DOI 10.15302/J-SSCAE-2026.02.020

生成式人工智能红队测试技术综述

吴世忠[1*]，周碧玉[2]，汤启瑞[2]，肖文杰[2]，张潇丹[2]，汤学海[2]，彭勇[1]，虎嵩林[2]

（1. 中关村实验室，北京 100194；2. 中国科学院信息工程研究所，北京 100085）

摘要：随着大语言模型等生成式人工智能技术的广泛应用，其伴生的安全风险日益凸显。红队测试作为主动发现和评估潜在威胁的关键手段，已在学术界、产业界及政策制定部门引起广泛重视。区别于现有综述多聚焦单一模型类型且偏重攻击方法归纳，本文将研究对象从大语言模型扩展至多模态模型、智能体及检索增强生成等多类系统，以“攻击生成”与“效果评判”为主线构建新的分类视角，系统梳理了国内外代表性产业实践。首先，厘清生成式人工智能红队测试的核心概念与风险分类体系；其次，从单轮与多轮交互视角梳理面向大语言模型的自动化红队攻击技术，并将攻击效果评判方法归纳为判别式、生成式与端侧三大类；进而，探讨多模态模型、智能体及检索增强生成系统所带来的新型安全威胁与测试范式创新；最后，深入剖析该领域在攻防技术持续演进及多方协同治理等方面面临的核心挑战，并对未来研究方向进行展望。本研究旨在为我国构建生成式人工智能安全测评体系、增强风险主动防御能力提供系统化技术参考与实践支撑。

关键词：红队测试；生成式人工智能；人工智能安全；大语言模型；自动化测试

中图分类号：TP18　　**文献标识码：**A

A Survey of Red Teaming for Generative Artificial Intelligence

Wu Shizhong[1*], Zhou Biyu[2], Tang Qirui[2], Xiao Wenjie[2], Zhang Xiaodan[2], Tang Xuehai[2], Peng Yong[1], Hu Songlin[2]

(1. Zhongguancun Laboratory, Beijing 100194, China; 2. Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100085, China)

Abstract: With the proliferation of generative artificial intelligence (AI) technologies, particularly large language models(LLMs), their associated safety risks have grown increasingly pressing. As a proactive safety methodology for discovering and evaluating potential threats, red teaming has attracted significant attention in academia, industry, and among policymakers. Distinct from existing reviews that mainly focus on a single model type and emphasize the categorization of attack methods, this study expands the scope to encompass multiple system types, including LLMs, multimodal models, agents, and retrieval-augmented generation (RAG). It establishes a novel classification framework centered on the dual dimensions of “attack generation” and “effect evaluation,” while systematically reviewing representative industrial practices. First, we clarify the core concepts and risk taxonomy of generative AI red teaming. Subsequently, from the perspectives of single-turn and multi-turn interactions, we survey automated red teaming attack techniques targeting LLMs, and categorize the corresponding attack effectiveness detection methods into three major classes: discriminative, generative, and edge-side detection. Furthermore, we explore the novel safety threats and innovative testing paradigms introduced by multimodal models, AI agents, and RAG systems. Finally, we examine the persistent challenges in this field, including

收稿日期：2026-03-02；**修回日期：**2026-03-23

通讯作者：*吴世忠，中关村实验室研究员，中国工程院院士，研究方向为信息技术安全性测试、网络安全风险评估；E-mail: wusz@cae.cn

资助项目：中国工程院咨询项目“网信领域人工智能安全防范战略研究”(2025-XBZD-08)

本刊网址：sscae.engineering.org.cn

the ongoing evolution of attack and defense methods and the imperative for multi-stakeholder governance, and outline directions for future research. This study aims to provide technical references and practical insights for China in constructing a safety assessment system for generative AI and enhancing proactive risk defense capabilities.
Keywords: red teaming; generative artificial intelligence (AI); AI safety; large language models; automated testing

一、前言

近年来，以大语言模型（LLM）和多模态大模型为代表的生成式人工智能技术取得重要突破，在开放域对话、复杂指令遵循、多模态内容生成等方面展现出卓越能力，有力推动了人工智能（AI）技术的规模化应用[1-3]。然而，其能力的快速提升也伴生出幻觉输出、有害内容生成、隐私泄露、对抗攻击等一系列严峻的安全风险[4-6]，不仅影响技术可信部署，更对公共治理、行业安全与社会伦理构成潜在挑战。包括图灵奖得主在内的多位顶尖学者曾多次公开警告，生成式人工智能安全问题已超越技术层面，成为关乎人类社会发展的重大挑战[7]；相关专家更是在联合声明中强调，将AI风险管控置于全球优先事项刻不容缓[8]。因此，系统性评估并防御此类安全风险，已成为当前AI发展进程中亟待应对的核心议题[9]。

生成式人工智能的安全风险根植于其核心特性。其一，大型神经网络固有的黑盒特性，使其决策逻辑复杂难溯、输出行为可解释性差，构成安全风险评估与漏洞排查的根本性障碍。即便采用上下文学习、思维链等主流推理范式，其内部作用机制仍不透明，模型处理复杂任务的推理路径与关键依据仍难以清晰呈现。虽然已有机制解释、探针分析、因果溯源等探索方向，但是目前仍难以形成对高风险输出稳定且可迁移的解释，无法及时支撑特定风险的前置防护设计。其二，模型参数与训练数据规模突破阈值后，易呈现“智能涌现”现象[10]，展现未预期的新能力，进一步加剧行为的不可预测性；而关于部分涌现是否与评测指标离散化、统计非线性有关[11]，其形成机理尚存争议。此类特性易导致潜在高风险能力或脆弱性在特定上下文、对抗式提示或多轮交互中被意外触发，进而绕过既有安全防护机制。黑盒决策机制与（真实或表观）能力涌现特性共同构成生成式人工智能安全治理的核心难题，对可解释性、可控性与行为可预测性提出了更高要求。

在此背景下，红队测试作为主动识别和评估AI系统安全风险的关键手段，正成为保障生成式人工智能安全的必要路径，其重要性在全球形成广泛共识。美国国家标准与技术研究院（NIST）发布的《Generative AI Profile（NIST AI 600-1）》[12]、欧盟的《人工智能法案》[13]等政策法规均明确建议或要求对AI系统开展红队测试；美国NIST、英国人工智能安全研究院（AISI）、日本AISI等多家国家级AI安全机构，以及Google、Meta、Anthropic等领先企业，普遍将其纳入模型发布前的标准流程。其中，Anthropic等公司将红队测试定位为对齐科学研究的关键实证工具，用于揭示模型潜在能力与价值观约束的冲突，以推动对齐技术迭代优化[14-15]。需特别指出的是，生成式人工智能红队测试在目标、流程与技术范式上显著区别于侧重标准化任务性能量化的AI基准测试[16-20]，亦有别于面向既定系统开展对抗性测试的传统网络安全红队[21-25]。其核心任务在于识别与刻画开放世界中模糊、动态且难以预见的安全失效行为[15]。这一本质差异决定了必须在攻击、评判技术与评估标准等层面，构建适配其特性的方法论体系[15,22-24,26]。

针对生成式人工智能红队测试研究激增但系统性综述匮乏的现状，本文旨在弥补现有工作局限于特定模型（如LLM[27-28]、生成式可视媒体[29]、多模态系统[30]）、缺乏全景梳理的不足，主要贡献如下：① 拓展研究边界，突破单一模型局限，首次将生成式智能体与检索增强生成等新型系统纳入综述研究；② 创新技术视角，基于攻击生成与效果评判维度解析自动化测试机制，归纳判别式、生成式与端侧三大范式及其演进逻辑；③ 整合实践路径，提炼国内外机构经验差异，为AI安全治理体系建设提供参考。

二、生成式人工智能红队测试：概念、流程与风险分类

（一）生成式人工智能红队测试概念

红队测试起源于军事对抗演习，旨在通过模拟真实攻击行为系统性评估防御体系的脆弱性，后逐

渐应用于网络安全等领域。与旨在验证模型性能上限的传统基准测试不同，红队测试属于对抗性评估，强调通过主动、系统化的攻击手段，识别系统在真实对抗环境中的潜在风险。

从理论定位来看，生成式人工智能红队测试可纳入安全测试理论的整体框架中进行理解。首先，在测试范式层面，生成式人工智能红队测试属于压力测试范畴，与验证功能达标的符合性测试不同，其核心在于探查系统在对抗条件下的失效模式与响应边界[24,26]。其次，在风险认知层面，生成式人工智能红队测试通过模拟真实攻击者策略，系统性揭示模型安全机制的响应阈值，将潜在未知风险转化为可识别、可评估的显性风险清单。基于上述理论定位，生成式人工智能红队测试在实践中必须同时满足三重属性：攻击行为的对抗性、测试过程的系统性，以及结果输出的可评估性。

基于上述理论框架审视当前实践，尽管全球主要机构已展开相关实践，但生成式人工智能红队测试的概念界定仍存在显著差异。如表 1 所示，不同机构的定义在侧重点与完备性上各有不同：部分定义（如 NIST、OpenAI）侧重于测试的结构化过程，但未明确强调“对手模拟”这一本质特征；另一些定义（如 Microsoft、Google）虽突出了模拟真实对手，却未系统涵盖对模型内在风险机制的探查。这种概念上的分散，导致实践容易陷入为测试而测试的误区，或仅关注表面漏洞而忽视深层风险机理。为整合现有共识中的核心要素，并与前述理论定位形成呼应，本文提出一个更具整合性与方法论指导意义的定义：生成式人工智能红队测试是在受控环境中通过对抗性攻击手段发现生成式人工智能系统中的漏洞和风险，以提升系统安全性的一种技术手段。该定义以对抗性为核心，强调通过主动施压暴

表 1　主要机构对生成式人工智能红队测试的定义与核心特征对比

机构	定义 / 描述	对手模拟	对抗方式	测试漏洞
NIST（美国）	一种结构化的测试工作，用于发现 AI 系统中的缺陷和漏洞，通常在受控环境中进行，并与 AI 开发者合作[31]	未明确提及	—	发现缺陷与漏洞
AISI（日本）	一种评估方法，从攻击者的视角检查 AI 安全应对结构和对策的有效性[32]	攻击者	攻击者视角评估	识别防御缺陷
OpenAI（美国）	红队测试是指使用人员或 AI 以结构化的方式探索新系统的潜在风险[33]	未明确提及	结构化探索	识别潜在风险
Meta（美国）	结合人工和 AI 红队演练，力求了解模型在应对不同类型对抗参与者和活动时的表现[34]	不同类型对抗者	人机结合演练	评估对抗下的表现 / 缺陷
Anthropic（美国）	红队测试是提高 AI 系统安全性的关键工具；它涉及对技术系统进行对抗性测试，以识别漏洞和潜在误用[15]	未明确提及	对抗性测试	识别漏洞与潜在误用
Google（美国）	谷歌公司的红队是一支由黑客组成的团队，他们模拟各种对手，从国家行为者到机会主义黑客再到内部威胁，以发现我们产品和服务中的漏洞[35]	各类对手	模拟对手演练	发现产品与服务的漏洞
NVIDIA（美国）	系统地测试包含 AI 模型的系统，以识别构成威胁或风险的漏洞和行为[36]	未明确提及	系统化测试	识别威胁性漏洞与行为
Microsoft（美国）	模拟真实世界对手及其工具、战术和程序，以识别风险、发现盲点并改善整体安全态势[37]	真实世界对手	模拟对手战术	识别风险与盲点
IBM（美国）	红队是一种交互式测试 AI 模型的方法，以防止有害行为，包括敏感数据的泄露和生成有害、有偏见或事实不准确的内容[38]	未明确提及	交互式测试	防止有害与偏见偏差输出
本文	生成式人工智能红队测试是在受控环境中通过对抗性攻击手段发现生成式人工智能系统中的漏洞和风险，以提升系统安全性的一种技术手段	对抗性攻击者	对抗性攻击	识别漏洞与安全风险，支持治理决策

露模型脆弱点；以系统性为框架，确保测试的可复现与可扩展；以可评估性为旨归，通过风险识别为模型治理提供决策依据。此概念框架为后文技术分析与讨论奠定了统一基础。

在人员配置上，生成式人工智能红队测试需要构建融合安全研究员、AI工程师、伦理学家和领域专家的跨学科团队[26,39]，以复现真实威胁环境的复杂性。例如，Anthropic公司的前沿红队通过与网络安全、生物安全等领域专家合作，前瞻性地发现了其前沿模型中“双用途”能力的早期预警信号[26]；OpenAI通过广泛邀请外部专家参与红队测试，显著提升了测试覆盖广度与探测深度[39]。在高风险应用场景（如医疗诊断、法律咨询）中，领域专家的参与对构建贴近实际情境的攻击用例、评估专业语境下的潜在风险尤为关键。

