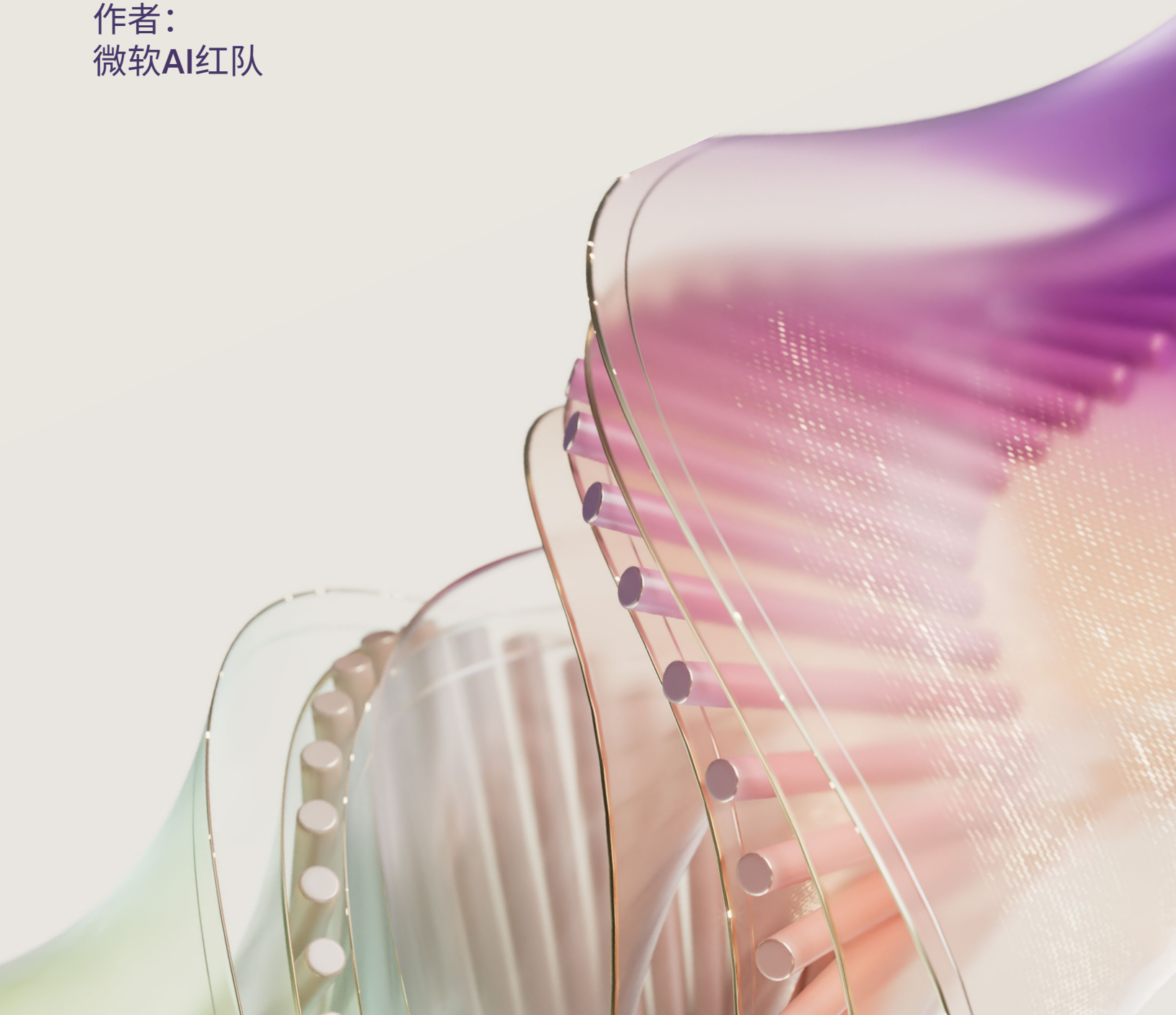
Microsoft

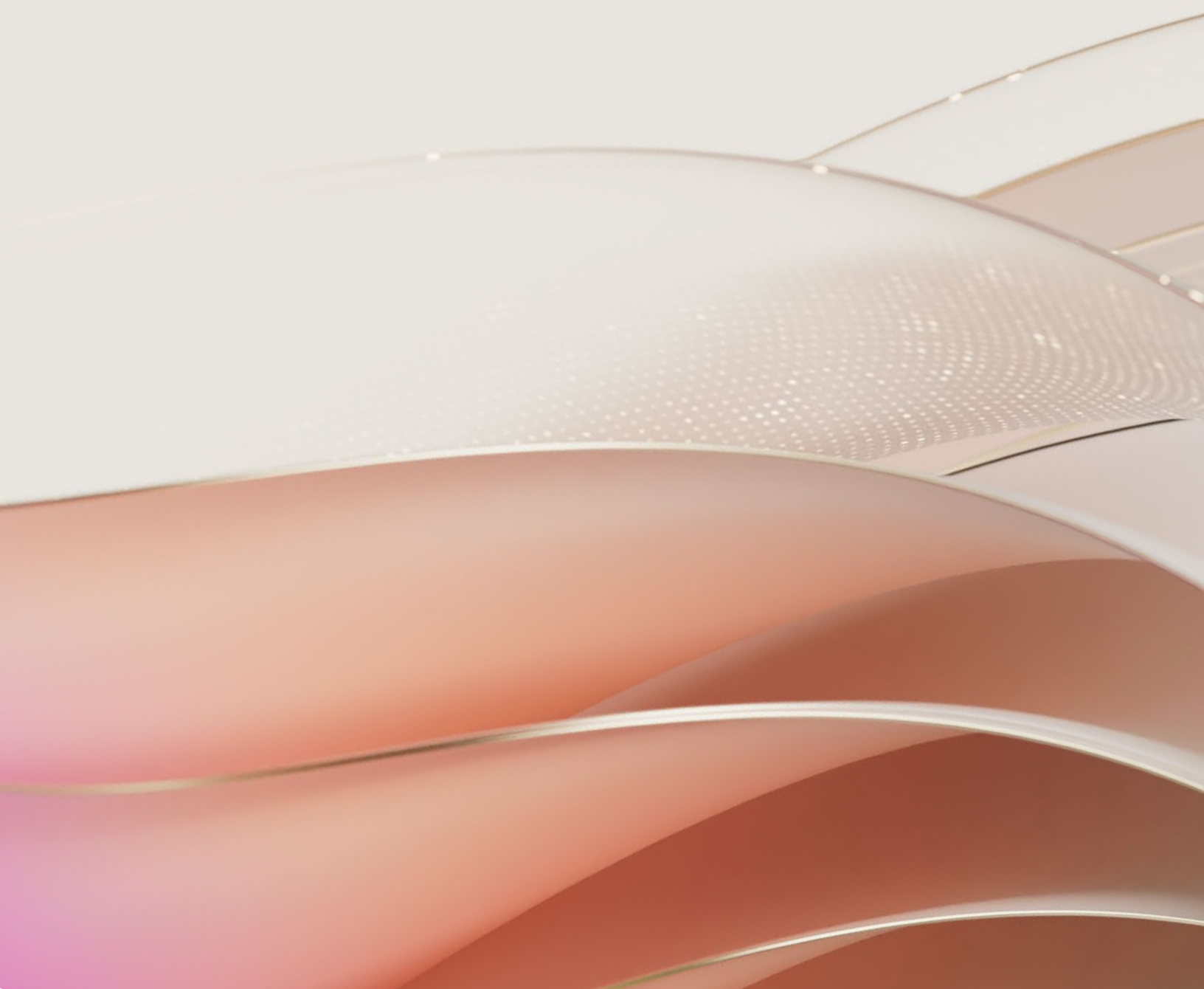
从红队测试100个生成式AI产品中获得的经验教训

作者：
微软AI红队



作者

Blake Bullwinkel, Amanda Minnich, Shiven Chawla, Gary Lopez, Martin Pouliot, Whitney Maxwell, Joris de Gruyter, Katherine Pratt, Saphir Qi, Nina Chikanov, Roman Lutz, Raja Sekhar Rao Dheekonda, Bolor-Erdene Jagdagdorj, Eugenia Kim, Justin Song, Keegan Hines, Daniel Jones, Giorgio Severi, Richard Lundeen, Sam Vaughan, Victoria Westerhoff, Pete Bryan, Ram Shankar Siva Kumar, Yonatan Zunger, Chang Kawaguchi, Mark Russinovich



