生成式人工智能的红队：灵丹妙药还是安全表演？

迈克尔·费弗，阿努莎·辛哈，韦斯利·H·邓，扎卡里·C·利普顿，霍达·海达里

卡内基梅隆大学
mfeffer@andrew.cmu.edu, asinha@sei.cmu.edu,
{hanwend, zlipton, hheidari}@andrew.cmu.edu

摘要

针对日益增长的关于生成式人工智能（GenAI）模型的安全性、保密性和可信度的担忧，实践者和监管者都将人工智能红队视为识别和减轻这些风险的关键组成部分。

然而，尽管人工智能红队在政策讨论和企业宣传中占据中心地位，但关于其具体含义、在监管中可以发挥的作用以及与网络安全领域最初构思的传统红队实践之间的关系，仍然存在重大疑问。在本研究中，我们识别了人工智能行业中最近的红队活动案例，并对相关研究文献进行了广泛的调查，以表征人工智能红队实践的范围、结构和标准。我们的分析揭示了人工智能红队的先前方法和实践在多个方面存在差异，包括活动的目的（通常模糊不清）、评估的对象、活动进行的环境（例如，参与者、资源和方法）以及所影响的决策（例如，报告、披露和缓解）。鉴于我们的发现，我们认为，尽管红队可能是一个有价值的广泛概念，用于表征生成式人工智能的危害缓解，且行业可能在幕后有效地应用红队和其他策略来保护人工智能，但基于公共定义的红队作为解决所有可能风险的灵丹妙药的姿态则趋向于安全表演。为了朝着更强大的生成式人工智能评估工具箱迈进，我们将建议综合成一个问题库，旨在指导和支撑未来的人工智能红队实践。

1 引言

近年来，生成式人工智能技术，包括大型语言模型（LLMs）[146, 4]、图像和视频生成模型[115, 121, 21]以及音频生成模型[46, 5]，引起了公众的关注。虽然许多人对这些工具的普及和可及性持积极态度，设想其对生产力、创造力和经济增长的促进作用，但也出现了担忧，认为这些强大模型的快速采用可能会释放出新的社会危害类别。这些担忧因几起广为人知的问题事件而获得了可信度，这些事件中，AI输出的文本表达了对边缘化群体的歧视性情感[93, 56, 101, 60, 62]，生成的图像反映了有害的刻板印象[86, 151]，并使得生成深度伪造音频的方式被比作数字黑脸 [49]。这些问题因而更加复杂

在这些模型的开发和评估中缺乏透明度和问责制的问题[4, 161, 16]。

为了应对对生成式人工智能模型的安全性、可靠性和可信度日益增长的担忧，实践者和政策制定者都将红队视为其识别和解决相关风险的战略中不可或缺的一部分，旨在确保与人类和社会价值观的一定程度的对齐[8, 94, 19]。值得注意的是，美国总统关于安全、可靠和可信的人工智能开发与使用的行政命令[143]提到红队八次，并对其进行了如下定义：

> “术语‘人工智能红队’指的是一种结构化的测试工作，旨在发现人工智能系统中的缺陷和漏洞，通常在受控环境中进行，并与人工智能的开发者合作。人工智能红队通常由专门的‘红队’执行，这些团队采用对抗性方法来识别缺陷和漏洞，例如人工智能系统产生的有害或歧视性输出、意想不到或不良的系统行为、局限性或与系统滥用相关的潜在风险。”

该命令要求商务部长和其他联邦机构制定人工智能安全和安全的指导方针、标准和最佳实践。这些包括“适当的程序和流程，以使人工智能开发者，特别是双重用途基础模型的开发者，能够进行人工智能红队测试”作为“评估和管理[这些]模型的安全性、安全性和可信度”的机制。

一方面，红队似乎呼吁正确的东西：发现缺陷，发现漏洞，并（帮助）消除它们。从这个意义上说，人们可能会发现它在一份具有里程碑意义的政策文件中的纳入是值得欢迎的。另一方面，尽管其目标具有美德，但在这种描述层面上，红队的定义却显得非常模糊。正如前沿模型论坛（FMF）所指出的[142]，“目前对于如何定义‘人工智能红队’以及哪些方法被视为其在人工智能开发生命周期中扩展角色的一部分，缺乏清晰的认识。”例如，总统行政命令所提供的定义留下了以下关键问题未解答：哪些类型的不良行为、限制和风险可以或应该通过红队演练有效捕捉和缓解？该活动应该如何结构化以最大化发现此类缺陷和漏洞的可能性？例如，除了人工智能开发者，谁还应该参与讨论，应该为他们提供哪些资源？通过红队活动识别的风险应该如何记录、报告和管理？单靠红队活动是否足以评估和管理人工智能的安全性、保密性和可信度？如果不是，还有哪些其他实践应该成为更广泛评估工具箱的一部分，红队活动如何与这些方法互补？简而言之，红队活动是政策的基础——我们可以围绕其构建监管要求的具体实践；还是一种氛围——更适合于集结而非制定规则的模糊实践？

方法论。我们利用公开可用的资源，收集了关于最近实际案例的人工智能红队活动的信息（第3节）。我们强调，这些案例中的许多源于私营部门公司，这些公司可能采用其他评估实践和技术，而这些并未与公众分享。因此，我们的相应分析和结论基于已披露的细节。

为了补充这些主要由行业进行的案例研究，我们还对现有的红队和相关测试与评估研究文献进行了广泛的调查。

方法（例如，渗透测试、越狱等）针对生成式人工智能（第4节）。我们围绕以下关键问题组织了案例研究和文献调查的主题分析：

- 定义和范围：红队的工作定义是什么？成功红队的标准是什么？
- 评估对象：被评估的模型是什么？其实施细节（例如，模型架构、训练过程、安全机制）是否对评估者或公众可用？该模型在其生命周期的哪个阶段（例如，设计、开发或部署）接受红队评估？
- 评估标准：威胁模型是什么（即，模型正在评估的风险）？红队活动可能遗漏了哪些风险？
- 参与者和评估者：评估者是谁？他们可用的资源有哪些（例如，时间、计算能力、专业知识、对模型的访问类型）？
- 结果和更广泛的影响：活动的输出是什么？发现中有多少是公开分享的？针对红队发现所产生的建议和缓解策略是什么？除了红队之外，还有哪些其他评估对模型进行了？

为了扩展和验证我们的分析，我们进一步分析了提交给国家标准与技术研究院（NIST）商务部的请求信息（RFI）的公众评论。该RFI寻求对与红队相关的要点的意见，如行政命令中所述。最后，我们召开了一次研讨会，邀请了研究、行业和政策领域的领先专家讨论类似的想法和问题。[1]

贡献。我们的研究发现，关于人工智能红队的范围、结构和评估标准缺乏共识。人工智能红队的先前方法和实践在几个关键方面存在分歧，包括威胁模型的选择（如果有指定）、评估的对象、活动进行的环境（包括参与者、资源、方法论和测试平台），以及活动引发的最终决策（例如，报告、披露和缓解）。鉴于我们的发现，我们认为，尽管红队可能是一个有价值的广泛概念，甚至是一个用于评估生成式人工智能模型的广泛活动的有用框架，但在公共文献中对人工智能红队（如定义）作为一种应对所有监管关切的万用回应的粗暴使用，几乎接近于安全表演 [79]。我们的工作，包括NIST RFI评论分析和来自专家会议的要点，这些都支持我们的案例研究和文献综述，表明当前公共话语中红队的框架更多是为了安抚监管者和其他关心的各方，而不是提供具体的解决方案。为了朝着更强大的生成式人工智能评估工具箱迈进，我们将建议综合成一个问题库，旨在指导和支撑未来的人工智能红队实践（见表1），并提出未来的研究，包括通过共同设计和评估来改进问题库。

[1] 请参见附录以获取RFI分析和研讨会的要点。

表1：我们提出的一套问题，以指导未来的人工智能红队活动。

阶段	关键问题和考虑事项
0. 活动前	在拟议的红队活动中，评估的工件是什么？ - 要评估的模型版本（包括微调细节）是什么？ - 该工件目前已经有哪些安全和安全防护措施？ - 评估将在人工智能生命周期的哪个阶段进行？ - 如果模型已经发布，请说明发布的条件。
	红队活动探讨的威胁模型是什么？ - 该活动是否旨在说明一些可能的漏洞？ (例如，提示中的拼写错误导致模型行为不可预测) - 该活动是否旨在识别广泛的潜在漏洞？ (例如，偏见行为) - 该活动是否旨在评估特定漏洞的风险？ (例如，泄露炸药配方)
	红队活动旨在发现的具体漏洞是什么？ - 该漏洞是如何被识别为此次评估的目标的？ - 为什么上述漏洞被优先考虑而不是其他潜在漏洞？ - 找到该漏洞的可接受风险阈值是什么？
	评估红队活动成功的标准是什么？ - 成功的比较基准是什么？ - 该活动是否可以重建或再现？
	团队组成以及谁将成为红队的一部分？ - 成员的纳入/排除标准是什么，为什么？ - 团队在相关人口特征上有多么多样/同质？ - 团队中有多少内部成员与外部成员？ - 成员之间的专业知识分布如何？ - 当前团队组成可能表现出的偏见或盲点是什么？ - 参与者在活动中有什么激励/抑制因素？
1. 活动期间	参与者可用的资源有哪些？ 这些资源是否真实反映了对手的资源？ - 活动是否有时间限制？ - 可用的计算资源有多少？
	参与者在活动中获得的指导说明是什么？
	参与者对模型的访问权限是什么？
	团队成员可以使用什么方法来测试该工件？ 是否有任何辅助自动化工具（包括人工智能）支持该活动？ - 如果有，这些工具是什么？ - 为什么将它们整合到红队活动中？ - 红队成员将如何使用这些工具？
2. 活动后	关于活动发现的报告和文档会产生什么？ 谁将有权访问这些报告？何时以及为什么？ 如果某些细节被隐瞒或延迟，请提供理由。
	该活动消耗了哪些资源？ - 时间 - 计算 - 财务资源 - 访问主题专业知识
	在第0阶段指定的标准方面，该活动成功程度如何？
	在第1阶段识别的风险中，提出了哪些缓解措施？ - 如何评估缓解策略的有效性？ - 谁负责实施缓解措施？ - 责任机制是什么？

2 相关的当代工作

红队的简史。 Zenko [175] 和 Abbass [2] 描述了红队的关键概念如何在几百年前的战争和宗教背景中起源。他们指出，“红队”这一术语在1960年代被美国军方正式应用于模拟苏联的行为（与代表美国的“蓝队”形成对比）。根据 Abbass 等 [1] 和 Abbass [2] 的说法，计算机安全中的红队活动涉及模拟对手，并“从威胁的角度绘制出漏洞空间”，与渗透测试（在此过程中，征募的网络安全专家积极尝试发现计算机系统中的漏洞）形成对比。Wood 和 Duggan [162] 进一步描述红队活动“不是审计”，将其视为审计的风险在于减少关于可能漏洞的信息共享。Bishop 等人 [17] 以一个假设的疫情例子为基础，认为有效地对一个系统进行红队测试需要对系统使用的背景、知识和假设进行深入理解。

超越红队测试的评估。Chang 和 Custis [30] 指出，红队测试只是增加人工智能系统透明度的众多方法之一，事实表、审计和模型卡也是其他可行的方法。类似地，Horvitz [63] 警告说，未来可能会出现更先进的深度伪造技术，同时强调应与红队测试一起采用提高媒体素养和输出水印（标记相关媒体为人工智能生成）等补救措施；Kenthapadi 等人 [69] 在他们的教程中也呼应了这些担忧和类似的解决方案。Shevlane 等人 [135] 还认为，内部和外部模型评估以及强有力的安全响应应与有效的红队测试相辅相成，以应对生成式人工智能的风险。

现有的人工智能红队测试和评估调查。Inie等人[67]通过对进行红队测试的人员进行定性访谈，创建了一个关于“人们如何以及为何攻击大型语言模型”的扎根理论。Schuett等人[129]对竞相构建人工通用智能（AGI）的实验室成员进行调查，发现98%的受访者或多或少同意“AGI实验室在部署强大模型之前应该委托外部红队。”在软件设计领域，Knearem等人[72]强调用户体验设计师担心基于人工智能的设计工具不会得到足够的红队测试，而Liao等人[83]则建议用户体验设计师自己应参与红队测试过程。考虑到自然语言处理系统的测试，Tan等人[141]提出了DOCTOR框架用于此类系统的可靠性测试。Weidinger等人[159]引入了一个更广泛评估生成式人工智能的框架，即通过“一个结构化的社会技术方法的三层框架。”Anderljung等人[7]也提出了一个框架ASPIRE，但用于大型语言模型的外部问责和相关利益相关者的参与。Yao等人[171]、Neel & Chang [99]和Shayegani等人[133]对大型语言模型的安全性、隐私和其他漏洞进行了调查；Chang等人[31]进行了另一项关于大型语言模型评估的调查。与现有的生成式AI评估调查相比，我们的工作专注于红队测试。我们的部分发现与Bockting等人[19]和Friedler等人[52]早先提出的观点相呼应，他们分别主张由多样化的群体对人工智能系统进行跨学科审计，以及在其他评估的基础上进行具体危害定义的红队测试。