根据测试执行方式和访问权限两大核心维度，可将生成式人工智能红队测试进一步划分为不同类别。其中，执行方式维度可分为基于专家经验的手工测试、依托工具实现规模化扫描的自动化测试、由智能体自主驱动的探索式测试三类；访问权限维度则分为白盒（完全掌握系统内部信息）、灰盒（部分系统信息可见）和黑盒（仅通过应用程序编程接口（API）交互）三类。两大维度相互补充，共同构成红队测试的方法论体系，为测试方案的科学设计提供基础框架。

（二）生成式人工智能红队测试流程

红队测试贯穿生成式人工智能系统的全生命周期，构成动态的安全防护闭环。开发期进行小规模测试以识别早期缺陷；发布前进行全面评估以验证防护机制的有效性；部署后通过社区参与和漏洞赏金等机制实现持续监测。这一全程贯通的体系，旨在推动系统安全能力的持续优化。

红队测试的有效性不仅取决于测试类型的选择，更依赖于一套结构化的实施流程，该流程通常涵盖四个核心阶段：测试规划、攻击执行、效果评估与系统改进[28,32,36]，如图1所示。在测试规划阶段，需明确测试目标、范围与资源分配，包括组建团队、准备测试环境及制定详细的测试计划。攻击执行阶段依据预设的风险场景，综合运用不同的访问权限和执行方式，生成并输入对抗性测试用例，以系统探测模型的潜在漏洞。强调威胁建模驱动的攻击设计，以应对高维、开放的输入空间以及模型对输入的极端敏感性[15,24,40]。效果评估阶段对模型输出的安全性及其潜在有害影响进行全面评判，其重点不仅在于判断单次攻击是否成功，更在于对输出内容有害程度的综合评估；该阶段产生的详细报告为后续系统改进提供直接依据。最后，系统改进阶段根据评估结果部署补丁、优化防护机制，并将测试数据纳入案例库用于持续训练和回归验证，从而形成“测试–评估–修复”的闭环迭代提升机制。其中，建立自动、准确的攻击效果检测机制，以高效评判攻击是否成功，是红队测试能否实现规模化和自动化的关键环节[41]。

在自动化红队测试场景下，上述攻击执行阶段可进一步抽象为一个由“攻击生成”与“效果评判”构成的“攻–判”闭环。攻击生成环节负责自动产生大量多样化的对抗性输入；效果评判环节则依据预设安全策略，对模型响应进行快速、客观且一致的分类。评判环节应超越简单的二元成败判断，扩展至对内容安全性、政策符合性等更宽泛有害影响的综合评估。实现高效、准确的自动化评判是红队测试规模化的核心挑战：一方面，自动化攻击可产生海量测试用例，依赖人工逐一评判在效率和成本上不可行；另一方面，自动化评判需具备高正确率和低误报率，以满足结果可靠性的严苛要

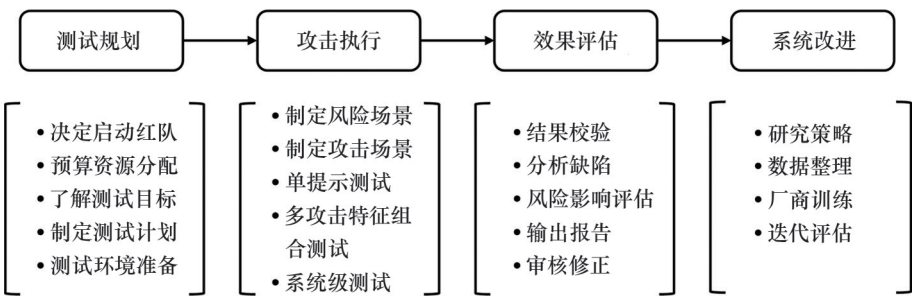

图1　红队测试流程图

求，避免因误报浪费专家复核资源或因漏报忽视严重漏洞。上述“攻–判”闭环不仅能系统挖掘模型漏洞，其持续迭代更为优化攻击策略与增强防御机制提供了数据驱动的基础，从而显著提升红队测试的整体效率与科学性。

（三）生成式人工智能内容安全风险分类体系

对生成式人工智能所引发的多样化安全风险进行系统分类，是有效开展风险评估与红队测试的前提。当前，相关风险分类框架可依据其目标与抽象程度，归纳为政策导向型、理论体系型和实践应用型三类。它们在功能上相互补充，共同构成了多层次、立体化的风险识别基础。

政策导向型框架。此类框架立足政策、伦理与社会规范，旨在确立宏观治理原则与合规方向，其核心是界定价值边界与社会责任，而非具体技术路径。例如，联合国教科文组织发布的《人工智能伦理问题建议书》通过确立价值观，以及涵盖数据政策、国际合作、环境与生态系统等11个领域的政策行动，为风险治理划定伦理基础[42]；《2025年国际人工智能安全报告》从全球治理视角系统区分了恶意使用风险、系统故障与系统性风险，并阐释其潜在的社会危害[43]。这类框架主要为后续具体风险的识别与管理提供顶层规范和目标指引。

理论体系型框架。此类研究致力于构建系统、完整的风险知识体系，强调分类的逻辑性与全面性，为学术研究提供理论基础。例如，部分研究将风险划分为歧视排斥、信息危害、虚假信息、恶意使用、人机交互危害及环境与社会经济影响六大类别[44]；此外，也有研究聚焦灾难性风险，归纳出恶意使用、AI竞赛、组织性风险与失控系统四类[45]；另有研究从风险受害主体[46-47]等新视角进行体系化梳理。全国信息安全标准化技术委员会发布的《人工智能安全标准体系（V1.0）》（征求意见稿）[48]也属于此类，围绕模型算法安全、数据安全、系统安全三类内生风险，以及网络域、现实域、认知域、伦理域四类应用风险构建了与国家网络安全标准体系衔接的整体框架，体现了在较高抽象层级上对风险的全面整合。这类方法的共同点是系统性强、理论价值高，但在落地时需结合具体场景进一步转化。

实践应用型框架。该类方法直接服务于具体安全任务，如红队测试、模型评估和内容过滤系统开发，强调可操作、可测量，通常形成可直接使用的风险列表、有害指令集和评估基准。例如，研究人员提炼出的11类典型风险并构建对抗测试指令集[49]；提出八大安全场景与七类问题，建立了易于量化的评估框架[50]；针对科学研究领域开发了SciGuard控制系统[51]；通过TrustAgent在多个阶段注入安全约束，实现了任务执行的全程安全控制[52]。这类框架实用性强，但其覆盖范围往往受限于已知用例，需与理论框架持续交互以保持更新。

综上所述，上述三类框架分别从治理、理论与实操层面构建了生成式人工智能风险分类的综合体系。在实践中，风险还可进一步从对象维度（数据、模型、系统层）[53-54]与类型维度（偏见、虚假信息、隐私侵犯等）[44,55]进行交叉细化，从而形成更立体、更具指导性的风险图谱，为红队测试与安全防御提供系统化依据。

三、大语言模型红队测试技术

LLM作为生成式人工智能的核心范式与技术底座，其应用最为广泛，相应的红队测试技术方法也最为成熟，具有显著的代表性。LLM红队测试技术正从早期依赖安全专家手工设计提示词的试探性攻击，向系统化、规模化的自动化范式演进，以应对日益复杂的安全威胁。在自动化红队测试体系中，攻击生成与效果评判构成紧密耦合的“攻–判”闭环，共同驱动测试流程的迭代与优化。本章将分别深入剖析自动化LLM红队测试中攻击生成与效果评判两大环节的核心技术路径、代表性方法及其面临的挑战。

（一）红队测试攻击技术

现有研究已从攻击阶段（如训练阶段攻击和推理阶段攻击）[40,56]、攻击者能力（如白盒、黑盒与灰盒）[56-57]、风险目标（如越狱、性能降级、隐私窃取）[24,57]及模型能力（如自回归特性、指令遵循与泛化能力）[58]等多个维度构建了分类体系，为理解红队攻击提供了宝贵的多维视角，但在攻击动态交互特性刻画方面尚存细化空间。为此，本文依据对话轮次将LLM红队攻击技术分为单轮与多轮两类[59]，并系统剖析两类交互模式下的技术特征与实施

路径。

1. 单轮红队攻击

单轮红队攻击旨在通过精心设计的提示词，在单次交互中直接突破模型的安全对齐机制以诱发有害输出。依据其实现原理，该类攻击可分为直接交互攻击和基于优化搜索的攻击两大类。

（1）直接交互攻击：该方法不依赖外部生成模型，而是基于预设启发式规则或固定模板，在单次对话中诱导目标模型绕过安全防护。其主要通过语义操纵、上下文诱导及输入变形等策略实施模板化或编码式攻击，具有实施直观、成本低、效率高等特点，适用于快速探测模型在特定输入模式下的安全脆弱性。

模板攻击通过预设提示词结构诱导模型输出违规内容，主要分为角色扮演与上下文示例两种典型路径。① 角色扮演攻击通过构造指令遵循与安全约束之间的认知冲突，引导LLM扮演单个或多个不受安全规则限制的虚拟角色（如“无所不能的AI”、“邪恶助手”），使其在遵循角色设定时淡化甚至忽略内置安全准则。典型案例如脱氧核糖核酸（DAN）系列攻击[60]及角色扮演越狱模板集[61]。该类方法依赖模型的指令遵循和角色模拟能力，无需访问模型内部信息，具有较高的攻击隐蔽性。② 上下文示例攻击则利用LLM的上下文学习能力，在提示词中嵌入若干“问题–有害回答”示例对，并在末尾附加目标请求，促使模型模仿示例中的违规回答模式[62-64]。该类攻击的有效性高度依赖于示例构造质量，对模型的上下文能力依赖性较强。

编码攻击通过对有害请求进行形式转换，利用安全对齐训练的数据分布局限性绕过内在安全机制。主要技术手段包括：① 低资源语言攻击，即将不安全请求转换为训练覆盖不足的语种（如将英语翻译为祖鲁语[65]，或以跨语言、语言混合方式实施攻击[66]）；② 密文与编码攻击，通过美国信息交换标准代码（ASCII）编码、加密、代码嵌入等方法对有害问题进行改写，诱导模型生成违规输出。典型案例包括将关键目标进行ASCII编码[67]、基于文本编码规则对有害输入加密[68]、将自然语言转换为代码输入[69]等。该类攻击的有效性主要取决于形式转换策略的隐蔽性以及安全对齐训练对非常规输入的覆盖程度，并对模型在分布外样本上的泛化与解析能力具有较强依赖。

（2）基于优化搜索的攻击：不同于依赖启发式策略的直接交互攻击，此类方法本质上是由模型驱动的。其核心思路是引入红队模型（通常为LLM或专门训练的攻击模型），将攻击生成视为可学习的优化任务。通过搜索、训练或迭代优化自动调整策略，从而生成能够动态绕过目标模型安全机制的对抗性提示，主要包含提示搜索与模型搜索两类技术路径。

提示搜索基于提示学习构建红队测试问题生成任务模板，使大模型具备自动生成测试问题的能力。典型案例包括强调开放生成与动态引导的测试用例生成方法[70]，以及基于预设风险场景分类体系与结构化模板的SafetyBench[5]。这类方法实现简单、效率较高，但生成质量和多样性高度依赖模板设计与基础模型能力。

模型搜索则以目标模型行为为反馈信号，将越狱攻击建模为优化搜索问题，通过在模型参数、输入或表示空间中寻找最优扰动以诱导违规输出。主要包括：① 模板生成，即基于越狱数据微调模型生成越狱提示，并采用变异和选择等机制迭代进化，生成更有效的隐蔽提示，再与种子问题组合实施攻击。典型案例包括AutoDAN[71]、Masterkey[72]和LLM-Fuzzer[73]；② 问题生成，通过监督微调或强化学习提升攻击模型对原始问题的改写能力。代表性工作包括LLM Stinger[74]、xJailbreak[75]和JailPO[76]，以及基于认知偏差的Dark Cite[77]和Cognitive Attack[78]；③ 对抗搜索，在模型表示空间中搜索使输出偏离安全对齐分布的对抗后缀。典型案例包括基于梯度信息直接搜索后缀的白盒方法地面指令制导（GCG）[79]及其变体I-GCG[80]和AmpleGCG[81]，基于约束与反馈迭代生成候选后缀的黑盒搜索方法[82]，以及通过训练本地代理模型以提升严格黑盒场景下搜索效率的代理搜索方法[83]。模型搜索方法虽能高效发现深层安全漏洞，但对模型访问权限和搜索策略依赖较强，且在语义自然性与跨模型迁移性方面仍受限制。

2. 多轮红队攻击

多轮红队攻击通过模拟人类多轮对话的多轮交互特性，利用LLM的上下文理解与逐步推理机制，在复杂交互中逐步诱导模型绕过安全防护。相较于单轮攻击，该类方法通常具备更高的隐蔽性与成功率。根据对实时交互上下文的依赖程度，可分为以

下两类。

（1）迭代式多轮攻击：通过反馈循环或任务分解，在多轮交互中逐步逼近攻击目标。其关键特征是每一轮的输入（提示或子问题）都基于前一轮的输出或评估结果动态调整，形成迭代优化过程。此类方法不依赖连续对话上下文的连贯性，更关注提示本身的有效性迭代。

