前沿模型论坛：什么是红队？

介绍

红队被频繁引用为开发安全和可靠的前沿模型及人工智能系统的常见技术。然而，目前对于如何定义“人工智能红队”以及在人工智能开发生命周期中其扩展角色的哪些方法被视为一部分，仍然缺乏清晰的认识。

在网络安全中，红队是一种模拟对系统进行现实攻击的技术，以测试漏洞并理解可能的对手能力和目标。我们在前沿模型的背景下定义“红队”为一种结构化过程，用于探测人工智能系统和产品，以识别有害能力、输出或基础设施威胁。与传统的“红队”类似，人工智能红队通常涉及主动识别整个系统中的缺陷和漏洞——包括数据、基础设施、应用程序——而不仅仅是模型输出。这是推动安全、可靠和可信赖的人工智能的重要工具，帮助团队识别系统的潜在风险，以便可以应用保护措施。这也是一个迭代过程，利用演练的结果和见解来指导风险的优先级和方法，实施缓解措施，然后重新进行评估以确定缓解措施的有效性。为了帮助理解当前在更广泛的人工智能红队框架下使用的一系列技术，剩余部分将概述FMF成员进行的已知演练的案例研究。

微软案例研究：红队测试必应聊天

背景

在2022年10月，微软意识到开放AI的新GPT-4模型。由于知道微软内部的团队会对将GPT-4集成到第一方和第三方产品中感兴趣，微软创建了一个跨职能的主题专家（SMEs）团队，以实验该模型并了解其能力和风险。为识别与必应聊天相关的风险，建立了后续的红队演练，这是微软首个集成GPT-4的产品。

方法论

对GPT-4基础模型的初轮红队演练是在没有微软额外缓解措施的情况下进行的。来自公司各个领域的20多位中小企业专家，涵盖法律、政策、人工智能、工程与研究、安全和负责任的人工智能等多种专业，齐聚一堂，探讨模型以识别其风险。除了直接对模型进行实验的团队外，该小组还向一小部分国家安全领域的中小企业专家开放了模型，并与他们进行了访谈，以更好地理解特定高风险领域的风险面。红队

对GPT-4原始模型的演练更加开放和探索，目标是识别尽可能多的风险和失败模式，确定进一步调查的风险领域，并实施早期缓解策略。

在Bing聊天开发并逐渐成熟的过程中，第二轮红队测试被启动。在这一轮红队测试中，来自公司各个部门的50多位中小企业专家聚集在一起，对集成了缓解措施的应用程序进行红队测试。我们采取了迭代的方法，每周进行红队测试冲刺，每周针对优先特性和风险领域进行红队测试，记录结果，并与相关的测量和缓解团队合作，确保红队测试结果为他们的下一步行动提供信息。虽然这一轮红队测试由于有了从红队测试基础GPT-4模型中得出的风险清单而更加有针对性，但团队仍然识别出了一系列新的风险，这导致了对新风险的重新优先排序以进行进一步调查。

结果

对Bing聊天的红队测试突显了迭代测试基础模型和下游应用程序的必要性。以开放的方式对基础模型进行红队测试，并且没有额外的缓解措施，使我们能够理解模型的风险表面和失败模式，识别进一步调查的领域，以及在应用程序开发中需要注意的事项。由于许多风险依赖于上下文和应用，红队在集成缓解措施的下游应用中是必要的。虽然每周的探索性和定性红队冲刺并不能替代大规模测量和缓解工作，但识别出的例子为创建大规模测量数据集提供了种子，帮助优先考虑和实施缓解策略，并帮助利益相关者做出更明智的决策。微软继续学习并改进我们为红队建立的流程。

开放AI案例研究：GPT-4的专家红队

背景

在2022年8月，开放AI开始招募外部专家进行红队测试并提供对GPT-4的反馈。红队已在语言模型中以多种方式应用：减少有害输出并利用外部专业知识进行领域特定的对抗性测试。开放AI的方法是迭代进行红队测试，从初步假设哪些领域可能是最高风险开始，测试这些领域，并根据需要进行调整。这也是一种迭代过程，因为开放AI在引入新的缓解和控制层时使用多轮红队测试。

方法论

开放AI招募了41名研究人员和行业专业人士——主要在公平性、对齐研究、行业信任与安全、虚假/误导信息、化学、生物风险、网络安全、核风险、经济学、人机交互、法律、教育和医疗保健等领域具有专业知识——以帮助更全面地理解GPT-4模型及其潜在部署风险。开放AI选择了

这些领域是基于多个因素，包括但不限于：在语言模型和AI系统中观察到的先前风险，以及开放AI观察到的用户对语言模型应用的兴趣增加的领域。参与这一红队过程的人员是根据他们在这些风险领域的先前研究或经验进行选择的。这些专家可以访问早期版本的GPT-4以及正在开发的缓解措施的模型。这使得在开发和完善过程中可以对模型和系统级缓解措施进行测试。

