Carnegie Mellon University
Software Engineering Institute

What Can Generative AI Red-Teaming Learn from Cyber Red-Teaming?

Anusha Sinha
James Lucassen
Keltin Grimes
Michael Feffer
Ellie Soto
Hoda Heidari
Nathan VanHoudnos

July 2025

TECHNICAL REPORT
CMU/SEI-2025-TR-006
DOI: 10.1184/R1/29410136

CERT Division and AI Division

https://www.sei.cmu.edu

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

DM25-0788

Table of Contents

Executive Summary **iv**

1 Introduction **1**

2 Prior Work **3**
2.1 Cyber Red-Teaming 3
2.2 Generative AI Red-Teaming 4

3 Methods **6**
3.1 Scoping 6
3.2 Planning 6
3.3 Searching 7
3.4 Screening 7
3.5 Eligibility 8
3.6 Synthesis 8
3.7 Limitations 8

4 Systematic Review of Generative AI Red-Teaming Approaches **10**
4.1 Results 10
4.1.1 Red-Teaming Objectives 10
4.1.2 Threat Models and Target Systems 11
4.1.3 Red-Teaming Stages 11
4.1.4 Tools and Techniques 12
4.2 Discussion 13

5 Synthesis of Key Themes in the Cyber Red-Teaming Literature **15**
5.1 Adversary Emulation 15
5.2 Operational Stages of Red-Teaming 15
5.3 Communication with Host Organizations 16
5.4 Comprehensiveness 16
5.5 Diverse and Open Source Tooling 17
5.6 Addressing Well-Known Vulnerabilities First 17
5.7 Standardized Manuals and Methodologies 18

6 Comparative Analysis of the Generative AI and Cyber Red-Teaming Literature **19**
6.1 Goals, Tooling, and Methodology 19
6.2 Comparison of Operational Red-Teaming Stages 20
6.3 Frameworks and Systemization 21

7 Recommendations **23**

8 References **25**

9 Appendix A: Systematic Review Methodology **48**
9.1 Appendix A.1: Search Terms 48
9.2 Appendix A.2: Screened Papers 49
9.2.1 Papers Screened from the Cyber Literature Review 49
9.2.2 Papers Screened from the Generative AI Literature Review 59
9.3 Appendix A.3: Extraction Templates 70
9.4 Appendix A.4: Final Paper List 72

9.4.1 Final Cyber Red-Teaming Paper List 72
9.4.2 Final Generative AI Red-Teaming Paper List 75

10 Appendix B: Operational Stages of Cyber Red-Teaming 83

List of Figures

Figure 1: Focus of Cyber and AI Literature 20

Figure 2: Papers Screened from the Cyber Literature Review at Each Stage 49

Figure 3: Papers Screened from the Generative AI Red-Teaming Review at Each Stage 59

Executive Summary

Red-teaming, a security practice rooted in adversarial emulation, has been widely applied across various domains, including cybersecurity and artificial intelligence (AI). This paper investigates the applicability of established cyber red-teaming methodologies to the evaluation of generative AI systems, addressing the growing need for robust security assessments in AI-driven applications. Through a pair of systematic literature reviews, we synthesize existing generative red-teaming approaches and analyze their alignment with established practices in cyber red-teaming.

Our analysis identifies key challenges in generative AI red-teaming, including inconsistencies in evaluation methodologies, limited threat modeling, and gaps in mitigation strategies. While generative AI red-teaming has made progress in identifying vulnerabilities through techniques such as jailbreaking and adversarial attacks, it lacks standardized frameworks for comprehensive security assessments. In contrast, cyber red-teaming employs well-established methodologies that emphasize adversary emulation, structured engagement stages with stakeholders, and detailed reporting, offering valuable insights for refining generative AI evaluations. Notably, generative AI red-teaming often prioritizes narrow measures of attack success over holistic security improvements, whereas cyber red-teaming integrates pre-engagement planning, post-exploitation analysis, and structured reporting into red-teaming processes to enhance outcomes. By incorporating practices from cyber red-teaming, generative AI red-teaming can evolve from isolated vulnerability identification to more systematic risk mitigation.

As generative AI continues to be deployed in critical domains, establishing rigorous and systematic red-teaming methodologies will be essential to ensuring its safe and reliable use. This paper concludes with recommendations for improving generative AI red-teaming practices, including the adoption of structured threat modeling techniques, the development of standardized evaluation metrics, and improved integration of red-teaming findings into risk mitigation efforts.

Abstract

Red-teaming, a security practice rooted in adversarial emulation, has been widely applied across various domains, including cybersecurity and artificial intelligence (AI). This paper investigates the applicability of established cyber red-teaming methodologies to the evaluation of generative AI systems, addressing the growing need for robust security assessments in AI-driven applications. Through a pair of systematic literature reviews, we synthesize existing generative AI red-teaming approaches and analyze their alignment with established practices in cyber red-teaming.

1 Introduction

Red-teaming is a security practice that involves emulating an attack by an adversary to identify vulnerabilities in the target system [CSRC 2015]. The practice has its roots in ancient military strategy, when commanders would employ adversarial thinking to find weaknesses in their plans [Tzu 2008]. The core idea of red-teaming is very generalizable and has been adopted in cybersecurity [Brangetto 2015], law enforcement [Meeham 2007], business strategy [Sun 2022], and artificial intelligence (AI) evaluation [Microsoft 2024]. This approach is particularly valuable in contexts where potential attack vectors are difficult to enumerate, such as when the risk surface is very broad [Teichmann 2023]. By proactively identifying security gaps, red-teaming helps organizations build more resilient systems, reducing the risk of successful attacks by malicious actors [CISA 2024].

As generative AI systems are increasingly integrated into high-stakes applications such as healthcare [Qiu 2024], finance [de Zarzà 2023], and national security [Gallagher 2024; Swanson 2024], the risks associated with their vulnerabilities grow more severe, and it becomes essential to ensure the reliability and robustness of these systems. Red-teaming has quickly become a critical tool for evaluating the security of generative AI systems [Ahmad 2024] because it enables researchers and developers to explore vulnerabilities such as the generation of harmful content [Boiko 2023, Burtell 2023, Ferrara 2024], susceptibility to adversarial attacks [Zou 2023], proliferation of algorithmic biases [Wei, X. 2025], and potential for misuse [Deshpande 2023]. The rapid pace of development and deployment of these systems further amplifies these risks, creating an urgent need to establish robust red-teaming practices [Bick 2024].

While generative AI red-teaming is increasingly relied upon for security evaluation, there are still significant gaps in our understanding of how to design, implement, and interpret these assessments effectively [Feffer 2024]. This raises an important question: To what extent can established methodologies from cyber red-teaming, which has a longer history of systematically assessing adversarial threats to complex software systems, inform the development of best practices for generative AI red-teaming?

Scholars have already begun exploring how best practices from cyber red-teaming can inform and improve generative AI red-teaming [Cranford 2023]. Red-teaming is a critical practice in the cybersecurity industry because it provides realistic assessments of security posture, mimicking the tactics and techniques of real-world adversaries [CISA 2024]. This allows organizations to strengthen their defenses, implement more effective mitigations, and refine incident response strategies before an actual attack occurs [Proofpoint 2024]. For example, the development of the OWASP Top 10 for Large Language Models (LLMs) builds upon the original OWASP Top Ten framework [OWASP 2024, 2025] and researchers have proposed a Coordinated Flaw Disclosure process for AI vulnerabilities [Cattell 2024] similar to the Coordinated Vulnerability Disclosure process used in cybersecurity [Householder 2020].

However, many of these efforts are in their early stages. There are minimal references to these efforts in the generative AI red-teaming literature, and there is little consensus on which lessons

from cyber red-teaming are most applicable to generative AI. Our goal in this paper is to provide a systematic overview of key differences between these fields and to comprehensively characterize the lessons that generative AI red-teaming can draw from cybersecurity to develop more rigorous and effective evaluation methodologies.

To structure our comparative analysis of generative AI and cyber red-teaming, we seek to first separately understand how evaluation exercises are conducted in each field before comparing activity in both fields. We focus on the following key research questions:

1. What is the current state of the art and practice for generative AI red-teaming?
2. What are the established best practices for red-teaming in cybersecurity?
3. What practices from cyber red-teaming could be used in generative AI red-teaming?

Motivated by these questions, we present two systematic reviews: a systematic review of state of the art generative AI red-teaming literature alongside a systematic review of the cyber red-teaming literature focused on surveys, best practices, and frameworks.

By synthesizing insights from cyber red-teaming, we provide a roadmap for improving generative AI red-teaming practices through a collection of recommendations for researchers and practitioners in the field. Our recommendations build on the following key differences that highlight the maturity gap between the two fields:

1. Cyber red-teaming exercises encompass more operational stages, target more attack surfaces, and emphasize realistic adversaries and threat models to a greater extent than generative AI red-teaming exercises.
2. Cyber red-teaming benefits from a more extensive array of off-the-shelf open source tools and authoritative manuals, many of which lack direct analogues in generative AI red-teaming.
3. Cyber red-teaming is more adept at leveraging well-documented and easily accessible vulnerabilities to improve the efficiency and focus of evaluations, employing specialized tools to this end that have yet to be developed for red-teaming generative AI.