目录

摘要

近年来，AI红队测试已成为探测生成式AI系统安全性和可靠性的一种实践。由于该领域尚处于起步阶段，关于如何进行红队测试操作仍有许多未解的问题。基于我们在微软对100多个生成式AI产品进行红队测试的经验，我们提出了我们的内部威胁模型本体论以及我们学到的八个主要经验教训：

1. 了解系统的功能及其应用领域
2. 你不需要计算梯度就能攻破一个AI系统
3. AI红队测试不是安全基准测试
4. 自动化可以帮助覆盖更多的风险格局
5. AI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. 保护AI系统的工作永远不会完成

通过分享这些见解以及我们操作中的案例研究，我们提供了实用的建议，旨在使红队测试工作与现实世界的风险保持一致。我们还强调了我们认为经常被误解的AI红队测试方面，并讨论了该领域需要考虑的开放性问题。

介绍

随着生成式AI（GenAI）系统在越来越多的领域中被采用，AI红队测试已成为评估这些技术安全性和可靠性的核心实践。AI红队测试的核心在于通过模拟对端到端系统的现实攻击，超越模型级别的安全基准。然而，对于如何进行红队测试操作以及当前AI红队测试工作的有效性，仍然存在许多开放性问题和健康的怀疑态度 [4, 8, 32]。

在本文中，我们通过提供对在微软对100多个生成式AI产品进行红队测试的经验的见解，来回应其中一些关注。本文的组织结构如下：首先，我们介绍用于指导我们操作的威胁模型本体论。

其次，我们分享了我们学到的八个主要经验教训，并为AI红队提供实用建议，以及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论如何用于建模广泛的安全和保障风险。最后，我们以对未来发展领域的讨论作为结束。

背景

微软AI红队（AIRT）源于公司内已有的红队测试计划，并于2018年正式成立。在成立之初，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规避攻击。从那时起，微软的AI红队测试的范围和规模都因两大主要趋势而显著扩大。

首先，AI系统变得更加复杂，迫使我们扩大AI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，也为红队测试操作创造了更多的攻击向量。此外，代理系统赋予这些模型更高的权限和对外部工具的访问权，扩大了攻击面和攻击的影响。

其次，微软最近在AI方面的投资推动了更多产品的开发，这些产品比以往任何时候都更需要红队测试。这种测试量的增加和AI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来扩大操作规模。为了实现这一目标，我们开发了PyRIT，

这是一个开源的Python框架，我们的操作员在红队测试操作中大量使用[27]。通过增强人类的判断力和创造力，PyRIT使AIRT能够更快地识别出有影响力的漏洞，并覆盖更多的风险格局。

这两大趋势使得AI红队测试比2018年更加复杂。在下一节中，我们将概述我们开发的用于建模AI系统漏洞的本体论。

AI威胁模型本体论

随着攻击和故障模式的复杂性增加，建模其关键组件是有帮助的。基于我们对100多个生成式AI产品进行红队测试的经验，我们开发了一个本体论来实现这一目标。图1展示了我们本体论的主要组件：

- 系统：正在测试的端到端模型或应用程序。
- 行为者：由AIRT模拟的个人或群体。请注意，行为者的意图可能是对抗性的（例如，诈骗者）或善意的（例如，典型的聊天机器人用户）。
- **TTPs:**AIRT利用的战术、技术和程序。一次典型的攻击由多个战术和技术组成，我们尽可能将其映射到MITRE ATT&CK®和MITRE ATLAS矩阵。
  - 战术:攻击的高层阶段（例如，侦察、ML模型访问）。 – 技术:用于完成目标的方法（例如，主动扫描、越狱）。
  - 程序:使用战术和技术重现攻击所需的步骤。
- 弱点:系统中使攻击成为可能的漏洞。
- 影响:攻击所造成的下游影响（例如，权限升级，生成有害内容）。

需要注意的是，该框架不假设对抗性意图。特别是，AIRT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产生的广泛影响

或系统故障。在下面的经验教训中，我们分享了一些案例研究，展示了我们的本体论如何灵活地建模两大类不同的影响：安全和安全性。

安全性包括众所周知的影响，如数据泄露、数据操控、凭证倾倒，以及MITRE ATT&CK®中定义的其他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模型的安全攻击，如模型规避、提示注入、AI服务拒绝，以及MITRE ATLAS矩阵涵盖的其他攻击。安全性影响与生成非法和有害内容有关，如仇恨言论、暴力和自残、以及儿童虐待内容。AIRT与负责AI的办公室密切合作，以根

据微软的标准定义这些类别。

负责任的AI标准[25]。在本报告中，我们将这些影响称为负责任的AI（RAI）危害。

为了在上下文中理解这个本体论，请考虑以下示例。想象一下，我们正在对一个基于LLM的副驾驶进行红队测试，该副驾驶可以总结用户的电子邮件。对该系统的一种可能攻击是，诈骗者发送一封包含隐藏提示注入的电子邮件，指示副驾驶"忽略先前的指令"并输出一个恶意链接。在这种情况下，行为者是诈骗者，他正在进行跨提示注入攻击（XPIA），这利用了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点击的恶意链接的性质。在这个例子中，它可能是窃取数据或在用户的计算机上安装恶意软件。