结果

外部红队演练识别了初步风险，这些风险促使了安全研究和在关键领域的进一步迭代测试。开放AI通过技术缓解措施、政策和执行杠杆减少了许多识别出的风险；然而，仍然存在一些风险。虽然这次早期的定性红队演练对于深入了解像GPT-4这样复杂的新模型非常有用，但它并不是对所有可能风险的全面评估，开放AI随着时间的推移继续了解这些及其他类别的风险。红队过程的结果已在GPT-4系统卡中总结并发布。

谷歌DeepMind案例研究：对谷歌DeepMind的Gopher模型进行对抗性探测

背景

随着语言模型能力的提升，开发能够在规模上发现潜在有害内容或漏洞的方法的兴趣也在增加。对抗性测试已成为一种"红队"方法，旨在通过自动或手动探测技术的组合发现模型中的有害内容或漏洞。

虽然手动技术进行对抗性测试可能有效，但结果会根据探测者的创造力而有所不同，并可能导致对模型评估中的关键安全疏漏。为了补充现有的手动对抗性测试AI系统的方法，我们的研究论文《用语言模型进行红队测试语言模型》介绍了利用"红队"语言模型（LM）生成多样化测试集的潜力，以评估目标语言模型的响应。

方法论

探测集中在GDM的Gopher（DPG）语言模型的"对话提示"变体，并采用三阶段方法识别导致模型失败的测试用例：

1.使用指定的"红队"语言模型生成测试用例，并通过自动评分模型生成的分数确认这些测试用例可能产生有害输出

2.使用"目标"语言模型，为每个选定的测试用例生成输出

3.使用自动评分模型识别导致有害输出的测试用例

还使用其他方法生成测试用例，利用红队语言模型，例如零样本采样和对成功对抗性问题的监督学习，以得出最终测试集。最终测试集用于对DPG模型进行红队测试，评估各种危害，包括攻击性内容、数据泄露、不当联系信息生成以及对群体的分布偏见。

结果

总体而言，这种方法论展示了如何利用语言模型有效且自动地发现其他语言模型中的问题行为，作为安全测试流程的一部分。所生成的各种方法的集合大约产生了500,000个引发对话的问题，这些问题引发了攻击性的回应。红色语言模型的问题在引发攻击性回应方面的表现与之前工作中人类撰写的对抗性示例相似或更好。不同的红队测试方法，如少量示例调优、监督学习和强化学习，能够在保持测试集多样性的同时增加问题的难度。

Anthropic案例研究：前沿威胁红队测试以确保人工智能安全

背景

前沿威胁红队测试需要投入大量精力，以发现人工智能模型可能创造或加剧国家安全风险的方式。为了使模型在国家安全领域创造或显著增加风险，它必须在各种任务和子任务中生成精确和可靠的信息。换句话说，单个令人担忧的模型输出孤立存在并不足以造成现实世界的伤害。

方法论

这个过程分为三个步骤：

1.我们与领域专家合作，定义（a）可能因人工智能能力的进步而加剧的高优先级威胁模型，（b）如果克服将显著增加国家安全风险的伤害障碍，以及（c）指示模型是否存在该能力的诊断任务。

2.主题专家广泛探测模型，以评估系统的能力是否根据预定义的威胁模型创造或加剧国家安全风险。

3.红队测试的见解为可重复的定量评估和缓解措施的发展提供了信息。

在2023年的六个月里，Anthropic与顶级生物安全专家进行了超过150小时的合作，评估我们的模型在帮助寻求通过生物手段造成伤害的恶意行为者方面的能力。这种对抗性测试使我们能够开发模型能力的定量评估

以及适当的缓解措施。我们教专家如何破解我们的模型，他们使用了一个定制的安全接口，而没有在我们公共部署中活跃的信任和安全监控及执行工具。

结果

我们发现了一些关键问题。首先，目前的前沿模型有时能够以专家级别产生复杂、准确、有用和详细的知识。在我们研究的大多数领域中，这种情况并不常见。在其他领域则是如此。然而，我们发现模型在规模扩大（即变大）时能力会增强的迹象。我们还认为，模型获得工具的访问权限可能会提升它们在生物学方面的能力。综合来看，我们认为未来几代大型语言模型如果没有适当的缓解措施，可能会加速恶意行为者相对于现有工具（例如搜索引擎）滥用生物学的努力。如果不加以缓解，我们担心这些风险是短期的，意味着它们可能在接下来的两到三年内实现。

除了技术缓解措施外，我们正在建立一个披露流程，以便实验室和其他利益相关者可以向其他相关方报告国家安全风险和缓解措施。最终，我们的目标是标准化人工智能系统的红队测试过程，而这一过程目前更多的是艺术而非科学。
一个强大且可重复的过程对于确保红队测试 (a) 准确反映模型能力，以及 (b) 建立一个不同模型可以有意义比较的共享基准至关重要。