We begin by presenting our findings from the systematic review of generative AI red-teaming research, outlining current approaches, challenges, and gaps in the field. We draw insights from a systematic review of academic research published in the previous calendar year (January–November 2024) and surveys covering work prior to 2024. Next, we synthesize key themes from the cyber red-teaming literature, highlighting the core principles, methodologies, and techniques that define effective cyber red-teaming from existing review literature and industry frameworks. We then present findings from a comparative analysis between these two syntheses to identify where generative AI red-teaming can benefit from cybersecurity's more mature practices. Finally, we present a set of recommendations for researchers and practitioners in generative AI red-teaming that can inform the development of more rigorous, structured, and effective red-teaming practices for generative AI, addressing existing gaps in evaluation methodologies, reporting standards, and adversarial testing frameworks. We hope that our contributions will foster more rigorous security assessments and support the development of comprehensive frameworks for evaluating and defending generative AI systems

2 Prior Work

2.1 Cyber Red-Teaming

Red-teaming is a foundational technique in cybersecurity, where it plays a critical role in assessing and enhancing the security of complex software systems [Abbass 2011]. Cyber red-teaming is a well-established field in academia, industry, and government, with surveys, manuals, books, and courses that define and describe best practices for maximizing red-team effectiveness (e.g., [Kalchenko 2018, Solisch 2022, Yadav 2014]).

Given the broad scope of software systems and diversity of potential adversaries, scholarly reviews in the professional community often focus on specific industry sectors where security needs and threat models differ significantly [Gbormittah 2024]. Others concentrate on specific network assets, such as cloud infrastructure, endpoint devices, or industrial control systems, tailoring red-teaming approaches to the unique vulnerabilities of each environment [Al-Ahmad 2019, Nutalapati 2020, Pozzobon 2018]. Additionally, red-teaming strategies often vary based on the rules of engagement, which define the scope, constraints, and ethical guidelines for conducting security assessments, ranging from full adversarial simulations with no prior knowledge (black-box testing) to cooperative assessments, where defenders are aware of and engaged in the process (white-box testing) [Shah 2014].

Among the reviews that cover the red-teaming space more broadly, there are notable inconsistencies in terminology and differing interpretations of best practices, likely reflecting variations in how red-teaming activities are conducted and analyzed across domains [Adam 2023, Nour 2023]. Our search for reviews on cyber red-teaming did not surface any meta-reviews or systematic studies that synthesize findings across multiple reviews, highlighting a gap in the field. We aim to address this gap to provide a more structured foundation for comparing cyber red-teaming with generative AI red-teaming, where methodologies remain even less standardized.

One highly cited resource in cyber red-teaming is MITRE's ATT&CK (Adversarial Tactics, Techniques, and Common Knowledge) framework, a comprehensive knowledge base that systematically categorizes the tactics, techniques, and procedures (TTPs) used by adversaries throughout the lifecycle of a cyber intrusion [MITRE 2024b]. The framework is structured into different tactics, which represent the overarching goals of an attack (e.g., initial access, privilege escalation, or data exfiltration), and techniques, which describe the specific methods adversaries use to achieve these goals (e.g., supply chain compromise, process injection, or exfiltration over web service). By providing a structured and repeatable approach to adversarial assessments, the ATT&CK framework helps security teams simulate real-world threats, assess vulnerabilities, and refine defensive measures in practice [Al-Sada 2025].

However, MITRE ATT&CK has limitations that are relevant for the red-teaming of AI systems. It does not fully account for victim systems that can dynamically respond to exploitation attempts and potentially alter the course of an attack [Al-Sada 2025]. Additionally, ATT&CK structures attacks in a largely linear progression, without explicitly modeling how adversaries may fluidly

shift between different stages, such as moving backward to re-establish access to a system or bypassing intermediate steps through novel exploit chains [Al-Sada 2025].

Though the framework has some limitations, the structured nature of ATT&CK provides a valuable scaffold for conducting systematic security analyses, offering a level of consistency and repeatability that is currently lacking in generative AI red-teaming. As we explore in the next section, efforts to develop a parallel framework for AI red-teaming could help bring similar methodological structure to the evaluation of generative AI systems, improving the field's ability to systematically identify, categorize, and mitigate threats.

2.2 Generative AI Red-Teaming

With the rapid advancement and deployment of generative AI, red-teaming has emerged as a key technique for identifying safety and security risks in these systems. Unlike cyber red-teaming, which focuses on infrastructure and software vulnerabilities, both our findings and prior work reveal that AI red-teaming primarily targets model behaviors, examining how generative models can be manipulated to produce harmful, biased, or otherwise unintended outputs [Feffer 2024].

A common red-teaming technique in generative AI security is *jailbreaking*, where researchers attempt to bypass or subvert the built-in safety mechanisms of AI models, allowing the generation of restricted, harmful, unhelpful, or unintended outputs [Wei, A. 2023]. This can involve exploiting vulnerabilities in promptly handling [Greshake 2023] or manipulating model responses to produce outputs that violate safety policies [Russinovich 2024]. Jailbreaking is closely related to generative AI red-teaming and may be considered a subset of red-teaming techniques [Feffer 2024]. While AI red-teaming efforts are growing in maturity, they remain relatively ad hoc compared to the more standardized and methodical practices in cybersecurity [Cattell 2024].

One effort to bring greater standardization to AI red-teaming is MITRE's ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) framework [MITRE 2024a], which aims to categorize and document real-world AI threats in a structured format similar to the highly cited ATT&CK framework in cybersecurity. ATLAS outlines tactics and techniques that adversaries may use to exploit AI systems, such as data poisoning, adversarial example generation, and model inversion attacks, offering a knowledge base to guide both offensive and defensive security research. However, despite its value in structuring known AI threats, ATLAS is still in its early stages of adoption and lacks the level of integration, tooling, and real-world validation that ATT&CK has in cybersecurity. Unlike ATT&CK, which is widely used in red-teaming operations and supported by a mature ecosystem of tools [Al-Sada 2025], ATLAS currently has limited tooling, few formalized case studies, and less industry-wide adoption [Feffer 2024]. Additionally, AI attacks are often more dependent on context and less deterministic than traditional cyber exploits, making it more difficult to develop standardized attack patterns and defenses.

As a result, while frameworks like ATLAS provide an important starting point for AI red-teaming, they have yet to reach the maturity needed for systematic adoption across industry and academia. Notably, our review finds that no generative AI red-teaming studies reference ATLAS, suggesting that its role in guiding AI security research is limited.

Analyses focusing specifically on AI red-teaming remain scarce. Existing reviews of generative AI security evaluation approaches cover the whole space of generative AI evaluation methods [Weidinger 2023] or broadly survey LLM research on privacy, security, and other vulnerabilities [Nguyen 2022; Yao, Y. 2024]. To date, only two other reviews have specifically cataloged AI red-teaming methodologies and the risk categories they address [Feffer 2024; Lin, L. 2025]. Feffer and colleagues provide an early overview of trends in generative AI red-teaming up until 2024, mapping attack techniques, target models, and evaluation criteria [Feffer 2024]. Lin and colleagues frame red-teaming research as a search problem and categorize attacks based on this framing [Lin, L. 2025]. Our work builds on and expands these findings by extending the analysis to more recent AI red-teaming efforts and introducing a comparative analysis with cyber red-teaming.

3 Methods

To conduct our two independent literature reviews, we followed the systematic review process outlined by Siddaway and colleagues, covering five stages: scoping, planning, searching, screening, and eligibility [Siddaway 2019].

- During the scoping stage, we defined the research questions and determined the scope of our reviews.
- In the planning stage, we established high-level inclusion and exclusion criteria, selected relevant databases, and designed a structured search strategy.
- The searching stage involved systematically querying academic databases, conference proceedings, and industry reports to gather relevant literature.
- In the screening stage, we reviewed titles and abstracts to filter studies based on relevance.
- In the eligibility stage, we conducted a full-text assessment to ensure alignment with our research objectives.

We conducted this process separately for our review of the cyber red-teaming literature and the generative AI red-teaming literature. The optional study quality stage was not included. After completing these stages, we conducted a comparative synthesis of the cybersecurity and AI red-teaming literature to identify similarities, differences, and emerging trends across both fields.

3.1 Scoping

For cyber red-teaming, we reviewed existing frameworks and academic literature published up to 2024, focusing on well-established reviews, guidelines, and frameworks. Our goal was to synthesize these established resources from a field that is relatively mature rather than investigate emerging techniques. We then conducted a meta-synthesis of these foundational reviews to extract key principles and common practices.

For AI red-teaming, we targeted academic literature from January to November 2024, focusing on works self-identifying as red-teaming or jailbreaking in their titles. We selected this date range to capture the latest methodologies and trends, which is important as the field of generative AI is rapidly evolving and because prior surveys have covered material before this date [Feffer 2024; Lin, L. 2025].

3.2 Planning

For each field, we used a keyword-pair approach consisting of two sets of search terms to select keywords for literature retrieval. For the cyber red-teaming literature, keywords in the first set included terms relating to red-teaming, such as “red-teaming,” “penetration testing,” “ethical hacking,” “vulnerability discovery,” and “cybersecurity assessment.” To identify review papers, we added a term from a second set of keywords, including “review,” “overview,” and “summary.” We used keywords from Feffer and colleagues for the AI red-teaming literature [Feffer 2024]. Keywords in the first set included terms relating to generative models, such as “GenAI” and

"LLM," while the second set of keywords consisted of terms relating to red-teaming, such as "red-teaming" and "jailbreaking." We list all search terms in Appendix A.1.

We included papers sourced from academic literature that provided complete bibliographic information and were written in English. We excluded papers if their primary topics were irrelevant to our research questions or if they exhibited low quality or poor legibility. In the cyber red-teaming review, we included only review papers. In contrast, for the AI red-teaming review, we included only primary literature that explicitly discussed real-world red-teaming activities.

To conduct a comprehensive literature review with the intended scope, we restricted keyword searches to titles instead of abstracts. Comprehensive searching of abstracts was out of scope for our screening budget, as the extensive body of cyber red-teaming research and explosion of interest in generative AI red-teaming presented thousands of papers to screen. To mitigate this, we considered either selecting the top few hundred hits from each search or performing a more restrictive title search to limit the number of hits. We chose to limit our search by title because it provided a higher quality selection of literature based on our initial sampling.