迭代式搜索：通过构建评估机制对生成提示进行评估打分，并反馈至攻击模型进行迭代优化，直至攻破目标模型防护。典型案例如PAIR[84]采用双模型协作与反馈迭代机制实现黑盒高效攻击，TAP[85]通过树状并行搜索与剪枝优化探索效率。PAP[86]引入社会心理学策略增强提示说服力，以及基于认知操纵的DeepInception[87]等方法，共同体现了多轮攻击中通过策略性交互设计与心理机制利用以逐步诱导模型失效的共性特征。

静态有害拆解：采用分步诱导的策略，将不安全的查询拆解为多个语义上无害或低风险的子查询，在多轮对话中依次提交。目标模型在处理每个子请求时可能无法识别其整体组合后的不安全意图，最终通过模型自身的上下文整合或用户的引导，形成完整的不安全输出。典型案例如FITD[88]基于心理学“先小后大”策略进行分步引导，SOT[89]系统总结了目的反转、关键词替换等四类分解范式，CoSafe数据集[90]揭示了利用指代现象渐进构建攻击路径的机制，以及Red Queen[91]通过将攻击伪装为风险预防以增强隐蔽性。该类方法共同体现了通过语义分解与多轮渐进引导规避单轮安全检测的核心特征。

（2）对话式多轮攻击：该类攻击高度依赖对话历史上下文，在交互过程中动态调整攻击策略。HARM[92]基于可扩展的细粒度风险分类体系自顶向下生成多样化的测试用例，借助人工红队多轮攻击数据，并结合微调策略、拒绝采样与强化学习等方法，使红队代理具备拟人化的多轮对抗探测能力；CoA[93]将攻击过程建模为类人审讯交互，通过多轮语义控制逐步逼近目标风险内容，在隐蔽性与攻击有效性之间取得平衡；MRJ-Agent[94]基于风险意图分解与心理诱导，将高风险查询拆解为多轮低风险子问题，由智能体根据反馈动态调整攻击轨迹；TROJail[95]则将多轮越狱构建为轨迹级强化学习问题，通过设计过程奖励直接优化最终攻击效果；MTSA[96]将多轮红队对抗建模为迭代优化过程，学习生成多轮对抗提示，并采用基于未来回报的多轮强化学习更新对话策略。总体而言，该类方法通过动态的交互调整，增强了攻击的隐蔽性、连续性与成功率，揭示了LLM在复杂交互环境中的深层脆弱性。

（二）自动化红队测试评判技术

评判技术的核心任务在于利用自动化内容分类器对攻击所诱发的模型输出进行安全风险评估，以准确判定攻击成效，从而为攻击策略的迭代提供决策依据。面对日益复杂的红队攻击，高精度、强泛化的LLM内容分类器已成为实现可靠、规模化、自动化评判的关键支撑，本节将系统梳理内容安全分类器的关键技术路径与代表性方法。

1. 判别式检测模型

判别式检测模型通过在标注数据上对Transformer架构的预训练模型进行监督微调，从而获得针对生成内容安全风险的分类能力。该类模型通过学习标注数据中与特定标签（如仇恨言论、暴力内容、虚假信息）相关的文本模式、语义特征及统计规律，建立从输入文本到预定义类别的映射关系，其输出通常为离散类别标签或连续置信度分数[97]。

判别式检测模型已形成多样化的系统、工具与基准。在商业化服务中，典型案例如在线内容审核工具Google Perspective API[98]采用判别式内容分类器，对毒性、侮辱等风险进行量化评分并支持多语言环境，但在应对对抗性字符扰动时存在局限[99]。OpenAI Moderation API基于生成式预训练变换器（GPT）系列模型构建专用的内容安全分类模型，可识别仇恨、骚扰、暴力、自残、色情、非法活动等多类安全风险[100-101]。微软研究院等机构参与提出的ToxiGen项目[97]则基于BERT模型构建了面向特定群体仇恨言论风险内容的判别式检测模型与数据集，揭示了训练数据质量对内容安全的关键影响。

在开源工具与评估基准方面，HarmBench[102]提供了配套的判别式内容安全分类器，针对非版权类行为基于Llama-2-13B-Chat微调实现二元分类，针对版权类行为则采用哈希匹配判别，并通过迭代蒸馏训练提升判断一致性。Do-Not-Answer[103]则提供了基于编码器式预训练模型微调的内容安全分类

器，支持对“指令–回复”文本进行二分类与六分类细粒度判定，涵盖拒绝回答、谨慎回应等多种策略类型。这些工具与基准共同为自动化红队测试中的效果评判提供了重要支持。

近年来，基于信息论视角的表示学习方法也可提升文本分类泛化能力。其中，监督对抗对比学习方法[104]在对比学习框架中引入对抗训练，利用监督对比学习生成高质量对抗样本。该方法能够学习具有类内扩散特性的结构化表示，有效利用标签级特征一致性并保留细粒度类内特征，从而显著提升模型泛化能力。在此基础上，进一步提出了结构化概率编码方法[105]，突破了传统的编码–解码框架，在统一的参数化模块中同时实现任务编码与预测，并引入类空间的结构化正则化技术，促进潜在空间中类别的均匀覆盖。这些方法在零样本分类任务上平均提升2.8%，达到了同期最优水平。

针对当前判别式检测模型对中文场景支持不足、依赖人工标注导致扩展性受限且缺乏专门评估基准的问题，近期研究提出了基于合成数据驱动的内容安全分类器。Libra[106]采用两阶段课程训练框架：第一阶段利用大规模高质量合成数据进行预训练，显著降低对昂贵人工标注的依赖并提升数据利用效率；第二阶段使用精选真实困难样本进行微调，以优化模型在边界案例上的表现。该工作同时构建并开源了首个专门面向中文大模型护栏性能评估的基准Libra-Test，覆盖7类关键有害场景，超过5700条融合真实、合成和翻译数据的标注样本，为全面评估中文大模型的安全防护能力提供了重要基础。实验表明，增加合成数据的规模能持续提升模型性能，且基于预训练的模型在不同规模的微调数据下均表现出更高的准确率。值得注意的是，仅含0.1 B参数的Libra-Tiny模型在有害信息检测任务上超越了多个主流开源大模型，展现出轻量化判别式模型的巨大潜力。

总体而言，判别式检测模型具有计算高效、部署成本低的优势，但其泛化能力有限，对训练分布外的新型攻击或需要深度推理的复杂攻击模式（如多轮诱导、隐蔽恶意）的识别能力尚且不足。

2. 生成式检测模型

生成式检测模型代表了与判别式检测并行的一种技术路线，其核心是直接利用LLM的语义理解与推理能力，通过生成自然语言判定的方式评估内容安全风险，即“LLM as Judge”。该方法可基于任务微调或提示工程，使LLM输出解释或评分，实现对输入与输出内容的语义级安全风险分类和审查。与侧重模式匹配的判别式方法相比，生成式路径在识别角色扮演、隐喻表达、多轮诱导等需深层语义理解的复杂与隐蔽攻击方面展现出更强的适应性。

生成式检测模型已在开源领域形成多项代表性应用。Meta推出的Llama Guard[107]基于Llama-2-7B微调，不仅能输出是否违规的判定，还能明确指认违反的具体安全策略（如仇恨言论、自我伤害等），显著提升了分类决策的透明性。谷歌的ShieldGemma[108]则基于Gemma-9B，专注于英语环境下四类主要安全风险的识别。StrongREJECT[109]构建了基于禁用指令集的越狱效果评判框架，通过量表驱动基于GPT-4o-mini的内容分类器对回复进行二元拒答判断与多维度评分，并提供了基于Gemma-2B的蒸馏版轻量化内容安全分类器。SALAD-Bench[110]则设计了覆盖多领域、多任务的三层级风险分类体系，在问答任务中采用经LoRA微调的Mistral-7B作为内容安全分类器，在多选任务中通过受控输出提示与正则解析实现自动评分。这些系统体现了生成式检测模型在细粒度分类、可解释输出与持续适应新型攻击方面的技术进展。

生成式检测模型的核心优势在于其深层的语义理解与推理能力，能够依据上下文进行接近人类认知的内容安全分析，不仅能做出安全性判定，还可提供解释性输出。在技术路径上，该类方法已从早期使用专门微调的中等规模模型，发展到直接调用与被测模型规模相当甚至更大的基础语言模型作为内容安全分类器。然而，生成式检测模型仍面临严峻挑战：计算开销大，在资源受限场景下实时性受限；模型本身存在幻觉或受对抗提示误导的风险；其“黑箱”特性虽在可解释性上有所改善，但完全透明可信的判决机制仍需探索。

3. 边缘端检测模型

边缘端检测模型旨在将内容安全判别能力本地化部署于用户终端设备（如智能手机、平板电脑、个人电脑、物联网设备），以实现低延迟的实时内容安全风险分类。随着OpenCLAW等智能体框架在端侧加速普及，用户与智能体的交互日趋频繁和深入，这也催生了对端侧红队测试的迫切需求。边

缘端检测模型通常基于大型基础模型，通过量化、知识蒸馏、剪枝及神经架构搜索等压缩与优化方法，转化为适应终端有限计算、内存与功耗约束的轻量版本，从而在保持可接受检测性能的同时满足边缘环境的实际部署需求。

Google 发布的 ShieldGemma 系列模型[108]是边缘内容安全分类器的代表性实践。该系列基于开源 Gemma 模型构建，并提供 2B、9B 与 27B 等不同规模版本。其中 ShieldGemma-2B 参数量仅为 20 亿，适配本地与边缘环境下的安全内容分类需求。尽管规模远小于同系列 ShieldGemma-9B/27B 等版本，官方基准评测显示 ShieldGemma-2B 在 ToxicChat 等安全风险识别数据集上优于 LlamaGuard-7B 等代表性基线模型[108]。

边缘检测模型旨在通过本地化部署实现低延迟响应与用户隐私保护。本地执行避免了数据传输带来的隐私风险，且不依赖网络连接，可提供实时检测。然而，受终端资源限制，此类模型在参数规模、计算能力与上下文长度上通常弱于云端模型，对复杂、语义微妙或需长程理解的内容检测能力有限。模型更新与大规模终端部署也面临工程挑战。未来需重点发展在极端压缩下保持判别能力的技术、高效轻量推理引擎以及安全可控的端云协同更新机制。

四、新型生成式人工智能系统红队测试技术

随着多模态大模型、智能体系统以及检索增强生成等新型生成式人工智能技术进入广泛应用与部署阶段，其特有的系统架构与交互模式也引入了新的安全挑战。这类系统在融合多源信息、执行自主决策以及与外部知识库动态交互的过程中，可能暴露出传统 LLM 所未涵盖的安全漏洞与攻击面。为此，本章将系统性地探讨针对上述新型生成式人工智能系统的红队测试方法与技术进展。

（一）多模态大模型的红队测试

多模态大模型因数据模态多样、模型结构复杂，面临特有的安全挑战，对其潜在安全风险进行系统性识别尤为关键。多模态攻击的核心在于利用模型在跨模态语义对齐与融合过程中的脆弱性，通过精心设计的输入扰动干扰其决策路径，从而诱导模型输出不安全内容。主要包括对抗性扰动与多模态提示注入两种典型方式。

对抗扰动通过对输入数据（如图像像素或文本词语）施加人眼难以察觉的微小修改，诱导模型生成不安全输出。此类方法多依赖基于梯度的噪声生成与优化技术[111]。典型案例包括通过模态转换与嵌入向量角偏差优化生成高效扰动的 CrossFire[112]、在黑盒场景下融合单模态与多模态扰动来生成对抗性样本从而揭示多种视觉–语言模型脆弱性的 VLAttack[113]；通过调整文本字体样式提升对抗性样本语义多样性与跨模态攻击迁移能力的字体样式增强迁移方法 TATM[114]。此外，也有研究人员[115]系统梳理了视觉–语言模型面临的对抗攻击、越狱、提示注入等威胁，并对相关资源与趋势进行了分析。

多模态提示注入是一种更为隐蔽的攻击方式，攻击者将恶意指令嵌入图像、音频等非文本模态中，伪装为看似正常的输入以操纵模型输出。模型处理此类输入时，隐藏的指令被激活，从而可能绕过安全防护机制，执行不安全操作。例如，提示注入攻击可将恶意指令嵌入到图像输入中，实现对模型的操控[116]；一些研究生成与指令相关的对抗性扰动，并将其隐藏在图像或音频中[117]，这些扰动对人不可察觉，却能诱导多模态模型输出攻击者预先设定的文本内容并执行相应指令。

（二）智能体系统的红队测试

基于生成式人工智能的智能体系统凭借其自主决策、环境交互与任务执行能力，正逐步应用于诸多关键领域。然而，其高度自主性、复杂决策逻辑以及与外部环境的深度耦合，也引入了新的安全威胁。早期研究表明，基于强化学习的智能体对对抗性输入具有显著脆弱性，例如，对策略网络施加微小扰动即可导致智能体在复杂环境中做出灾难性决策[118]。随着模型上下文窗口的持续扩展，一种更为隐蔽的风险形式——长上下文隐蔽指令注入日益凸显，攻击者可将恶意指令嵌入文档、邮件等外部数据源，利用智能体对长文本的无差别处理实现行为劫持。