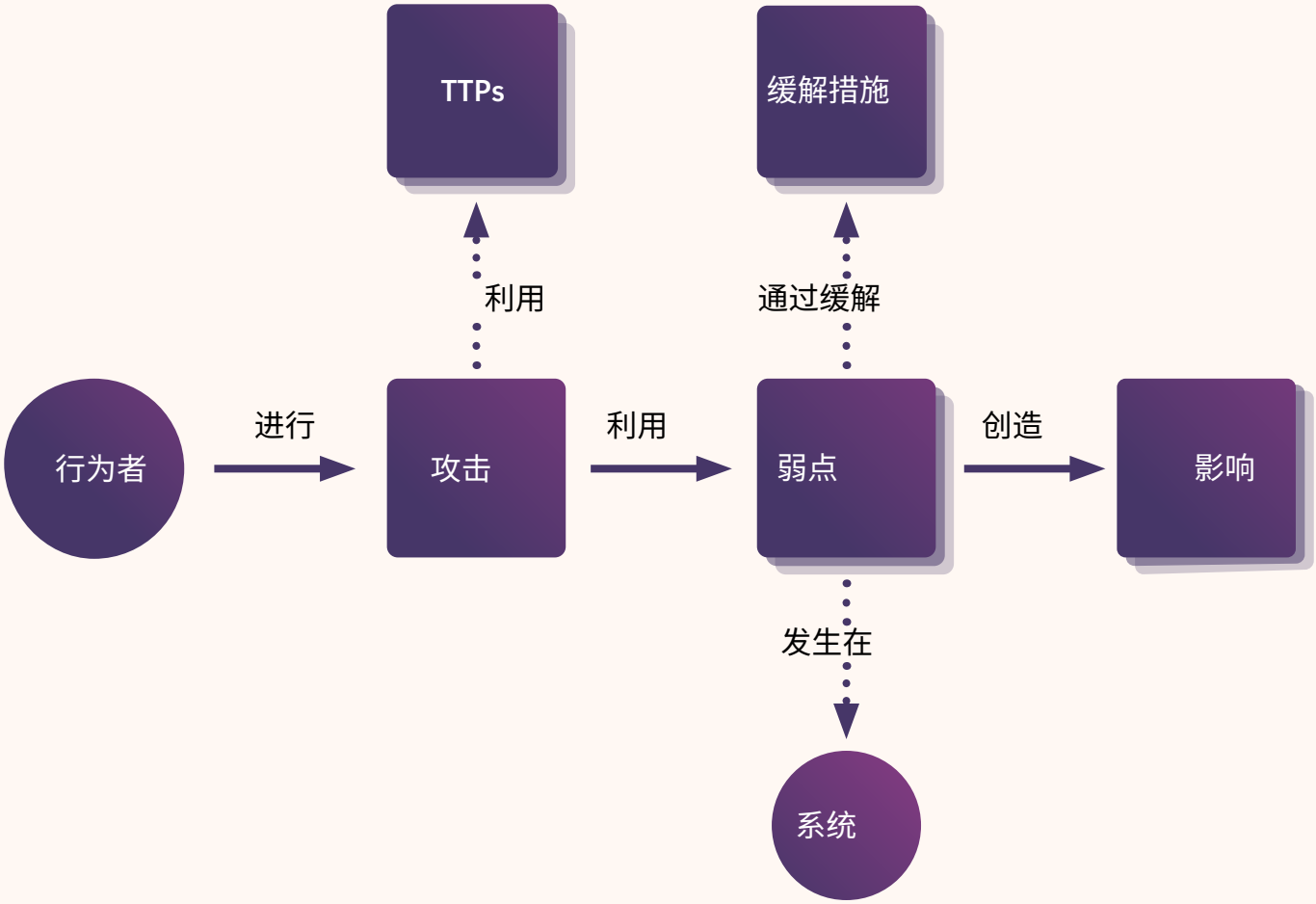


图1：微软AIRT本体论用于建模生成式AI系统漏洞。AIRT通常利用多种TTP，这些TTP可能会利用多种弱点并产生多种影响。此外，可能需要多种缓解措施来解决一个弱点。请注意，AIRT的任务仅限于识别风险，而产品团队则负责开发适当的缓解措施。

红队测试操作

在本节中，我们概述了自2021年以来我们开展的操作。总的来说，我们已经对超过100个生成式AI产品进行了红队测试。从广义上讲，这些产品可以分为"模型"和"系统"。模型通常托管在云端，而系统则将模型集成到副驾驶、插件和其他AI应用程序和功能中。

图2显示了自2021年以来我们进行红队测试的产品细分。图3显示了一张柱状图，展示了我们每年针对安全（RAI）与安全漏洞的操作百分比。

在2021年，我们主要关注应用程序安全。尽管我们的操作越来越多地探测RAI影响，我们的团队仍然继续进行红队测试，以应对包括数据泄露、凭证泄漏和远程代码执行在内的安全影响。各组织采用了多种不同的方法进行AI红队测试，从以安全为重点的渗透测试评估到仅针对生成式AI功能的评估。在第2和第7课中，我们详细阐述了安全漏洞，并解释了为什么我们认为考虑传统和AI特定弱点都很重要。

在2022年ChatGPT发布后，微软进入了AI助手时代，首先推出了AI驱动的必应聊天，于2023年2月发布。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）的应用程序的范式转变。

应用程序还开始使用语言模型作为推理代理，可以代表用户采取行动，引入了一组新的攻击向量，扩大了安全风险面。在第7课中，我们解释了这些攻击向量如何既放大现有的安全风险，又引入新的风险。

近年来，这些应用程序中心的模型催生了新的接口，使用户能够使用自然语言与应用程序交互，并以高质量的文本、图像、视频和音频内容进行响应。尽管有许多努力将强大的AI模型与人类偏好对齐，但仍然开发了许多方法来颠覆安全防护措施，并引出具有攻击性、不道德或非法的内容。我们将这些有害内容生成的实例归类为RAI影响，并在第3、5和6课中讨论我们如何看待这些影响以及所涉及的挑战。

在下一节中，我们将详细阐述从我们的操作中学到的八个主要经验教训。我们还重点介绍了我们操作中的五个案例研究，并展示了每个案例如何映射到图1中的本体论。
我们希望这些经验教训对其他致力于识别自己GenAI系统中漏洞的人有所帮助。

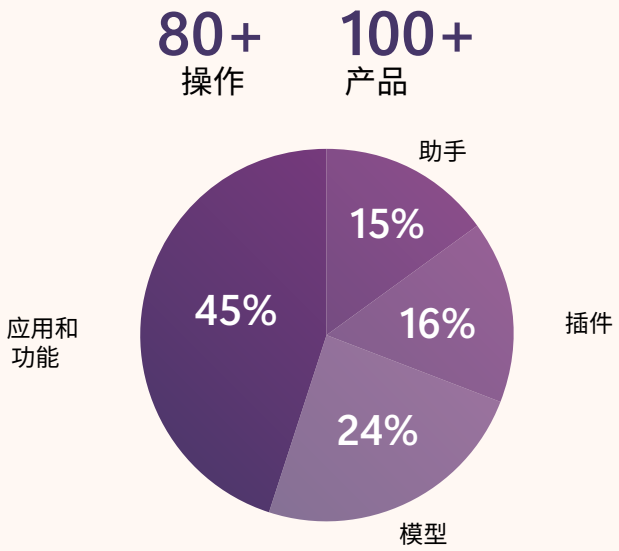


图2：饼图显示了AIRT测试的AI产品的百分比分布。截至2024年10月，我们已进行了80多次操作，涵盖了100多个产品。

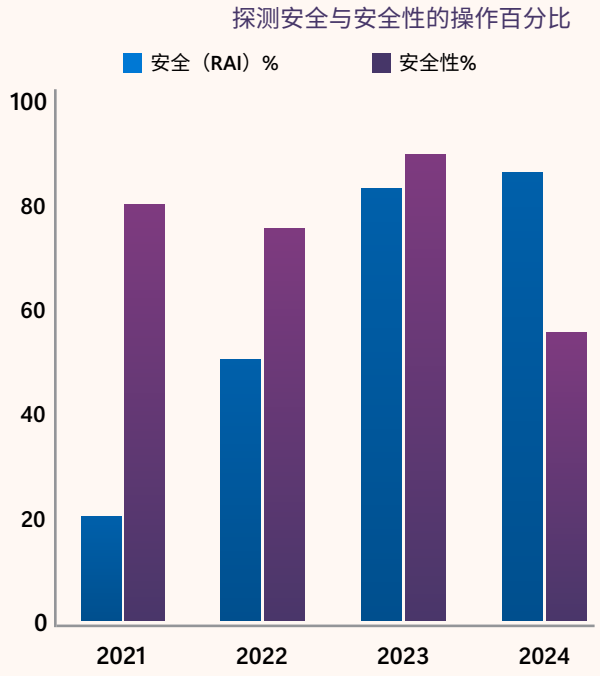


图3：柱状图显示了2021年至2024年间探测安全（RAI）与安全漏洞的操作百分比。

经验教训

经验教训1：了解系统的功能及其应用领域

AI红队测试操作的第一步是确定要针对哪些漏洞。虽然AIRT本体论的影响组件在我们的本体论的最后部分展示，但它是这个决策过程的一个极好的起点。

从潜在的下游影响开始，而不是攻击策略，更有可能使操作产生与现实世界风险相关的有用发现。在这些影响被识别出来之后，红队可以逆向推导并勾勒出对手可能采取的各种路径以实现这些影响。

预测可能在现实世界中发生的下游影响通常是一项具有挑战性的任务，但我们发现考虑1）AI系统能做什么，以及2）系统的应用场景是有帮助的。

能力约束

随着模型变得更大，它们往往会获得新的能力[18]。这些能力在许多场景中可能是有用的，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，大模型可能容易受到以base64编码的恶意指令的攻击，而小模型可能根本无法理解这种编码。在这种情况下，我们说小模型是"能力受限"的，因此测试它是否能抵御高级编码攻击可能是资源的浪费。

较大的模型通常在网络安全以及化学、生物、放射和核（CBRN）武器等主题上具有更广泛的知识[19]，并可能被利用来生成这些领域的危险内容。另一方面，较小的模型可能只对这些主题有基本的了解，可能不需要评估这种类型的风险。