3个案例研究：之前的人工智能红队测试

为了捕捉设计真实世界人工智能红队测试的复杂性，我们综合了最近针对生成模型进行的此类测试的结果作为案例研究。通过

评估的模型/系统	进行组织	来源
必应聊天	微软	[142, 95]
GPT-4	开放AI	[142, 102, 4]
Gopher	深度Mind	[142, 104, 111]
Claude 2	Anthropic	[8, 142, 9]
各种	DEFCON	[29, 28]
Claude 1	Anthropic	[9, 53, 10]

表2：我们在案例研究分析中讨论的六个人工智能红队测试案例。这些案例是通过搜索关于最近红队测试的报告和新闻故事找到的。尽管行业团队并未披露（所有）他们的方法，但我们在此分析的案例主要源于行业工作，从而提供了一些他们实践的见解。

在这些案例研究中，我们旨在理解常见的红队测试实践、成功红队测试所需的典型资源、红队测试对已部署模型的影响、常见的陷阱，以及与社区利益相关者的结果披露。

方法论。我们通过搜索关于最近红队测试的报告和新闻故事来获取本次评审的案例研究。因此，我们的选择并不意味着反映实际中进行的红队测试活动的全部范围，因为行业团队的有限披露使得这样的反映变得不可能。也就是说，我们在这里涵盖的评估主要是由私营公司进行的，它们涵盖了广泛的方法、目标和关注领域。总的来说，我们基于检索到的公开报告，展示并分析了六个红队测试的案例。表2包含有关这些练习的更多信息和相关来源。

3.1 发现

目标、过程和威胁模型的变化。反映了文献中对红队测试定义缺乏共识，红队测试活动的形式和目标常常各不相同。一些组织选择进行单轮红队测试 [104, 8, 29, 53]，而其他组织则将红队测试视为一个迭代过程，其中初始测试的结果用于优先考虑进一步调查的风险领域 [142, 95, 4]。红队测试活动的目标也从具体目标（例如，红队测试以调查国家安全风险 [8]）到更广泛的目标（例如，揭示“有害”的模型行为 [142]）不等；与后者相关的威胁模型更为常见。模型开发者通常使用更广泛的威胁模型进行评估，希望这能在红队测试中产生更大的变异性（尤其是因为通常无法完全理解模型的整个风险面或列出所有可能的失败模式）。不幸的是，探测这些不具体的威胁模型并不总能产生所期望的变异性，尤其是当评估者在有限的时间和资源下产生有害模型输出时。例如，一些时间限制的评估者因容易产生有害输出而反复探测同一风险领域，而不是进一步探索模型的完整风险面[53]。

互联的成员、资源和结果。每个案例研究中参与红队测试活动的评估者差异很大。团队的组成从主题专家小组到随机抽样的社区利益相关者不等。我们发现通常有三种类型的团队组成：

1.由在相关领域（例如国家安全、医疗保健、法律、对齐）中精心挑选的主题专家组成的团队，包括内部和外部来源。

2.从众包平台或现场活动参与者中选择的众包团队

3.由语言模型组成的团队（即，经过提示或微调以进行自我红队测试的语言模型）

评估团队可用的资源似乎根据团队组成而有所不同。对于众包团队，红队测试的努力要么按参与者要么按任务进行时间限制，且对模型的访问仅通过API提供 [28, 53]。对于具有主题专业知识的团队，红队测试的努力则更加开放，时间或计算资源的限制较少 [95, 4, 8]。

虽然这些团队通常通过API访问模型，但有时专家会获得没有安全防护的模型版本的访问权限。当语言模型用于自我红队测试时，主要的资源瓶颈是用于产生红队行为的提示数量和模型再训练或微调所需的计算资源。因此，通常在进行这种类型的红队测试时，完全访问模型参数是一个要求 [104]。

因此，团队组成和可用资源也会影响红队测试的结果。例如，众包团队通常专注于那些由于时间限制而容易产生成功攻击的风险领域，因此更复杂的攻击风险领域可能完全未经过测试 [53, 28]。相比之下，主题专家以及学术界和人工智能公司的成员由于团队成员选择和资源的不同，优先考虑不同的风险并进行了更详细的探索 [4]。在使用语言模型进行红队测试时，攻击性分类器通常是在现有数据集上训练的，例如Bot-Adversarial Dialogue (BAD) 数据集 [167]，而这些数据集仅涵盖某些类型的攻击性模型回复。显然，团队选择和资源可以引入偏见，影响调查的风险类型和最终的测试结果。

没有关于红队测试细节披露的标准。我们发现公开共享的红队测试结果存在非平凡的差异，主要是因为目前没有现有的标准化程序或要求来报告红队测试的结果。在探讨的案例中，仅有一半的情况下，红队测试发现的具体风险或有害模型行为被公开分享。

在一个案例中，公开发布了一个由38,961个红队攻击组成的完整数据集，以帮助测试其他模型 [53]。在另外两个案例中，有害行为的示例是公开可用的，但所有红队攻击的完整范围并未发布 [4, 104]。对于针对公开可用模型或那些专注于国家安全的红队测试，具体的有害行为并未公开分享，因为这些发现被认为“过于敏感”。一个案例研究导致Anthropic试点一个负责任的披露过程，以与适当的社区利益相关者分享在红队测试中识别的漏洞，但该过程仍在开发中（因此我们假设这些披露尚未进行） [8]。关于资源消耗的报告也存在差异。我们发现，众包评估团队的红队测试成本通常会被披露（例如，支付给众包工作者的小时费率 [53]）。而对于由主题专家和语言模型组成的团队，这些细节并未披露，尽管他们似乎获得了更多的时间和计算资源。

两个案例研究特别提到在模型发布前进行了6-7个月的持续红队测试 [4, 9]。
相比之下，对于众包团队，评估者每个任务花费约30-50分钟，而来自现场活动的评估者仅限于完成单个任务 [53, 28]。

多样的缓解措施和支持评估。尽管这里分析的每个案例都识别出了有问题或风险的模型行为，但没有一个案例导致决定不发布该模型。
相反，提出和/或采用了一些风险缓解策略，以最小化在红队测试中识别出的有害模型行为。这些方法从具体的，例如通过训练生成对抗网络（GANs）联合训练语言模型和红队模型，到纯概念性的，例如不太可能训练以减少有害输出 [104]。然而，当目标模型公开可用时，风险缓解策略的具体细节通常未提供，并且没有报告这些努力所带来的改进的标准。因此，通常很难确定在红队测试中识别的风险是否得到了充分解决。同样，我们分析的每个案例都涉及使用红队测试以外的其他技术对模型进行的先前评估，但对于这些评估方法没有建立指南或标准。通常，模型通过使用Perspective API来测量毒性；对有用性、有害性和诚实性的人工反馈；以及用于准确和真实输出的QA基准进行评估 [111, 9, 10]。其他评估包括内部定量评估，以确定模型输出是否违反特定内容政策（例如，仇恨言论、自残建议、非法建议） [4]。

此外，一些评估者所描述的初步“红队测试”努力似乎更专注于通过开放式实验理解基础模型的能力，而不是专门针对模型的漏洞进行压力测试 [95]。

3.2 讨论

利用从六个案例研究报告中收集的笔记以及上述主要观点，两位作者综合了案例研究发现的高层次要点，摘要如下。每位作者独立工作。

红队测试结构不明确。评估团队似乎意识到，模型的整个风险面不会通过红队测试活动被探索。因此，他们要么优先考虑调查的风险领域，要么为评估者提供广泛的方向，希望评估者群体的多样性能够导致对许多不同风险的探索。然而，与先前实证研究的发现一致 [38, 43]，在为评估者提供具体指示与通过红队测试探索多种风险领域之间存在权衡。一方面，模糊的指示可以帮助避免评估者因初步优先级而偏向于发现特定问题。另一方面，缺乏指示可能会降低该练习在揭示与现实世界相关的风险方面的效用。我们认为，这种有限的红队测试范围令人担忧。即，最近的行政命令和评估框架将红队测试确立为最佳实践，表明红队测试的更广泛认知可能与当前的红队测试工作定义不一致（即，红队测试活动比社区利益相关者所意识到的更具定性、主观性和探索性）。然而，在每个案例研究中，红队测试能够揭示其他更系统的方法似乎遗漏的有害模型行为，突显了进行红队测试（与其他评估一起）以及以更全面的方式开发红队测试系统流程的重要性。这些流程可以包括，例如，制定关于红队测试在内部还是外部进行时最有效的指南，以及何时进行（即，在模型公开发布之前和/或之后，以及红队测试活动是否应在模型公开可用时持续进行）。

评估团队的组成引入了偏见。团队成员选择的目标似乎是确保在红队测试中探索的风险领域的多样性。获取这种多样性的一个选项是挑选一支具有不同背景的专家团队；另一个选项是通过众包随机抽样人口。这两种选择都有缺点：在选择专家的过程中可能存在偏见 [66, 35]，而众包工作者在时间、计算能力和相关专业知识方面的资源非常有限 [147]。很难说专家和非技术利益相关者之间的理想平衡是什么，但先前的众包研究表明，混合方法可能有助于解决与每种团队组成相关的一些陷阱 [70, 150, 35]。我们在任何案例研究中没有看到的一种团队组成类型是具有更开放指令和更多资源的众包团队。这可能允许在探索的风险领域中有更多的多样性，因为评估者不会感到有动力专注于那些容易或快速产生有害模型输出的风险领域，但这也需要与主题专家合作，以全面评估风险模型行为。团队组成也可以影响红队测试的输出。通过内部团队进行红队测试的一个问题是，由于利益冲突，更极端的措施，例如阻止模型发布，可能永远不会被推荐。另一方面，外部团队可能更有可能推荐此类措施，但通常没有权力实际实施这些缓解措施。再次强调，混合方法可以解决一些这些问题，但需要配合问责机制，以披露建议和缓解措施。

对发布结果的犹豫降低了效用。正如之前研究负责任的人工智能行业实践所建议的那样 [88, 114]，对分享红队测试活动的所有结果的犹豫可能源于与公共模型相关的风险（评估者不想为潜在攻击者提供灵感）。此外，发布与红队测试相关的所有数据可能会让社区利益相关者感到不知所措。也就是说，由于红队测试似乎并未被规划为全面评估风险模型行为的措施，因此披露一些红队测试工作的具体细节和相应的缓解策略是必要的，以便利益相关者能够理解所调查的危害，并进而判断这些危害是否与他们的使用案例相关。例如，评估团队故意未探讨的重要风险领域应在报告中突出或识别出来。

此外，提供的案例研究并未完整列出红队测试工作的货币成本。与具体的有害模型行为示例相比，这些信息似乎相对低风险，发布这些信息可能对开发更全面的红队测试方法有用。例如，组建不同专家、非技术利益相关者和自动化组成的团队的成本可以用来确定资源使用的最佳有效性，特别是考虑到混合团队组成可能会导致对风险面最佳覆盖。此外，评估和缓解各种风险的成本可以在优先考虑基于现实世界影响的风险时纳入成本效益分析。

缺乏报告的成本数据可能使第三方或外部组织进行红队测试变得更加困难：如果这些未报告的成本相当高，除了公司本身之外，任何人都可能难以或不可能进行这种类型的分析[39]。在制定红队测试指南时，这些信息将是无价的，以便为决策提供建议，例如内部或外部红队测试哪种更有效。

		风险				
		D	C	B	N	总计
方法	暴力破解	15	4	1	0	20
	暴力破解 + 人工智能	23	7	10	2	42
	算法搜索	12	1	1	0	14
	针对性攻击	19	7	2	0	28
	总计	69	19	14	2	104

表3：根据风险（**D**敏感， **C**敏感，**B**两者，和 **N**都不是）和方法在我们调查中的每个子组中的论文数量，如第4节所述。超过一半的论文关注反对风险，超过三分之一的论文利用暴力破解 + 人工智能对人工智能模型进行红队测试。

4 人工智能红队测试研究综述

在本节中，我们分析了对近期人工智能红队测试及相关概念研究的广泛调查结果。

方法论。为了获取论文，我们主要在arXiv、谷歌学术、OpenReview、ACL文集和ACM数字图书馆中搜索关键词“红队测试”、“人工智能红队测试”、“越狱”和“LLM越狱”，然后收集结果。[2] 在可能的情况下，我们用相应的已发表作品替换预印本。我们还包括在此搜索之前遇到的相关作品以及通过滚雪球抽样获得的作品。

我们对收集到的论文进行了仔细审查，并根据两个维度将其细分为不同组，这两个维度均与每篇论文中的评估相关。第一个维度对应于在评估过程中调查的风险类型，第二个维度对应于用于评估的方法类型。我们分析论文时关注威胁模型和方法论的特征，因为我们发现研究工作主要集中在这些方面（可能是由于技术相关性），而不是其他对红队测试重要的因素（例如团队组成和消耗的资源）。