在商业实践中，Microsoft 365 Copilot 等智能体产品被披露存在由提示注入引发的信息泄露风险，凸显了实际部署中的安全隐患[119]。谷歌公司将智能

体系统的核心安全风险归纳为恶意行为与数据泄露两类：前者指智能体执行非预期或违反安全策略的动作，后者涉及用户隐私或敏感信息的不当暴露[120]。鉴于智能体系统的高风险属性，开展系统化、规范化的安全测评显得尤为迫切。当前，根据测评对象的功能定位与运行环境，智能体系统红队测试的研究与实践主要围绕以下两类展开：通用智能体与操作系统智能体（OS-Agent）。

1. 通用智能体测试

通用智能体旨在执行广泛的、开放域任务，其测评工作已从早期依赖形式化问答的静态评估转向基于高保真环境仿真的动态测试，以更真实地暴露其在复杂、多变环境中的安全隐患。典型案例如苏黎世联邦理工学院的AgentDojo框架[121]，通过构建可动态替换工具与任务的沙盒基准，实现在4个应用场景下提示注入等攻击的可复现测试与防御效果量化；ASB测评基准[122]围绕10类应用场景，采用多种攻防方法构建了400项安全测评任务；OpenAgentSafety[123]在模拟真实工具环境下，结合规则基础评判与LLM评判，对多轮多用户任务进行安全评估。

此外，智能体在特定情境下可能表现出代理性错位风险。Anthropic的研究发现，在特定激励与威胁设定下，模型驱动的智能体可能出现勒索、泄密等内部威胁行为[124]。据此，Anthropic提出了面向“破坏–监控”对抗场景的安全评测框架SHADE-Arena[125]，通过设置在表面良性可解决任务中隐含恶意副任务的双重目标场景，全面评估智能体在双重目标驱动下的行为倾向与风险。

在多智能体系统中，安全风险可通过智能体间交互被放大，形成“群体越狱”等复杂威胁。例如，史密斯探员攻击表明，一张对抗性图像即可引发百万级智能体系统中的链式越狱扩散[126]。这种具有传染性的越狱现象表明，多智能体系统中单一安全漏洞可能引发灾难性后果，对整体系统防护提出了更高要求。部分研究人员[127]进一步将多智能体协作中的风险归纳为协调失误、冲突与合谋三类失败模式，指出其根源在于智能体间复杂的动态交互机制，而非单个智能体本身的缺陷。

2. 操作系统智能体测试

OS-Agent是在操作系统层面运行的软件代理，其核心功能在于直接管理与监控操作系统及其关联的软硬件环境，通常部署于云服务、容器及虚拟化环境中，实现对操作系统底层资源的精细化管控。当前，针对OS-Agent的安全评测已从传统的个人电脑端扩展至移动设备端，且普遍采用沙箱环境仿真的方式进行。

OS-Harm[128]是首个面向图形界面操作系统智能体的安全基准，构建了“任务构建–环境构建–状态数据采集–结果研判”的系统化测评流程。该基准采用OS-World框架搭建沙箱环境，基于11款桌面应用构建150项沙盒任务，从用户滥用、即时注入攻击与模型不当行为三个风险维度量化违规率，并采用LLM结合提示词从任务完成度与违规率两方面综合评估安全性。

RiOSWorld[129]是首个面向真实操作系统、支持多模态输入且覆盖全面风险类型的计算机使用智能体（CUA）安全基准。该基准利用492项真实桌面任务，系统评估模型在用户恶意指令与环境陷阱双重风险下的表现，结果表明，当前多数CUAs在安全性上尚未达到可信水平。

RedTeamCUA[130]在混合沙盒环境下评估由Web提示注入引发的CUA安全风险，创新性地将传统Web红队攻击与操作系统级攻击链相结合，诱使CUA将恶意操作误认为是任务的必要步骤，从而构建了首个Web-OS混合沙盒红队框架，专门用于端到端间接提示注入攻击测试。该研究基于180项高仿真任务与10类对抗模板构建了包含864个样本的RTC-Bench数据集，测试显示6个主流CUAs均未能有效防御此类攻击。

随着移动设备端侧大模型智能体的广泛部署，其安全评估需求日益凸显。Mobile LLM Agents[131]首次针对手机端大模型智能体系统开展安全评估，发现其存在跨语言、用户界面与系统权限三层攻击面，可能引发隐私泄露、任务劫持甚至设备恢复出厂设置等严重后果。该研究同时开源了自动化测评框架AgentScan，为后续安全评估提供了基础工具支持。

（三）检索增强生成系统的红队测试

检索增强生成（RAG）系统通过引入外部知识库，提升了大语言模型在时效性与事实准确性方面的表现，但其“检索–生成”分离的架构也引入了新的攻击面[132]。本文聚焦于RAG系统知识库与检

索过程两个关键环节，分别介绍对应的红队测试攻击手段：知识库投毒与对抗扰动。

知识库投毒是当前针对RAG系统最具威胁的攻击方式之一。攻击者通过向外部知识库中注入精心构造的虚假或恶意数据，污染信息源，使系统基于被篡改内容生成错误或有害回答。例如，PoisonedRAG[133]通过优化方法生成对抗性文档，能有效诱导RAG模型针对特定查询输出攻击者预设的恶意答案。

除污染知识源外，攻击者还可对检索过程实施对抗性干扰。该技术通过扰动查询输入，破坏检索模型的相关性排序与匹配逻辑，致使其返回不相关或误导性的文档，从而损害RAG系统信息获取的可靠性。研究表明，检索模型对查询的微小变化极为敏感，简单的词汇调整即可显著影响其排序结果[134]。

五、生成式人工智能红队测试实践

在生成式人工智能快速发展的背景下，红队测试已成为保障模型安全性、可靠性与合规性的关键机制。然而，各国在技术发展水平、法律监管环境及产业生态上的差异，导致红队测试实践呈现多元化格局。一方面，国际层面正逐步构建涵盖规范框架、法律法案至跨境互认的治理路径；另一方面，科技企业与研究机构亦结合自身产品特性与风险模型，探索多样化的红队测试方法论与工具。同时，我国正立足本土监管与产业需求，构建适配大模型的评估方法与红队测试体系。总体而言，政策推动、实践探索与本土化创新相互交织，共同推动红队测试的制度化与常态化发展。基于此，本节将依次梳理政府政策推进、国内外主流实践与方法，以揭示其发展脉络、共性规律与差异挑战。

（一）生成式人工智能红队测试的全球政策与监管推进

围绕全球主要国家与地区在生成式人工智能红队测试方面的政策设计、监管要求与制度化建设路径展开梳理，分析其治理逻辑、实施机制及共同趋势。

美国政策框架。NIST在AI风险管理方面扮演重要角色。NIST于2023年1月发布了首个《人工智能风险管理框架（AI RMF 1.0）》，这是一个跨行业的自愿性框架，旨在帮助组织识别和管理AI系统相关安全风险，并将可信赖性考量融入其设计、开发、使用和评估全过程[135]。该框架强调通过对AI进行全生命周期的风险评估，以提高其安全性和可靠性。

日本方法指南。日本AISI于2024年9月25日发布了《AI安全红队测试方法指南》初版（Version 1.0），为AI系统红队测试提供了基础方法框架[32]。该指南明确红队测试旨在从攻击者视角识别系统脆弱点与防御短板，以提升整体安全性，并强调其在评估抗攻击能力、发现防护盲区、跟踪复杂攻击技术演进及保障系统安全使用方面的重要作用。指南将测试按知识基础（黑盒、白盒、灰盒）、环境（生产、暂存、开发）与执行方式（自动化、人工、智能体驱动）进行分类，特别指出需关注提示注入、提示泄露等AI特有攻击类型，并建议在专业领域中引入领域专家以应对复杂风险。

欧盟法案要求。欧盟《人工智能法案》于2024年通过，旨在通过基于风险的规则促进可信AI发展[13]。该法案要求高风险AI系统提供者在投放市场前进行模型评估，包括对抗性测试，以识别并减轻歧视性、误导性、不安全或欺骗性输出等潜在风险，确保系统稳健可靠。对于具有系统性风险的通用AI模型，提供商需执行并记录对抗性测试，以识别并降低系统性风险。

国际互认机制。2024年11月20日，美国商务部和国务院在旧金山联合举办国际人工智能安全研究网络（INASI）首次会议，宣告该组织成立。10个初始成员通过联合使命宣言，旨在汇聚全球技术专家，促进对AI安全风险及缓解措施的共识，推动最佳实践应用与共享。INASI合作聚焦四大领域：① 加速AI安全研究；② 合作制定模型测试与评估最佳实践；③ 促进形成先进AI系统测试解释等通用方法论；④ 增强全球包容性，确保信息共享。INASI发布联合声明评估先进AI系统风险，强调风险评估应具备可操作性、透明性、全面性、多方参与性、可迭代性及可重复性六项原则[136]。该声明基于《布莱切利宣言》《首尔联合声明》及其他AI安全倡议成果，呼吁成员实施统一风险评估指导方法，并为推进全球协调一致与互操作性制定路线图。

（二）国内外主流红队测试实践

OpenAI红队测试体系。OpenAI的红队测试实践始于对早期LLM的内部风险评估，并逐步发展为系统化、跨学科的安全测评体系。在GPT-3与Codex阶段，团队主要通过内部定性评估识别潜在滥用风险（如虚假信息、偏见、不安全代码）[137-139]。DALL-E 2发布时，首次引入外部专家参与红队测试，将红队测试确立为模型对齐与安全的重要环节[33,140]。2023年，通过组建OpenAI红队网络，招募涵盖生物安全、网络安全、社会科学等领域的全球专家，建立长期协作机制，以探测传统基准难以覆盖的复杂风险场景[39,139]。在方法论上，采用人工红队与自动化测试相结合的模式：外部专家提供多样化攻击视角与深入反馈，同时利用GPT-4等模型生成合成测试数据，实现规模化检测[139]。实践表明，红队测试有效揭示了多项关键漏洞，如在DALL-E 3中识别出可越狱生成违规图像的脆弱点，相关发现推动了过滤机制与自动化防御措施的增强[139,141]。尽管其方法论与部分测试结果已通过研究报告和系统卡片公开，但完整的工程化红队工具链尚未开源。

Anthropic合作测试模式。Anthropic的红队测试实践始于2022年中期，初期聚焦于识别并降低语言模型的有害输出，通过对抗性测试提升模型安全性[15]。随着模型能力演进，测试范围逐步扩展至生物安全、网络安全及自主AI等领域，旨在为国家层面的安全风险提供早期预警[26]。在组织机制上，建立了由内部前沿红队主导、联合国际专家网络的协作体系，涵盖生物安全专家、网络安全分析师及语言学家等多学科背景，并与新加坡资讯通信媒体发展管理局、AI Verify Foundation及NIST等机构开展合作[15,26]。在方法论上，该公司融合领域专项测试、多语言红队与自动化评估，开发了宪法分类器以应对通用越狱攻击[14,142]。自动评估表明，宪法分类器将模型越狱成功率从86%降至不足5%[142]。此外，相关研究还覆盖智能体滥用风险分析[142]，并探索实施ASL-3（Anthropic安全与对齐等级3）安全标准，以规范模型在高风险任务中的行为准则[143]。

Google社会技术框架。Google针对前沿AI系统安全组建了跨学科红队，融合网络安全、威胁情报与机器学习专业能力，通过高保真对抗模拟识别模型及系统脆弱性[35,144]。其在《安全AI框架》指导下，构建了覆盖提示注入、训练数据提取、模型后门等攻击手法的威胁分类体系[145]。DeepMind提出的社会–技术红队框架通过参数化指令生成可复现的攻击模板，提升了测试的覆盖率与严谨性，并引入人口统计特征匹配与多视角仲裁流程以提升评估公平性[144]。此外，谷歌公司还提出“AI-for-AI”范式，利用语言模型生成攻击提示，同时通过AI支持的红队测试（AART）构建自动化数据生成管道，实现多样化、文化敏感的对抗性评估[70,146]。目前，上述方法已应用于Gemini、Gemma等核心AI产品的安全开发流程，但相关实现仍多为研究原型或内部工具，尚未形成完整的工程化红队平台。

Meta上下文驱动红队体系。Meta的红队测试最初依赖专家手动设计攻击场景，自Llama 3起才系统性纳入模型对齐流程中。该公司提出了一种半自动化的攻击提示生成与防御框架，通过专家设计的高风险提示训练模型，构建了覆盖敏感主题的SAP数据集，用于安全微调与对抗性评估[147]。在组织上，（生成式）人工智能红队联合内部多职能团队及外部专家开展红队评估，并展示了针对Llama 3的端到端评估流程，包括高风险领域的外部专家增益测试与闭环防护策略[148]。在方法论层面，团队通过对抗模型与目标模型的多轮博弈生成强攻击提示，并将失配样本纳入安全微调数据以降低违规率[149]。同时，Llama Guard系列模型被用作输入/输出过滤与分类的护栏，CyberSecEval基准则用于量化模型在网络攻防及不安全内容生成方面的风险[107,150]。上述方法已工程化应用于Llama 3的开发全周期，并通过Purple Llama等社区项目开源了部分评估基准与工具，以支持第三方复现与协同改进。