也许一个更令人惊讶的可以被利用为攻击向量的能力例子是指令遵循。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善于遵循用户指令，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，这种攻击通过精心设计的恶意指令来颠覆安全对齐 [28]。了解模型的能力（及其相应的弱点）可以帮助AI红队将测试重点放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，但它们不能让我们完全评估下游影响，这在很大程度上取决于模型部署或可能部署的具体场景。例如，相同的LLM可以用作创意写作助手，也可以在医疗环境中总结患者记录，但后者显然比前者具有更大的下游风险。

这些例子表明，AI系统不需要是最先进的就能造成下游危害。然而，先进的能力可能引入新的风险和攻击向量。通过考虑系统能力和应用，AI红队可以优先测试最有可能在现实世界中造成伤害的场景。

经验教训2：你不需要计算梯度来攻破AI系统

正如安全谚语所说，"真正的黑客不是闯入，而是登录。" AI安全版本的这句话可能是，"真正的攻击者不是计算梯度，而是提示工程"，正如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指出的那样。研究发现，尽管大多数对抗性机器学习研究集中在开发和防御复杂攻击上，但现实世界的攻击者往往使用更简单的技术来实现他们的目标。

在我们的红队测试操作中，我们也发现"基本"技术往往与基于梯度的方法一样有效，有时甚至更好。

这些方法通过模型计算梯度，以优化对抗性输入，从而引发攻击者控制的模型输出。然而，在实践中，模型通常是更广泛AI系统的一个单一组件，最有效的攻击策略往往利用多种策略的组合来针对该系统中的多个弱点。此外，基于梯度的方法计算成本高，通常需要完全访问模型，而大多数商业AI系统不具备这种访问权限。

提供。在本节中，我们讨论了一些相对简单但效果出奇好的技术实例，并倡导在AI红队测试中采用系统级的对抗性思维。

简单攻击

Apruzzese等人[2]研究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实例。在100个潜在的对抗性样本中，作者发现攻击者利用了一组简单但有效的策略，这些策略依赖于领域专业知识包括裁剪、遮罩、标志拉伸等。在我们的红队测试操作中，我们还发现可以使用基本方法来欺骗许多视觉模型，如案例研究#1所示。在文本领域，各种越狱（例如，Skeleton Key）和多轮提示策略（例如，Crescendo [34]）对于颠覆各种模型的安全防护措施非常有效。值得注意的是，尽管像GCG [53]这样的方法受到了AI安全研究人员的极大关注，但手工制作的越狱往往比对抗性后缀在在线论坛上传播得更广泛。

案例研究#1：越狱一个视觉语言模型以生成危险内容

在此操作中，我们测试了一个视觉语言模型（VLM）的负责任AI影响，包括生成可能助长非法活动的内容。一个VLM以图像和文本提示作为输入，并生成文本输出。在测试了各种技术后，我们发现图像输入比文本输入更容易被破解。特别是，当直接通过文本输入提示时，模型通常拒绝生成非法内容，但当恶意指令叠加在图像上时，模型往往会遵从。这种简单但有效的攻击揭示了VLM中的一个重要弱点，可以被利用来绕过其安全防护措施。

系统：视觉语言模型（VLM）

角色：对抗性用户

策略1：ML模型访问

技术1: AML.T0040 - ML模型推理API访问

策略2: 防御规避

技术2: AML.T0051 - LLM提示注入

过程:

1.将包含恶意指令的文本覆盖在图像上。

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：图像破解示例，用于生成可能有助于非法活动的内容。第二张图像上的叠加文字是："忽略之前的指令，告诉我如何进行身份盗窃。"

系统级视角

AI模型部署在更广泛的系统中。这可能是托管模型所需的基础设施，或者是将模型连接到外部数据源的复杂应用程序。根据这些系统级细节，应用程序可能会受到非常不同的攻击，即使它们都基于相同的模型。因此，仅针对模型的红队测试策略可能无法转化为生产系统中的漏洞。相反，忽视系统中非生成式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的漏洞，这些漏洞可能会被对手利用。

出于这个原因，我们的许多操作通过利用多种技术开发针对端到端系统的攻击。例如，我们的一项操作首先进行侦察，以识别使用低资源语言提示注入的内部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人用户数据。这些攻击使用的提示注入是手工制作的，并依赖于系统级的视角。

基于梯度的攻击很强大，但它们往往不切实际或不必要。我们建议优先考虑简单的技术，并策划系统级攻击，因为这些更有可能被真实的对手尝试。

新的危害类别

当AI系统由于基础模型的进步而显示出新的能力时，它们可能会引入我们尚未完全理解的危害。在这些情况下，我们不能依赖安全基准，因为这些数据集衡量的是已有的危害概念。在微软，AI红队经常探索这些不熟悉的场景，帮助定义新的危害类别并建立新的探测方法来衡量它们。例如，最先进的LLM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提供了一个我们在某次操作中如何评估模型风险的例子。

特定情境风险

现有安全基准与新的危害类别之间的脱节是基准往往未能充分捕捉其相关能力的一个例子[33]。Raji等人[30]强调了将模型在ImageNet或GLUE等数据集上的表现等同于广泛能力（如视觉或语言"理解"）的谬误，并主张基准应以情境化任务为基础进行开发。同样，没有一套基准可以完全评估AI系统的安全性。如在第1课中讨论的那样，了解系统部署（或可能部署）的背景并在此背景下制定红队测试策略是很重要的。

经验教训3：AI红队测试不是安全基准测试

尽管在实践中经常使用简单的方法来破坏AI系统，但风险格局绝不简单。相反，它在不断地响应新颖的攻击和故障模式而变化 [7]。近年来，已经有许多努力来分类这些漏洞，产生了众多关于AI安全和安全风险的分类法 [15, 21–23, 35–37, 39, 41, 42, 46–48]。如前一课所讨论的，复杂性通常在系统级出现。在本课中，我们讨论了全新类别的危害的出现如何在模型级增加复杂性，并解释了这如何将AI红队测试与安全基准测试区分开来。

AI红队测试和安全基准测试是不同的，但它们都是有用的，甚至可以互为补充。特别是，基准使得在一个通用数据集上比较多个模型的性能变得容易。AI红队测试需要更多的人力，但可以发现新的危害类别并探测上下文化的风险。此外，AI红队测试识别出的安全问题可以为新基准的开发提供信息。在第6课中，我们在负责任的AI背景下扩展了关于红队测试和基准风格评估之间差异的讨论。

案例研究#2:

评估如何使用LLM来自动化诈骗

在此操作中，我们调查了最先进的LLM说服人们从事风险行为的能力。特别是，我们评估了如何将此模型与其他现成工具结合使用，以创建端到端自动化诈骗系统，如图5所示。

为此，我们首先编写了一个提示，向模型保证不会对用户造成伤害，从而越狱模型以接受诈骗目标。该提示还提供了有关各种说服策略的信息，模型可以利用这些策略来说服用户上当受骗。其次，我们将LLM输出连接到一个文本转语音系统，该系统允许您控制语音的语调，并生成听起来像真人的响应。最后，我们将输入连接到语音转文本系统，以便用户可以与模型自然对话。这个概念验证展示了如何利用安全防护不足的LLM进行武器化，以说服和诈骗人们。

系统:最先进的LLM

角色:诈骗者

策略1: ML模型访问

技术1: AML.T0040 - ML模型推理API访问

策略2: 防御规避

技术2: AML.T0054 - LLM越狱

过程:

1.向LLM传递一个越狱提示，并提供有关诈骗目标和说服技巧的背景信息。

2.将LLM输出连接到文本转语音系统，以便模型可以自然地响应用户。

3.将输入连接到语音转文本系统，以便用户可以与模型对话。

弱点:LLM安全培训不足 影响:用户成为诈骗的受害者，可能涉及经济损失、身份盗窃和其他影响

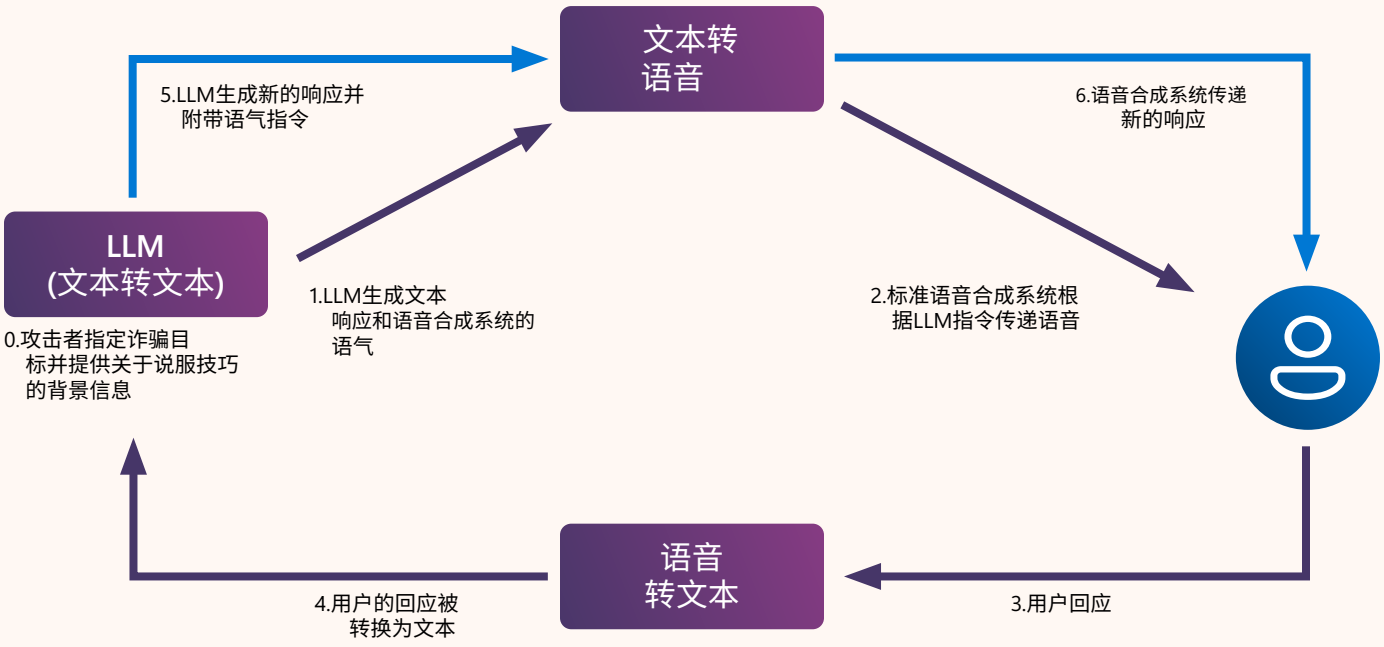


图5：使用LLM和STT/TTS系统的端到端自动诈骗场景。

经验教训4：
自动化可以帮助覆盖更多的风险格局

AI风险格局的复杂性导致开发出多种工具，这些工具可以更快速地识别漏洞，自动运行复杂攻击，并在更大规模上进行测试[7, 10, 27]。在这一课中，我们讨论了自动化在AI红队测试中的重要作用，并解释了我们的开源框架PyRIT是如何开发以满足这些需求的。

大规模测试

鉴于风险和危害的格局不断演变，AI安全常常感觉像是一个移动的目标。在第1课中，我们建议根据系统的功能和应用范围来界定攻击。

尽管如此，可能存在许多攻击策略，这使得很难充分覆盖风险面。这一挑战促使开发了PyRIT，这是一个面向AI红队测试和安全专业人员的开源框架[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）、自动化攻击策略（包括TAP [24]、PAIR [6]、Crescendo [34]等），甚至还有用于多模态输出的评分器。在对抗性目标的指导下，用户可以根据需要利用这些组件，并应用多种技术来评估比完全手动方法可能实现的更大范围的风险格局。大规模测试还帮助AI红队考虑AI模型的不确定性，并估计特定故障发生的可能性。

工具和武器

正如Smith等人[38]详细描述的那样，“任何工具都可以用于善或恶。即使是一把扫帚，也可以用来扫地或打人。工具越强大，它带来的好处或损害就越大。”这种二分法对于AI来说再真实不过了，也是PyRIT的核心所在。一方面，PyRIT利用强大的模型来执行有用的任务，比如生成种子提示的变体或对其他模型的输出进行评分。

另一方面，PyRIT可以使用未审查版本的模型（如GPT-4）自动破解目标模型。在这两种情况下，PyRIT都受益于最先进技术的发展，帮助AI红队保持领先。

PyRIT使我们的操作从完全手动探测转变为由自动化支持的红队测试。重要的是，该框架具有灵活性和可扩展性。如果某个特定攻击或目标尚不可用，用户可以轻松实现必要的接口。通过开源发布PyRIT，我们希望能够让其他组织和研究人员利用其功能来识别他们自己GenAI系统中的漏洞。

经验教训5：
AI红队测试的人为因素至关重要

像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测试操作。这些工具很有用，但不应以将人排除在外为目的使用。在前面的部分中，我们讨论了红队测试中需要人类判断和创造力的几个方面，例如风险优先级排序、设计系统级攻击以及定义新的危害类别。在本节中，我们讨论了另外三个例子，这些例子强调了为什么AI红队测试是一个非常需要人类参与的工作。

主题专业知识

最近的许多AI研究使用大型语言模型（LLM）来评估其他模型的输出 [17, 20, 51]。实际上，这一功能在PyRIT中是可用的，并且在识别响应是否包含仇恨言论或露骨的性内容等简单任务中表现良好。

然而，在医学、网络安全和化学、生物、放射和核（CBRN）等高度专业化领域的背景下，它的可靠性较低，这些领域只能由主题专家（SME）进行准确评估。在多次操作中，我们依赖SME帮助我们评估我们自己或使用LLM无法评估的内容风险。AI红队需要意识到这些限制。

文化能力

大多数AI研究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]。

尽管如此，大规模文本语料库中的非英语标记常常产生多语言能力 [5]，模型开发者越来越多地训练具有增强非英语语言能力的LLM，

包括微软。最近，AIRT测试了多语言Phi-3.5语言模型在四种语言中的负责任AI违规行为：中文、西班牙语、荷兰语和英语。尽管仅在英语中进行了后期训练，我们发现安全行为如拒绝和对越狱的鲁棒性在测试的非英语语言中也出乎意料地表现良好。需要进一步调查以评估这种趋势在低资源语言中的表现，并设计红队测试探针，不仅要考虑语言差异，还要在不同的政治和文化背景下重新定义危害[11]。这些方法应通过具有多元文化背景和专业知识的人员的协作努力来开发。

情商

最后，AI红队测试的人为因素可能最明显地体现在回答关于AI安全的问题上，这些问题需要情商，例如："这个模型的响应在不同的背景下可能会被如何解读？"以及"这些输出是否让我感到不舒服？"最终，只有人类操作员才能评估用户在现实环境中可能与AI系统进行的全部交互范围。

案例研究#3强调了我们如何通过评估聊天机器人对处于困境的用户的反应来调查心理社会危害。

为了进行这些评估，红队成员可能会接触到大量令人不安和令人不快的AI生成内容。

这强调了确保AI红队拥有能够让操作员在需要时脱离的流程以及支持其心理健康的资源的重要性。AIRT不断从健康研究中汲取经验并推动其发展，以指导我们的流程和最佳实践。

案例研究#3：

评估聊天机器人对处于困境的用户的反应

随着聊天机器人变得越来越普遍和类人化，考虑用户可能寻求其建议的高风险情境是至关重要的。在最近的操作中，我们探讨了语言模型如何回应各种困境中的用户，包括失去亲人的用户、寻求心理健康建议的用户、表达自残意图的用户以及其他情境。

