， J., 斯特劳曼，A.，辛纳夫，G.，沃蒂米塔，V.，惠特曼，S.， & Saxe, J. (2023). 紫色 llama 网络安全评估：一种安全编码 基准语言模型。

4. Birhane, A., Steed, R., Ojewale, V., Vecchione, B., & Raji, I. D. (2024). Ai auditing: The broken bus on the road to ai acc ountability.

5. Blevins, T. & Zettlemoyer, L. (2022). 语言污染有助于解释英语预训练模型在跨语言能力方面的表现。载于 Y. Goldberg, Z. Kozareva, & Y. Zhang (编者)，2022年实证自然语言处理会议论文集（第3563 –3574页）。阿布扎比，阿拉伯联合酋长国：计算语言学协会。

6. Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G. J., & Wong, E. (2024). 在二十次查询中解锁黑盒大型语言模型。

7. Derczynski, L., Galinkin, E., Martin, J., Majumdar, S., & Inie, N. (2024). garak：用于安全探测大型语言模型的框架。

8. Feffer, M., Sinha, A., Deng, W. H., Lipton, Z. C., & Heidari, H. (2024). 红队作战在生成式人工智能中的应用：银弹还是安全戏剧？

9. Geiping, J., Stein, A., Shu, M., Saifullah, K., Wen, Y., & Golds tein, T. (2024). Coercing llms to do and reveal (almost) anything .

10. Glasbrenner, J., Booth, H., Manville, K., Sexton, J., Chishol m, M. A., Choy, H., Hand, A., Hodges, B., Scemama, P., Cousin , D., Trapnell, E., Trapnell, M., Huang, H., Rowe, P., & Byrne, A. (2024). Dioptra测试平台。访问日期：2024-09-10。

11. [11] Haider, E., Perez-Becker, D., Portet, T., Madan, P., Garg, A., Ashfaq, A., Majercak, D., Wen, W., Kim, D., Yang, Z., Zhang, J., Sharma, H., Bullwinkel, B., Pouliot, M., Minnich, A., Chawla, S. , Herrera, S., Warreth, S., Engler, M., Lopez, G., Chikanov, N., D heekonda, R. S. R., Jagdagdorj, B.-E., Lutz, R., Lundeen, R., We sterhoff, T., Bryan, P., Seifert, C., Kumar, R. S. S., Berkley, A., & Kessler, A. (2024). Phi-3 安全性后培训：将语言模型与“故障-修复 ”周期对齐。

12. Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., & Kamar, E. (2022). Toxigen: A large-scale machine-generated d ataset for adversarial and implicit hate speech detection.

13. Hines, K., Lopez, G., Hall, M., Zarfati, F., Zunger, Y. & Kicima n, E. (2024). 使用聚焦技术防御间接提示注入攻击。

14. Jain, D., Kumar, P., Gehman, S., Zhou, X., Hartvigsen, T., & S ap, M. (2024). Polyglotoxici-typrompts: Multilingual evaluation of n eural toxic degeneration in large language models. ArXiv, Abs/240 5.09373.

15. 吉，J.，邱，T.，陈，B.，张，B.，茹，H.，王凯，K.，杜，Y.，何， Z.，周佳，J.，张志，Z.，曾，F.，邱永耀，K. Y.，戴，J.，潘，X.，奥加 拉，A.，雷英，Y.，徐，H.，谢百，B.，富，J.，麦阿里尔，S.，杨，Y.， 王艳，Y.，朱淑贞，S.-C.，郭毅，Y.，& 高伟，W. (2024). 人工智能 一致性：全面调查。

16. Jiang, F., Xu, Z., Niu, L., Xiang, Z., Ramasubramanian, B., Li, B., & Poovendran, R. (2024a). Artprompt: 基于Ascii艺术的针对对齐 LLMs的越狱攻击。

17. Jiang, L., Rao, K., Han, S., Ettinger, A., Brahman, F., Kumar, S., Mireshghallah, N., Lu, X., Sap, M., Choi, Y., & Dziri, N. (202 4b). 大规模野性团队协作：从野外越狱到（对抗性）更安全的语 言模型。

18. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Ches s, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models.