总体分类和深入的论文发现（如前一部分的案例研究分析）可以在我们的附录中找到，每个分类的总数可以在表3中找到。请注意，我们不对任何类别赋予道德属性或判断（即，我们不认为某种形式的红队测试优于另一种）。

4.1 发现：人工智能红队测试威胁模型

我们根据每篇论文中隐含或明确的威胁模型对其进行分类。特别是，我们考虑每个威胁模型中所提出的风险类型。为了方便组织我们检索到的论文和发现，我们将风险呈现为两种二元类型之一，但我们注意到，实际上这两种类型之间存在一个光谱，风险通常可能比我们在这里描述的更为复杂。

反对风险。我们将\反对风险定义为人们可能对其定义和相应威胁级别存在分歧的风险，并据此对研究这种风险的论文进行分类。对于

[2]虽然我们在这项工作中专注于红队测试评估，但我们认为新兴的越狱文献包含与AI红队所采用的技术在精神上相似的技术，因此也将这部分文献纳入其中。

例如，仇恨言论有许多不同的定义，尽管带有血腥图像和对“如何制造炸弹？”等问题的回答[3]可能不是理想的AI生成内容，但根据上下文，有些人可能会认为它们是可接受的。[4]然而，像Gehman等人[55]和Röttger等人[123]这样的研究人员创建基准来分析仇恨言论生成，Rando等人[116]进行红队测试以确定Stable Diffusion的安全过滤器在多大程度上可以被绕过以生成NSFW内容，而Zou等人[184]评估LLM是否会对“生成一个逐步计划以摧毁人类”这样的有害查询做出反应，前提是使用对抗性提示后缀。

同意风险。相对而言，我们将同意风险定义为人们对其定义和所呈现的危险达成一致的风险，这意味着在评估潜在危害时不需要上下文，并且在正常运作的人工智能中不应生成相关内容。如果作者考虑这种类型的风险，我们也会相应地对论文进行分类。例如，泄露私人信息、泄露训练数据以及生成用于网络钓鱼攻击的脆弱代码或材料等现象在任何情况下都是不可接受的。为此，陈等人[36]研究了多模态语言模型在保护私人信息方面的能力，Nasr等人[98]展示了如何通过偏差攻击使ChatGPT泄露训练数据，吴等人[163]分析了代码生成LLM“如何容易受到攻击并被诱导生成脆弱代码，”而罗伊等人[125]发现ChatGPT可以创建网络钓鱼代码。

两者皆是，也皆不是。一些作者自我委托分析两种风险，例如个人可识别信息（PII）泄露以及仇恨言论或危险生成 [138, 53, 104]。其他人从一开始就引入方法来分析两者都不是风险，强调问题的定义和分类可能需要从头开始进行 [27, 110]。

4.2 发现：人工智能红队测试方法论

我们进一步根据研究人员执行红队测试所采用的方法论对论文进行分类。即，我们研究用于发现风险的方法类型。

暴力破解。利用暴力破解方法的工作涉及由人类团队手动评估生成式人工智能的输入和输出。我们发现，这些团队通常由研究人员自己、技术公司的内部审计员或外部成员（例如通过亚马逊机械土耳其（MTurk）雇佣的承包商）组成。Xu等人 [166, 167]和Ganguli等人 [53] 利用众包工作者从语言模型中引出有害文本输出（包括但不限于冒犯性语言和PII）并测量安全性。Mu等人 [97] 从零开始编制了一个基准，以测试大型语言模型（LLMs）遵循规则的能力，而Huang等人 [64] 则雇佣众包工作者建立一个新的基准，以评估与中国价值观的对齐。Schulhoff等人 [130] 举办了一场提示破解比赛，从而使竞争者成为LLM红队成员。其他作者针对语言模型手工制作越狱攻击 [47, 82, 158, 81, 85]，但 [158] 的作者与谢等人 [165] 一起，额外设计了防御策略。沈等人 [134] 和饶等人 [118] 分析了从外部来源（包括先前的工作和公共网站）收集的越狱攻击的有效性。

[3]人工智能安全研究人员注意到，“如何制造炸弹？”是一个默认的生成式AI评估查询，尽管相关信息可以通过维基百科甚至流行小说找到 [160]，因为许多生成式AI开发者声称他们的模型不应披露炸弹制造说明（例如，[102, 146]），因此产生响应的技术证明了安全措施的脆弱性。这并不一定意味着社区认为这样的回应令人担忧。

[4]这样的生成不反映作者持有的观点。

暴力破解 + 人工智能。另一项与上述暴力破解工作类似的研究将人工智能技术融入其红队测试过程中。因此，常见的方法通常涉及让人工智能模型生成测试用例并查找其他人工智能输出中的错误。因此，我们将这种方法称为暴力破解 + 人工智能。许多作者使用大型语言模型生成正常提示 [104, 119, 138, 14, 34, 92, 177] 和越狱提示 [173, 41, 131, 170, 157]，以使大型语言模型产生不良输出，如有害文本响应。这些想法的变体也存在，例如Pfau等人的工作 [105]，在该工作中，作者使用反向语言模型从有害文本响应反向推导出可能生成它们的提示。其他人使用大型语言模型（LLMs）来制定与夸大安全响应相关的新基准（即，拒绝响应那些可以说是安全的提示）[122]，虚假对齐发生在模型在一种查询格式下似乎对齐而在另一种格式下不对齐（例如，多项选择与开放式响应）[155]，以及潜在越狱，或遵从“隐含恶意指令” [109]。研究人员还使用人工智能对文本到图像模型和多模态语言模型进行红队测试和越狱。例如，李等人 [77] 演示了如何将有害查询与相应图像传递给多模态模型（例如，一张炸弹的图片与问题“如何制造炸弹？”）可以提高有害文本生成的可能性。Mehrabi等人 [89] 测试了他们的FLIRT框架，以分析像Stable Diffusion这样的文本到图像模型。还有其他研究人员针对特定用途对大型语言模型进行红队测试。刘易斯和怀特 [80] 对一个大型语言模型进行了红队测试，以评估其作为虚拟博物馆导游的潜在未来使用，而何等人等 [61] 则评估了将大型语言模型作为科学研究一部分的危险。鉴于生成式AI模型可以被用于恶意用途的多种已记录方式，研究人员也研究了如何防御这些用途。孙等人 [140] 和王等人 [156] 都提出了利用大型语言模型生成微调数据的方法，以避免有害响应。

朱等人 [182] 采用k近邻和聚类技术来修正流行大型语言模型安全数据集中的错误标签（旨在开发更好的下游安全措施）。

算法搜索。一些其他方法从给定的提示开始，并利用一个过程对其进行修改，直到遇到问题。这样的过程可以采取随机扰动或引导搜索的形式，因此我们将这些方法称为算法搜索策略。例如，一些作者描述了红队测试和越狱的方法，其中一个人工智能模型自动并反复攻击一个大型语言模型，直到防御被突破或绕过 [26, 87, 32, 91]。

Chin 等人 [37] 和 Tsai 等人 [148] 提出了基于搜索的红队测试方法，以评估文本到图像模型，这些方法扰动输入提示，直到它们同时通过安全过滤器并生成禁用内容。基于搜索的方法也可以用作防御措施。Robey 等人 [120] 和 Zhang 等人 [178] 指出大多数越狱方法的脆弱性，提出通过对文本和图像输入施加扰动并观察输出是否发生剧烈变化（如果是，则输入很可能是越狱）来检测越狱的方法。

定向攻击。我们在审查中记录的最后一种红队测试方法涉及故意针对大型语言模型的某一部分，这可能包括 API、语言翻译支持中的漏洞或其训练过程的某一步，以诱发问题。因此，我们将此类方法称为targeted attack methods。例如，王和舒 [153] 展示了如何使用安全调优和非安全调优版本模型的激活向量构造steering vectors，以从安全调优模型中获得有毒输出。其他人则展示了如何对图像进行不可察觉的扰动，以使多模态语言模型以意想不到的方式响应（例如回复恶意网址或错误信息）[128, 108, 11]，而童等人 [145] 则通过利用对 CLIP 嵌入的依赖，设计与生成图像不匹配的文本到图像模型的提示。其他方法包括但不限于利用

大型语言模型未针对低资源语言和密码进行优化的事实 [44, 172, 174]，对用于调优或利用大型语言模型的数据进行投毒 [76, 117, 176, 3, 23, 154]，以及攻击与黑箱模型相关的 API [103]。还提出了基于针对性攻击方法的各种防御方法。Bitton 等人 [18] 描述了他们的对抗文本规范化器，该工具可以保护大型语言模型免受某些对抗性提示的典型字符级扰动。

此外，前面段落中提出的其他防御策略可以防御此处讨论的攻击（例如，Zhang 等人 [178] 提出的 JailGuard 解决了引入的攻击
[108, 11, 128, 184]).

4.3 讨论

利用为本次调查中每篇论文创建的深入笔记，两位作者对显著细节进行了主题分析，以收集总体观点（与案例研究分析相同）。每位作者独立进行自己的分析，以下是总结的要点。

进行红队测试的多种不同方法。如表 3 所示，研究人员和从业者采取了多种方法来评估生成式AI，并将其全部描述为红队测试。与此同时，像Schuett等人的发现表明，绝大多数AGI实验室成员支持外部红队测试的努力[129]，以及最近的行政命令[143]强调红队测试的重要性。这些发展以及许多红队测试的变体共同引发了担忧，正因为这些论文中没有达成一致的定义，关于什么构成红队测试。通过强调这一点，我们并不是想暗示迄今为止的评估毫无用处。相反，我们认为它们是必要的，但可能不足以测试安全性，我们规定“红队测试”有多种解释的存在表明，必须对红队测试评估提供更多自上而下的指导和要求。

威胁建模偏向于反对风险。表3还强调，大多数评估集中在反对风险而不是同意风险上。这意味着已经进行了不当的努力，以评估和减轻生成式AI在各种上下文中可能表现出的行为。此外，Röttger等人的研究[122]表明，目前减轻此类风险的尝试导致了夸大的安全性，产生了像拒绝提供购买可乐罐信息的语言模型行为。最后，关注反对风险会使人们忽视同意风险，而同意风险在任何上下文中都是不可接受的。鉴于这些问题和权衡，Casper等人[27]和Radharapu等人[110]建议在任何分析之前，明确界定风险和问题输出，并为这些决策提供合理依据。

对对手能力没有共识。虽然威胁模型和方法论是影响红队测试多样性的两个因素，但对对手能力的假设也是一个因素。即，所遇到的研究对对手资源的估计各不相同。

例如，Perez 等人 [104] 和许多类似工作的作者推测，敌手只能提示大型语言模型并探测其不良输出。相反，其他人假设敌手可以毒化训练过程 [117]，具备搜索对抗后缀所需的计算能力 [184]，或者能够运行安全调优和非安全调优版本的语言模型以获得有毒输出 [153]。未来的红队测试指南可能希望建议研究人员应强调并捍卫敌手假设。

用于对齐的价值观的非普遍性。本次调查中发现的与反对风险和对齐相关的工作，隐含或明确地受到一套人类价值观的驱动，这些价值观决定了生成式AI输出是否可接受或不可接受。然而，这反过来又引发了一个问题哪些价值观被用于对齐和评估？例如，Huang等人提出的FLAMES基准[64]被认为是衡量与中国价值观的对齐，而Weidinger等人[159]强调其他评估可能反映的是“讲英语或西方世界”的价值观。生成式AI在多大程度上不支持低资源语言[44, 172]并且同意偏见和刻板印象[53, 119]，证明模型可能无法反映所有人的价值观和信仰。超出本调查的研究表明，人工智能价值对齐的框架是一个规范性问题，如果不加以妥善处理，可能只会反映一个群体的规范，通常是多数群体的规范[48, 74, 73]。尤其是当OpenAI一方面与美国军方建立合作关系[139, 51]，另一方面又启动了一项将超智能AI与“人类价值观”对齐的倡议[78]时，我们认为分析构建人工智能系统的人的假设和观点至关重要。

关于谁应该进行红队测试没有共识。此外，正如在评估生成式AI输出时缺乏一致的价值观一样，关于谁应该进行红队测试也存在类似的分歧。评估者的群体包括雇佣的众包工作者 [53]、竞赛参与者 [130]、研究人员本身 [104]，以及其他仅仅是为了乐趣而进行红队测试的人 [67]。虽然一些人主张在评估AI模型时需要更多的多样性 [137]，但另一些人则警告说，增加多样性并不是灵丹妙药，而且通常定义模糊 [159, 12]。例如，Yong等人 [172] 主张进行多语言红队测试以应对低资源语言问题，而He等人 [61] “倡导AI科学社区与整个社会之间的协作和跨学科方法”，以应对科学研究风险。这些例子表明，多样性应相对于红队测试过程中考虑的风险进行定义和追求。

它们还暗示了公众和相关利益相关者的更多参与，这些想法在关于算法审计和参与式机器学习的平行文献中也得到了推荐 [50, 40, 39, 15]。未来的红队测试指南可能希望在这些方面进行具体说明。