3.3 Searching

We searched Google Scholar using keyword pairs. For cyber red-teaming, we combined one "cyber red-teaming" keyword with one "review" keyword. For AI red-teaming, we combined "genAI" keywords with "AI red-teaming" keywords. Queries were formatted as *intitle:"keyword 1" AND intitle:"keyword 2"* to retrieve papers with relevant titles. This search yielded 471 cyber red-teaming papers and 455 AI red-teaming papers.

3.4 Screening

We applied inclusion and exclusion criteria in the following order:

1. Deduplication by title
2. Removal of entries without valid links
3. Removal of entries without publication dates
4. Removal of non-English entries

The remaining titles and abstracts were manually screened for relevance. The following are some common exclusions:

- Cyber: geotechnical engineering papers, primary papers, pedagogy-focused papers, automated red-teaming frameworks, non-cybersecurity contexts, and poorly legible papers
- AI: defense-focused papers, review papers, and papers not focused on generative AI systems

We provide more detail on the papers screened in Appendix A.2.

Inter-rater reliability was tested with two reviewers. An initial sample of 20 papers from each set showed 72.5% agreement (29/40). After discussing criteria and borderline cases, a second sample received 85% agreement (34/40).

3.5 Eligibility

We read the full texts of the remaining papers and eliminated papers based on additional exclusion criteria. For the cybersecurity papers, we excluded additional papers based on language, topic relevance, or legibility. For the generative AI papers, we excluded 5 additional papers due to focus misalignment or quality issues.

Our final selections included 42 cyber red-teaming papers and 99 AI red-teaming papers. We list the final papers in Appendix A.4. Despite starting with a nearly equal number of initial search results, the final selection contained more than twice as many AI papers. This discrepancy likely stems from two key factors: the relative age of the literature and differences in screening challenges. Cyber red-teaming has a longer history, which led to a higher proportion of older or inaccessible papers. Of the papers screened out for broken links, 164 were cyber papers, compared to just 1 AI paper. Additionally, our keyword-based search had a higher false positive rate for cyber red-teaming, as many retrieved papers focused on broader cybersecurity topics rather than red-teaming specifically. We analyzed all selected papers using an extraction template with standardized questions to ensure consistent data collection across all papers and reviewers. Full extraction templates can be found in Appendix A.3.

To provide quantitative results, we grouped variations in phrasing. For example, "recon," "perform reconnaissance," and "gather information" were grouped under "reconnaissance" unless context indicated distinct meanings, such as differentiating between public information gathering and internal target exploitation.

3.6 Synthesis

After completing the individual reviews, we conducted a comparative synthesis of the cybersecurity and AI red-teaming literature. This synthesis focused on identifying similarities and differences in methodologies, best practices, and emerging trends across the two fields. By juxtaposing the established frameworks from cybersecurity with the evolving practices in AI red-teaming, we highlighted areas of convergence and divergence. This comparison provided insights into how traditional red-teaming principles are being adapted or challenged by the unique requirements of generative AI systems. We detail findings from this synthesis in the following sections.

3.7 Limitations

Our literature reviews only cover academic papers, which leaves out resources such as corporate white papers, blog posts, manuals, books, or community forums. For cybersecurity, much red-teaming activity happens behind closed doors in industry or government, and the same is likely increasingly true for generative AI as models are increasingly used in production systems. Despite this, many of the cyber reviews we analyzed are informed by corporate or government red-teaming results. On the generative AI side, red-teaming is still a relatively new practice for both academia and industry, and our impression is that activities in both sectors are at similar stages of maturity and development. Additionally, our overarching findings align with previous analyses of non-academic generative AI red-teaming [Feffer 2024].

Our specific search terms may also have biased our results towards jailbreaking. In particular, 69 of the 99 AI papers were scraped via the ‘jailbreak’ keyword. This could be viewed as a systematic bias towards jailbreaking papers. However, we believe explicitly including ‘jailbreak’ as a search term fairly captures the AI community’s view on red-teaming. Our search terms follow Feffer et al., who found red-teaming and jailbreaking to have quite similar motivations and techniques [Feffer 2024]. Of the 22 papers found via the ‘red teaming’ keyword, all aim to get the model to answer harmful questions, produce toxic outputs, or generate inappropriate images, which fall under the umbrella of jailbreaking, further reinforcing the similarity.

4 Systematic Review of Generative AI Red-Teaming Approaches

Our review of recent generative AI red-teaming literature found a mix of both promising themes and areas for improvement. The shortcomings included a focus on only a small number of specific attack surfaces and methodologies, lack of consideration for informing mitigations, and inconsistencies and potential issues in evaluation methodologies. These gaps suggest areas where best practices from cyber red-teaming could inform improvements to generative AI red-teaming methods.

4.1 Results

4.1.1 Red-Teaming Objectives

Across all 99 red-teaming activities analyzed in our work, the primary objective of red-teaming was to force the model to elicit harmful, unwanted, or policy-violating outputs. Of these, 85 focused on jailbreaking (inducing responses to malicious queries or performing harmful actions), while others aimed at generating toxic (7) or biased (2) outputs, degrading model performance (4), or evading detection (1). One study also leveraged jailbreak methods to leak training data [Zhang, H. 2024]. In 84 cases, the primary goal of the red-teaming activity was to maximize the Attack Success Rate (ASR), or the percentage of cases that produced a harmful or unwanted output, according to some evaluation criteria. Other common goals included score-based approaches to measuring harm (12), diversity of successful attacks (11), relevance of generation to original prompt (7), and toxicity level (4). This strong emphasis on ASR highlights a tendency to optimize attack efficacy, often without thorough consideration of the real-world impact of these vulnerabilities.

Additionally, some studies characterized harms using OpenAI's usage policy categories (13) [OpenAI 2025] or the MLCommons AI Safety taxonomy. The latter was referenced in 2 studies, with an additional 12 studies making use of Llama Guard models [Chi 2024], which are built around this taxonomy [Vidgen 2024]. Many studies also drew harmful prompts for red-teaming from open source datasets or benchmarks, such as AdvBench (39) [Zou 2023], HarmBench (8) [Mazeika 2024], or the Anthropic HH-RLHF dataset (8) [Bai 2022]. The usage of these frameworks and datasets demonstrates recent efforts to standardize harm evaluation.

However, despite these efforts, there is limited discussion on how these red-teaming activities contribute to mitigating vulnerabilities in practical settings. Among the studies analyzed, only 34 focused on evaluating actual defenses of generative AI based attacks, with just 12 demonstrating effective mitigation of their own red-teaming approaches. While 36 papers recommended mitigations to their introduced attacks, only 18 rigorously tested the effectiveness of proposed mitigations. This highlights a significant gap between attack development and meaningful defensive progress, suggesting that current generative AI red-teaming research is more focused on identifying and exploiting vulnerabilities than on developing and validating robust defense mechanisms to address them.

4.1.2 Threat Models and Target Systems

Although only 17 papers explicitly defined their threat model as white- or black-box, our analysis revealed that 72 papers relied solely on black-box attacks, while 26 utilized white-box attacks. Moreover, 11 of these white-box attack papers incorporated transfer attacks (e.g., developing a white-box attack then transferring it to black-box settings). Most attacks (93) operated on direct, valid inputs to the model or system. Of the remaining cases, 5 poisoned a Retrieval Augmented Generation (RAG) database so that malicious inputs would automatically be retrieved and placed into future model inputs, and 1 fine-tuned a model on malicious datasets. External red-teamers were recruited in only 3 studies [Deng, D. 2024, Dominique 2024, Weidinger 2024], while in almost all cases, red-teamers were simply the authors experimenting with their method and baselines.

Text-based LLMs were the primary targets (79 studies), but some papers explored multimodal models, including LLMs with image input (9), audio input (2), video input (1), and image output (1). Five studies examined text-to-image models, one paper used both text-to-image models and LLMs with image output, and another investigated text-to-motion models. This distribution suggests that while text-based models remain the dominant focus, there is increasing interest in evaluating vulnerabilities in multimodal systems, which could prevent novel risks as well as challenges for red teams.

Among the 65 papers targeting closed-source models, OpenAI's GPT series was the most tested (64), followed by Anthropic's Claude (18) and Google's Gemini (18). The prevalence of OpenAI's models in red-teaming research suggests that these systems are considered key benchmarks in AI security evaluations, likely due to their widespread adoption and integration into various applications. Because these models do not provide access to their internal architectures or training data, red-teaming efforts to evaluate their vulnerabilities have largely relied on black-box attack strategies. Conversely, open-source models were used in 82 studies, with frequent targets including Meta's Llama (70), LMSYS' Vicuna (29), Mistral (22), and Alibaba's Qwen (19). This trend highlights the importance of open source LLMs in security research, as they offer accessibility for thorough testing and reproducibility of red-teaming methodologies. However, the open nature of these models can also raise concerns about the broader dissemination of potential vulnerabilities and transfer attacks.

4.1.3 Red-Teaming Stages

Using the MITRE ATLAS framework, we categorized red-teaming activities to understand how different attack techniques are applied and which aspects remain underexplored. Our findings indicate that privilege escalation and defensive evasion were the most frequently employed tactics, primarily through jailbreaking and prompt injection. Reconnaissance, resource development, machine learning (ML) attack staging, and impact were also applied universally across all studies, often implicitly (e.g., all reviewed papers discuss related red-teaming research in their introduction and/or a 'Related Work' section, but we see no explicit mention of reconnaissance activities).

More sophisticated attack stages, such as persistence, execution, and credential access, were rarely explored. Despite the extensive focus on privilege escalation, only a few studies (3) investigated persistence mechanisms, such as poisoning RAG databases to maintain access or exploring the

long-term effects of jailbreaking on model responses. Similarly, execution-based techniques, such as scripting or plugin compromise, were absent from all studies. Only one study examined exfiltration by investigating training data leakage, highlighting the lack of research into potential data extraction risks. While all studies demonstrated some form of impact analysis, typically through the erosion of model integrity, none analyzed real-world consequences or downstream harms, suggesting that red-teaming research often stops at demonstrating vulnerabilities rather than assessing their broader implications.