NVIDIA自动化探测框架。NVIDIA将传统信息安全融入大模型开发生命周期，由攻防对抗安全专家与数据科学家组成的AI红队从攻击者视角评估机器学习系统的安全风险[151]。同时，通过生成式红队挑战等社区合作机制引入更广泛的外部参与[152]。英伟达公司在公开技术博客中讨论了提示注入与分词器实现安全等生成式系统风险：其中，提示注入可借助插件与工具链扩展攻击面，诱发远程代码执行、服务器端请求伪造或结构化查询语言注入等后果；而分词器初始化验证不足则可能破坏输入/输出与模型内部计算之间的一致性，从而破坏应用完整性[153-154]。在方法论与工具层面，英伟达公司开源了

插件化架构的自动化漏洞扫描器 Garak，支持对幻觉、数据泄露、越狱等风险进行跨模型系统化探测[155]。同时，NVIDIA 通过 NeMo Guardrails 等可编程护栏机制提升 LLM 应用的安全可控性，并利用 Llama Guard 等模型提供面向对话场景的输入 / 输出安全分类能力[107,156]。结合 Garak 的自动化探测与相关护栏机制，英伟达公司展示了从自动化风险探测、护栏配置到持续评测的安全工程流程，体现了红队发现向可运营防护实践衔接的工程思路[155-156]。

Microsoft 智能体测试工具。Microsoft 自 2018 年起组建 AI 红队，其测试范围从初期的传统安全漏洞与模型对抗攻击评估，逐步扩展至覆盖价值观影响等更广泛的安全范畴[37,157]。截至 2024 年 10 月，该团队已完成对上百款生成式人工智能产品的系统性测试，通过对抗性模拟持续提升系统的安全与鲁棒性[41]。Microsoft 的 AI 红队以内部团队为核心，成员涵盖 17 种语言背景的安全专家，并与网络安全、偏见检测和隐私等领域的外部专家网络协作，通过生态系统伙伴开展全球性、多文化的风险评估与价值对齐工作[37,157]。在方法论与工具方面，Microsoft 结合自动化对抗模拟与专家测试，开源了红队测试工具包 PyRIT[158]。该工具包提供可复用的提示库与攻击策略编排能力，支持在多类风险维度上进行自动化红队测试，并可集成 TAP、PAIR 等自动化攻击策略以生成多样化测试样例[158]。截至 2025 年 4 月，Microsoft 在 Azure AI Foundry 平台中上线 AI 红队智能体公测版，深度集成 PyRIT 能力，实现对模型与应用的全自动风险扫描、攻击成功率量化以及测试过程的实时记录与追踪[159]。

上海人工智能实验室 OpenRT 红队框架。OpenRT 是上海人工智能实验室提出的开源红队测试框架，旨在为多模态大模型构建统一、模块化的安全评测基础设施[160]。针对当前红队工具分散、测试单一且难以规模化的问题，OpenRT 通过模块化设计将模型、数据集、攻击方法、判别器与评估指标解耦，借助统一编排器实现高并发、可扩展的自动化红队测试。平台集成了 37 种攻击方法，覆盖黑盒与白盒场景，支持单轮与多轮对话、多智能体协同以及视觉 – 文本跨模态攻击，可对目标模型进行高吞吐、可扩展的安全探测。在评估机制上，OpenRT 结合规则判别与大模型判别，从攻击成功率、效率、隐蔽性与多样性等多维度进行量化分析。基于大规模实验，该框架系统揭示了主流模型在自适应与多轮攻击下仍存在显著安全风险。

星河红队测试平台（Galaxy）。Galaxy 由中国科学院信息工程研究所研发，提供覆盖风险场景生成、红队攻击模拟和多模型协同检测的全流程安全评估解决方案。平台基于预定义的安全风险类别自动生成单轮与多轮对话测试样例，通过越狱攻击和对话引导策略对目标模型进行主动探测，以揭示安全薄弱点；随后融合大模型与轻量模型的检测结果，依据多维指标输出量化评分并生成综合测评报告。平台构建了“八大类安全风险场景、二十项具体风险点”评测体系，重点关注意识形态偏差、违法犯罪引导及内容合规性等方向。在技术实现上，采用权威引用驱动攻击和两阶段训练框架等技术提升测评深度与模型鲁棒性。此外，Galaxy 平台通过“产学研”协同机制，组织生成式人工智能安全大赛，推动业内多方协作与技术创新。

（三）国内外主流红队测试实践对比分析

表 2 从组织形式、方法论、公开性及本地化特点等方面，对比了国内外主流红队测试体系建设上的主要实践路径。对比表明，国际科技企业与我国研究机构均形成了较为系统的实践路径：国际机构多采用内部核心红队结合外部专家网络，并搭配自动化工具链的模式，注重人工与 AI 驱动红队结合，并通过论文、基准与部分工具开源实现可复现研究（如 OpenAI、Anthropic、Meta、Microsoft）。国内机构（如上海人工智能实验室、中国科学院信息工程研究所）则以研究所主导、“产学研”协同为主要特征，重点建设面向国产大模型的自动化风险评测平台，体现出更强的本土化适配与产业化落地特色。二者共同推动了生成式人工智能安全评估能力的持续演进。

六、讨论与展望

生成式人工智能的颠覆性潜力，将其安全与治理问题推升为关乎技术与社会稳健发展的根本性挑战。在此背景下，红队测试已从一项可选的技术实践，升格为确保其可信发展的关键治理支柱。本综述通过系统梳理现有技术脉络，凝练出一个核心论断：红队测试的角色必须从孤立的、后期的技术检

查点，演进为贯穿生成式人工智能全生命周期的、跨学科的治理基础设施。然而，当前实践多局限于技术团队主导的发布前测试，侧重于点状漏洞挖掘，难以有效应对金融、生物安全等复杂场景中的系统性风险。因此，构建一个融合技术、伦理、法律与社会科学的协同治理框架，逐渐成为重要研究议题。面向这一转型目标，本节将从测试范围、测试对象、测试边界与安全理论四个维度，系统阐述其未来的演进路径。

（一）测试范围：超越内容安全，迈向可信评估

红队测试的评估范畴正经历一次根本性跃迁：从早期聚焦于狭义的内容安全，迈向涵盖真实性、公平性、隐私性与鲁棒性的系统性可信评估。这一跃迁的深层动因在于，生成式人工智能模型正从被动的信息生成工具，深度演变为主动的关键决策参与者，其潜在风险的复杂性与系统性已远超有害内容本身。为响应这一趋势，学术界与工业界已构建如TrustLLM在内的多维评价体系[161]。因此，红队测试的核心使命需要被重新定义：从诱导显性有害内容，升级为系统性地探查与揭示模型在多维可信目标下的结构性风险与失范行为，从而为构建可解释、可操作的生成式人工智能风险治理体系提供坚实基础。

（二）测试对象：跨越认知模型，通向具身系统

生成式人工智能技术范式的迭代，直接驱动着红队测试对象的演进，其路径呈现出三个清晰的递进层次。

其一，从单一模态到多模态的拓展。测试对象已从处理纯文本的大语言模型，扩展至能够同时处理并生成文本、图像、音频乃至视频的多模态生成模型。这一拓展要求红队测试发展跨模态对抗策略，以系统评估模型在多源异构信息输入下的认知一致性与安全鲁棒性。

其二，从静态模型到动态智能体的跨越。测试

表2　国内外主流红队测试实践对比分析表

组织机构	组织形式	核心方法论	开源工具 / 资源	典型特点
OpenAI	内部红队+外部红队网络专家协作	人工红队与自动化攻击生成结合；AI辅助样本扩充与风险场景覆盖	方法论和研究报告公开；未提供完整开源工具	跨学科专家参与；将红队嵌入模型开发生命周期；实施人工与自动化混合流程
Anthropic	内部前沿红队+国际专家合作	宪法分类器；领域特定红队测试与合成数据评估结合	研究论文与实验框架部分公开	多语言测试覆盖；越狱防护与长期风险评估；实施高风险任务评估
Google/ DeepMind	跨学科红队+安全研究团队	STAR社会技术红队框架；AI红队实现AI-for-AI自动化生成	方法与框架论文公开；无完整工程化开源工具	采用社会–技术结合方法；实施参数化与系统性评估；发布学术研究导向的方法论
Meta	内部AI红队+外部专家网络	多轮自动化红队；结合人工与模型驱动的对抗生成与评估	Purple Llama仓库；CyberSecEval基准；Llama Guard守护器	攻防一体化（紫队）思路；开放基准与评估体系；支持社区协同改进机制
NVIDIA	内部AI红队+开发者社区合作	自动化漏洞扫描（Garak）；工程化安全护栏（NeMo Guardrails）；“先测后防”流水线	Garak工具；NeMo Guardrails框架（均已开源）	提供开发与运行阶段集成的工程化工具；依托社区驱动持续迭代
Microsoft	内部AI红队+云生态伙伴	自动化对抗模拟与人工探测结合；PyRIT框架支持模型安全红队评测	PyRIT开源工具；AI红队智能体（Azure集成）	平台化红队体系；支持企业级自动化测试；多文化价值对齐与云端集成
上海人工智能实验室	研究机构主导+开源社区协作	模块化自动化红队；多模态、多轮与多智能体攻击；规则判别与大模型判别结合评估	OpenRT开源红队框架	架构具备模块化、高并发、可扩展特性；覆盖黑盒 / 白盒与跨模态攻击
中国科学院信息工程研究所	研究所主导+“产学研”合作	风险场景自动生成；多模型协同检测与在线攻防演练	Galaxy红队平台（部分框架开放）	构建本土化风险分类体系；适配国产大模型；产业协同与合规导向

焦点正从模型的输出安全性，根本性转移至以大模型为核心的自主智能体。其核心在于评估智能体在长周期任务中的目标一致性、工具调用合规性与行为伦理的系统性，这需要构建具备领域知识依赖的动态测试场景，以检验其在真实任务闭环中的行为可靠性。

其三，从数字空间到物理世界的延伸。以“视觉–语言–动作”模型为代表的具身智能体，将测试场域从虚拟环境锚定至物理世界。红队测试必须关注其在传感器噪声、执行偏差及环境不确定性等扰动下的行为鲁棒性与失效模式，并建立虚拟仿真与真实环境之间的有效评估桥接机制。

（三）测试边界：突破人工智能红队，融合网络测试

随着生成式人工智能模型深度集成于网络服务生态，其红队测试的边界正从算法层，结构性延伸至网络服务层。这种边界的消融体现为攻击面与测试方法的双重融合。其一，是攻击面的双向叠加：当AI模型能力通过API、插件等接口开放时，安全评估便不再局限于提示注入等原生风险，而必须涵盖其作为网络服务所继承的全部传统攻击面。其二，是测试方法的链式融合：一次完整的深度渗透，必须将AI原生的对抗性提示与传统的网络漏洞挖掘技术无缝串联，形成连贯的混合攻击链。因此，AI红队与网络红队的人为分野已失去现实意义。未来的红队测试，必须将具备智能与网络双重属性的融合系统视为测试新常态，并对测试者的跨域知识体系与一体化评估能力提出根本性要求。

（四）安全理论：穿透现象描述，深入机理解释

当前，大模型安全红队测试正面临一场深刻的范式鸿沟：日益丰富的攻击和评测实践探索与严重滞后的内在机理认知之间的矛盾日益尖锐。这一理论赤字从根本上制约了测试的系统性与前瞻性，使我们仍停留在“知其然，而不知其所以然”的阶段。例如，安全对齐缘何导致模型能力的性能坍塌？稀疏的安全神经元如何构建防御，又如何被精准绕过？这种对内在机理的黑盒状态，使当前的攻防博弈在很大程度上仍是一种经验主义的试错，而非科学主义的预判。

因此，红队测试的未来范式升级，其关键在于实现从现象观测到机理探源的深刻转型。其核心在于解构安全对齐与模型能力的竞争机制、精绘关键安全回路的实现图谱、揭示对抗扰动在模型内部的传播路径。唯有构建于深层次的机理解释能力之上，红队测试方能超越“猫鼠游戏”的被动循环，从根源上设计出更具穿透力的攻击范式，指导更为稳健的对齐策略，最终演化为兼具预见性、解释性与指导性的安全治理支柱。

（五）小结

综上所述，生成式人工智能红队测试的未来发展将呈现系统性、前瞻性与融合性并进的总体趋势：测试范围从单一内容安全迈向多维可信评估，测试对象从静态模型向动态具身系统延伸，测试边界突破算法局限并与网络安全深度融合，测试理论从经验观测向机理解释跃升。这四大维度的协同演进，共同指向构建一个覆盖全生命周期、贯通算法与网络空间、融合技术治理的红队测试新范式，为生成式人工智能的稳健发展提供坚实的科学支撑与制度保障。