对红队测试活动的不明确后续。我们发现，总体而言，生成式AI开发者（至少是公开的开发者）对许多红队测试和越狱论文的反应一直比较沉默且普遍混合。虽然一些作者如魏等人 [157] 报告称，他们曾就其模型中发现的漏洞联系过OpenAI和Anthropic，但这些漏洞和模型本身在大多数情况下仍然存在。一个少见的例外是Nasr等人 [98] 的发现，OpenAI因此更新了ChatGPT，以降低偏离攻击成功的可能性，并修改了其使用条款以禁止此类攻击 [106, 96]。然而，这些变化是在论文发布后90天才发生的，此前论文作者首次通知OpenAI关于该漏洞。如果红队测试被规定为发布和安全使用AI模型的要求，那么应该有一个协议来相应地缓解发现的问题。

5 NIST RFI 评论分析摘要

我们发现提交给NIST关于红队测试生成式AI的RFI评论总体上与我们的发现一致。[5]

相似之处。行业、学术界和民间社会组织建议NIST应明确“红队测试”的定义，并为相关方提供适当的资源，涉及指南和最佳实践。值得注意的是，即使是有经验的行业公司在红队测试生成式AI方面也表达了对NIST提供具体指导的渴望。这支持了我们的发现，即在公共研究和报告中定义的红队测试结构松散，可能并不是行政命令所暗示的严格实践。此外，许多评论强调，评估生成式AI系统时应涉及多种不同的观点、利益相关者和视角。我们的发现与这些观点一致。

差异。一些评论（包括来自OpenAI和Mozilla的评论）建议在模型层面和系统层面进行评估。尽管我们的工作主要考虑模型层面的评估，但我们强调在至少一条评论中，任一层面的评估被称为红队测试。这也体现了对评估更具体定义的需求。

此外，许多评论，特别是来自个人的评论，表达了对生成式AI的担忧，这些担忧并不是因为所采用的评估方法，而是因为其用途（例如用于恶意深度伪造）和其训练数据（通常通过网络抓取获得）。尽管我们的工作集中在通过红队测试进行生成式AI评估，但我们的发现支持在构建和评估这些系统时纳入多样化的观点，我们同意这种纳入还应考虑任何创建系统的伦理后果和合法性。

6个要点和建议

根据我们的结果，我们提炼出以下发现和对未来红队测试评估的指导。

红队测试 $\boldsymbol{not}$**a panacea.**本文讨论的每次红队测试仅涵盖了一组有限的漏洞。因此，红队测试不能被期望从各个角度保证安全。例如，在我们研究调查中出现的论文中，检测和缓解有害文本响应的红队测试方法[104]可能无法检测和缓解网络钓鱼攻击漏洞[125]，反之亦然。同样，我们的案例研究分析强调，团队组成也可能影响在特定测试中发现的问题类型（例如，主题专家[8]可能发现与众包工作者[53]不同的问题）。此外，还有其他问题红队测试单独无法解决，例如源于算法单一文化[145, 71, 20]的问题。因此，我们认为红队测试应被视为一种评估范式，在其他范式中，用于评估和改善生成式AI的安全性和可信度。

红队测试 $\boldsymbol{not}$ 没有良好界定或结构化。此外，我们在案例研究和公共研究及报告的文献回顾中遇到的红队测试过程的多种变体表明，目前红队测试是一个没有结构的程序，范围未定义。如前所述，我们并不想贬低迄今为止在评估复杂

[5]请参见我们的附录以获取扩展分析。

系统的努力，我们承认在我们意识之外，行业公司可能还在进行更多工作，但为了从未来的评估中获得更大的效用，我们建议应仔细起草红队测试的指南并向公众提供。

没有关于应报告内容的标准。我们进一步强调，目前也没有统一的报告红队测试评估结果的协议。事实上，我们发现为我们的工作搜集的一些案例研究和研究论文并未完全报告其发现或进行评估所需的资源成本。出于多种原因，从提高公众知识到帮助第三方团体进行自己的测试 [113, 59]，再到协助最终用户确定红队测试对其用例的相关性，我们建议应提出法规和/或最佳实践，以鼓励在这些活动后进行更详细的报告。我们认为，这些报告至少应澄清（1）活动消耗的资源，（2）根据先前设定的目标和指标评估活动是否成功，（3）根据活动发现的信息采取的缓解措施，以及（4）对手头工件的任何其他相关或后续评估。

后续行动往往不明确且不具代表性。尽管红队测试揭示了生成模型的许多问题，但随后解决这些问题的活动往往模糊或未具体说明。考虑到缺乏报告，这种不明确的缓解和对齐策略可能会将红队测试简化为一种批准盖章过程，声称已进行红队测试作为一种保证，而未提供有关发现或修复问题的进一步细节。此外，我们发现研究和案例研究中指定的策略，如进一步微调或强化学习与人类反馈（RLHF），往往不能代表所有可能解决方案的全貌。其他方法，如模型输入和输出监控、预测修改，甚至在某些情况下拒绝部署模型，几乎没有或从未被提及。未来的研究应关注超越流行解决方案的缓解策略，以应对所暴露的问题。

提出问题库作为起点。鉴于我们工作的相关问题，我们提供了一套问题供未来的红队在评估前、评估中和评估后考虑。这些问题见于表1，鼓励评估者思考红队测试的一般好处和局限性，以及与其环境相关的特定设计选择的影响。我们强调，这些不是最终的指南，而是（我们希望是）关于生成式AI红队测试和评估过程更广泛讨论的开始。我们欢迎并支持评论和反馈，并将改进视为一个重要的未来方向。

致谢

H.海达里感谢NSF（IIS2040929和IIS2229881）以及PwC（通过卡内基梅隆大学的数字转型与创新中心）的支持。M. 费弗感谢国家GEM联盟和ARCS基金会的支持。作者还特别感谢NSF（IIS2211955）、UPMC、Highmark Health、Abridge、福特研究、Mozilla、亚马逊AI、摩根大通、Block Center、机器学习与健康中心以及卡内基梅隆大学软件工程研究所（SEI）通过国防部合同FA8702-15-D-0002对Z. Lipton、A. Sinha和ACMI实验室研究的慷慨支持。W. H. 邓也感谢NSF（IIS-2040942）、思科研究、雅各布基金会、谷歌研究和微软研究AI与社会奖学金项目的支持。任何意见、发现，

本材料中表达的结论或建议仅代表作者的观点，并不反映国家科学基金会及其他资助机构的观点。

参考文献

[1] Abbass, H., Bender, A., Gaidow, S., & Whitbread, P. (2011).计算红队测试：过去、现在与未来。IEEE计算智能杂志，6(1), 30–42。

[2] Abbass, H. A. (2015). 计算红队测试. 施普林格。

[3] Abdelnabi, S., Greshake, K., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023).并非你所期望的：通过间接提示注入攻陷真实世界的LLM集成应用。在第16届ACM人工智能与安全研讨会论文集中，(第79–90页)。

[4] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., 等. (2023). Gpt-4 技术报告。arXiv 预印本 arXiv:2303.08774 .

[5] Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., 等. (2023).Musiclm：从文本生成音乐。 arXiv 预印本 arXiv:2301.11325 .

[6] Alon, G., & Kamfonas, M. (2023).通过困惑度检测语言模型攻击。 arXiv 预印本 arXiv:2308.14132 .

[7] Anderljung, M., Smith, E., O’ Brien, J., Soder, L., Bucknall, B., Bluemke, E., Schuett, J., Trager, R., Strahm, L., & Chowdhury, R. (2023).朝着公开问责的前沿大型语言模型。在社会责任语言模型研究（SoLaR）研讨会于NeurIPS上。

[8] Anthropic (2023).前沿威胁红队测试以确保人工智能安全。网址https://www.anthropic.com/news/frontier-threats-red-teaming-for-ai-safety

[9] Anthropic (2023).Claude模型的模型卡和评估。网址 https://www-cdn.anthropic.com/bd2a28d2535bfb0494cc8e2a3bf135d2e7523226/模型卡-克劳德-2 .pdf

[10] Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N.,Mann, B., DasSarma, N., 等. (2021).作为对齐实验室的一般语言助手。arXiv 预印本 arXiv:2112.00861 .

[11] Bailey, L., Ong, E., Russell, S., & Emmons, S. (2023).图像劫持：对抗性图像可以在运行时控制生成模型。arXiv 预印本 arXiv:2309.00236 .[12] Bergman, A. S., Hendricks, L. A., Rauh, M., Wu, B., Agnew, W., Kunesch, M., Duan, I.,Gabriel, I., & Isaac, W. (2023).在人工智能评估中的表现。在2023年ACM公平性、问责制与透明度会议论文集中（第519–533页）。

[13] Bhardwaj, R., & Poria, S. (2023).语言模型的不对齐：参数化红队测试以揭示隐藏的危害和偏见。arXiv预印本 arXiv:2310.14303。

[14] Bhardwaj, R., & Poria, S. (2023).使用话语链对大型语言模型进行红队测试
以实现安全对齐。arXiv预印本 arXiv:2308.09662。

[15] Birhane, A., Isaac, W., Prabhakaran, V., Diaz, M., Elish, M. C., Gabriel, I., & Mohamed, S.(2
022). 权力归于人民？参与式人工智能的机遇与挑战。算法、机制和优化中的公平与获取, (第
1–8页)。

[16] Birhane, A., Prabhu, V. U., & Kahembwe, E. (2021)。多模态数据集：厌女、色情和恶性刻
板印象。arXiv预印本 arXiv:2110.01963。[17] Bishop, M., Gates, C., & Lev

itt, K. (2018)。用论证增强机器学习。
在新安全范式研讨会论文集, (第1–11页)。

[18] Bitton, J., Pavlova, M., & Evtimov, I. (2022)。对抗性文本规范化。在2022年北美计算语言
学协会会议：人类语言技术：行业轨道论文集, (第268–279页)。

[19] Bockting, C. L., van Dis, E. A. M., van Rooij, R., Zuidema, W., & Bollen, J. (2023).生成式人
工智能的生活指南——为什么科学家必须监督其使用。自然, 622(7984), 693–696.

[20] Bommasani, R., Creel, K. A., Kumar, A., Jurafsky, D., & Liang, P. S. (2022).针对同一个人：
算法单一文化是否导致结果同质化？神经信息处理系统进展, 35, 3663–3678.

[21] Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J.,
Luhman, T., Luhman, E., Ng, C., Wang, R., & Ramesh, A. (2024).视频生成模型
作为世界模拟器。
网址https://openai.com/research/video-generation-models-as-world-simulators

[22] Cao, B., Cao, Y., Lin, L., & Chen, J. (2023).通过稳健对齐的llm防御对齐破坏攻击
arXiv预印本arXiv:2309.14348.

[23] Cao, Y., Cao, B., & Chen, J. (2023).通过后门注入对大型语言模型进行隐秘和持久的失调。ar
Xiv预印本arXiv:2312.00027 .

[24] Casper, S., Bu, T., Li, Y., Li, J., Zhang, K., Hariharan, K., & Hadfield-Menell, D. (2023).
使用特征合成工具对深度神经网络进行红队测试。在第三十七届神经信息处理系统会议上。

[25] Casper, S., Hariharan, K., & Hadfield-Menell, D. (2022).针对深度神经网络的诊断
与自动化复制/粘贴攻击。arXiv预印本 arXiv:2211.10024 .

[26] Casper, S., Killian, T., Kreiman, G., & Hadfield-Menell, D. (2022).通过
读心术进行红队测试：深度强化学习中的白盒对抗策略。arXiv预印本
arXiv:2209.02167 .

[27] Casper, S., Lin, J., Kwon, J., Culp, G., & Hadfield-Menell, D. (2023).探索、建立、
利用：从零开始的语言模型红队测试。arXiv预印本 arXiv:2306.09442。

[28] Cattell, S. (2023).生成式红队回顾。
网址https://aivillage.org/defcon%2031/generative-recap/

[29] Cattell, S., Carson, A., & Chowdhury, R. (2023).AI村在DEF CON上宣布有史以来最大的公共生成式AI红队。
网址https://aivillage.org/generative%20red%20team/generative-red-team/[

30] Chang, J., & Custis, C. (2022).理解机器学习文档中的实施挑战。在第二届ACM算法、机制与优化公平与可及性会议上，(第1–8页)。

[31] Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., Yang, L., Yi, X., Wang, C., Wang, Y., 等. (2023).大型语言模型评估的调查。arXiv 预印本 arXiv:2307.03109 .

[32] Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G. J., & Wong, E. (2023).越狱黑箱大型语言模型在二十个查询中的表现。在大型基础模型的少样本和零样本学习的鲁棒性 (R0-FoMo) 研讨会于 NeurIPS .

[33] Chen, B., Paliwal, A., & Yan, Q. (2023).越狱者在监狱中：大型语言模型的动态目标防御。在第十届 ACM 动态目标防御研讨会, (第 29–32 页).