Beyond planning and conducting attacks, red-teaming also involves reporting and disclosure, which play a crucial role in mitigating identified risks. Our analysis reveals inconsistencies and gaps in reporting practices. While all reviewed studies resulted in publicly available research papers, only 42 released their research code without restrictions, while 7 stated intentions to publish code but did not provide accessible links, and 2 released code with access restrictions. Transparency regarding resource consumption was particularly lacking. Many (82) papers did not provide details on the costs and resources consumed by their red-teaming exercises, and of those that did, only 3 specified both the time and monetary costs. Similarly, only 11 papers shared datasets produced during their red-teaming activities, 4 of which were only available by request. Most notably, only 8 papers reported engaging in responsible disclosure of the vulnerabilities they identified, all of which involved disclosures to LLM providers.

Overall, our analysis of the operational stages of generative AI red-teaming reveals that it tends to focus on narrow objectives, often emphasizing specific attack techniques rather than comprehensive security assessments. Moreover, most efforts in this space come from individual researchers and academic institutions rather than dedicated security teams or red-teaming professionals, contributing to the fragmented and exploratory nature of the field. While this is a nascent and rapidly evolving area, the lack of structured frameworks and collaboration across sectors suggests significant room for growth.

4.1.4 Tools and Techniques

Across red-teaming activities, we see a wide variety of tools and approaches applied with a small number of commonly used resources. Generally, automated methods dominated harm evaluation (79 cases), with limited reliance on manual assessment. Only 5 studies used manual methods exclusively, while 8 combined manual and automated evaluations and 7 lacked sufficient details. The preference for leveraging advanced LLMs for evaluating generated content, such as GPT-4 (19), GPT-3.5 (11), GPT-4o (10), Llama Guard models (12), and various BERT-like models (10), reflects an increasing reliance on automated classifiers. However, the variability in evaluation criteria used with these automated approaches does not eliminate concerns about consistency and reproducibility across studies.

Methods for generating attacks varied significantly, with 31 studies employing manual inputs or single-shot LLM-generated prompts, 27 refining prompts iteratively through LLMs, 21 leveraging optimization techniques, and 18 implementing LLM-guided search methods. The predominance of single-turn attack strategies (78 studies) over multi-turn (13 studies) and multi-agent interactions (2 studies) suggests that more work is needed to explore adversarial dynamics in complex, long-term engagements.

Popular baselines for red-teaming included GCG (30) [Zou 2023], PAIR (26) [Chao 2024b], AutoDAN (15) [Liu, X. 2024b], DeepInception (9) [Li, X. 2024c], GPTFuzzer (9) [Yu, J. 2024a], and TAP (9) [Mehrotra 2025]. The widespread reliance on these baselines suggests that red-teaming research is beginning to consolidate around a core set of methodologies. Though further innovation is needed to address emerging and evolving threats, our results indicate a growing emphasis on measuring and quantifying safety risks.

4.2 Discussion

Our findings reveal a significant gap between theoretical red-teaming efforts and practical AI security improvements. While many studies successfully expose vulnerabilities, they often fail to analyze how these weaknesses could be exploited in real-world adversarial settings, nor do they contribute to meaningful mitigations. Most research remains focused on short-term vulnerabilities rather than long-term system compromise, limiting its impact on broader security considerations. Future work should expand beyond immediate attack efficacy by exploring, for example, more persistent threats, credential access risks, or the long-term security implications of red-teaming activities. By incorporating more comprehensive threat models, research could better inform proactive defenses and mitigation strategies.

One of the most striking patterns in the literature is the overwhelming emphasis on jailbreak attacks. These attacks undoubtedly uncover vulnerabilities that are both new [Sheng 2019] and pose real risks to organizations and individuals [Weidinger 2021], but they are not the only relevant risk vector. For example, we saw no red-teaming studies of training data poisoning [Carlini 2024a, Rando 2024b], privacy leakage [Nasr 2023], or model stealing [Carlini 2024b], which could readily be integrated with red-teaming approaches. Additionally, traditional cyber techniques, such as gaining unauthorized account access, man in the middle attacks, denial of service attacks, and replay attacks, remain largely unexplored in generative AI red-teaming. Expanding red-teaming methodologies to include these vectors could provide a more holistic assessment of generative AI risks.

The dominant focus on maximizing ASR in jailbreaking attacks further raises concerns about evaluation practices. While ASR measures how often a model bypasses safeguards, it does little to indicate the actual severity of harms posed by a successful attack. In addition, research has shown that automated metrics can have low agreement with human evaluators when assessing the success of jailbreak attacks [Mazeika 2024, Souly 2025]. Both Mazeika and colleagues and Souly and colleagues introduce improved LLM-based classifiers to address these shortcomings, though we only see limited adoption of these methods (5 and 3 uses, respectively). Though issues are more prevalent with automated evaluations, even human evaluators can have inconsistent judgements in some cases [Thomas 2025]. Furthermore, Mazeika and colleagues find that many of the common harms evaluated can easily be accomplished through an online search and emphasize the importance of *differentially* harmful behavior—behaviors for which the jailbreaking attack method is likely the easiest pathway to accomplishing that harm. These findings suggest two key areas for improvement: the need for more rigorous and standardized evaluation procedures and a shift toward prioritizing risks where generative AI systems create novel or amplified threats. Considering more practical threat models and adversary goals may help refine methodologies in a way that better aligns with real-world security challenges.

Another gap in the literature is the discussion of legal protections for good-faith AI red-teaming, or so-called 'safe harbors,' which has been a recent area of concern [Longpre 2024]. In general, there is significant ambiguity about the extent to which AI red-teaming research is allowed by various organizations. Some organizations extend traditional cybersecurity responsible disclosure policies or bug-bounty programs to include AI research [Meta 2024, OpenAI 2023, Vela 2023]. Other organizations lack clear policies for red-teaming. For example, we were unable to determine whether AI red-teaming research was explicitly allowed under the usage policies for models developed by Anthropic and Mistral [Anthropic 2025, Mistral 2025]. We also see only two references to the usage of ethics committees [Liu, Y. 2024a; Weidinger 2024]. This gap in legal and ethical considerations raises concerns about the extent to which red-teaming research can be conducted safely and transparently. Without explicit policies or institutional backing, researchers may face legal uncertainties or ethical dilemmas when engaging in adversarial testing. Future work should explore frameworks for responsible red-teaming, ensuring that research efforts are both legally protected and aligned with ethical best practices.

The prevalence of vulnerabilities in generative AI models also raises questions about reporting and reproducibility. Given that jailbreak vulnerabilities have been well-established, even prior to our search period [Ganguli 2022, Perez 2022, Wei, A. 2023], and effective jailbreak prompts can readily be found online [Chao 2024a], the potential harm of publicizing an exploit is limited. All literature reviewed was publicly and freely available, and many authors publicly released code or datasets to replicate their red-teaming activities, which may be helpful for future red-teaming research or mitigation strategies. However, responsible disclosure was rarely performed, which may be because these vulnerabilities are well known.

As the capabilities of generative AI models increase, the potential harm from jailbreak attacks also increases, yet few papers demonstrated potentially effective mitigations for their exploits. In fact, the primary goal of many of the papers reviewed was to maximally *exploit* this vulnerability (to maximize ASR). Recent research has suggested that mitigating specific, highly effective exploits does not translate well to unseen attacks, so pursuing other red-teaming goals, such as attack diversity, may be more effective [Lee 2024]. Expanding beyond a narrow focus on known vulnerabilities and integrating a wider range of adversarial techniques may enhance the long-term security impact of AI red-teaming research.

Overall, while generative AI red-teaming research has made significant strides in identifying vulnerabilities, there remain critical gaps in evaluation, threat modeling, legal protections, and mitigation strategies. Addressing these gaps will be essential in ensuring that red-teaming efforts contribute not only to identifying weaknesses but also to strengthening generative AI security in practice.

5 Synthesis of Key Themes in the Cyber Red-Teaming Literature

Our meta-review of the cyber red-teaming literature reviews and surveys revealed a number of key themes in best practices for cyber red-teaming. These themes include adversary emulation, clear and well-structured communication, comprehensiveness in attack coverage, diverse and open source tools, leveraging "low-hanging fruit," and applying standardized manuals and methodologies. Each of these elements plays a role in the effectiveness of cyber red-teaming efforts.

5.1 Adversary Emulation

A central tenet of cyber red-teaming is adversary emulation, a practice where red teams seek to simulate real-world cyberattacks as accurately as possible. The review literature and surveys reflect a strong consensus on this approach. Out of the 22 papers reviewed that provided an explicit definition of red-teaming or penetration testing, 18 defined it in terms of adversary emulation. For example, Teichmann & Boticiu describe red-teaming as "a complete simulation of a cyber-attack in which experts use various tools and techniques to mimic the attack," while Gbormittah refers to penetration testing as "a technique for assessing system security through simulated cyberattacks" [Teichmann 2023, Gbormittah 2024]. Emulating realistic attack behaviors allows red teams to identify vulnerabilities that genuine threat actors are likely to exploit. Because these vulnerabilities mirror the pathways real attackers would take, they represent the most immediate and consequential risks to an organization. Addressing these vulnerabilities not only strengthens defenses against known threats but can also uncover and protect against emerging attack techniques.