七、结语

生成式人工智能的迅猛发展，将其安全治理推至至关重要的战略位置。在此背景下，红队测试已从辅助性的评估工具，演进为与模型能力同步迭代的核心对抗范式，成为理解并驾驭大模型安全边界的关键路径。本文系统性地梳理了生成式人工智能红队测试的概念内涵、流程方法与风险分类体系，重点分析了针对大语言模型的单轮与多轮攻击技术、判别式与生成式评判模型及其边缘化部署策略，并探讨了面向多模态、智能体等新型生成式人工智能系统的红队测试挑战与国内外实践。研究表明，红队测试已逐步形成从攻击生成到攻效评判的完整技术链条，并呈现出自动化程度提升、测试对象拓展和应用场景延伸的发展趋势。

然而，面对生成式人工智能模型能力的快速演进和安全威胁形态的持续变化，当前生成式人工智能红队测试在可解释性、泛化能力、测试效率和工程化落地等方面仍存在不足。着眼未来，有必要融合认知科学、形式化验证等跨学科理论，深入揭示攻击与防御的内在机理，发展轻量化、可迁移的攻

击生成与攻效评判技术，构建动态、鲁棒且具备可解释性的安全测评体系，推动红队测试由经验驱动的漏洞发现，向机理认知驱动的系统性安全评估范式演进。我国在生成式人工智能红队测试领域探索形成的系统化路径与解决方案，可为全球应对AI安全挑战提供重要的技术参考与实践范例。

利益冲突声明

本文作者在此声明不存在任何利益冲突或财务冲突。

Received date: March 2, 2026; **Revised date:** March 23, 2026

Corresponding author: Wu Shizhong is a research fellow from Zhongguancun Laboratory, and a member of the Chinese Academy of Engineering. His major research fields include information technology security testing and cybersecurity risk assessment. E-mail: wusz@cae.cn

Funding project: Chinese Academy of Engineering project “Strategic Research on Artificial Intelligence Safety Prevention in the Cyber and Information Domain” (2025-XBZD-08)

参考文献

[1] OpenAI，Achiam J，Adler S，et al. GPT-4 technical report[PP/OL]. V6. arXiv（2024-03-04）[2026-02-10]. https://doi.org/10.48550/arXiv.2303.08774.

[2] Touvron H，Lavril T，Izacard G，et al. LLaMA：Open and efficient foundation language models[PP/OL]. arXiv（2023-02-27）[2026-02-10]. https://doi.org/10.48550/arXiv.2302.13971.

[3] Team G，Anil R，Borgeaud S，et al. Gemini：A family of highly capable multimodal models[PP/OL]. V5. arXiv（2025-05-09）[2026-02-10]. https://doi.org/10.48550/arXiv.2312.11805.

[4] Wei A，Haghtalab N，Steinhardt J. Jailbroken：How does LLM safety training fail?[C]//The Thirty-Seventh Annual Conference on Neural Information Processing Systems，2023：80079-80110.

[5] Zhang Z X，Lei L Q，Wu L D，et al. SafetyBench：Evaluating the safety of large language models[C]//The 62nd Annual Meeting of the Association for Computational Linguistics，2024：15537-15553.

[6] 李希陶，吴江，郑庆华，等. 大语言模型越狱攻击：模型、根因及其攻防演化[J]. 中国科学：信息科学，2025，55（6）：1372-1405.
Li X T，Wu J，Zheng Q H，et al. Jailbreaking large language models：Models，origins，and evolution of attacks and defenses[J]. Scientia Sinica Informationis，2025，55（6）：1372-1405.

[7] Bengio Y，Hinton G，Yao A，et al. Managing extreme AI risks amid rapid progress[J]. Science，2024，384（6698）：842-845.

[8] Center for AI Safety. Statement on AI risk[EB/OL]. （2023-05-30）[2026-02-04]. https://aistatement.com/.

[9] 杨伟平，程豪杰，周百顺，等. M3-SafetyBench：多领域多场景多维度的大语言模型安全评估体系[J]. 中国科学：信息科学，2025，55（11）：2923-2940.
Yang W P，Cheng H J，Zhou B S，et al. M3-SafetyBench：A comprehensive benchmark for evaluating the safety of large language models across multiple domains，scenarios，and dimensions[J]. Scientia Sinica Informationis，2025，55（11）：2923-2940.

[10] Wei J，Tay Y，Bommasani R，et al. Emergent abilities of large language models[PP/OL]. V2. arXiv（2022-10-26）[2026-02-10]. https://doi.org/10.48550/arXiv.2206.07682.

[11] Schaeffer R，Miranda B，Koyejo S. Are emergent abilities of large language models a mirage?[PP/OL]. V2. arXiv（2023-05-22）[2026-02-10]. https://doi.org/10.48550/arXiv.2304.15004.

[12] National Institute of Standards and Technology. Artificial intelligence risk management framework：Generative artificial intelligence profile：NIST AI 600-1[R]. Gaithersburg：National Institute of Standards and Technology，2024：18，50.

[13] Smuha N A. Regulation 2024/1689 of the eur. Parl. & council of June 13，2024（EU artificial intelligence act）[J]. International Legal Materials，2025，64（5）：1234-1381.

[14] Bai Y T，Kadavath S，Kundu S，et al. Constitutional AI：Harmlessness from AI feedback[PP/OL]. arXiv（2022-12-15）[2026-02-10]. https://doi.org/10.48550/arXiv.2212.08073.

[15] Anthropic. Challenges in red teaming AI systems[EB/OL]. （2024-06-12）[2026-02-10]. https://www.anthropic.com/research/challenges-in-red-teaming-ai-systems.

[16] Reuel A，Hardy A，Smith C，et al. BetterBench：Assessing AI benchmarks，uncovering issues，and establishing best practices[C]//The Thirty-Eighth Annual Conference on Neural Information Processing Systems，2024：21763-21813.

[17] Eriksson M，Purificato E，Noroozian A，et al. Can we trust AI benchmarks? An interdisciplinary review of current issues in AI evaluation[C]//The AAAI/ACM Conference on AI，Ethics，and Society，2025：850-864.

[18] 中国信息通信研究院人工智能研究所，人工智能关键技术和应用评测工业和信息化部重点实验室. 大模型基准测试体系研究报告（2024年）[R]. 北京：中国信息通信研究院，2024：2.
Artificial Intelligence Research Institute of China Academy of Information and Communications Technology，Ministry of Industry and Information Technology Key Laboratory of Artificial Intelligence Key Technologies and Applications Evaluation. Large model benchmarking system research report（2024）[R]. Beijing：China Academy of Information and Communications Technology，2024：2.

[19] Eriksson M，Purificato E，Noroozian A，et al. AI benchmarks：Interdisciplinary issues and policy considerations[C]//ICML Workshop on Technical AI Governance（TAIG），2025：1-13.

[20] Raji I D，Bender E M，Paullada A，et al. AI and the everything in the whole wide world benchmark[PP/OL]. arXiv（2021-11-26）[2026-02-10]. https://doi.org/10.48550/arXiv.2111.15366.

[21] Scarfone K，Souppaya M，Cody A，et al. Technical guide to information security testing and assessment：NIST Special Publication 800-115[R]. Gaithersburg：National Institute of Standards and Technology，2008：2-1-2-5，5-2-5-6.

[22] Strom B E，Applebaum A，Miller D P，et al. MITRE ATT&CK：Design and philosophy：MP180360R1[R]. McLean：The MITRE Corporation，2020：3.

[23] Applebaum A，Miller D，Strom B，et al. Intelligent，automated red team emulation[C]//The 32nd Annual Conference on Computer Security Applications，2016：363-373.

[24] European Central Bank. TIBER-EU framework：How to implement the European framework for threat intelligence-based ethical red

teaming[EB/OL].（2025-02-11）[2026-08-18]. https://www.ecb.europa.eu/pub/pdf/other/ecb.tiber_eu_framework_2025~b32eff9a10.en.pdf.

[25] Sinha A, Lucassen J, Grimes K, et al. What can generative AI red-teaming learn from cyber red-teaming?[R]. Pittsburgh: Software Engineering Institute, Carnegie Mellon University, 2025: 10-12, 19-20.

[26] Anthropic. Progress from our frontier red team[EB/OL].（2025-03-19）[2026-01-20]. https://www.anthropic.com/news/strategic-warning-for-ai-risk-progress-and-insights-from-our-frontier-red-team.

[27] Abuadbba A, Moore K, Goel D, et al. From promise to peril: Rethinking cybersecurity red and blue teaming in the age of LLMs[J]. IEEE Security & Privacy, 2026, 24(2): 53-63.

[28] Jabbar M S, Al-Azani S, Alotaibi A, et al. Red teaming large language models: A comprehensive review and critical analysis[J]. Information Processing & Management, 2025, 62(6): 104239.

[29] 段俊贤, 刘思雨, 关霁洋, 等. 生成式可视媒体鉴别与安全[J]. 中国科学: 信息科学, 2025, 55(8): 1925-1949.
Duan J X, Liu S Y, Guan J Y, et al. Survey on generative visual media detection and security[J]. Scientia Sinica Informationis, 2025, 55(8): 1925-1949.

[30] Lin L Z, Mu H L, Zhai Z N, et al. Against the Achilles' heel: A survey on red teaming for generative models[J]. Journal of Artificial Intelligence Research, 2025, 82: 687-775.

[31] Booth H, Souppaya M, Vassilev A, et al. Secure software development practices for generative AI and dual-use foundation models: An SSDF community profile: NIST SP 800-218A[R]. Gaithersburg: National Institute of Standards and Technology, 2024: 22.

[32] Japan AI Safety Institute. Guide to red teaming methodology on AI safety（version 1.00）[EB/OL].（2024-09-25）[2026-08-18]. https://aisi.go.jp/assets/pdf/ai_safety_RT_v1.00_en.pdf.

[33] OpenAI. Advancing red teaming with people and AI[EB/OL].（2024-11-21）[2026-01-20]. https://openai.com/index/advancing-red-teaming-with-people-and-ai/.

[34] Meta AI. Expanding our open source large language models responsibly[EB/OL].（2024-07-23）[2026-01-20]. https://ai.meta.com/blog/meta-llama-3-1-ai-responsibility/.

[35] Google. Google's AI red team: The ethical hackers making AI safer[EB/OL].（2023-07-19）[2026-01-20]. https://blog.google/innovation-and-ai/technology/safety-security/googles-ai-red-team-the-ethical-hackers-making-ai-safer/.

[36] NVIDIA. Defining LLM red teaming[EB/OL].（2025-02-25）[2026-01-20]. https://developer.nvidia.com/blog/defining-llm-red-teaming/.

[37] Microsoft. Microsoft AI red team building future of safer AI[EB/OL].（2023-08-07）[2026-01-20]. https://www.microsoft.com/en-us/security/blog/2023/08/07/microsoft-ai-red-team-building-future-of-safer-ai/.

[38] IBM. What is red teaming for generative AI[EB/OL].（2024-04-11）[2026-01-20]. https://research.ibm.com/blog/what-is-red-teaming-gen-AI.

[39] OpenAI. OpenAI red teaming network[EB/OL].（2023-09-19）[2026-01-20]. https://openai.com/index/red-teaming-network/.

[40] Verma A, Krishna S, Gehrmann S, et al. Operationalizing a threat model for red-teaming large language models（LLMs）[J/OL]. Transactions on Machine Learning Research, 2024-07-21. https://openreview.net/forum?id=sSAp8ITBpC.

[41] Verma A, Krishna S, Gehrmann S, et al. Operationalizing a threat model for red-teaming large language models（LLMs）[PP/OL]. V2. arXiv（2025-07-10）[2026-01-18]. https://doi.org/10.48550/arXiv.2407.14937.

[42] United Nations Educational, Scientific, Cultural Organization. Recommendation on the ethics of artificial intelligence[M]. Paris: UNESCO（United Nations Educational, Scientific and Cultural Organization）, 2022: 25-38.

[43] Bengio Y, Mindermann S, Privitera D, et al. International AI safety report: DSIT 2025/001[R/OL]. London: Department for Science, Innovation and Technology, 2025（2025-01-29）[2026-01-18]. https://www.gov.uk/government/publications/international-ai-safety-report-2025.

[44] Weidinger L, Uesato J, Rauh M, et al. Taxonomy of risks posed by language models[C]//The 2022 ACM Conference on Fairness, Accountability, and Transparency（FAccT '22）, 2022: 214-229.

[45] Hendrycks D, Mazeika M, Woodside T. An overview of catastrophic AI risks[PP/OL]. V6. arXiv（2023-10-09）[2026-01-20]. https://doi.org/10.48550/arXiv.2306.12001.

[46] Derczynski L, Kirk H R, Balachandran V, et al. Assessing language model deployment with risk cards[PP/OL]. arXiv（2023-03-31）[2026-01-18]. https://doi.org/10.48550/arXiv.2303.18190.