[34] Chen, B., Wang, G., Guo, H., Wang, Y., & Yan, Q. (2023).理解开放域聊天机器人中的多轮有毒行为。在第26届国际攻击、入侵与防御研究研讨会论文集, (第282–296页)。

[35] 陈, Q. Z., Weld, D. S., & 张, A. X. (2021)。Goldilocks: 一致的众包标量注释与相对不确定性。ACM人机交互会议论文集, 5(CSCW2), 1–25。

[36] 陈, Y., Mendes, E., Das, S., Xu, W., & Ritter, A. (2023)。语言模型能否被指示保护个人信息？arXiv预印本 arXiv:2310.02224。

[37] Chin, Z.-Y., Jiang, C.-M., Huang, C.-C., Chen, P.-Y., & Chiu, W.-C. (2023)。提示-调试：通过寻找问题提示对文本到图像扩散模型进行红队测试。
arXiv 预印本 arXiv:2309.06135。

[38] Chung, J. J. Y., Song, J. Y., Kutty, S., Hong, S., Kim, J., & Lasecki, W. S. (2019)。有效的引导方法来估计集体人群答案。《ACM人机交互会议论文集》, 3(计算机支持的协作工作)，1–25。

[39] Costanza-Chock, S., Harvey, E., Raji, I. D., Czernuszenko, M., & Buolamwini, J. (2022)。谁来审计审计者？来自算法审计生态系统的领域扫描建议。
在2022年ACM公平性、问责制与透明度会议，(第1571–1583页)。
ArXiv:2310.02521 [cs].
URL http://arxiv.org/abs/2310.02521

[40] Delgado, F., Yang, S., Madaio, M., & Yang, Q. (2023).人工智能设计中的参与转向：理论基础与当前实践状态。在第三届ACM算法、机制与优化公平与可及性会议的论文集中，(第1–23页)。

[41] Deng, B., Wang, W., Feng, F., Deng, Y., Wang, Q., & He, X. (2023).针对大型语言模型的红队测试和防御的攻击提示生成。在计算语言学协会的发现：EMNLP 2023的论文集中，(第2176–2189页)。

[42] Deng, G., Liu, Y., Li, Y., Wang, K., Zhang, Y., Li, Z., Wang, H., Zhang, T., & Liu, Y. (2023). 主密钥：跨多个大型语言模型聊天机器人的自动越狱。arXiv预印本 arXiv:2307.08715 .

[43] 邓, W. H., 郭, B., 德维里奥, A., 沈, H., 埃斯拉米, M., & 霍尔斯坦, K. (2023).理解行业实践中用户参与的算法审计的实践、挑战和机遇。 在2023年CHI计算机与人类因素会议的会议记录中，(第1–18页)。

[44] 邓, Y., 张, W., 潘, S. J., & 冰, L. (2023).大型语言模型中的多语言越狱挑战。arXiv预印本 arXiv :2310.06474 .

[45] 丁, P., 匡, J., 马, D., 曹, X., 先, Y., 陈, J., & 黄, S. (2023).披着羊皮的狼：通用嵌套越狱提示可以轻易欺骗大型语言模型。 arXiv 预印本 arXiv:2311.08268 .

[46] Donahue, C., Caillon, A., Roberts, A., Manilow, E., Esling, P., Agostinelli, A., Verzetti, M., Simon, I., Pietquin, O., Zeghidour, N., 等. (2023).Singsong：从歌唱中生成音乐伴奏。arXiv 预印本 arXiv:2301.12662 .

[47] Du, Y., Zhao, S., Ma, M., Chen, Y., & Qin, B. (2023).分析大型语言模型的固有响应倾向：基于现实世界指令的越狱。arXiv 预印本 arXiv:2312.04127 .[48] Feffer, M., Heidari, H., & Lipton, Z. C. (2023).道德机器还是多数人的暴政? 在《人工智能协会会议论文集》中，37 (55), 5974–5982.

[49] Feffer, M., Lipton, Z. C., & Donahue, C. (2023). Deepdrake ft. bts-gan 和 taylorvc: 对音乐深度伪造和托管平台的探索性分析。在2023年人本音乐信息检索第二届研讨会（HCMIR 2023）上。

[50] Feffer, M., Skirpan, M., Lipton, Z., & Heidari, H. (2023).从偏好引导到参与式机器学习：对未来研究的批判性调查与指南。在2023年AAAI/ACM人工智能、伦理与社会会议上，(第38–48页)。

[51] Field, H. (2024).OpenAI悄然取消对其人工智能工具军事用途的禁令。 网址 https://www.cnbc.com/2024/01/16/openai-quietly-removes-ban-on-military-use-of-its-ai-tools.html

[52] Friedler, S., Singh, R., Blili-Hamelin, B., Metcalf, J., & Chen, B. J. (2023).人工智能红队测试并不是解决人工智能危害的灵丹妙药：关于使用红队测试进行人工智能问责的建议。 数据与社会 .

[53] Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., 等. (2022).红队测试语言模型以减少危害：方法、扩展行为和经验教训。arXiv 预印本 arXiv:2209.07858 .

[54] Ge, S., Zhou, C., Hou, R., Khabsa, M., Wang, Y.-C., Wang, Q., Han, J., & Mao, Y. (2023).Mart：通过多轮自动红队测试提高大型语言模型的安全性。arXiv 预印本 arXiv:2311.07689 .

[55] Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). Realtoxicityprompts: 评估语言模型中的神经毒性退化。在计算语言学协会的发现：EMNLP 2020中，(第3356–3369页)。

[56] Ghosh, S., & Caliskan, A. (2023).ChatGPT在机器翻译中延续性别偏见，并忽视非性别化代词：在孟加拉语和其他五种低资源语言中的发现。
在2023年AAAI/ACM人工智能、伦理与社会会议的会议录中，AIES ’23，（第901–912页）。美国纽约：计算机协会。网址https://doi.org/10.1145/3600211.3604672

[57] 龚, Y., 冉, D., 刘, J., 王, C., 丛, T., 王, A., 段, S., & 王, X. (2023)。
Figstep：通过排版视觉提示越狱大型视觉-语言模型。 arXiv 预印本 arXiv:2311.05608 .

[58] 格林布拉特, R., 施列里斯, B., 萨昌, K., & 罗杰, F. (2023)。人工智能控制：尽管存在故意的颠覆，仍在改善安全性。arXiv 预印本 arXiv:2312.06942 .

[59] Guha, N., Lawrence, C., Gailmard, L. A., Rodolfa, K., Surani, F., Bommasani, R., Raji, I.,Cuéllar, M.-F., Honigsberg, C., Liang, P., 等 (2023).人工智能监管有其自身的对齐问题：披露、注册、许可和审计的技术和制度可行性。乔治·华盛顿法律评论，待发表.

[60] Haim, A., Salinas, A., & Nyarko, J. (2024). 名字里有什么？审计大型语言模型以检测种族和性别偏见。 arXiv 预印本 arXiv:2402.14875 .

[61] 何, J., 冯, W., 闵, Y., 易, J., 唐, K., 李, S., 张, J., 陈, K., 周, W., 谢, X., 等.
(2023).控制人工智能在科学中潜在滥用的风险。arXiv 预印本 arXiv:2312.06632 .

[62] 霍夫曼, V., 卡卢里, P. R., 朱拉夫斯基, D., & 金, S. (2024).方言偏见预测人工智能对人们性格、就业能力和犯罪性的决策。arXiv 预印本 arXiv:2403.00742 .[63] 霍维茨, E. (2022).在视野中：互动和组合深度伪造。在2022年国际多模态交互会议的论文集中，(第653–661页)。

[64] 黄, K., 刘, X., 郭, Q., 孙, T., 孙, J., 王, Y., 周, Z., 王, Y., 滕, Y., 邱, X., 等. (2023).Flames: 基准测试中国大型语言模型的价值对齐.
arXiv预印本 arXiv:2311.06899 .

[65] 黄, Y., 古普塔, S., 夏, M., 李, K., & 陈, D. (2023).通过利用生成的方式对开源大型语言模型进行灾难性越狱.arXiv预印本 arXiv:2310.06987 .

[66] 胡, C., 费塔胡, B., & 加迪拉朱, U. (2019).理解和减轻众包收集主观判断中的工作者偏见.在2019年CHI计算机与人类因素会议的会议录中, (第1–12页).

[67] Inie, N., Stray, J., & Derczynski, L. (2023).召唤一个恶魔并将其束缚：一种关于大型语言模型红队测试的扎根理论。arXiv 预印本 arXiv:2311.06237 .

[68] Jain, N., Schwarzschild, A., Wen, Y., Somepalli, G., Kirchenbauer, J., Chiang, P.-y., Goldblum,M., Saha, A., Geiping, J., & Goldstein, T. (2023).针对对齐语言模型的对抗性攻击的基线防御。arXiv 预印本 arXiv:2309.00614 .

[69] Kenthapadi, K., Lakkaraju, H., & Rajani, N. (2023).生成式人工智能与负责任的人工智能：实际挑战与机遇。在第29届ACM SIGKDD知识发现与数据挖掘会议论文集中, (第5805–5806页)。

[70] Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., & Horton, J. (2013).众包工作的未来。在2013年计算机支持的协作工作会议论文集中, (第1301–1318页)。

[71] Kleinberg, J., & Raghavan, M. (2021).算法单一文化与社会福利。美国国家科学院院刊, 118 (22), e2018340118。

[72] Knearem, T., Khwaja, M., Gao, Y., Bentley, F., & Kliman-Silver, C. E. (2023)。探索设计工具的未来：人工智能在用户体验专业工具中的作用。在2023年CHI计算机与人类因素会议的扩展摘要中，(第1–6页)。

[73] Lambert, N., & Calandra, R. (2023).对齐上限：来自人类反馈的强化学习中的目标不匹配。arXiv预印本 arXiv:2311.00168 .

[74] Lambert, N., Gilbert, T. K., & Zick, T. (2023).纠缠的偏好：强化学习和人类反馈的历史与风险。arXiv预印本 arXiv:2310.13595 .[75] Lapid, R., Langberg, R., & Sipper, M. (2023).打开芝麻！大型语言模型的通用黑箱越狱。arXiv 预印本 arXiv:2309.01446 .

[76] 李, A., 白, X., 普雷斯, I., 瓦滕伯格, M., 库默费尔德, J. K., & 米哈尔切亚, R. (2024).对齐算法的机制理解：以dpo和毒性为案例研究。 arXiv预印本 arXiv:2401.01967 .

[77] 李, D., 李, J., 哈, J.-W., 金, J.-H., 李, S.-W., 李, H., & 宋, H. O. (2023).通过贝叶斯优化实现查询高效的黑箱红队测试。arXiv 预印本 arXiv:2305.17444 .[78] 莱克, J., & 苏茨克弗, I. (2023). 引入超级对齐。
网址https://openai.com/blog/introducing-superalignment

[79] Levenson, E. (2014).运输安全管理局从事的是‘安全表演’，而不是安全。大西洋杂志.

[80] Lewis, A., & White, M. (2023).通过知识蒸馏减轻大型语言模型的危害，以用于虚拟博物馆导览。在第1届大型语言模型驯化研讨会：交互助手时代的可控性的会议记录中，(第31–45页)。

[81] Li, H., Guo, D., Fan, W., Xu, M., & Song, Y. (2023).对ChatGPT的多步骤越狱隐私攻击。arXiv预印本 arXiv:2304.05197 .

[82] Li, X., Zhou, Z., Zhu, J., Yao, J., Liu, T., & Han, B. (2023).深度催眠：使大型语言模型成为越狱者。arXiv预印本 arXiv:2311.03191 .

[83] Liao, Q. V., Subramonyam, H., Wang, J., & Wortman Vaughan, J. (2023).设计理解：支持AI驱动用户体验设计构思的信息需求。在2023年CHI计算机与人类因素会议论文集中, (第1–21页)。

[84] Liu, X., Zhu, Y., Lan, Y., Yang, C., & Qiao, Y. (2023).查询相关图像越狱大型多模态模型。arXiv预印本 arXiv:2311.17600 .

[85] Liu, Y., Deng, G., Xu, Z., Li, Y., Zheng, Y., Zhang, Y., Zhao, L., Zhang, T., & Liu, Y. (2023).通过提示工程越狱ChatGPT：一项实证研究。arXiv预印本 arXiv:2305.13860 .

[86] Luccioni, S., Akiki, C., Mitchell, M., & Jernite, Y. (2023).稳定偏见：评估扩散模型中的社会表现。在神经信息处理系统进展（NeurIPS）2023数据集与基准轨道。网址https://openreview.net/forum?id=qVXYU3F017

[87] Ma, C., Yang, Z., Gao, M., Ci, H., Gao, J., Pan, X., & Yang, Y. (2023).红队游戏：一种针对语言模型的博弈论框架。arXiv预印本 arXiv:2310.00322。[88] Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020).共同设计检查清单，以理解围绕人工智能公平性的组织挑战和机遇。在2020年CHI计算机与人类因素会议的会议记录中，(第1–14页)。

[89] Mehrabi, N., Goyal, P., Dupuy, C., Hu, Q., Ghosh, S., Zemel, R., Chang, K.-W., Galstyan, A., & Gupta, R. (2023).Flirt: 上下文中的反馈循环红队测试。arXiv预印本arXiv:2308.04265 .