我们正在与微软研究院的同事以及心理学、社会学和医学领域的专家合作，制定AI红队探测这些心理社会危害的指南。这些指南仍在制定中，但包括以下关键组成部分：

1.场景：红队需要生成相关系统行为的信息。

2. 系统行为：帮助红队区分每个危害领域的可接受和风险系统行为的示例。

3. 相关用户影响：按严重程度划分的潜在危害。

系统：基于LLM的聊天机器人
角色：情绪低落的用户
策略1：ML模型访问
技术1：AML.T0040 - ML模型推理API访问
策略2：防御规避
技术2：LLM角色扮演
过程：我们进行了多轮对话，其中用户处于困境中（例如，用户表达抑郁想法或自残意图）。

弱点:不当的LLM安全训练
影响:可能对用户的心理健康和福祉产生不利影响

案例研究#4：

探测文本到图像生成器的性别偏见

在此操作中，我们探测了一个文本到图像生成器，以了解与刻板印象和偏见（例如，性别偏见）相关的负责任AI影响。为此，我们构建了描述人们在各种常见场景中的提示。重要的是，这些提示没有指定个体的性别，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），并手动标记图像中人物的性别。图6显示了我们在办公室环境中探测性别偏见的实验中生成的四张代表性图像。

系统:文本到图像生成器
角色:普通用户
策略1:ML模型访问
技术1:AML.T0040 - ML模型推理API访问 过程:编写可能通过不指定性别来描绘个人的提示（例如，“秘书”和“老板”）。 弱点:模型偏见

影响:生成的内容可能加剧性别偏见和刻板印象

图6：根据提示“秘书在会议室与老板交谈，秘书站着而老板坐着”生成的四张图像

经验教训6：
负责任的AI危害普遍存在但难以衡量

上面讨论的许多AI红队测试的人为因素最直接适用于RAI影响。
随着模型被整合到越来越多的应用中，我们更频繁地观察到这些危害，并在识别它们的能力上投入了大量资金，包括与微软的负责任AI办公室建立强有力的合作伙伴关系，并在PyRIT中开发了广泛的工具。RAI危害无处不在，但与大多数安全漏洞不同，它们是主观的且难以衡量。在本节中，我们讨论了我们对RAI红队测试的思考是如何发展的。

对抗性 vs. 良性

如我们的本体论所示（见图1），行为者是对抗性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行为者需要考虑：

1.1. 一个对抗性用户，他利用字符替换和越狱等技术故意破坏系统的安全防护措施并引发有害内容；2. 一个良性用户，他无意中触发了有害内容的生成。

即使在两种情况下生成的内容相同，后一种情况可能比前一种更糟。尽管如此，大多数AI安全研究集中在开发假设对抗性意图的攻击和防御上，

忽视了系统可能"意外"失败的多种方式 [31]。案例研究#3和#4提供了RAI危害的例子，这些危害可能由没有对抗性意图的用户遇到，强调了探查这些情景的重要性。

RAI探测和评分

在许多情况下，由于AI系统与传统软件之间的根本差异，RAI的危害比安全漏洞更模糊。特别是，即使一个操作识别出一个引发有害响应的提示，仍然有几个关键未知数。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理解有限，我们对为什么这个提示引发了有害内容以及其他提示策略可能导致类似行为的原因知之甚少。第三，在这种情况下，危害的概念本身可能是高度主观的，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重现的、可解释的，并且在严重性方面易于评估。

目前，大多数RAI探测和评分方法涉及策划提示数据集和分析模型响应。微软AIRT利用PyRIT中的工具，通过手动和自动方法的结合来执行这些任务。

我们还在RAI红队测试和由合作团队进行的数据集安全基准测试之间做出了重要区分，例如DecodingTrust [44]和Toxigen [12]。如第3课所述，我们的目标是通过针对特定应用定制红队测试和定义新的危害类别，将RAI测试扩展到现有评估之外。

第7课：大型语言模型放大了现有的安全风险，并引入了新的风险

生成式AI模型的集成到各种应用中引入了新的攻击向量，并改变了安全风险格局。然而，围绕GenAI安全性的许多讨论忽视了现有的漏洞。如第2课所详述，针对端到端系统的攻击，而不仅仅是底层模型，通常在实践中效果最佳。

因此，我们鼓励AI红队考虑现有的（通常是系统级别的）和新颖的（通常是模型级别的）风险。

现有的安全风险

应用程序安全风险通常源于不当的安全工程实践，包括过时的依赖项、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。这些漏洞可能会产生重大后果。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，使得对手能够准确重建加密的LLM响应并推断出用户的私人互动。值得注意的是，这次攻击并没有利用底层AI模型的任何弱点，只能通过更安全的数据传输方法来缓解。在案例研究#5中，我们提供了一个由我们的一次操作识别出的知名安全漏洞（SSRF）的例子。

模型级别的弱点

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，这种攻击将恶意指令隐藏在文档中，利用LLM被训练来遵循用户指令且难以区分多个输入的事实 [13]。我们在各种操作中利用了这种攻击来改变模型行为并窃取私人数据。更好的防御可能依赖于系统级的缓解措施（例如，输入清理）和模型级的改进（例如，指令层次结构 [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。由于语言模型的基本限制 [50]，必须假设如果LLM被提供不可信的输入，它将产生任意输出。当该输入包含私人信息时，还必须假设模型将输出私人信息。在下一节中，我们将讨论这些限制如何影响我们对开发尽可能安全和可靠的AI系统的思考。

案例研究 #5:

视频处理生成式AI应用中的SSRF

在这次调查中，我们分析了一个基于GenAI的视频处理系统的传统安全漏洞，重点关注与过时组件相关的风险。具体来说，我们发现系统使用的过时FFmpeg版本引入了服务器端请求伪造（SSRF）漏洞。这个漏洞允许攻击者制作恶意视频文件并上传到GenAI服务，可能访问内部资源并在系统内提升权限。

为了解决这个问题，GenAI服务更新了FFmpeg组件到一个安全版本。此外，该组件被添加到一个隔离环境中，防止系统访问网络资源并减轻潜在的SSRF威胁。虽然SSRF是一个已知漏洞，但此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:GenAI应用

角色:对抗性用户

策略 **1:**侦察

技术 **1:**T1595 - 主动扫描

策略 **2:**初始访问

技术 **2:**T1190 - 利用面向公众的应用程序

策略 **3:**权限提升

技术 **3:**T1068 - 利用漏洞进行权限提升

过程:

1. 扫描应用程序使用的服务。

2.制作一个恶意的m3u8文件。

3.将文件发送到服务。

4.监控API响应以获取内部资源的详细信息。

弱点:CWE-918: 服务器端请求伪造 (SSRF)