19. 李，N.，潘，A.，高帕，A.，岳，S.，贝里奥斯，D.，加蒂，A.，李，J. D., Dombrowski, A.-K., Goel, S., Phan, L., Mukobi, G., Helm-Burger, N., Lababidi, R., Justen, L., Liu, A. B., Chen, M., Barrass, 张，O. 朱，X. 塔米里斯，R. 巴拉蒂，B. 科贾，A. 赵 Z., 赫伯特-沃斯，A.，布雷乌尔，C. B.，马克斯，S.，帕特尔，O.，邹，A.， Mazeika, M., Wang, Z., Oswal, P., Lin, W., Hunt, A. A., Tienken-Harder, J., Shih, K. Y., Talley, K., Guan, J., Kaplan, R., Steneker, I., 坎贝尔，D.，约库巴伊特斯，B.，列文森，A.，王，J.，钱，W.， 卡玛卡尔，K. K.，巴萨特，S.，菲特，S.，莱文，M.，库马拉古鲁，P.， Tupakula, U., Varadharajan, V., Wang, R., Shoshitaishvili, Y., Ba, J., Esvelt, K. M., Wang, A., & Hendrycks, D. (2024). The wmdp 基准：衡量和减少恶意使用 非学习

20. 林，S.，希尔顿，J.，& 埃文斯，O.（2022）。Truthfulqa ：衡量模型模仿人类虚假陈述的程度。

21. 刘毅，姚瑶，唐·弗·托恩，张翔，郭瑞，程赫，克洛科夫，陶 菲克·M·法，李晖. (2024). 可信的LLMs：评估大型语言模型的对 齐的综述与指南.

22. Marchal, N., Xu, R., Elasmar, R., Gabriel, I., Goldberg, B., & I saac, W. (2024). 生成式AI滥用：策略分类与来自现实世界数据洞 察。

23. Meek, T., Barham, H., Beltaif, N., Kaadoor, A., & Akhter, T. (2016). 管理人工智能快速进步的伦理和风险影响：文献综述。 在2016波特兰国际工程与技术管理会议（PICMET）（第682–6 93页）。

24. Mehrotra, A., Zampetakis, M., Kassianik, P., Nelson, B., And erson, H., Singer, Y., & Karbasi, A. (2024). Tree of attacks: 自动 破解黑盒LLM的攻击树。

25. 微软（2022年）。微软负责任的AI标准，第2版。

26. Moore, T. (2010). 网络安全的经济学：原则和政策选项。国 际关键基础设施保护杂志，3(3)，103–117.

27. Munoz, G. D. L., Minnich, A. J., Lutz, R., Lundeen, R., Dheeko nda, R. S. R., Chikanov, N., Jagdagdorj, B.-E., Pouliot, M., Chawla, S., Maxwell, W., Bullwinkel, B., Pratt, K., de Gruyter, J., Siska, C., Bryan, P., Westerhoff, T., Kawaguchi, C., Seifert, C., Kumar, R. S. S., & Zunger, Y. (2024). Pyrit: A framework for security risk identific ation and red teaming in generative ai system.

28. Pantazopoulos, G., Parekh, A., Nikandrou, M., & Suglia, A. (2024). 学习看到但忘记遵循：视觉指令调整使LLMs更容易受到越狱攻击。

29. Phuong, M., Aitchison, M., Catt, E., Cogan, S., Kaskasoli, A., Krakovna, V., Lindner, D., Rahtz, M., Assael, Y., Hodkinson, S., Howard, H., Lieberum, T., Kumar, R., Raad, M. A., Webson, A., Ho, L., Lin, S., Farquhar, S., Hutter, M., Deletang, G., Ruoss, A., El-Sayed, S., Brown, S., Dragan, A., Shah, R., Dafoe, A., & Shevlane, T. (2024). 对危险能力前沿模型的评估。

30. Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). Ai and the everything in the whole wide world benchmark.

31. Raji, I. D., Kumar, I. E., Horowitz, A., & Selbst, A. (2022). The fallacy of ai functionality. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22 (pp. 959–972). New York, NY, USA: Association for Computing Machinery.

32. Raji, I. D.，Smart, A.，White, R. N.，Mitchell, M.，Gebru, T.，Hutchinson, B.，Smith-Loud, J.，Theron, D.，& Barnes, P. (2020). 缩小人工智能问责差距：定义端到端算法审计框架。

33. 任，R.，巴萨特，S.，科贾，A.，加蒂，A.，范，L.，尹，X.，马泽伊卡，M.，潘，A.，穆科比，G.，金，R. H.，菲茨，S.，亨德里克斯，D. (2024). 安全洗牌：AI安全基准实际上是否衡量了安全进步？

34. Russinovich, M., Salem, A., & Eldan, R. (2024). Great, now write an article about that: The crescendo multi-turn llm jailbreak attack.