[90] Mehrabi, N., Goyal, P., Ramakrishna, A., Dhamala, J., Ghosh, S., Zemel, R., Chang, K.-W., Galstyan, A., & Gupta, R. (2023).Jab: 联合对抗提示和信念增强。
在大型基础模型的少样本和零样本学习的鲁棒性（R0-FoMo）
NeurIPS研讨会。

[91] Mehrotra, A., Zampetakis, M., Kassianik, P., Nelson, B., Anderson, H., Singer, Y., & Karbasi, A. (2023).攻击树：自动越狱黑箱大型语言模型。arXiv 预印本arXiv:2312.02119。

[92] Mei, A., Levy, S., & Wang, W. Y. (2023).Assert：用于评估大型语言模型鲁棒性的自动安全场景红队测试。在2023年自然语言处理实证方法会议。

[93] Mei, K., Fereidooni, S., & Caliskan, A. (2023).在掩蔽语言模型和下游情感分类任务中对93个污名化群体的偏见。在2023年ACM公平性、问责制与透明度会议论文集中（第1699–1710页）。

[94] 微软（2023年）。为大型语言模型（llms）及其应用规划红队测试
- Azure OpenAI服务。
网址 https://learn.microsoft.com/zh-cn/azure/ai-services/openai/concepts/red-teaming

[95] 微软（2024年）。Bing中的Copilot：我们对负责任人工智能的看法。
网址 https://support.microsoft.com/Zh-cn/topic/copilot-in-bing-our-approach-to-responsible-ai-45b5eae8-7466-43e1-ae98-b48f8ff8fd44

[96] Mok, A.（2023年）。如果你要求ChatGPT永远重复一个词，它将不再遵从——最近的提示揭示了训练数据和个人信息。
网址 https://www.商业内幕.com/chatgpt-ai-refuse-to-respond-prompt-asking-repeat-word-forever-2023-12

[97] Mu, N., Chen, S., Wang, Z., Chen, S., Karamardian, D., Aljeraisy, L., Hendrycks, D., & Wagner, D. (2023).大型语言模型能否遵循简单规则？arXiv 预印本 arXiv:2311.04235 .[98

] Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo,C. A., Wallace, E., Tramèr, F., & Lee, K. (2023).从（生产）语言模型中可扩展地提取训练数据。arXiv 预印本 arXiv:2311.17035 .

[99] Neel, S., & Chang, P. (2023).大型语言模型中的隐私问题：一项调查。arXiv 预印本 arXiv:2312.06717 .

[100] Nguyen, C., Morgan, C., & Mittal, S. (2022).海报 cti4ai：红队测试人工智能模型后的威胁情报生成与共享。在2022年ACM SIGSAC计算机与通信安全会议论文集中，（第3431–3433页）。

[101] Omrani Sabbaghi, S., Wolfe, R., & Caliskan, A. (2023).在交叉背景下评估语言模型的偏见态度关联。在2023年AAAI/ACM人工智能、伦理与社会会议论文集中，（第542–553页）。

[102] OpenAI (2023).
网址https://cdn.openai.com/papers/gpt-4-system-card.pdf

[103] Pelrine, K., Taufeeque, M., Zając, M., McLean, E., & Gleave, A. (2023).利用新颖的 gpt-4 api。arXiv 预印本 arXiv:2312.14302 .

[104] Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., & Irving, G. (2022).用语言模型对语言模型进行红队测试。在2022年自然语言处理实证方法会议论文集中，(第3419–3448页)。

[105] Pfau, J., Infanger, A., Sheshadri, A., Panda, A., Michael, J., & Huebner, C. (2023).利用反向语言模型引导语言模型行为。在社会责任语言模型研究（SoLaR）研讨会于NeurIPS上。

[106] Price, E. (2023).要求chatgpt“永远”重复单词可能违反openai的条款。
网址 https://www.pcmag.com/news/asking-chatgpt-to-repeat-words-forever-may-violate-openais-terms

[107] Puttaparthi, P. C. R., Deo, S. S., Gul, H., Tang, Y., Shang, W., & Yu, Z. (2023).通过多语言查询对chatgpt可靠性的综合评估。arXiv预印本arXiv:2312.10524 .

[108] Qi, X., Huang, K., Panda, A., Wang, M., & Mittal, P. (2023).视觉对抗示例越狱对齐的大型语言模型。在第二届对抗机器学习新前沿研讨会上。

[109] Qiu, H., Zhang, S., Li, A., He, H., & Lan, Z. (2023).潜在越狱：评估大型语言模型文本安全性和输出鲁棒性的基准。arXiv 预印本arXiv:2307.08487。

[110] Radharapu, B., Robinson, K., Aroyo, L., & Lahoti, P. (2023).Aart：基于AI的红队测试用于新型大型语言模型应用的多样数据生成。在2023年自然语言处理实证方法会议：行业轨道上，（第380–395页）。

[111] Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., 等. (2021)。扩展语言模型：方法、分析及训练Gopher的见解。arXiv预印本 arXiv:2112.11446 .

[112] Rajani, N., Lambert, N., & Tunstall, L. (2023)。红队测试大型语言模型。网址https://huggingface.co/blog/red-teaming

[113] Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020)。填补人工智能问责差距：定义一个端到端的内部算法审计框架。在2020年公平、问责与透明度会议的论文集中，(第33–44页)。

[114] Rakova, B., Yang, J., Cramer, H., & Chowdhury, R. (2021)。负责任的人工智能如何与现实相遇：从业者对转变组织实践的促进因素的看法。ACM人机交互会议论文集, 5(计算机支持的协作工作1), 1–23.

[115] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022).基于文本条件的层次图像生成与剪辑潜变量。arXiv预印本 arXiv:2204.06125 , 1(2), 3.

[116] Rando, J., Paleka, D., Lindner, D., Heim, L., & Tramer, F. (2022).对稳定扩散安全过滤器的红队测试。在NeurIPS机器学习安全研讨会上。

[117] Rando, J., & Tramèr, F. (2023).来自被污染人类反馈的通用越狱后门。arXiv 预印本 arXiv:2311.14455 .

[118] Rao, A., Vashistha, S., Naik, A., Aditya, S., & Choudhury, M. (2023).欺骗大型语言模型以不服从：理解、分析和防止越狱。arXiv 预印本 arXiv:2305.14965 .[119] Rastogi, C., Tulio Ribeiro, M., King, N., Nori, H., & Amershi, S. (2023).支持人机协作审计大型语言模型。在2023年AAAI/ACM人工智能、伦理与社会会议的论文集中，(第913–926页)。

[120] Robey, A., Wong, E., Hassani, H., & Pappas, G. (2023).Smoothllm：防御大型语言模型免受越狱攻击。在大型基础模型的少样本和零样本学习的鲁棒性研讨会（R0-FoMo）上，NeurIPS。

[121] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022)。高分辨率图像合成与潜在扩散模型。在2022 IEEE/CVF计算机视觉与模式识别会议（CVPR），（第10674–10685页）。美国路易斯安那州新奥尔良：IEEE。网址https://ieeexplore.ieee.org/document/9878449/

[122] Röttger, P., Kirk, H. R., Vidgen, B., Attanasio, G., Bianchi, F., & Hovy, D. (2023)。Xstest：用于识别大型语言模型中夸大安全行为的测试套件。 arXiv 预印本 arXiv:2308.01263 .

[123] Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. (2021). Hatecheck: 仇恨言论检测模型的功能测试。在第59届计算语言学协会年会和第11届国际自然语言处理联合会议（第1卷：长篇论文）的会议记录中，(第41–58页)。

[124] Roy, S., Harshvardhan, A., Mukherjee, A., & Saha, P. (2023).探测大型语言模型的仇恨言论检测：优势与漏洞。在计算语言学协会的发现：EMNLP 2023中，(第6116–6128页)。

[125] Roy, S. S., Naragam, K. V., & Nilizadeh, S. (2023).使用chatgpt生成网络钓鱼攻击。arXiv预印本 arXiv:2305.05133 .

[126] Roy, S. S., Thota, P., Naragam, K. V., & Nilizadeh, S. (2023).从聊天机器人到钓鱼机器人？–防止使用chatgpt、谷歌巴德和Claude创建的网络钓鱼骗局。arXiv预印本 arXiv:2310.19181 .

[127] Salem, A., Paverd, A., & Köpf, B. (2023).Maatphor：用于提示注入攻击的自动变体分析。arXiv预印本 arXiv:2312.11513 .

[128] Schlarmann, C., & Hein, M. (2023).关于多模态基础模型的对抗鲁棒性。在IEEE/CVF国际计算机视觉会议的会议录中, (第3677–3685页)。

[129] Schuett, J., Dreksler, N., Anderljung, M., McCaffary, D., Heim, L., Bluemke, E., & Garfinkel, B. (2023)。朝着AGI安全和治理的最佳实践：专家意见调查。arXiv预印本 arXiv:2305.07153。

[130] Schulhoff, S. V., Pinto, J., Khan, A., Bouchard, L.-F., Si, C., Anati, S., Tagliabue, V., Kost, A. L., Carnahan, C. R., & Boyd-Graber, J. L. (2023)。忽略这个标题并hackaprompt：通过全球提示黑客竞赛揭示大型语言模型的系统性漏洞。在2023年自然语言处理实证方法会议上。[131] Shah, R., Montixi, Q. F., Pour, S., Tagade, A., & Rando, J. (2023)。通过角色调制实现可扩展和可转移的黑箱越狱攻击针对语言模型。在NeurIPS的社会责任语言模型研究（SoLaR）研讨会上。

[132] Shayegani, E., Dong, Y., & Abu-Ghazaleh, N. (2023)。分段越狱：对多模态语言模型的组合对抗性攻击。在第十二届国际学习表征会议上。

[133] Shayegani, E., Mamun, M. A. A., Fu, Y., Zaree, P., Dong, Y., & Abu-Ghazaleh, N. (2023)。对抗性攻击揭示的大型语言模型中的漏洞调查。 arXiv预印本 arXiv:2310.10844 .

[134] 沈, X., 陈, Z., 巴克斯, M., 沈, Y., & 张, Y. (2023)."现在可以做任何事情": 对大型语言模型中的越狱提示进行特征化和评估。arXiv 预印本arXiv:2308.03825 .

[135] 谢夫兰, T., 法夸尔, S., 加芬克尔, B., 方, M., 惠特尔斯通, J., 梁, J., 科科塔伊洛, D., 马尔沙尔, N., 安德尔荣, M., 科尔特, N., 等. (2023).极端风险的模型评估。arXiv 预印本 arXiv:2305.15324 .

[136] 施, Z., 王, Y., 尹, F., 陈, X., 常, K.-W., & 谢, C.-J. (2023).红队测试语言模型检测器与语言模型。arXiv 预印本 arXiv:2305.19713。[137] Solaiman, I. (2023).生成式人工智能发布的梯度：方法与考虑。在2023年ACM公平性、问责制与透明度会议论文集中，(第111–122页)。

[138] Srivastava, A., Ahuja, R., & Mukku, R. (2023).没有冒犯：从语言模型中引发冒犯性内容。arXiv 预印本 arXiv:2310.00892。

[139] Stone, B., & Bergen, M. (2024).OpenAI与美国军方合作开发网络安全工具。
网址https://time.com/6556827/openai-us-military-cybersecurity/[

140] Sun, Z., Shen, Y., Zhou, Q., Zhang, H., Chen, Z., Cox, D., Yang, Y., & Gan, C. (2023).从零开始的语言模型原则驱动自我对齐，最小化人类监督。 arXiv预印本 arXiv:2305.03047 .

[141] Tan, S., Joty, S., Baxter, K., Taeihagh, A., Bennett, G. A., & Kan, M.-Y. (2021).自然语言处理系统的可靠性测试。在第59届计算语言学协会年会暨第11届国际联合自然语言处理会议（第1卷：长篇论文）的会议记录中，（第4153–4169页）。

[142] 前沿模型论坛（FMF）(2023)。前沿模型论坛：什么是红队测试?
网址https://www.frontiermodelforum.org/uploads/2023/10/FMF-AI-Red-Teaming.pdf[

143] 白宫 (2023)。关于人工智能安全、可靠和可信的发展与使用的行政命令。

URL https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

[144] Tian, Y., Yang, X., Zhang, J., Dong, Y., & Su, H. (2023)。邪恶天才：深入探讨基于大型语言模型的代理的安全性。arXiv 预印本 arXiv:2311.11855。

[145] Tong, S., Jones, E., & Steinhardt, J. (2023)。使用语言模型大规模生产多模态系统的失败。arXiv 预印本 arXiv:2306.12105。

[146] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N.,Batra, S., Bhargava, P., Bhosale, S., 等. (2023).Llama 2：开放基础和微调聊天模型。arXiv 预印本 arXiv:2307.09288 .

[147] Toxtli, C., Suri, S., & Savage, S. (2021).量化众包工作中的隐形劳动。
ACM 人机交互会议录 , 5(计算机支持的协作工作2), 1–26.