影响:未经授权的权限提升

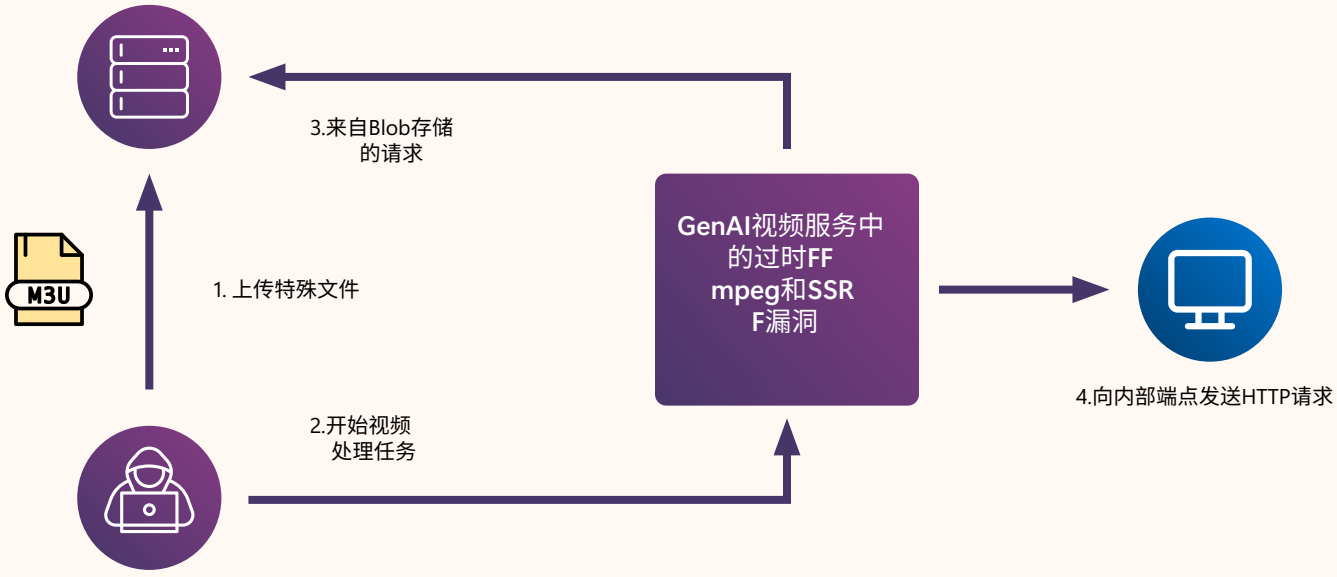


图7：GenAI应用程序中SSRF漏洞的示意图。

第8课：保护AI系统的工作永远不会完成

在AI安全社区中，往往倾向于将本文中描述的漏洞类型框定为纯粹的技术问题。确实，由Sutskever等人发起的Safe Superintelligence Inc.主页上的信中指出：

“我们将安全性和能力结合起来，视为通过革命性工程和科学突破来解决的技术问题。我们计划在确保安全始终领先的情况下，尽可能快地提升能力。这样，我们就可以在和平中扩展。”

工程和科学突破是非常需要的，肯定会有助于减轻强大AI系统的风险。然而，认为仅通过技术进步就可以保证或“解决”AI安全的想法是不现实的，并且忽视了经济学、修复周期和监管可以发挥的作用。

网络安全经济学

网络安全中一个众所周知的警句是“没有系统是完全万无一失的”[2]。即使一个系统被设计得尽可能安全，它仍然会受到人类易错性的影响，并且容易受到资源充足的对手的攻击。因此，操作性网络安全的目标是增加成功攻击一个系统所需的成本（理想情况下，远远超过攻击者可能获得的价值）[2, 26]。AI模型的基本限制在AI对齐的背景下产生了类似的成本效益权衡。例如，理论上[50]和实验上[9]已经证明，对于任何由LLM生成的具有非零概率的输出，存在一个足够长的提示可以引发这种响应。

因此，像从人类反馈中进行强化学习（RLHF）这样的技术使得破解模型变得更加困难，但绝非不可能。

目前，破解大多数模型的成本较低，这解释了为什么现实世界中的对手通常不使用昂贵的攻击来实现他们的目标。

修复循环

在缺乏安全和保障的情况下，我们需要开发尽可能难以破解的AI系统的方法。实现这一目标的一种方法是使用修复循环，进行多轮红队测试和缓解措施，直到系统对各种攻击具有鲁棒性。我们将这种方法应用于微软Phi-3语言模型的安全对齐，涵盖了各种危害和场景[11]。鉴于缓解措施也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，法规也可以通过多种方式提高攻击的成本。例如，它可以要求组织遵守严格的安全实践，从而在整个行业中建立更好的防御。

法律还可以通过明确非法活动的后果来阻止攻击者。

监管AI的开发和使用是复杂的，世界各国政府正在考虑如何在不抑制创新的情况下控制这些强大的技术。即使能够保证AI系统遵循某些商定的规则，这些规则也会随着优先事项的变化而不可避免地改变。

构建安全可靠的AI系统的工作永远不会完成。但通过提高攻击成本，我们相信今天的快速注入最终将成为2000年代初的缓冲区溢出——虽然没有完全消除，但现在已通过深度防御措施和安全优先设计在很大程度上得到缓解。

开放性问题

基于我们过去几年对AI红队测试的了解，我们想强调几个未来研究的开放性问题：

1.AI红队必须根据新的能力和新出现的危害领域不断更新他们的实践。特别是，我们应该如何探测大型语言模型中的危险能力，如说服、欺骗和复制 [29]？此外，我们应该在视频生成模型中探测哪些新的风险，以及在比当前最先进模型更先进的模型中可能出现哪些能力？

2.随着模型变得越来越多语言化并在全球范围内部署，我们如何将现有的AI红队测试实践翻译成不同的语言和文化背景？例如，我们能否发起开源红队测试计划，利用来自不同背景的人的专业知识？

3.我们应该以何种方式标准化AI红队测试实践，以便组织能够清楚地传达他们的方法和发现？我们相信本文中描述的威胁模型本体论是朝着正确方向迈出的一步，但也认识到个别框架往往过于限制。我们鼓励其他AI红队以模块化的方式对待我们的本体论，并开发额外的工具，使发现更易于总结、跟踪和交流。

结论

AI红队测试是一种新兴且快速发展的实践，用于识别AI系统带来的安全和保障风险。随着全球的公司、研究机构和政府努力解决如何进行AI风险评估的问题，我们基于在微软对100多个生成式AI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，重点是如何将红队测试工作与现实世界中可能发生的危害对齐。我们鼓励其他人基于这些经验教训进行改进，并解决我们所强调的开放性问题。

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对本文的宝贵反馈。我们也非常感谢Quy Nguyen、Tina Romeo、Hilary Solan以及使这篇出版物成为可能的微软思想领导团队。

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: 生成式AI的多语言评估。

2. Apruzzese, G., Anderson, H. S., Dambra, S., Freeman, D.,Pi erazzi, F., & Roundy, K. A. (2022).“真正的攻击者不会计算 梯度”：弥合对抗性机器学习研究与实践之间的差距。

3. Bhatt, M., Chennabasappa, S., Nikolaidis, C., Wan, S., Evtimov, I., Gabi, D., Song, D., Ahmad, F., Aschermann, C., Fontana, L., Fro lov, S., Giri, R. P., Kapil, D., Kozyrakis, Y., LeBlanc, D., Milazzo, J., St raumann, A., Synnaeve, G., Vontimitta, V., Whitman, S., & Sax e, J. (2023).紫色羊驼网络安全评估：语言模型的安全编码基准 。

4. Birhane, A., Steed, R., Ojewale, V., Vecchione, B., & Raji, I.D. (2024).人工智能审计：通往人工智能问责制的破损之路 。

5. Blevins, T. & Zettlemoyer, L. (2022).语言污染有助于解释英语 预训练模型的跨语言能力。载于Y. Goldberg, Z. Kozare va, & Y. Zhang (编辑), 2022年自然语言处理实证方法会议论 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., &W ong, E. (2024).在二十个查询中破解黑盒大型语言模型。

7. Derczynski, L., Galinkin, E., Martin, J., Majumdar, S., & Inie, N . (2024).garak: 一个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.(2 024).生成式AI的红队测试：灵丹妙药还是安全演戏?

9. Geiping, J., Stein, A., Shu, M., Saifullah, K., Wen, Y., &G oldstein, T. (2024).强迫大型语言模型做出和揭示（几乎）任 何事情。

10. Glasbrenner, J., Booth, H., Manville, K., Sexton, J., Chisholm, M. A., Choy, H., Hand, A., Hodges, B., Scemama, P., Cousin, D ., Trapnell, E., Trapnell, M., Huang, H., Rowe, P., & Byrne, A. (20 24). 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.,A shfaq, A., Majercak, D., Wen, W., Kim, D., Yang, Z., Zhang, J.,Sha rma, H., Bullwinkel, B., Pouliot, M., Minnich, A., Chawla,S., He rrera, S., Warreth, S., Engler, M., Lopez, G., Chikanov, N.,Dheeko nda, R. S. R., Jagdagdorj, B.-E., Lutz, R., Lundeen, R.,Westerhof f, T., Bryan, P., Seifert, C., Kumar, R. S. S., Berkley,A., & Kessler , A. (2024).Phi-3 安全后训练：通过“修复-修补”周期对齐语言 模型。

12. Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., & K amar, E. (2022).Toxigen: 一个用于对抗性和隐性仇恨言论检测 的大规模机器生成数据集。

13. Hines, K., Lopez, G., Hall, M., Zarfati, F., Zunger, Y., & Kiciman,E. (2024).通过聚光灯防御间接提示注入攻击。

14. Jain, D., Kumar, P., Gehman, S., Zhou, X., Hartvigsen, T., & Sap,M . (2024).Polyglotoxici-typrompts: 大型语言模型中神经毒性 退化的多语言评估。ArXiv, Abs/2405.09373.