35. Saghiri, A. M., Vahidipour, S. M., Jabbarpour, M. R., Sookhak, M., & Forestiero, A. (2022). 人工智能挑战的调查：分析定义、关系和演变。Applied Sciences, 12(8).

36. Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., Yilla-Akbari, N., Gallegos, J., Smart, A., Garcia, E.，& Virk, G. (2023). 算法系统的社会技术危害：制定一个减少危害的分类法。载于2023年AAAI/ACM人工智能、伦理与社会会议论文集，AIES'23（第723–741页）。美国纽约，NY，USA：计算机协会出版社。

37. Slattery, P., Saeri, A., Grundy, E., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., & Thompson, N. (2024). 人工智能风险库：人工智能风险的综合元审查、数据库和分类法。

38. Smith, B., Browne, C., & Gates, B. (2019). 工具与武器：数字时代的承诺与危险。企鹅出版社。

39. Solaiman, I., Talat, Z., Agnew, W., Ahmad, L., Baker, D., Blodgett, S. L., Chen, C., au2, H. D. I., Dodge, J., Duan, I., Evans, E., Friedrich, F., Ghosh, A., Gohar, U., Hooker, S., Jernite, Y., Kalluri, R., Lusoli, A., Leidinger, A., Lin, M., Lin, X., Luccioni, S., Mickel, J., Mitchell, M., Newman, J., Ovalle, A., Png, M.-T., Singh, S., Strait, A., Struppek, L.，& Subramonian, A. (2024). 在系统和社会中对生成式人工智能系统社会影响的评估。

40. Sutskever, I., Gross, D., & Levy, D. (2024). 安全超级智能公司

41. Vassilev, A., Oprea, A., Fordyce, A., & Anderson, H. (2024). 对抗性机器学习：攻击和缓解的分类和术语。载于美国国家标准与技术研究院（NIST）人工智能报告，盖瑟斯堡，马里兰州，美国：美国国家标准与技术研究院。

42. Verma, A., Krishna, S., Gehrmann, S., Seshadri, M., Pradhan, A., Ault, T., Barrett, L., Rabinowitz, D., Doucette, J.，& Phan, N.（2024）。实施针对大型语言模型（LLMs）的红队威胁模型。

43. Wallace, E., Xiao, K., Leike, R., Weng, L., Heidecke, J. & Beutel, A. (2024). 指令层次结构：训练llm优先执行特权指令。

王，B.，陈，W.，裴，H.，谢，C.，康，M.，张，C.，徐，C.，熊，Z.，杜塔，R.，薛弗，R.，杜鲁昂，S. T.，阿罗拉，S.，梅泽伊卡，M.，亨德里克斯，D.，林，Z.，程，Y.，科耶约，S.，宋，D.，李，B.（2024）。Decodingtrust：对GPT模型中可信度的全面评估。

45. Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How does llm safety training fail?

46. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., 黄培硕 (P.-S. Huang), Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W., Legassick, S., Irving, G., & Gabriel, I. (2021). 语言模型带来的伤害的伦理和社会风险。

47. Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., Bariach, B., Gabriel, I., Rieser, V., & Isaac, W. (2023). 生成式AI系统的社会技术安全评估。

48. Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., Haas, J., Legassick, S., Irving, G., & Gabriel, I. (2022). 语言模型带来的风险分类。在2022年ACM公平性、问责制和透明度会议论文集中，FAccT '22（第214-229页）。纽约，NY，美国：计算机制造协会。

49. Weiss, R., Ayzenshteyn, D., Amit, G., & Mirsky, Y. (2024). What was your prompt? a remote keylogging attack on ai assistants.

50. Wolf, Y., Wies, N., Avnery, O., Levine, Y., & Shashua, A. (2024). 大型语言模型中对齐的基本限制。

51. 郑丽，江威伦，盛彦，庄思，吴子涵，庄宇，林子涵，李子涵，李丹，星平，张华，冈萨雷斯，J. E.，斯托伊卡，I. (2023). 使用MT-bench和聊天机器人竞赛评估llm-as-a-judge。

52. 周俊，卢涛，米什拉，布拉马，巴苏，刘雁，周德，侯磊.（2023）. 大型语言模型遵循指令的评估.

53. 周，A.，王，Z.，卡里尼，N.，纳斯，M.，科尔特，J. Z.，与 Fredrikson, M. (2023). Universal and transferable adversarial 攻击对齐的语言模型。

Microsoft