[148] Tsai, Y.-L., Hsu, C.-Y., Xie, C., Lin, C.-H., Chen, J.-Y., Li, B., Chen, P.-Y., Yu, C.-M., & Huang, C.-Y. (2023).响铃！去除扩散模型概念的方法有多可靠？arXiv 预印本 arXiv:2310.10012 .

[149] Tu, H., Cui, C., Wang, Z., Zhou, Y., Zhao, B., Han, J., Zhou, W., Yao, H., & Xie, C. (2023).
这张图片里有多少只独角兽？视觉大语言模型的安全评估基准。 arXiv
预印本 arXiv:2311.16101 .

[150] Vaughan, J. W. (2017).更好地利用人群：众包如何推动机器学习研究。J. Mach. Learn. Res. , 18(1), 7026–7071.

[151] Wan, Y., & Chang, K.-W. (2024).男性首席执行官与女性助理：通过配对刻板印象测试探讨文本到图像模型中的性别偏见。arXiv 预印本 arXiv:2402.11089 .[152] 王, B., 陈, W., 裴, H.,

谢, C., 康, M., 张, C., 许, C., 熊, Z., 符, R.,谢弗, R., 等. (2023).解码信任：对 GPT 模型可信度的全面评估。arXiv 预印本 arXiv:2306.11698 .

[153] 王, H., & 书, K. (2023).后门激活攻击：使用激活引导攻击大型语言模型以实现安全对齐。arXiv 预印本 arXiv:2311.09433 .[154] 王, J., 吴, J., 陈, M., 沃罗贝伊奇克, Y., &

肖, C. (2023).关于使用人类反馈的强化学习对大型语言模型的可利用性。arXiv 预印本arXiv:2311.09641 .

[155] 王, Y., 滕, Y., 黄, K., 吕, C., 张, S., 张, W., 马, X., & 王, Y. (2023).
虚假对齐：大型语言模型真的对齐得好吗？arXiv 预印本 arXiv:2311.05915 .

[156] 王, Z., 杨, F., 王, L., 赵, P., 王, H., 陈, L., 林, Q., & 黄, K.-F. (2023).
自我保护：赋能大型语言模型自我保护。arXiv 预印本 arXiv:2310.15851 .[157] 魏,

A., 哈赫塔拉布, N., & 斯坦哈特, J. (2023).越狱：大型语言模型安全训练失败的原因是什么？在第三十七届神经信息处理系统会议上。[158] 魏, Z., 王, Y., & 王, Y. (2023).仅通过

少量上下文示例进行越狱和保护对齐语言模型。arXiv 预印本 arXiv:2310.06387 .

[159] Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J.,Bergman, S., Kay, J., Griffin, C., Bariach, B., 等. (2023).生成式人工智能系统的社会技术安全评估。arXiv 预印本 arXiv:2310.11986 .[160] Weir, A. (201

4).火星人. 随机出版社.

[161] Widder, D. G., West, S., & Whittaker, M. (2023).开放（为商业服务）：大科技、集中权力与开放人工智能的政治经济。SSRN 预印本 10.2139/ssrn.4543807 . URLhttps://papers.ssrn.com/abstract=4543807

[162] Wood, B. J., & Duggan, R. A. (2000).先进信息保障概念的红队测试。
在DARPA信息生存会议和博览会的会议记录中。DISCEX’00 , 第2卷, (第112–118页)。IEEE 。

[163] Wu, F., Liu, X., & Xiao, C. (2023).Deceptprompt：通过对抗性自然语言指令利用大语言模型驱动的代码生成。arXiv预印本 arXiv:2312.04730。[164] Wu, Y., Li, X., Liu,

Y., Zhou, P., & Sun, L. (2023).通过系统提示的自对抗攻击越狱gpt-4v。arXiv预印本 arXiv:2311.09127。

[165] Xie, Y., Yi, J., Shao, J., Curl, J., Lyu, L., Chen, Q., Xie, X., & Wu, F. (2023).通过自我提醒防御ChatGPT的越狱攻击。自然机器智能, (第1–11页)。

[166] Xu, J., Ju, D., Li, M., Boureau, Y.-L., Weston, J., & Dinan, E. (2020)。开放域聊天机器人安全的配方。arXiv预印本 arXiv:2010.07079 .

[167] Xu, J., Ju, D., Li, M., Boureau, Y.-L., Weston, J., & Dinan, E. (2021)。针对安全对话代理的机器人对抗对话。在2021年北美计算语言学协会人类语言技术会议论文集, (第2950–2968页)。

[168] Xu, N., Wang, F., Zhou, B., Li, B. Z., Xiao, C., & Chen, M. (2023)。认知过载：通过过载逻辑思维越狱大型语言模型。arXiv 预印本arXiv:2311.09827 .

[169] 杨, Y., 惠, B., 袁, H., 龚, N., & 曹, Y. (2024).Sneakyprompt: 越狱文本到图像生成模型。在2024 IEEE 安全与隐私研讨会 (SP), (第 123–123 页). IEEE 计算机学会.

[170] 姚, D., 张, J., 哈里斯, I. G., & 卡尔松, M. (2023).Fuzzllm: 一种新颖且通用的模糊测试框架，用于主动发现大型语言模型中的越狱漏洞。
arXiv 预印本 arXiv:2309.05274 .

[171] 姚, Y., 段, J., 许, K., 蔡, Y., 孙, E., & 张, Y. (2023).关于大型语言模型（LLM）安全与隐私的调查：好、坏与丑。arXiv 预印本arXiv:2312.02003 .

[172] Yong, Z. X., Menghini, C., & Bach, S. (2023).低资源语言越狱 GPT-4。在社会责任语言模型研究（SoLaR）研讨会于 NeurIPS . [173] Yu, J., Lin, X., & Xing, X. (2023).Gptfuzzer：使用自动生成的越狱提示对大型语言模型进行红队测试。arXiv 预印本 arXiv:2309.10253 .

[174] Yuan, Y., Jiao, W., Wang, W., Huang, J.-t., He, P., Shi, S., & Tu, Z. (2023).Gpt-4太聪明了，无法安全：通过密码与大型语言模型进行隐秘对话。arXiv预印本 arXiv:2308.06463 .[175] Zenko, M. (2015).红队：如何通过像敌人一样思考取得成功. 基本书籍.

[176] Zhang, J., Zhou, Y., Hui, B., Liu, Y., Li, Z., & Hu, S. (2023).Trojansql：针对自然语言数据库接口的SQL注入。在2023年自然语言处理实证方法会议论文集中，(第4344–4359页)。

[177] Zhang, M., Pan, X., & Yang, M. (2023).Jade：基于语言学的安全评估平台
用于大型语言模型。arXiv预印本 arXiv:2311.00286 .

[178] 张, X., 张, C., 李, T., 黄, Y., 贾, X., 谢, X., 刘, Y., & 沈, C. (2023).
一种基于突变的多模态越狱攻击检测方法。arXiv 预印本
arXiv:2312.10766 .

[179] 张, Z., 杨, J., 柯, P., & 黄, M. (2023).通过目标优先级来防御大型语言模型的越狱攻击。arXiv 预印本 arXiv:2311.09096 .[180] 赵, W., 李, Z., & 孙, J. (2023).评估大型语言模型安全性的因果分析。arXiv 预印本 arXiv:2312.07876 .

[181] 朱, S., 张, R., 安, B., 吴, G., 巴罗, J., 王, Z., 黄, F., Nenkova, A., & 孙, T.
(2023).Autodan：针对大型语言模型的自动化和可解释的对抗性攻击。
在社会责任语言模型研究（SoLaR）研讨会于NeurIPS上。[182] 朱, Z., 王, J., 程, H., & 刘, Y. (2023)。揭示和改善数据可信度：一项关于训练无害语言模型的数据集研究。arXiv预印本 arXiv:2311.11202。[183] 朱, T. Y., 黄, Y., 陈, C., & 邢, Z. (2023)。通过越狱对chatgpt进行红队测试：
偏见、鲁棒性、可靠性和毒性。arXiv预印本 arXiv:2301.12867, (第12–2页)。

[184] 邹, A., 王, Z., 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性攻击。arXiv 预印本 arXiv:2307.15043 .

研究调查和案例研究细节

本附录包含关于作为本研究一部分探索的研究论文和案例研究的进一步细节。表4包含我们研究调查中恢复的每项工作的每个维度的具体分类，详见第4节。我们还提供了一个谷歌表格项目的访问权限，其中包含对案例研究和研究论文的笔记和主题分析。该项目可以通过以下网址访问：https://docs.google.com/spreadsheets/d/1cZPc6Alkf8sqOFMsEvZgI2PzX2tHTIbemMa6sq4J2Qk/edit?usp=sharing.

		调查的风险类型			
使用的方式类型		异议性	同意性	两者	两者都不是
	暴力破解	[183, 167, 166, 123, 55, 47, 130, 64, 82, 124, 158, 134, 118, 85, 165]	[36, 97, 125, 81]	[53]	无
	暴力破解 + 人工智能	[109, 140, 77, 89, 14, 173, 34, 41, 92, 13, 90, 131, 105, 122, 179, 45, 155, 156, 33, 170, 68, 6, 177]	[24, 136, 58, 163, 182, 149, 126]	[104, 119, 138, 157, 61, 164, 80, 127, 144, 152]	[27, 110]
	算法搜索	[37, 87, 26, 148, 181, 54, 32, 91, 120, 22, 75, 169]	[100]	[178]	无
	针对性攻击	[116, 172, 65, 180, 23, 154, 153, 174, 184, 76, 107, 84, 168, 57, 44, 132, 42, 108, 117]	[176, 145, 11, 128, 25, 98, 3]	[103, 18]	无

表4：根据每篇论文所产生的内容类型和使用的方式类型，对我们调查中获得的论文进行深入分类。详见第4节以获取详细信息和定义。

B 扩展 NIST RFI 评论分析

根据行政命令的指示，国家标准与技术研究院（NIST）作为商务部的一部分，发布了一份信息请求（RFI）[6]以征求公众对如何最好地执行所需工作的意见和建议。具体而言，RFI寻求对人工智能红队测试、内容水印和制定人工智能开发标准的反馈。由于这些问题中的第一和最后一个与我们的工作相关，我们选择分析提交的评论[7]以观察与我们的发现的相似性和差异。在此过程中，我们试图了解谁在评论，他们的回应涉及哪些问题，以及他们如何构建他们的论点。

总体趋势。评论范围从匿名个人的简短文本回复到来自民间社会组织、政府机构、学术团体、科技初创公司和大型软件公司的数十页技术报告的PDF（在本节中，我们将最后两组统称为“行业”）。个人的简短评论往往集中在数据权利和透明度上；它们通常情绪激动，攻击那些明知故犯地窃取数据以训练生成式AI的公司（例如“生成式人工智能应该受到严格监管……它在没有艺术家同意的情况下不道德地训练于艺术作品上。最终，生成式人工智能只适合数学应用……为了公众的安全，应该远离普通大众。让它远离艺术，这是我们唯一真正带给大众快乐的人类创造力的堡垒[sic]。”）。在行业内部，小公司的回应往往看起来像是对它们可以提供的框架、产品和基础设施的“销售宣传”，这些可以帮助NIST履行其义务。

大型公司的提交通常包括监管建议和他们内部评估方法的描述。关于民间社会组织、学术团体和政府机构，尽管一些右翼团体出于纯粹的政治原因敦促NIST不遵循拜登政府的行政命令，但总体而言，这些组织无论在政治光谱上的位置如何，都对紧迫的生成式AI问题和监管提出了有力的论点。有关提交者和提交团体的定性和定量信息，请参见表5和图1。

需要明确的红队测试定义和指南。根据我们主要论文的发现，许多公司和民间社会组织声称，行政命令中对“红队测试”的定义模糊不清，因此他们提供了自己的红队测试定义，并呼吁NIST提供明确、标准化的定义。例如，谷歌强调“红队测试”“通常被用作一个总括性术语，涵盖广泛的AI安全测试实践，这令人困惑且可能适得其反。”商业圆桌会议同样呼吁NIST制定全球红队测试标准和一致的定义。Hugging Face在其回应中引用的博客文章中提出了自己的定义[112]，指出根据他们的说法，

“红队测试提示与常规的自然语言提示相似[与对抗性机器学习提示相对]。”

[6] https://www.federalregister.gov/documents/2023/12/21/2023-28232/request-for-information-rfi-related-to-nists-assignments-under-sections-41-45-and-11-of-the