5.2 Operational Stages of Red-Teaming

We identified 24 papers in our review that broke down the cyber red-teaming process into stages. We selected the most frequently mentioned stages and ordered them chronologically to identify the major stages of a cyber red-teaming engagement. These stages were pre-engagement (10 mentions), threat modeling (7), reconnaissance (15), scanning (8), vulnerability analysis (5), initial access (8), maintaining access (7), exploitation (15), post-exploitation (7), and reporting (15). These stages are defined in more detail in Appendix B and the counts of papers discussing each stage are presented in Figure 1. Many of these stages map to one or more ATT&CK tactics. We found that out of the 14 ATT&CK tactics, only Resource Development and Defense Evasion were not encompassed by these ten main operational stages of red-teaming [MITRE 2024b]. The high coverage of adversary tactics underscores the field's emphasis on realistic attack simulation.

The stages we identified also describe activities beyond the scope of ATT&CK, as they encompass activities that occur outside of the attack process itself. The pre-engagement and threat modeling stages take place before an attack begins, while the post-exploitation and reporting stages occur after its conclusion. These additional stages play a crucial role in effective red-teaming by ensuring that engagements are strategically designed, aligned with realistic threat models, and yield actionable insights. Pre-engagement planning and threat modeling enhance the relevance

and impact of assessments, while thorough post-exploitation analysis and reporting facilitate meaningful security improvements and mitigation strategies.

5.3 Communication with Host Organizations

Effective cyber red-teaming extends beyond revealing technical exploits and requires close coordination with the host organization. Communication is important at many operational stages of red-teaming, but the literature reveals that it is particularly crucial during pre-engagement, threat modeling, post-exploitation, and reporting.

Before an engagement begins, discussions with the host organization help define the scope of the exercise, set rules of engagement, and establish legal considerations, such as liability waivers and non-disclosure agreements, to protect both evaluators and the host organization [ISECOM 2012, Modesti 2024, Vasenius 2022]. During the threat modeling phase, red teams work with the host organization to determine likely attacker profiles, potential host organization asset compromises, and system vulnerabilities [ISECOM 2012; Modesti 2024; Liu, B. 2012].

Effective communication during the post-exploitation and reporting stages is equally critical. During the post-exploitation phase, the red team demonstrates system compromise and notifies the host organization according to the rules of engagement [ISECOM 2012, Modesti 2024, Parveen 2023]. During the reporting phase, effective communication ensures that findings are effectively documented and that remediation strategies are prioritized based on real risk assessments [Altulaihan 2023, Shah 2014, Vasenius 2022]. Without clear and well-structured communication with the red teams, the host organization has no way to learn from the findings of the red team. This may leave critical vulnerabilities unaddressed and significantly diminish the value of the red-teaming engagement. Effective reporting not only highlights the vulnerabilities discovered but also provides the organization with a roadmap for mitigating risks and strengthening defenses. Without it, even the most productive red-teaming exercises that uncover relevant and realistic vulnerabilities will not translate into meaningful security improvements.

5.4 Comprehensiveness

Cyber red-teaming is distinguished by its broad coverage of attack surfaces. We identified 27 papers in the cyber red-teaming literature that mentioned one or more attack surfaces considered during cyber red-teaming. Many (8) of these attack surfaces were mentioned by 3 or more papers. The most common attack surfaces were *network* (mentioned 10 times), *social* (9), *application* (6), *web application* (6), *mobile* (4), *wireless* (4), *internet of things* (4), and *physical* (3). A broad approach ensures that organizations are prepared for a wide variety of attack vectors, rather than narrowly focusing on one domain. This breadth of coverage is essential for realistic adversary emulation, as real attackers take the path of least resistance and are not constrained by artificial scope boundaries. If an organization secures most attack surfaces but neglects one, adversaries will inevitably exploit the weakest link. To address this reality, cyber red teams must evaluate every available attack surface to ensure a holistic security assessment. By identifying gaps in coverage, red teams help organizations defend against the full range of adversarial threats, rather than just isolated vulnerabilities.

5.5 Diverse and Open Source Tooling

The review literature highlights an extensive range of tools available for cyber red-teaming, with 418 different tools mentioned by name. The most popular were *nmap* (12 mentions), *Metasploit* (10), *Nessus* (9), *Wireshark* (8), *Kali Linux* (8), *Burpsuite* (7), *ZAP* (6), *Acutenix* (6), and *Nikto* (6). These tools spanned a variety of different categories including *fuzzers* (10), *static analysis* (8), *dynamic analysis* (8), *network scanners* (4), *vulnerability scanners* (4), *information gathering tools* (3), *password crackers* (3), *wireless tools* (3), and *web application tools* (3). Some of these categories of tools correspond to techniques, such as *fuzzing*, a technique which uses malformed inputs to discover bugs [Pargaonkar 2023]. Other categories of tools correspond to stages of the red-teaming process, such as *vulnerability scanners* or *information gathering tools*. Finally, some categories correspond to attack surfaces, such as *network scanners* or *wireless tools*. This diversity of tooling corresponds to the diversity of techniques, stages, and attack surfaces involved in cyber red-teaming and plays a key role in supporting comprehensive adversary emulation.

Based on a random sample of 50 of these tools, the vast majority (86%) of tools mentioned in the review literature are open source. The open source nature of these tools offers significant advantages, particularly in keeping pace with current and emerging threats. Open source security tools benefit from broad, global contributions, allowing researchers and practitioners to continuously refine techniques, add new functionalities, and address novel attack methods as they arise [Khan 2012]. Open source tools also enable greater transparency and customization, allowing organizations to tailor their specific security evaluations to specific threat landscapes [Russo 2016].

5.6 Addressing Well-Known Vulnerabilities First

The literature reveals that several common strategies in cyber red-teaming are based on targeting well-known vulnerabilities first, before moving on to more sophisticated attack techniques. For example, industry-recognized vulnerability lists, such as the OWASP Top Ten and the SANS Top 25, are often used as starting points for red teams or as checklists of the most likely places to find vulnerabilities [HackerOne 2020]. These vulnerability lists received 11 and 3 mentions in the literature respectively.

Vulnerability scanners take this one step further by using heuristics and pattern matching to automatically identify likely instances of known common vulnerabilities, and they are widely used in cyber red-teaming. Some of the most frequently mentioned cyber red-teaming tools, such as *Nessus* (9), *Wireshark* (8), *Burpsuite* (7), *ZAP* (6), *Acutenix* (6), and *Nikto* (6), include vulnerability scanning capabilities. Exploit frameworks extend the automatic targeting of well-known vulnerabilities even further, by not only automatically scanning for vulnerabilities but also automating the process of exploiting them as well. The second most popular cyber red-teaming tool, *Metasploit* (10) is an exploit framework.

These popular techniques all rely on an accumulated body of knowledge about common vulnerabilities and exploits. Vulnerability lists require an agreed-upon taxonomy of vulnerabilities. Automatic vulnerability scanners and exploit frameworks do not analyze targets from first principles; instead, they operate using a predefined checklist of heuristics and patterns known to be associated with vulnerabilities, gleaned from the experience of manual red-teaming. Without this

accumulated knowledge, both vulnerability scanners and exploit frameworks would be far less effective, as red teams would need to rediscover and analyze each vulnerability from scratch.

This preference for targeting known vulnerabilities aligns with real-world attacker behavior. Threat actors typically seek the path of least resistance, leveraging existing exploit frameworks and pre-built attack tools rather than developing new exploits from scratch. By incorporating "low-hanging fruit" into their assessments, red teams can provide organizations with actionable insights into critical security gaps that require immediate attention.

Once the "low-hanging fruit" is exhausted, cyber red teams proceed to more sophisticated vulnerabilities and exploits, but these tools and techniques provide a structured and well-understood "opening theory" for cyber red teams as well as automation of common opening moves. This structured approach ensures that red teams can efficiently identify and exploit critical vulnerabilities, making their overall assessment process more effective and scalable.

5.7 Standardized Manuals and Methodologies

The literature indicates that cyber red-teaming benefits from well-documented methodologies and certification programs that help standardize best practices. Our review identified several widely used manuals, including the Open Source Security Testing Methodology Manual (OSSTMM) [Herzog 2010], the Penetration Testing Execution Standard (PTES) [ISECOM 2012], the Payment Card Industry Data Security Standard (PCI-DSS) [PCI 2024], and the OWASP Testing Guide (OTG) [OWASP 2020]. These frameworks provide structured guidance on conducting thorough security assessments, promoting consistency across engagements. In addition to manuals, we also identified eight professional certifications mentioned in the literature that validate the skills of evaluators and reinforce industry-wide best practices [Fuchs 2019].

Despite the existence of widely cited manuals and certifications, the field does not have broad consensus on a single optimal framework or credential. Different organizations and practitioners may favor different methodologies based on their specific needs, regulatory requirements, or areas of expertise. However, the repeated citation of certain manuals and certifications across the literature suggests a degree of consistency, if not consensus, in how cyber red-teaming is approached. While no single standard dominates, there is a shared understanding of best practices that guides red-teaming engagements. This level of consistency helps create a common language and baseline for assessments, ensuring that red teams operate within a structured and methodologically sound framework even in the absence of universal agreement.

6 Comparative Analysis of the Generative AI and Cyber Red-Teaming Literature

We have separately reviewed the cyber red-teaming and generative AI red-teaming literature. We now compare the two, with a focus on practices that generative AI red-teaming can adapt from the established best practices in cyber red-teaming.

6.1 Goals, Tooling, and Methodology

While the cyber red-teaming literature consistently identifies eight common attack surfaces, generative AI red-teaming collectively investigates only three. Most generative AI attacks (93) operated on direct, valid inputs to the model or system. Of the remaining cases, 5 poisoned a RAG database so that malicious inputs would automatically be retrieved and placed into future model inputs, and 1 fine-tuned a model on malicious datasets. Furthermore, although cyber red-teaming frequently emphasizes adversary emulation, only 17 out of 99 generative AI red-teaming papers explicitly defined their threat model. This finding underscores an area for improvement in methodological rigor within this domain.