[47] Yu Y M, Liu Y R, Zhang J, et al. Understanding generative AI risks for youth: A taxonomy based on empirical data[PP/OL]. V2. arXiv（2025-02-25）[2026-01-20]. https://doi.org/10.48550/arXiv.2502.16383.

[48] 全国网络安全标准化技术委员会秘书处. 人工智能安全标准体系（V1.0）（征求意见稿）[R]. 北京: 全国网络安全标准化技术委员会秘书处, 2025: 1.
Secretariat of National Information Security Standardization Technical Committee. Artificial intelligence security standard system（version 1.0）（draft for comments）[R]. Beijing: Secretariat of National Information Security Standardization Technical Committee, 2025: 1.

[49] Qi X Y, Zeng Y, Xie T H, et al. Fine-tuning aligned language models compromises safety, even when users do not intend to![C]//The Twelfth International Conference on Learning Representations, 2024.

[50] Sun H, Zhang Z X, Deng J W, et al. Safety assessment of Chinese large language models[PP/OL]. arXiv（2023-04-20）[2026-01-20]. https://doi.org/10.48550/arXiv.2304.10436.

[51] He J Y, Feng W T, Min Y S, et al. Control risk for potential misuse of artificial intelligence in science[PP/OL]. arXiv（2023-12-11）[2026-01-20]. https://doi.org/10.48550/arXiv.2312.06632.

[52] Hua W Y, Yang X J, Jin M Y, et al. TrustAgent: Towards safe and trustworthy LLM-based agents[C]// The 2024 Conference on Empirical Methods in Natural Language Processing, 2024: 10000-10016.

[53] Papernot N, McDaniel P, Sinha A, et al. SoK: Security and privacy in machine learning[C]//2018 IEEE European Symposium on Security and Privacy（EuroS&P）, 2018: 399-414.

[54] Yao Y F, Duan J H, Xu K D, et al. A survey on large language

model（LLM）security and privacy：The good，the bad，and the ugly[J]. High-Confidence Computing，2024，4(2)：100211.

[55] Liu H C，Wang Y Q，Fan W Q，et al. Trustworthy AI：A computational perspective[J]. ACM Transactions on Intelligent Systems and Technology，2023，14(1)：1-59.

[56] Vassilev A，Oprea A，Fordyce A，et al. Adversarial machine learning：A taxonomy and terminology of attacks and mitigations：NIST AI 100-2e2023[R]. Gaithersburg：National Institute of Standards and Technology，2024：6-13.

[57] Shayegani E，Al Mamun M A，Fu Y，et al. Survey of vulnerabilities in large language models revealed by adversarial attacks[PP/OL]. arXiv（2023-10-16）[2026-01-20]. https：//doi.org/10.48550/arXiv.2310.10844.

[58] OpenAI. GPT-4o system card[EB/OL].（2024-08-08）[2026-01-18]. https：//cdn.openai.com/gpt-4o-system-card.pdf.

[59] Purpura A，Wadhwa S，Zymet J，et al. Building safe GenAI applications：An end-to-end overview of red teaming for large language models[C]//The 5th Workshop on Trustworthy NLP（TrustNLP 2025），2025：335-350.

[60] Shen X Y，Chen Z Y，Backes M，et al. "Do anything now"：Characterizing and evaluating in-the-wild jailbreak prompts on large language models[C]//The 2024 ACM SIGSAC Conference on Computer and Communications Security，2024：1671-1685.

[61] Liu Y，Deng G L，Xu Z Z，et al. A hitchhiker's guide to jailbreaking ChatGPT via prompt engineering[C]//The 4th International Workshop on Software Engineering and AI for Data Quality in Cyber-Physical Systems/Internet of Things，2024：12-21.

[62] Wang J X，Liu Z C，Park K H，et al. Adversarial demonstration attacks on large language models[PP/OL]. V2. arXiv（2023-10-14）[2026-01-20]. https：//doi.org/10.48550/arXiv.2305.14950.

[63] Zhou X Y，Qiang Y，Zade S Z，et al. Hijacking large language models *via* adversarial in-context learning[PP/OL]. V3. arXiv（2025-05-29）[2026-08-18]. https：//doi.org/10.48550/arXiv.2311.09948.

[64] Anil C，Durmus E，Panickssery N，et al. Many-shot jailbreaking[C]//The Thirty-Eighth Annual Conference on Neural Information Processing Systems，2024：129696-129742.

[65] Yong Z X，Menghini C，Bach S H. Low-resource languages jailbreak GPT-4[PP/OL]. V2. arXiv（2024-01-27）[2026-01-20]. https：//doi.org/10.48550/arXiv.2310.02446.

[66] Li J，Liu Y，Liu C Y，et al. A cross-language investigation into jailbreak attacks in large language models[PP/OL]. arXiv（2024-01-30）[2026-01-20]. https：//doi.org/10.48550/arXiv.2401.16765.

[67] Jiang F Q，Xu Z C，Niu L Y，et al. ArtPrompt：ASCII art-based jailbreak attacks against aligned LLMs[C]//The 62nd Annual Meeting of the Association for Computational Linguistics，2024：15157-15173.

[68] Yuan Y L，Jiao W X，Wang W X，et al. GPT-4 is too smart to be safe：Stealthy chat with LLMs *via* cipher[C]//The Twelfth International Conference on Learning Representations，2024.

[69] Ren Q B，Gao C，Shao J，et al. CodeAttack：Revealing safety generalization challenges of large language models *via* code completion[C]//The 62nd Annual Meeting of the Association for Computational Linguistics，2024：11437-11452.

[70] Perez E，Huang S，Song F，et al. Red teaming language models with language models[C]//The 2022 Conference on Empirical Methods in Natural Language Processing，2022：3419-3448.

[71] Liu X G，Xu N，Chen M H，et al. AutoDAN：Generating stealthy jailbreak prompts on aligned large language models[C]//The Twelfth International Conference on Learning Representations，2024：1-21.

[72] Deng G L，Liu Y，Li Y K，et al. MASTERKEY：Automated jailbreaking of large language model chatbots[C]//The 2024 Network and Distributed System Security Symposium（NDSS），2024：1-18.

[73] Yu J H，Lin X W，Yu Z，et al. LLM-fuzzer：Scaling assessment of large language model jailbreaks[C]//The 33rd USENIX Security Symposium（USENIX Security 24），2024：4657-4674.

[74] Jha P，Arora A，Ganesh V. LLM stinger：Jailbreaking LLMs using RL fine-tuned LLMs（Student Abstract）[C]//The AAAI Conference on Artificial Intelligence，2025：29393-29395.

[75] Lee S，Ni S W，Wei C，et al. xJailbreak：Representation space guided reinforcement learning for interpretable LLM jailbreaking[PP/OL]. V2. arXiv（2025-01-30）[2026-01-20]. https：//doi.org/10.48550/arXiv.2501.16727.

[76] Li H Y，Ye J W，Wu J，et al. JailPO：A novel black-box jailbreak framework *via* preference optimization against aligned LLMs[C]//The AAAI Conference on Artificial Intelligence，2025：27419-27427.

[77] Yang X K，Tang X H，Han J Z，et al. The dark side of trust：Authority citation-driven jailbreak attacks on large language models[PP/OL]. arXiv（2024-11-18）[2026-01-20]. https：//doi.org/10.48550/arXiv.2411.11407.

[78] Yang X K，Zhou B Y，Tang X H，et al. Exploiting synergistic cognitive biases to bypass safety in LLMs[C]//The AAAI Conference on Artificial Intelligence，2026：2200-2208.

[79] Zou A，Wang Z F，Carlini N，et al. Universal and transferable adversarial attacks on aligned language models[PP/OL]. V2. arXiv（2023-12-20）[2026-01-20]. https：//doi.org/10.48550/arXiv.2307.15043.

[80] Jia X J，Pang T Y，Du C，et al. Improved techniques for optimization-based jailbreaking on large language models[C]//The Thirteenth International Conference on Learning Representations，2025：1-22.

[81] Liao Z Y，Sun H. AmpleGCG：Learning a universal and transferable generative model of adversarial suffixes for jailbreaking both open and closed LLMs[C]//The First Conference on Language Modeling，2024：1-33.

[82] Guo X G，Yu F X，Zhang H，et al. COLD-attack：Jailbreaking LLMs with stealthiness and controllability[C]//The 41st International Conference on Machine Learning，2024：16974-17002.

[83] Sitawarin C，Mu N，Wagner D，et al. PAL：Proxy-guided black-box attack on large language models[PP/OL]. arXiv（2024-02-15）[2026-01-20]. https：//doi.org/10.48550/arXiv.2402.09674.

[84] Chao P，Robey A，Dobriban E，et al. Jailbreaking black box large language models in twenty queries[C]//2025 IEEE Conference on Secure and Trustworthy Machine Learning（SaTML），2025：23-42.

[85] Mehrotra A，Zampetakis M，Kassianik P，et al. Tree of attacks：Jailbreaking black-box LLMs automatically[C]//The Thirty-Eighth Annual Conference on Neural Information Processing Sys-

tems，2024：61065-61105.

[86] Zeng Y，Lin H P，Zhang J W，et al. How johnny can persuade LLMs to jailbreak them：Rethinking persuasion to challenge AI safety by humanizing LLMs[C]//The 62nd Annual Meeting of the Association for Computational Linguistics，2024：14322-14350.

[87] Li X，Zhou Z K，Zhu J N，et al. DeepInception：Hypnotize large language model to be jailbreaker[PP/OL]. V5. arXiv（2024-11-28）[2026-01-20]. https://doi.org/10.48550/arXiv.2311.03191.

[88] Weng Z X，Jin X L，Jia J Y，et al. Foot-in-the-door：A multi-turn jailbreak for LLMs[C]//The 2025 Conference on Empirical Methods in Natural Language Processing，2025：1939-1950.

[89] Zhou Z H，Xiang J Y，Chen H P，et al. Speak out of turn：Safety vulnerability of large language models in multi-turn dialogue[PP/OL]. V2. arXiv（2024-10-30）[2026-01-20]. https://doi.org/10.48550/arXiv.2402.17262.

[90] Yu E X，Li J，Liao M，et al. CoSafe：Evaluating large language model safety in multi-turn dialogue coreference[C]//The 2024 Conference on Empirical Methods in Natural Language Processing，2024：17494-17508.

[91] Jiang Y F，Aggarwal K，Laud T，et al. Red Queen：Exposing latent multi-turn risks in large language models[C]//The 63rd Annual Meeting of the Association for Computational Linguistics，2025：25554-25591.

[92] Zhang J C，Zhou Y，Liu Y X，et al. Holistic automated red teaming for large language models through top-down test case generation and multi-turn interaction[C]//The 2024 Conference on Empirical Methods in Natural Language Processing，2024：13711-13736.

[93] Yang X K，Zhou B Y，Tang X H，et al. Chain of attack：Hide your intention through multi-turn interrogation[C]//The 63rd Annual Meeting of the Association for Computational Linguistics，2025：9881-9901.

[94] Wang F X，Duan R J，Xiao P，et al. MRJ-agent：An effective jailbreak agent for multi-round dialogue[PP/OL]. V2. arXiv（2025-01-07）[2026-01-20]. https://doi.org/10.48550/arXiv.2411.03814.

[95] Xiong X Q，Li O X，Liu Z，et al. TROJail：Trajectory-level optimization for multi-turn large language model jailbreaks with process rewards[C]//The 64th Annual Meeting of the Association for Computational Linguistics，2026：48086-48109.

[96] Guo W Y，Li J，Wang W Y，et al. MTSA：Multi-turn safety alignment for LLMs through multi-round red-teaming[C]//The 63rd Annual Meeting of the Association for Computational Linguistics ，2025：26424-26442.

[97] Hartvigsen T，Gabriel S，Palangi H，et al. ToxiGen：A large-scale machine-generated dataset for adversarial and implicit hate speech detection[C]//The 60th Annual Meeting of the Association for Computational Linguistics，2022：3309-3326.

[98] Lees A，Tran V Q，Tay Y，et al. A new generation of perspective API：Efficient multilingual character-level transformers[C]//The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining，2022：3197-3207.

[99] Hosseini H，Kannan S，Zhang B S，et al. Deceiving Google's perspective API built for detecting toxic comments[PP/OL]. arXiv（2017-02-27）[2026-01-20]. https://doi.org/10.48550/arXiv.1702.08138.

[100] Markov T，Zhang C，Agarwal S，et al. A holistic approach to undesired content detection in the real world[C]//The AAAI Conference on Artificial Intelligence，2023：15009-15018.

[101] OpenAI. Moderation[EB/OL]. [2026-01-22]. https://platform.openai.com/docs/guides/moderation.

[102] Mazeika M，Phan L，Yin X W，et al. HarmBench：A standardized evaluation framework for automated red teaming and robust refusal[C]//The 41st International Conference on Machine Learning，2024：35181-35224.