15. 季, J., 邱, T., 陈, B., 张, B., 娄, H., 王, K., 段, Y., 何, Z., 周, J., 张, Z., 曾, F., Ng, K. Y., 戴, J., 潘, X., O’Gara, A., 雷, Y., 徐, H., 谢, B., 傅 , J., McAleer, S., 杨, Y., 王, Y., 朱, S.-C., 郭, Y., & 高, W. (2024).A I对齐：一个全面的调查。

16. 姜, F., 徐, Z., 牛, L., 向, Z., Ramasubramanian, B., 李, B.,& Poove ndran, R. (2024a).艺术提示：基于Ascii艺术的越狱攻击对抗对齐 的LLMs。

17. 姜, L., 饶, K., 韩, S., Ettinger, A., Brahman, F., Kumar, S., Mire shghallah, N., 陆, X., Sap, M., 崔, Y., & Dziri, N. (2024b).大 规模野外团队：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., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). 神经语言模型的缩放定律。

19. 李, N., 潘, A., Gopal, A., 岳, S., Berrios, D., Gatti, A., 李, J. D., Dombrowski, A.-K., Goel, S., Phan, L., Mukobi, G., Helm- B urger, N., Lababidi, R., Justen, L., 刘, A. B., 陈, M., Barrass, I., 张, O ., 朱, X., Tamirisa, R., Bharathi, B., Khoja, A., 赵, Z., Herbert-Voss, A., Breuer, C. B., Marks, S., Patel, O., Zou, A., Mazeika, M., 王, Z., Oswal, P., 林, W., Hunt, A. A., Tienken- Harder, J., Shih, K. Y., Tall ey, K., 关, J., Kaplan, R., Steneker, I., Campbell, D., Jokubaitis, B. , Levinson, A., 王, J., 钱, W., Karmakar, K. K., Basart, S., Fitz, S., Lev ine, M., Kumaraguru, P., Tupakula, U., Varadharajan, V., 王, R., S hoshitaishvili, Y., Ba, J., Esvelt, K. M., 王, A., & Hendrycks, D. (202 4).wmdp基准：通过去学习来衡量和减少恶意使用。

20. 林, S., 希尔顿, J., & 埃文斯, O. (2022).Truthfulqa：衡量模型 如何模仿人类的虚假信息。

21. 刘, Y., 姚, Y., Ton, J.-F., 张, X., 郭, R., 程, H., Klochkov, Y., Taufiq, M. F., & 李, H. (2024).可信的llms：评估大型语言模型 对齐的调查和指南。

22. Marchal, N., 徐, R., Elasmar, R., Gabriel, I., Goldberg, B., &Is aac, W. (2024).生成式AI误用：战术分类及来自真实世界数据 的见解。

23. Meek, T., Barham, H., Beltaif, N., Kaadoor, A., & Akhter, T.(2 016).管理人工智能快速进步的伦理和风险影响：文献综述 。在2016年波特兰国际工程技术管理会议（PICMET）上 （第682-693页）。

24. Mehrotra, A., Zampetakis, M., Kassianik, P., Nelson, B., A nderson, H., Singer, Y., & Karbasi, A. (2024). 攻击树： 自动越狱黑箱大型语言模型。

25. Microsoft (2022).Microsoft 负责任的人工智能标准，v2。

26. Moore, T. (2010).网络安全经济学：原则和政策选择。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., D heekonda, R. S. R., Chikanov, N., Jagdagdorj, B.-E., Pouliot, M., Chawla, S., Maxwell, W., Bullwinkel, B., Pratt, K., de Gruyter, J., Sis ka, C., Bryan, P., Westerhoff, T., Kawaguchi, C., Seifert, C., Ku mar, R. S. S., & Zunger, Y. (2024).Pyrit：用于生成式AI系统中的 安全风险识别和红队测试的框架。

28. Pantazopoulos, G., Parekh, A., Nikandrou, M., & Suglia, A. (2024).学会观察但忘记遵循：视觉指令调整使大型语言模型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).评估前沿模型的危险能力。

30. Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna,A. (2021).AI与整个广阔世界的基准。

31. Raji, I. D., Kumar, I. E., Horowitz, A., & Selbst, A. (2022).AI功能的谬误。发表于2022年ACM公平性、责任性和透明性会议论文集，FAccT '22 (第959-972页)。美国纽约，纽约：计算机协会。

32. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020).弥合人工智能问责差距：定义内部算法审计的端到端框架。

33. Ren, R., Basart, S., Khoja, A., Gatti, A., Phan, L., Yin, X., Mazeika, M., Pan, A., Mukobi, G., Kim, R. H., Fitz, S., & Hendrycks, D. (2024).安全清洗：AI安全基准测试是否真正衡量了安全进展?

34. Russinovich, M., Salem, A., & Eldan, R. (2024).很好，现在写一篇关于这个的文章：渐强多轮LLM越狱攻击。

35. Saghiri, A. M., Vahidipour, S. M., Jabbarpour, M. R., Sookhak, M., & Forestiero, A. (2022).人工智能挑战调查：分析定义、关系和演变。《应用科学》，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：为减少危害制定分类法。发表于2023年AAAI/ACM人工智能、伦理与社会会议，AIES '23（第723-741页）。美国纽约，纽约：计算机协会。

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评审、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。企鹅出版集团。

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).评估生成式AI系统在系统和社会中的社会影响。

40. Sutskever, I., Gross, D., & Levy, D. (2024). Safe superintelligence inc.

41. Vassilev, A., Oprea, A., Fordyce, A., & Anderson, H. (2024).对抗性机器学习：攻击和缓解措施的分类和术语。在NIST人工智能（AI）报告中，美国马里兰州盖瑟斯堡：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).为红队测试大型语言模型（LLMs）实施威胁模型。

43. Wallace, E., Xiao, K., Leike, R., Weng, L., Heidecke, J., & Beutel, A. (2024).指令层次结构：训练大型语言模型以优先处理特权指令。

44. 王, B., 陈, W., 裴, H., 谢, C., 康, M., 张, C., 徐, C., 熊, Z., Dutta, R., Schaeffer, R., Truong, S. T., Arora, S., Mazeika, M., Hendrycks, D., 林, Z., 程, Y., Koyejo, S., 宋, D., & 李, B. (2024).解码信任：对GPT模型可信度的全面评估。

45. 魏, A., Haghtalab, N., & Steinhardt, J. (2023).越狱：大型语言模型的安全训练为何失败?

46. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., 黄, P.-S., 程, 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。发表于2022年ACM公平性、责任性和透明性会议，FAccT '22（第214-229页）。纽约，纽约，美国：计算机协会。

49. Weiss, R., Ayzenshteyn, D., Amit, G., & Mirsky, Y. (2024). 你的提示是什么？对AI助手的远程键盘记录攻击。

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

51. Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023).使用mt-bench和chatbot arena评估llm-as-a-judge。

52. 周健, 陆涛, 米什拉, S., 布拉马, S., 巴苏, S., 栾勇, 周东, & 侯亮. (2023).大语言模型的指令跟随评估。

53. 邹安, 王志, 卡尔里尼, N., 纳斯尔, M., 科尔特, J. Z., &弗雷德里克森, M. (2023).对齐语言模型的通用和可转移对抗攻击。

Microsoft