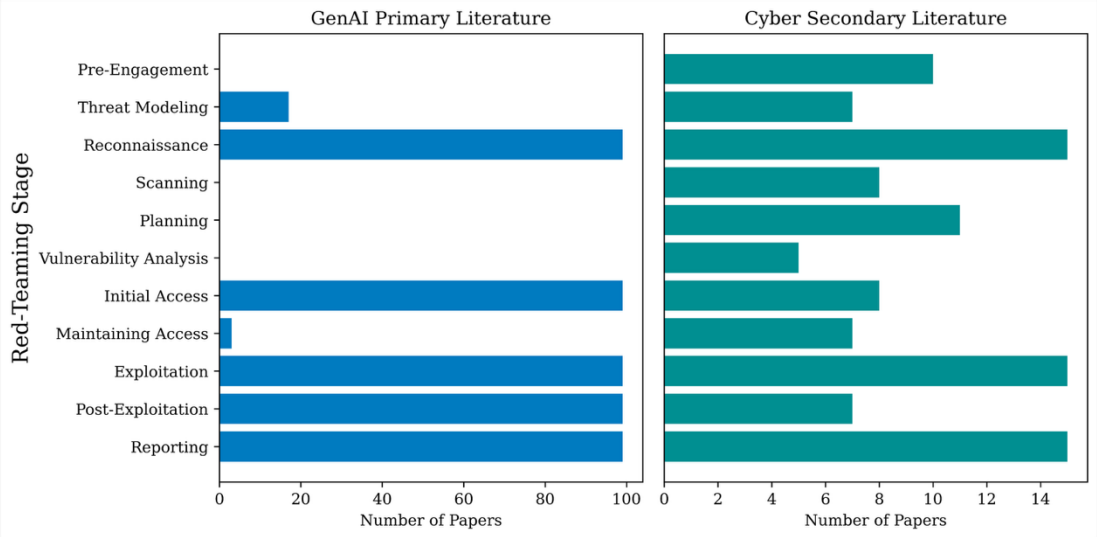
The objectives of these red-teaming approaches also differ significantly. In cyber red-teaming, objectives are collaboratively determined with host organizations, balancing threat prioritization with engagement costs to ensure that identified vulnerabilities are both relevant and actionable. In contrast, generative AI red-teaming primarily seeks to elicit restricted outputs from generative models, typically assessed through attack success rates or severity metrics. While this approach facilitates rapid experimentation and benchmarking, it lacks structured prioritization. Without systematic threat prioritization, generative AI red-teaming risks identifying vulnerabilities that may be misaligned with actual security risks faced by host organizations.

Cyber red-teaming relies on mature, well-developed tools such as Kali Linux and Metasploit, whereas generative AI red-teaming predominantly utilizes research codebases, open source models, and Python-based evaluation tools. While cyber red-teaming references at least nine distinct tool categories, generative AI red-teaming is largely confined to dynamic analysis. Although some generative AI papers cite AI fuzzing algorithms, these function more as jailbreaking utilities rather than conventional cyber fuzzing tools.

Another key distinction lies in comprehensiveness. Generative AI red-teaming typically focuses on a narrower attack surface and a limited range of adversarial tactics, primarily those outlined in MITRE ATLAS. In contrast, cyber red-teaming evaluates entire systems, including interdependencies across different components. While cyber red-teaming enables a holistic security assessment, it necessitates specialized expertise. The more constrained scope of generative AI red-teaming allows for rapid, targeted evaluations but limits its applicability to broader system security.

6.2 Comparison of Operational Red-Teaming Stages

To contextualize generative AI red-teaming within cyber red-teaming frameworks, we map generative AI methodologies onto the established cyber red-teaming stages. The analysis of generative AI red-teaming research reveals a strong focus on later-stage processes when compared to the cyber red-teaming stages outlined in Figure 1, highlighted by the fact that every paper includes an attack on a target system and an analysis of red-teaming outcomes (Post-Exploitation). No generative AI papers discuss formal engagement planning or rules of engagement, and while all implicitly assume a threat model, only 17 explicitly define it. All generative AI papers perform some level of reconnaissance by gathering information from the internet, yet no explicit mentions of automated scanning techniques are found. Planning is assumed but not systematically described, and no explicit discussions of systematic vulnerability assessments are present. Initial access is typically achieved through direct prompting, RAG poisoning, or adversarial fine-tuning, while maintaining access is rarely discussed.



Most generative AI research focuses on the exploitation and post-exploitation stages of red-teaming identified in the cyber literature, while early-stage processes like pre-engagement, scanning, and vulnerability analysis are largely absent. Reconnaissance is implicitly conducted by all studies through the "prior work" section of each publication but is generally conducted more comprehensively in cyber red-teaming activities.

Figure 1: Focus of Cyber and AI Literature

While all generative AI papers engage in some form of reporting, differences in reporting practices further distinguish between the two fields. Cyber red-teaming follows established disclosure practices, wherein findings are confidentially reported to host organizations before public dissemination. Most cyber engagements do not result in publicly accessible documentation, even post-mitigation. In contrast, generative AI red-teaming findings are typically published in academic papers or preprints without prior notification to impacted entities. While this practice accelerates research progress, it lacks the structured vulnerability disclosure frameworks found in cybersecurity. Notably, Cattell et al. have proposed a vulnerability disclosure process for AI systems that could address this gap [Cattell 2024]. However, no references to this framework are found in

existing generative AI red-teaming literature, indicating a critical need for formalized disclosure protocols. A more structured approach, such as Coordinated Vulnerability Disclosure (CVD), could ensure that vulnerabilities affecting multiple stakeholders are responsibly communicated prior to public release [Householder 2024].

In the future, the implications of various reporting strategies may need to be weighed carefully. Researchers should engage in responsible disclosure and allow relevant parties to mitigate vulnerabilities or specific exploits prior to publication. Open source models, however, further complicate the picture, as risks from misuse cannot be mitigated once the models have been released, so publication at any point may be dangerous. Placing more focus on potential mitigations may also be useful, such as testing promising recent mitigations like Circuit Breakers [Zou 2025] or Latent Adversarial Training [Sheshadri 2024]. Two papers evaluated the Circuit Breakers defense, and in both cases, it successfully defended against the attacks.

Finally, while we do encourage the red-teaming of closed-source systems, it is important to remember that closed-source systems pose reproducibility challenges [Rando 2025]. There is evidence that common closed-source LLMs are updated silently over time, reducing the effectiveness of prior attacks [Chao 2024a]. The use of open source tools and standardized evaluation frameworks are crucial for reproducibility. We observe consistent usage of open source models, but evaluation methodologies are often not directly comparable across papers. Advancing responsible disclosure practices alongside improvements in reproducibility across studies will be crucial for strengthening AI security in a rapidly evolving threat landscape.

6.3 Frameworks and Systemization

Cyber red-teaming benefits from well-established frameworks, including vulnerability scanners, exploit databases, and structured training programs. These methodologies streamline engagements by prioritizing widely recognized vulnerabilities before exploring more sophisticated attack vectors. Certifications further standardize skills and ensure methodological consistency across practitioners. Generative AI red-teaming lacks similar systematization but has begun to develop structured resources, such as attack prompt datasets and vulnerability classification frameworks. However, these resources primarily serve diagnostic rather than offensive testing purposes. Establishing standardized methodologies, best practices, and training programs could significantly enhance the rigor and impact of generative AI red-teaming.

Manuals and certifications provide structured guidance to red teams, delineating the red-teaming process into discrete stages and outlining key objectives and recommended methodologies for each phase. They serve as comprehensive references, integrating best practices accumulated over time. While there is no single authoritative set of manuals or guides universally adopted across the field, the widespread usage of resources, such as the OWASP Top Ten, indicates a level of consensus and standardization that is not present in generative AI red-teaming [OWASP 2024]. In parallel, professional certifications define areas of specialized expertise within red-teaming, specifying requisite skills and employing structured curricula to systematically impart and assess expertise. Collectively, these elements contribute to a highly structured and standardized process. Aspiring cyber red-teamers have access to formalized training courses that equip them with the

necessary competencies, including proficiency in widely recognized tools and techniques that target common vulnerabilities effectively.

The generative AI red-teaming community has yet to establish comparable mechanisms, but initial steps towards systematization are evident. The development of successful attack prompt datasets, such as JailbreakBench [Chao 2024a] and RealToxicityPrompts [Gehman 2020], enables researchers to evaluate system vulnerabilities against a repository of previously documented exploits. However, these datasets function primarily as diagnostic tools rather than offensive capabilities akin to vulnerability scanners or exploit frameworks. Additionally, resources such as the OWASP Top 10 for LLM Applications provide guidance on common vulnerabilities specific to LLM-based systems [OWASP 2025], resembling the broader vulnerability lists used in cybersecurity. As the field progresses, further refinement and formalization of methodologies, training programs, and standardized assessment frameworks will be crucial to establishing generative AI red-teaming as a rigorous and systematic discipline.

Ultimately, cyber red-teaming offers well-developed methodologies that could inform and enhance generative AI red-teaming, particularly in structured engagement planning, vulnerability prioritization, responsible disclosure mechanisms, and standardized methodologies. As the field of generative AI security matures, integrating these practices will be essential to ensuring the effectiveness, ethical responsibility, and long-term sustainability of generative AI red-teaming efforts.