[103] Wang Y X，Li H N，Han X D，et al. Do-not-answer：A dataset for evaluating safeguards in LLMs[PP/OL]. V2. arXiv（2023-09-04）[2026-01-20]. https://doi.org/10.48550/arXiv.2308.13387.

[104] Hu D，Bao Y N，Wei L W，et al. Supervised adversarial contrastive learning for emotion recognition in conversations[C]//The 61st Annual Meeting of the Association for Computational Linguistics，2023：10835-10852.

[105] Hu D，Wei L W，Liu Y X，et al. Structured probabilistic coding [C]//The AAAI Conference on Artificial Intelligence，2024：12491-12501.

[106] Chen Z Y，Yu H M，Wu X，et al. Libra：Large Chinese-based safeguard for AI content[C]//Natural Language Processing and Chinese Computing：14th National CCF Conference，2025：567-580.

[107] Inan H，Upasani K，Chi J F，et al. Llama guard：LLM-based input-output safeguard for human-AI conversations[PP/OL]. arXiv（2023-12-07）[2026-01-20]. https://doi. org/10.48550/arXiv.2312.06674.

[108] Zeng W J，Liu Y C，Mullins R，et al. ShieldGemma：Generative AI content moderation based on gemma[PP/OL]. V2. arXiv（2024-08-04）[2026-01-20]. https://doi.org/10.48550/arXiv.2407.21772.

[109] Souly A，Lu Q Y，Bowen D，et al. A StrongREJECT for empty jailbreaks[C]//The Thirty-Eighth Annual Conference on Neural Information Processing Systems，2024：125416-125440.

[110] Li L J，Dong B W，Wang R H，et al. SALAD-bench：A hierarchical and comprehensive safety benchmark for large language models[C]//The 62nd Annual Meeting of the Association for Computational Linguistics，2024：3923-3954.

[111] Luo H C，Gu J D，Liu F Y，et al. An image is worth 1000 lies：Adversarial transferability across prompts on vision-language models[C]//The Twelfth International Conference on Learning Representations，2024.

[112] Dou Z H，Hu X，Yang H B，et al. Adversarial attacks to multimodal models[C]//The 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis，2024：35-46.

[113] Yin Z Y，Ye M C，Zhang T R，et al. VLATTACK：Multimodal adversarial attacks on vision-language tasks *via* pre-trained models [C]//The Thirty-Seventh Annual Conference on Neural Information Processing Systems，2023：52936-52956.

[114] Cheng H，Xiao E J，Yang J Y，et al. Transfer attack for bad and good：Explain and boost adversarial transferability across multimodal large language models[C]//The 33rd ACM International Conference on Multimedia，2025：5010-5019.

[115] Ye M，Rong X K，Huang W K，et al. A survey of safety on large vision-language models：Attacks，defenses and evaluations[PP/OL]. arXiv（2025-02-14）[2026-01-20]. https://doi. org/10.48550/

arXiv.2502.14881.

[116] Gong Y H，Ran D S，Liu J M，et al. Figstep：Jailbreaking large vision-language models *via* typographic visual prompts[C]//The AAAI Conference on Artificial Intelligence，2024：18748-18756.

[117] Bagdasaryan E，Hsieh T Y，Nassi B，et al. Abusing images and sounds for indirect instruction injection in multi-modal LLMs[PP/OL]. V4. arXiv（2023-10-03）[2026-01-20]. https://doi.org/10.48550/arXiv.2307.10490.

[118] Huang S，Papernot N，Goodfellow I，et al. Adversarial attacks on neural network policies[PP/OL]. arXiv（2017-02-08）[2026-01-20]. https://doi.org/10.48550/arXiv.1702.02284.

[119] Microsoft. CVE-2025-32711：M365 copilot information disclosure vulnerability[EB/OL].（2025-06-11）[2026-01-18]. https://msrc.microsoft.com/update-guide/vulnerability/CVE-2025-32711.

[120] Díaz S，Kern C，Olive K. Google's approach for secure AI agents：An introduction[EB/OL].（2025-05）[2026-01-18]. https://storage.googleapis.com/gweb-research2023-media/pubtools/1018686.pdf.

[121] Debenedetti E，Zhang J，Balunovic M，et al. AgentDojo：A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents[C]//The Thirty-Eighth Annual Conference on Neural Information Processing Systems，2024：82895-82920.

[122] Zhang H，Huang J，Mei K，et al. Agent security bench（ASB）：Formalizing and benchmarking attacks and defenses in LLM-based agents[C]//The Thirteenth International Conference on Learning Representations，2025：1-36.

[123] Vijayvargiya S，Soni A B，Zhou X H，et al. OpenAgentSafety：A comprehensive framework for evaluating real-world AI agent safety[C]//The Fourteenth International Conference on Learning Representations，2026：1-26.

[124] Lynch A，Wright B，Larson C，et al. Agentic misalignment：How LLMs could be insider threats[PP/OL]. V2. arXiv（2025-10-16）[2026-01-20]. https://doi.org/10.48550/arXiv.2510.05179.

[125] Kutasov J，Sun Y Q，Colognese P，et al. SHADE-arena：Evaluating sabotage and monitoring in LLM agents[PP/OL]. V2. arXiv（2025-07-08）[2026-01-20]. https://doi.org/10.48550/arXiv.2506.15740.

[126] Gu X M，Zheng X S，Pang T Y，et al. Agent S mith：A single image can jailbreak one million multimodal LLM agents exponentially fast[C]//The 41st International Conference on Machine Learning，2024：16647-16672.

[127] Hammond L，Chan A L，Clifton J，et al. Multi-agent risks from advanced AI[PP/OL]. arXiv（2025-02-19）[2026-01-18]. https://doi.org/10.48550/arXiv.2502.14143.

[128] Kuntz T，Duzan A，Zhao H，et al. OS-harm：A benchmark for measuring safety of computer use agents[C]//The Thirty-Ninth Annual Conference on Neural Information Processing Systems，2025：50396-50427.

[129] Yang J Y，Shao S，Liu D R，et al. RiOSWorld：Benchmarking the risk of multimodal computer-use agents[C]//The Thirty-Ninth Annual Conference on Neural Information Processing Systems，2025：9322-9369.

[130] Liao Z Y，Jones J，Jiang L X，et al. RedTeamCUA：Realistic adversarial testing of computer-use agents in hybrid web-OS environments[C]//The Fourteenth International Conference on Learning Representations，2026：1-46.

[131] Wu L X，Wang C，Liu T M，et al. From assistants to adversaries：Exploring the security risks of mobile LLM agents[PP/OL]. V2. arXiv（2025-05-20）[2026-01-20]. https://doi.org/10.48550/arXiv.2505.12981.

[132] Liang X，Niu S M，Li Z Y，et al. SafeRAG：Benchmarking security in retrieval-augmented generation of large language model[C]//The 63rd Annual Meeting of the Association for Computational Linguistics，2025：4609-4631.

[133] Zou W，Geng R P，Wang B H，et al. PoisonedRAG：Knowledge corruption attacks to retrieval-augmented generation of large language models[C]//The 34th USENIX Security Symposium，2025：3827-3844.

[134] Perçin S，Su X，Syed Q S，et al. Investigating the robustness of retrieval-augmented generation at the query level[C]//The Fourth Workshop on Generation，Evaluation and Metrics（GEM^2），2025：439-457.

[135] Tabassi E. Artificial intelligence risk management framework（AI RMF 1.0）：NIST AI 100-1[R]. Gaithersburg：National Institute of Standards and Technology，2023：1-5.

[136] International Network of AI Safety Institutes. Joint statement on risk assessment of advanced AI systems[EB/OL].（2024-11-20）[2026-01-18]. https://www.nist.gov/document/joint-statement-risk-assessment-advanced-ai-systems-international-network-aisis.

[137] Brown T B，Mann B，Ryder N，et al. Language models are few-shot learners[C]//The Thirty-Fourth Annual Conference on Neural Information Processing Systems，2020：1877-1901.

[138] Chen M，Tworek J，Jun H，et al. Evaluating large language models trained on code[PP/OL]. V2. arXiv（2021-07-14）[2026-01-20]. https://doi.org/10.48550/arXiv.2107.03374.

[139] Ahmad L，Agarwal S，Lampe M，et al. OpenAI's approach to external red teaming for AI models and systems[PP/OL]. arXiv（2025-01-24）[2026-01-20]. https://doi.org/10.48550/arXiv.2503.16431.

[140] OpenAI. DALL·E 2 Preview：Risks and limitations[EB/OL].（2022-04-11）[2026-01-18]. https://github. com/openai/dalle-2-preview/blob/main/system-card.md.

[141] OpenAI. DALL·E 3 system card[R]. San Francisco：OpenAI，2023：9-13.

[142] Sharma M，Tong M，Mu J，et al. Constitutional classifiers：Defending against universal jailbreaks across thousands of hours of red teaming[PP/OL]. arXiv（2025-01-31）[2026-01-20]. https://doi.org/10.48550/arXiv.2501.18837.

[143] Anthropic. Activating AI safety level 3 protections[EB/OL].（2025-05-22）[2026-01-18]. https://www. anthropic. com/news/activating-asl3-protections.

[144] Weidinger L，Mellor J F J，Pegueroles B G，et al. STAR：Socio-Technical approach to red teaming language models[C]//The 2024 Conference on Empirical Methods in Natural Language Processing，2024：21516-21532.

[145] Fabian D，Crisp J. Why red teams play a central role in helping organizations secure AI systems[EB/OL].（2023-07）[2026-01-18]. https://services.google.com/fh/files/blogs/google_ai_red_team_digital_final.pdf.

[146] Radharapu B，Robinson K，Aroyo L，et al. AART：AI-assisted

red-teaming with diverse data generation for new LLM-powered applications[C]//The 2023 Conference on Empirical Methods in Natural Language Processing:Industry Track,2023:380-395.

[147] Deng B Y, Wang W J, Feng F L, et al. Attack prompt generation for red teaming and defending large language models[C]//The 2023 Conference on Empirical Methods in Natural Language Processing,2023:2176-2189.

[148] Grattafiori A, Evtimov I, Bitton J, et al. Taming the beast: Inside the LLaMA 3 red team process[EB/OL]. 2024[2026-01-18]. https://www.youtube.com/watch?v=UQaNjwLhAmo.

[149] Ge S Y, Zhou C T, Hou R, et al. MART: Improving LLM safety with multi-round automatic red-teaming[C]//The 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2024: 1927-1937.

[150] Bhatt M, Chennabasappa S, Nikolaidis C, et al. Purple llama CyberSecEval: A secure coding benchmark for language models[PP/OL]. arXiv (2023-12-07)[2026-01-20]. https://doi.org/10.48550/arXiv.2312.04724.

[151] NVIDIA. NVIDIA AI red team: An introduction[EB/OL]. (2023-06-14)[2026-01-20]. https://developer.nvidia.com/blog/nvidia-ai-red-team-an-introduction.

[152] NVIDIA. Strength in numbers: NVIDIA and generative red team challenge unleash thousands to vet security at DEF CON[EB/OL]. (2023-08-10)[2026-01-20]. https://blogs.nvidia.com/blog/nvidia-generative-red-team-challenge/.

[153] NVIDIA. Securing LLM systems against prompt injection[EB/OL]. (2023-08-03)[2026-01-20]. https://developer.nvidia.com/blog/securing-llm-systems-against-prompt-injection/.

[154] NVIDIA. Secure LLM tokenizers to maintain application integrity[EB/OL]. (2024-06-27)[2026-01-20]. https://developer.nvidia.com/blog/secure-llm-tokenizers-to-maintain-application-integrity/.

[155] Derczynski L, Galinkin E, Martin J, et al. Garak: A framework for security probing large language models[PP/OL]. arXiv (2024-06-16)[2026-01-20]. https://doi.org/10.48550/arXiv.2406.11036.

[156] Rebedea T, Dinu R, Sreedhar M N, et al. NeMo guardrails: A toolkit for controllable and safe LLM applications with programmable rails[C]//The 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2023: 431-445.

[157] Microsoft. Responsible AI transparency report: How we build, support our customers, and grow[R]. Redmond: Microsoft Corporation, 2024: 9.

[158] Munoz G D L, Minnich A J, Lutz R, et al. PyRIT: A framework for security risk identification and red teaming in generative AI system[PP/OL]. arXiv (2024-10-01)[2026-01-20]. https://doi.org/10.48550/arXiv.2410.02828.

[159] Microsoft. Introducing AI red teaming agent: Accelerate your AI safety and security journey with Azure AI Foundry[EB/OL]. (2025-04-04)[2026-02-08]. https://devblogs.microsoft.com/foundry/ai-red-teaming-agent-preview/.

[160] Wang X, Chen Y H, Li J C, et al. OpenRT: An open-source red teaming framework for multimodal LLMs[PP/OL]. V2. arXiv (2026-01-10)[2026-01-20]. https://doi.org/10.48550/arXiv.2601.01592.

[161] Huang Y, Sun L C, Wang H R, et al. Position: TrustLLM: Trustworthiness in large language models[C]//The 41st International Conference on Machine Learning, 2024: 20166-20270.