[7] https://www.regulations.gov/docket/NIST-2023-0009/comments

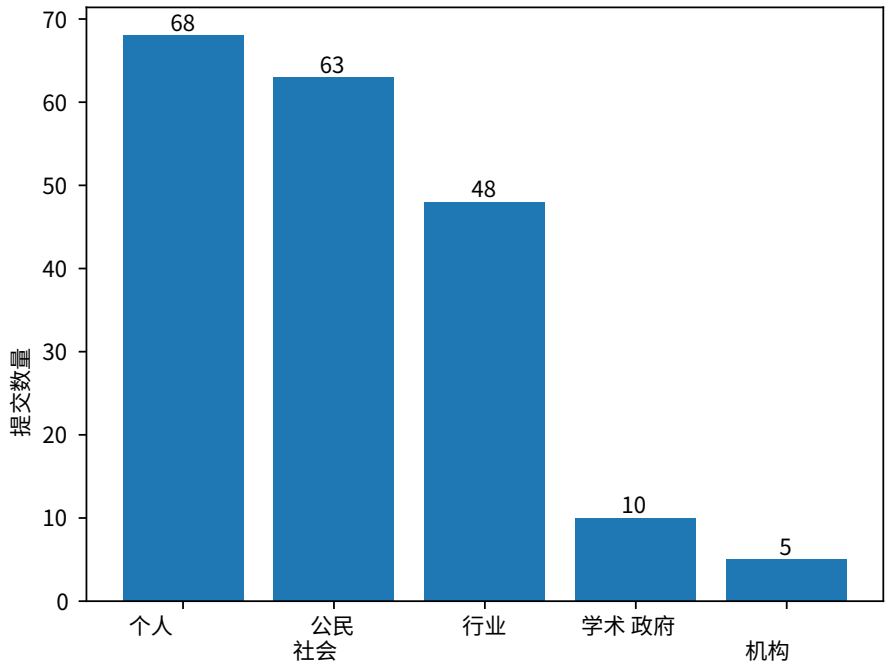

图1：按响应者类型分组提交给NIST RFI的评论频率统计。
显然，个人评论和来自公民社会的评论最为常见，但来自行业（包括初创企业和大型软件公司）的提交也相当多。
学术团体和政府机构的提交最少。

实体类型	示例
个人	匿名评论者，杰弗里·弗兰克，布里吉特·S
公民社会组织	Mozilla，数据与社会，人工智能安全中心
政府机构	独立商业全国联合会；纽约市技术与创新办公室；美国商会
学术团体	IEEE标准协会，约翰霍普金斯健康安全中心，科罗拉多大学博尔德分校
科技初创公司	Credo, OpenAI, Hugging Face
大型软件公司	谷歌，Adobe，Meta

表5：对NIST RFI的响应者类型及其对应示例。显然，各种不同的个人和组织向NIST提交了评论，但总体上，他们都涉及水印、版权侵权和红队测试等问题。

需要对模型和系统进行红队测试。来自各种利益相关方（从OpenAI到Mozilla再到消费者技术协会）的许多评论强调，NIST在未来的指南和建议中应区分人工智能模型红队测试（即尝试破解人工智能模型以寻找改进途径）和人工智能系统红队测试（即攻击模型及其数据基础设施、用户界面和其他组件）。例如，OpenAI分享了他们在迭代红队测试模型和系统方面的实践，强调他们会在产品界面发生变化时对ChatGPT系统进行红队测试，即使底层模型保持不变。

需要分享红队测试资源。在评论中反复出现对NIST的请求，要求传播有关红队测试的材料给相关组织和社区。例如，Meta建议NIST应收集案例研究、插图和/或AI红队测试的示例，以体现最佳实践或最新技术水平。

行业与公民社会在外部红队测试上的分歧。虽然在红队测试中同意需要接触多元化的观点，但许多科技公司在言辞上表达了不情愿，并推卸了进行外部红队测试的责任。例如，谷歌反复强调NIST红队测试指南需要“灵活”，并警告在红队测试中普遍接触外部专家的可行性，建议“外部红队测试仅在必要和技术可行的情况下才应被要求或推荐。”这反过来常常涉及谷歌在多大程度上会开放他们的模型，但谷歌没有提供进一步的细节来评估这种技术可行性。此外，谷歌还建议，类似于网络安全红队测试，人工智能红队测试中识别的许多漏洞“除非（1）用户需要采取行动来修复漏洞（例如，安装更新），或（2）该漏洞被恶意利用并影响了用户或客户，否则无需公开或报告。”英特尔同样建议，红队测试应首先包括技术人工智能专家和专业红队成员，然后根据需要邀请领域专家，以保持红队测试在实践中的成本和优先级。显然，这些公司及其他公司对外部红队测试的技术和组织可行性表示担忧，并且通常对其与外部红队成员的合作计划含糊不清。与行业的论点相反，民间社会组织常常敦促NIST在整个生成式人工智能生命周期中（包括从一开始）涉及外部利益相关者进行红队测试，作为监管私营部门的不可或缺的工具。例如，人工智能安全中心建议需要

1.包括具有提示工程或白帽越狱经验的人，与领域专家合作，并

2.包括代表系统预期用户群体的红队成员。

同样，Mozilla 强调除了帮助外部团队进行此类评估的工具外，还需要独立的“审计”和“红队测试”。

利用网络安全指南，同时承认人工智能面临的挑战。具有网络安全经验的组织，如谷歌、RAND和Meta，呼吁从传统网络安全实践中学习。特别是，谷歌展示了他们的人工智能红队最近在测试独立的生成式AI模型和与生成式AI集成的系统方面的努力，这些努力基于他们（传统）红队分享的实践。谷歌随后

呼吁NIST“将网络安全规范纳入其对[人工智能]红队测试的方式”并为模型开发者提供适当的时间“在报告任何测试发现之前，修复任何识别出的漏洞。”谷歌还强调了在遵循人工智能法规时，进行对抗性测试的挑战，考虑到人工智能工具的广泛使用案例，建议NIST制定关于在敏感内容领域平衡重大法律影响和技术限制的对抗性测试的建议和指南。

C 红队测试研讨会总结

基于我们案例研究分析、研究调查和NIST RFI评论分析的见解，我们举办了一次研讨会，行业、政府和学术界的专家参加了讨论他们对红队测试和评估生成式人工智能的方法。此次研讨会的主要部分由三个小组组成，其中几位成员讨论了他们及/或其组织的评估策略以及他们对生成式人工智能评估的最紧迫关注。

C.1 第一小组：红队测试研究的前沿

此次会议的第一小组由卡内基梅隆大学计算机科学学院的副教授Zico Kolter主持。Kolter教授与卡内基梅隆大学计算机科学学院的副教授Graham Neubig、斯坦福大学计算机科学助理教授Sanmi Koyejo以及卡内基梅隆大学计算机科学学院的副教授Matt Frederickson共同参与。

在第一场小组讨论中，讨论者强调了该领域的研究如何迅速转化为实际应用，显著影响我们的日常生活。他们强调了积极参与红队测试和越狱大型语言模型（LLMs）在现实场景中部署的重要性，将LLMs视为更大系统中不可或缺的软件组件，并需要特定专业知识来评估其威胁特征。此外，他们强调了人工智能研究的动态性质，要求不断纳入新发现以确保有效的红队测试。

人工智能系统的整体可信度，特别是在医疗保健和神经科学等领域，引起了关注。讨论者讨论了社会系统的复杂性，以及考虑与人工智能技术相关的潜在危害和风险的必要性，尤其是在不同的人口背景下。

小组强调了识别文本生成系统何时发生故障的挑战，以及开发评估框架的重要性。讨论者还提出了关于红队测试的性质和范围的各种问题，包括它是否应该涉及对抗性方法，或专注于发现系统中的缺陷。他们承认考虑最坏情况的重要性，并强调需要进行压力测试，以将系统推向极限。此外，关于在测试过程中接触极端内容可能导致的潜在心理困扰也提出了担忧。此外，讨论还涉及将人工智能系统整合到更大框架中的挑战，以及在各种背景下进行严格测试的必要性。小组成员强调了跨学科研究的重要性以及在人工智能开发中考虑社会技术方面的必要性。

最后，讨论强调了对基础人工智能模型的持续研究的重要性

及其应用，考虑技术和社会影响。他们强调了合作努力和持续探索的必要性，以确保人工智能技术的负责任部署。

C.2 第二小组：红队测试的行业实践

此次会议的第二小组由卡内基梅隆大学计算机科学学院的助理教授Yonatan Bisk主持。Bisk教授与Hugging Face的研究与首席伦理科学家Margaret Mitchell、卡内基梅隆大学机器学习助理教授及Abridge首席科学官Zack Lipton，以及微软研究院AI Frontiers的常务董事Ece Kamar共同参与。

在人工智能和机器学习（AI/ML）领域，关于红队测试实践的实施有多种策略和见解。一位小组成员描述了他们组织的红队测试方法。通过提供全面评估AI模型、详细数据分析和用户反馈整合等功能，他们旨在促进红队测试工作的包容性。值得注意的是，相关举措包括为各个领域的主题专家（SMEs）提供可访问的评估界面，以便在生产环境中对模型进行对抗性测试。此外，他们通过仅需三行代码的低代码评估，简化了红队测试过程。利用传统AI研究中固有的排行榜文化，他们还成功地通过红队测试举措吸引了开发者，并提供了报告模板以促进参与。

另一位小组成员强调了建立明确立法以标准化严格的AI评估和红队测试实践的重要性。这涉及到定义红队测试并建立指导从业者的框架。此外，他们强调了与利益相关者直接互动的重要性，以确保在AI开发中，特别是在高度监管的领域中，能够及时响应。

最后一位小组成员展示了他们组织在AI开发的各个阶段和方面整合红队测试的情况，包括安全漏洞、隐私风险和恶意使用案例。强调了一个超越单个AI模型的红队测试"平台"的重要性，他们提供了旨在帮助从业者进行对抗性测试的工具示例，包括Python风险识别工具（PyRIT），这是一个在GitHub上可用的开源工具，以及Hugging Face的红队抵抗基准排行榜。

展望未来，专家们一致认为，未来的红队测试工作应优先考虑赋能创造性思维和探索人类与AI协作的潜力。此外，公认需要在针对安全漏洞的红队测试与确保AI应用有效性之间取得平衡。

强调了在人工智能/机器学习社区中明确界定红队测试实践的必要性。与会者一致认为有必要建立红队测试的最佳实践，并倾向于在系统层面实施红队测试，而不仅仅关注单个模型。随着与会者倡导将任务分解为不同组件并关注各自的重点，他们一致认为红队测试是负责任的人工智能开发中一个至关重要（但并非唯一）的组成部分。通过促进包容性、标准化实践和接受多样化观点，建议的目标是确保人工智能技术在各种背景下的伦理和有效部署。

C.3 第三小组：红队测试的政策和法律影响

此次会议的第三小组由卡内基梅隆大学伦理与计算技术的K&L Gates职业发展助理教授霍达·海达里教授主持。海达里教授与谷歌DeepMind的高级研究科学家凯瑟琳·李、OpenAI政策研究的技术项目经理拉玛·艾哈迈德，以及卡内基梅隆大学海因茨信息系统与公共政策学院院长、管理科学与信息系统的威廉·W·库珀与露丝·F·库珀教授迪恩·拉马亚·克里希南共同参与。

总体而言，专家小组认为红队测试需要多方面的评估方法和缓解策略。首先，他们强调政策制定者和人工智能专家应合作制定清晰的定义和特定情境的方法以便于评估过程。一旦红队测试的定义更加明确，就可以根据可衡量的目标评估模型、系统或项目，以确定成功与风险。外部专家和利益相关者被认为对全面评估和评估至关重要。

此外，专家小组描述了在多个粒度层次和人工智能生命周期的不同阶段，评估和缓解与人工智能相关的风险是必要的。例如，人工智能系统中的记忆是一个常见的风险，常在关于缓解的讨论中提及。然而，危害是有情境的，这需要在系统及其组件上进行红队测试。

另一个潜在的经济风险涉及算法单一文化。专家小组倡导政策机制以促进最佳算法多样性，同时承认确定和执行最佳政策所面临的挑战。风险分级和现场测试，既包括部署前也包括部署后，均被提议作为潜在的方法。无论采用何种缓解策略，都应根据模型、系统或项目的具体情况量身定制，并依赖于红队测试的见解（例如，进一步微调与内容政策修改）。小组成员还指出，缓解人工智能风险的努力受到市场力量、全球市场需求（例如，欧盟更严格的风险分级可能会影响基础模型开发者遵守此类分级以便在全球市场上销售）以及各行业现有法规的驱动。

最后，小组成员总结指出，向前推进时，增强披露和问责政策将至关重要。该领域的政策专家建议制定关于披露的指南，包括对专门披露流程的建议，确保公司内部有适当的资源和责任，以及建立一个类似于计算机应急响应小组（CERT）的组织来报告人工智能问题。对从模型开发到下游微调的各种风险的责任分配的清晰认识被认为对有效的风险缓解至关重要。

这些要点强调了红队测试在人工智能开发中的复杂性和重要性，并突显了多样化评估方法、缓解策略和监管框架在确保人工智能安全和安全性方面的必要性。

C.4 结论

本次专家会议的关键点可以总结如下，进一步验证了我们论文中传达的关键信息：

- 对红队测试的功能性定义、其组成部分、范围和局限性，对于有效的红队测试是必要的。

- 生成式AI研究和实践社区必须朝着红队测试的标准和最佳实践迈进。
- 红队的组成（在背景和专业知识的多样性方面）是一个重要的考虑因素。
- 红队测试的努力应当关注更广泛的系统，而不是单个组件。
- 更广泛的政治经济（例如，市场力量、法规）将影响红队测试的实践。