7 Recommendations

This systematic review reveals key differences between cyber red-teaming and generative AI red-teaming, highlighting opportunities for the latter to mature by leveraging best practices from the former. To support this growth, we propose the following actionable recommendations for researchers and practitioners engaged in generative AI security research and development:

1. **Incorporate realistic threat models**. Adopt adversary-focused frameworks and consider real-world implications, including financial stakes, to improve the relevance of generative AI red-teaming. Moving beyond simplistic and often unreliable metrics like Attack Success Rate will lead to more impactful and realistic threat assessments.
2. **Expand attack surface considerations**. To enhance generative AI red-teaming, expand the focus beyond direct model inputs and RAG databases to include diverse attack vectors such as data pipelines, deployment environments, and user interfaces. This broader scope will enable more comprehensive security evaluations.
3. **Integrate cyber operational stages**. Integrate critical stages from cyber red-teaming, such as pre-engagement planning, detailed threat modeling, and robust reporting. Including responsible disclosure practices will also better align generative AI red-teaming with established security protocols to enhance its effectiveness.
4. **Ensure actionable mitigations**. Generative AI red-teaming efforts should prioritize generating insights that translate into concrete security improvements rather than repeatedly demonstrating jailbreak feasibility. Borrowing from cybersecurity's iterative testing and feedback loops will help establish more effective mitigation pathways.
5. **Bridge the gap between evaluators and model developers**. In cybersecurity, goals are often refined through direct engagement with system hosts. AI evaluators should adopt similar practices, improving dialogue with model developers to tailor security assessments.
6. **Develop open source tooling**. Invest in creating accessible, well-supported open source tools tailored for generative AI red-teaming. Solutions analogous to popular cyber tools, such as Metasploit and Wireshark, will streamline security evaluations, lower barriers for new practitioners, and can be used alongside cyber-specific tooling to uncover synergistic vulnerabilities.
7. **Diversify red-teaming techniques**. Expand beyond the current focus on dynamic analysis, either by directly incorporating methods like static analysis and automated vulnerability scanning or by developing dynamic analysis methods to achieve similar outcomes. This diversification will deepen the analytical capabilities and effectiveness of generative AI red-teaming.
8. **Enhance automation for scalability**. Invest in automated tools for generative AI red-teaming to improve scalability and reproducibility of efforts, reducing reliance on manual testing.
9. **Standardize vulnerability identification**. Establish structured exploit frameworks and standardized vulnerability lists for generative AI systems, akin to Metasploit and building on initial work in this space (e.g., OWASP Top Ten for LLMs). Large benchmarking datasets

alone often fail to address this need, as they do not provide systematic methodologies or tools for identifying and exploiting vulnerabilities. These resources will improve efficiency and ensure consistent identification of common vulnerabilities.

10. **Develop authoritative manuals and guidelines**. Formulate standardized manuals and methodologies for generative AI red-teaming drawing from cybersecurity's OSSTMM and PTES. These guidelines will promote consistency, establish best practices, and elevate the field's maturity.

Progress across these ten recommendations would constitute a significant advancement in the effectiveness of generative AI red-teaming, helping to ensure the safe deployment of AI systems.

8 References

URLs are valid as of the publication date of this report.

[Abbass 2011]
Abbass, H. et al. Computational Red Teaming: Past, Present and Future. *IEEE Computational Intelligence Magazine*. Volume 6. Issue 1. January 17, 2001. Pages 30–42. https://doi.org/10.1109/MCI.2010.939578

[Aboelfotoh 2019]
Aboelfotoh, S. F. & Hikal, N. A. A Review of Cyber-Security Measuring and Assessment Methods for Modern Enterprises. *International Journal on Informatics Visualization*. Volume 3. Number 2. 2019. https://joiv.org/index.php/joiv/article/view/239

[Adam 2023]
Adam, H. M.; Widyawan; & Putra, G. D. A Review of Penetration Testing Frameworks, Tools, and Application Areas. Pages 319-324. In *2023 IEEE 7th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE).* 2023. https://doi.org/10.1109/ICITISEE58992.2023.10404397

[Ahmad 2024]
Ahmad, L.; Agarwal, S.; Lampe, M.; & Mishkin, P. *OpenAI's Approach to External Red Teaming for AI Models and Systems*. OpenAI. 2024. http://cdn.openai.com/papers/openais-approach-to-external-red-teaming.pdf

[Al-Ahmad 2019]
Al-Ahmad; A. S., Kahtan; H., Hujainah, F.; & Jalab, H. A. Systematic Literature Review on Penetration Testing for Mobile Cloud Computing Applications. *IEEE Access.* Volume 7. November 29, 2019. Pages 173524–173540. https://doi.org/10.1109/ACCESS.2019.2956770

[Aldauiji 2022]
Aldauiji, F.; Batarfi, O.; & Bayousef, M. Utilizing Cyber Threat Hunting Techniques to Find Ransomware Attacks: A Survey of the State of the Art. *IEEE Access*. Volume 10. June 8, 2022. Pages 61695–61706. https://doi.org/10.1109/ACCESS.2022.3181278

[Al-Sada 2025]
Al-Sada, B.; Sadighian, A.; & Oligeri, G. MITRE ATT&CK: State of the Art and Way Forward. *ACM Computing Surveys*. Volume 57. Issue 1. October 7, 2024. Pages 1–37. https://doi.org/10.1145/3687300

[Altayaran 2021]
Altayaran, S. A. & Elmedany, W. Integrating Web Application Security Penetration Testing into the Software Development Life Cycle: A Systematic Literature Review. Pages 671-676. In *2021 International Conference on Data Analytics for Business and Industry (ICDABI)*. October 2021. https://doi.org/10.1109/ICDABI53623.2021.9655950

[Altulaihan 2023]
Altulaihan, E. A.; Alismail, A.; & Frikha, M. A Survey on Web Application Penetration Testing. *Electronics*. Volume 12. Issue 5. March 4, 2023. Page 1229. https://doi.org/10.3390/electronics12051229

[Amayuelas 2024]
Amayuelas, A. et al. MultiAgent Collaboration Attack: Investigating Adversarial Attacks in Large Language Model Collaborations via Debate. Pages 6929–6948. In *Findings of the Association for Computational Linguistics: EMNLP*. November 2024. DOI: 10.18653/v1/2024.findings-emnlp.407. https://aclanthology.org/2024.findings-emnlp.407/

[Anthropic 2025]
Anthropic. *Responsible Disclosure Policy*. February 14, 2025. https://www.anthropic.com/responsible-disclosure-policy

[Bai 2022]
Bai, Y. et al. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv* [preprint]. April 12, 2022. https://doi.org/10.48550/arXiv.2204.05862

[Beaman 2022]
Beaman, C. et al. Fuzzing Vulnerability Discovery Techniques: Survey, Challenges and Future Directions. *Computers & Security*. Volume 120. July 2022. Page 102813. https://doi.org/10.1016/j.cose.2022.102813

[Bick 2024]
Bick, A.; Blandin, A.; & Deming, D. J. The Rapid Adoption of Generative AI. *National Bureau of Economic Research*. September 2024. https://www.nber.org/papers/w32966

[Boiko 2023]
Boiko, D. A.et al. Emergent Autonomous Scientific Research Capabilities of Large Language Models. *arXiv* [preprint]. April 2023. http://arxiv.org/abs/2304.05332

[Bowen 2024]
Bowen, D. et al. Data Poisoning in LLMs: Jailbreak-Tuning and Scaling Laws. *arXiv* [preprint]. December 2024. https://doi.org/10.48550/arXiv.2408.02946

[Brangetto 2015]
Brangetto, P. et al. Cyber Red Teaming. Organisational, Technical and Legal Implications in a Military Context. *NATO Cooperative Cyber Defence Centre of Excellence.* 2015. https://ccdcoe.org/library/publications/cyber-red-teaming-organisational-technical-and-legal-implications-in-a-military-context/

[Briggs 2018]
Briggs, J. et al. Survey of Layered Defense, Defense in Depth, and Testing of Network Security. In *Selected Readings in Cybersecurity.* Cambridge Scholars Publishing. Pages 105–119. 2018. ISBN: 1-5275-1641-5.

[Burtell 2023]
Burtell, M. & Woodside, T. Artificial Influence: An Analysis Of AI-Driven Persuasion. *arXiv* [preprint]. March 2023. http://arxiv.org/abs/2303.08721

[Carlini 2024a]
N. Carlini et al. Poisoning Web-Scale Training Datasets is Practical. Pages 407-425. In *2024 IEEE Symposium on Security and Privacy (SP)*. May 2024. https://doi.ieeecomputersociety.org/10.1109/SP54263.2024.00179

[Carlini 2024b]
Carlini, N. et al. Stealing Part of a Production Language Model. *arXiv* [preprint]. July 2024. https://doi.org/10.48550/arXiv.2403.06634

[Cattell 2024]
Cattell, S. et al. Coordinated Flaw Disclosure for AI: Beyond Security Vulnerabilities. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Volume 7. Issue 1. Pages 267-280. https://ojs.aaai.org/index.php/AIES/article/view/31635

[Chang 2024]
Chang, Z. et al. Play Guessing Game with LLM: Indirect Jailbreak Attack with Implicit Clues. *arXiv* [preprint]. https://doi.org/10.48550/arXiv.2402.09091

[Chao 2024a]
Chao, P. et al. JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models. *arXiv* [preprint]. October 2024. https://doi.org/10.48550/arXiv.2404.01318

[Chao 2024b]
Chao, P. et al. Jailbreaking Black Box Large Language Models in Twenty Queries. Pages 23-42. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. April 2024. https://doi.ieeecomputersociety.org/10.1109/SaTML64287.2025.00010

[Chen 2022]
Chen, L. et al. A Survey on Threat Hunting: Approaches and Applications. Pages 340-344. In *2022 7th IEEE International Conference on Data Science in Cyberspace (DSC).* July 2022. DOI: 10.1109/DSC55868.2022.00053. https://ieeexplore.ieee.org/document/9900201

[Chen 2024]
Chen, Z. et al. AgentPoison: Red-teaming LLM Agents via Poisoning Memory or Knowledge Bases. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*. September 2024. https://neurips.cc/virtual/2024/poster/94715

[Chi 2024]
Chi, J. et al. Llama Guard 3 Vision: Safeguarding Human-AI Image Understanding Conversations. *Meta*. November 2024. https://ai.meta.com/research/publications/llama-guard-3-vision-safeguarding-human-ai-image-understanding-conversations/

[CISA 2024]
CISA. *Enhancing Cyber Resilience: Insights from CISA Red Team Assessment of a US Critical Infrastructure Sector Organization.* CISA. November 2024. https://www.cisa.gov/news-events/cybersecurity-advisories/aa24-326a

[Cong 2024]
Cong, N. T. et al. An Overview of Static and Dynamic Analysis in Application Security Testing. *Journal of Military Science and Technology.* Volume 99. Number 99. November 2024. Pages 1–11. https://doi.org/10.54939/1859-1043.j.mst.99.2024.1-11

[Cranford 2023]
Cranford, J. Red Team vs. Blue Team: What's the Difference? *CrowdStrike*. April 16, 2023. https://www.crowdstrike.com/en-us/cybersecurity-101/advisory-services/red-team-vs-blue-team/

[CSRC 2015]
CSRC. Glossary: Red Team. *CSRC*. 2015. https://csrc.nist.gov/glossary/term/red_team

[Dang 2024]
Dang, P. et al. DiffZOO: A Purely Query-Based Black-Box Attack for Red-teaming Text-to-Image Generative Model via Zeroth Order Optimization. *arXiv* [preprint]. August 2024. https://doi.org/10.48550/arXiv.2408.11071

[de Zarzà 2023]
de Zarzà, I. et al. Optimized Financial Planning: Integrating Individual and Cooperative Budgeting Models with LLM Recommendations. *AI*. Volume5. Issue 1. 2024. Pages 91–114. https://www.mdpi.com/2673-2688/5/1/6

[Deng, D. 2024a]
Deng, D. et al. AdversaFlow: Visual Red Teaming for Large Language Models with Multi-Level Adversarial Flow. *IEEE Transactions on Visualization and Computer Graphics*. Volume 31. Issue 1. September 2024. Pages 492–502. https://doi.org/10.1109/TVCG.2024.3456150

[Deng, G. 2024b]
Deng, G. et al. Pandora: Jailbreak GPTs by Retrieval Augmented Generation Poisoning. *arXiv* (preprint). February 13, 2024. https://doi.org/10.48550/arXiv.2402.08416

[Deshpande 2023]
Deshpande, A. et al. Toxicity in ChatGPT: Analyzing Persona-assigned Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Pages 1236–1270. December 2023. DOI: 10.18653/v1/2023.findings-emnlp.88. https://aclanthology.org/2023.findings-emnlp.88/

[Dominique 2024]
Dominique, B. et al. Prompt Templates: A Methodology for Improving Manual Red Teaming Performance. *CHI 2024*. May 2024. https://research.ibm.com/publications/prompt-templates-a-methodology-for-improving-manual-red-teaming-performance

[Dong 2024]
Dong, Y. et al. Harnessing Task Overload for Scalable Jailbreak Attacks on Large Language Models. *arXiv* [preprint]. October 2024. https://doi.org/10.48550/arXiv.2410.04190

[Doumbouya 2024]
Doumbouya, M. K. B. et al. h4rm3l: A Dynamic Benchmark of Composable Jailbreak Attacks for LLM Safety Assessment. *arXiv* [preprint]. September 2024. https://doi.org/10.48550/arXiv.2408.04811

[Eceiza 2021]
Eceiza, M. et al. Fuzzing the Internet of Things: A Review on the Techniques and Challenges for Efficient Vulnerability Discovery in Embedded Systems. *IEEE Internet of Things Journal*. Volume 8. Issue 13. July 2021. Pages 10390–10411. https://doi.org/10.1109/JIOT.2021.3056179

[Feffer 2024]
Feffer, M. et al. Red-Teaming for Generative AI: Silver Bullet or Security Theater? Pages 421-437. In *Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES 2024)*. October 2024. DOI: https://doi.org/10.1609/aies.v7i1.31647. https://ojs.aaai.org/index.php/AIES/article/view/31647

[Ferrara 2024]
Ferrara, E. GenAI Against Humanity: Nefarious Applications of Generative Artificial Intelligence and Large Language Models. *Journal of Computational Social Science*. Volume 7. Issue 1. February 2024. Pages 549–569. https://doi.org/10.1007/s42001-024-00250-1

[Fuchs 2019]
Fuchs, M. & Lemon, J. *SANS 2019 Threat Hunting Survey: The Differing Needs of New and Experienced Hunters*. SANS Institute. October 2019. https://www.sans.org/media/analyst-program/2019-threat-hunting-survey-differing-experienced-hunters-39220.pdf

[Gallagher 2024]
Gallagher, S. K. et al. Assessing LLMs for High Stakes Applications. Pages 103–105. In *Proceedings of the 46th International Conference on Software Engineering: Software Engineering in Practice*. April 2024. https://doi.org/10.1145/3639477.3639720

[Ganguli 2022]
Ganguli, D. et al. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. *arXiv* [preprint]. November 2022. https://doi.org/10.48550/arXiv.2209.07858

[Gbormittah 2024]
Gbormittah, E. A Systematic Literature Review on Cyberwarfare and State-Sponsored Hacking: Penetration Testing Insights. *TechRxiv* [preprint]. August 2024. https://www.techrxiv.org/doi/full/10.36227/techrxiv.172373675.55159205

[Gehman 2020]
Gehman, S. et al. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. Pages 3356–3369. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. November 2020. DOI: 10.18653/v1/2020.findings-emnlp.301. https://aclanthology.org/2020.findings-emnlp.301/

[Ghaffarian 2018]
Ghaffarian, S. M. & Shahriari, H. R. Software Vulnerability Analysis and Discovery Using Machine-Learning and Data-Mining Techniques: A Survey. *ACM Computing Surveys*. Volume 50. Issue 4. August 2017. Pages 1-36. https://doi.org/10.1145/3092566

[Gibbs 2024]
Gibbs, T. et al. Emerging Vulnerabilities in Frontier Models: Multi-Turn Jailbreak Attacks. *arXiv* [preprint]. August 29, 2024. https://doi.org/10.48550/arXiv.2409.00137

[Greshake 2023]
Greshake, K. et al. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. Pages 79–90. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*. November 2023. https://doi.org/10.1145/3605764.3623985

[Großmann 2015]
Großmann, J. & Seehusen, F. Combining Security Risk Assessment and Security Testing Based on Standards. In *Risk Assessment and Risk-Driven Testing.* F. Seehusen et al [editors]. Springer International Publishing. November 13, 2015. Pages 18–33. ISBN 978-3-319-26416-5. https://doi.org/10.1007/978-3-319-26416-5_2

[Gu 2024]
Gu, X. et al. Agent Smith: A Single Image Can Jailbreak One Million Multimodal LLM Agents Exponentially Fast. Pages 16647–16672. In *ICML'24: Proceedings of the 41st International Conference on Machine Learning*. July 2024. https://dl.acm.org/doi/10.5555/3692070.3692731

[HackerOne 2020]
HackerOne. A Starters Guide to Pentesting with OWASP. *Hacker101.* July 23, 2020. https://www.hacker101.com/sessions/pentest_owasp.html

[Han 2024]
Han, V. T. Y.; Bhardwaj, R.; & Poria, S. Ruby Teaming: Improving Quality Diversity Search with Memory for Automated Red Teaming. *arXiv* [preprint]. June 2024. https://doi.org/10.48550/arXiv.2406.11654

[Hardy 2024]
Hardy, A. F. et al. ASTPrompter: Weakly Supervised Automated Language Model Red-Teaming to Identify Low-Perplexity Toxic Prompts. *arXiv* [preprint]. July 2024. https://doi.org/10.48550/arXiv.2407.09447

[Herzong 2010]
Herzog, P. OSSTMM 3. *Institute for Security and Open Methodologies.* 2010. https://www.isecom.org/OSSTMM.3.pdf

[Hong 2024]
Hong, Z.-W. et al. Curiosity-Driven Red-teaming for Large Language Models. *arXiv* [preprint]. February 2024. https://doi.org/10.48550/arXiv.2402.19464

[Householder 2024]
Householder, A. et al. *Lessons Learned in Coordinated Disclosure for Artificial Intelligence and Machine Learning Systems*. Software Engineering Institute, Carnegie Mellon University. August 2024. https://insights.sei.cmu.edu/library/lessons-learned-in-coordinated-disclosure-for-ai-and-ml-systems/

[Householder 2020]
Householder, A.; Wassermann, G.; Manion, A.; & King, C. *CERT® Guide to Coordinated Vulnerability Disclosure*. CMU/SEI-2017-sr-022. Software Engineering Institute, Carnegie Mellon University. 2020. https://doi.org/10.1184/R1/12367340.V1

[Hu 2024]
Hu, K. et al. Efficient LLM Jailbreak via Adaptive Dense-to-Sparse Constrained Optimization. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*. December 2024. https://papers.nips.cc/paper_files/paper/2024/hash/29571f8fda54fe93631c41aad4215abc-Abstract-Conference.html

[Huang, X. 2024]
Huang, X. et al. Medical MLLM is Vulnerable: Cross-Modality Jailbreak and Mismatched Attacks on Medical Multimodal Large Language Models. Pages 3797-3805. In *Proceedings of the AAAI Conference on Artificial Intelligence*. April 2025. DOI: 10.1609/aaai.v39i4.32396. https://ojs.aaai.org/index.php/AAAI/article/view/32396

[Huang, Y. 2024]
Huang, Y. et al. Perception-Guided Jailbreak against Text-to-Image Models. Pages 26238-26247. In *Proceedings of the AAAI Conference on Artificial Intelligence*. April 2025. DOI: 10.1609/aaai.v39i25.34821. https://ojs.aaai.org/index.php/AAAI/article/view/34821

[ISECOM 2012]
ISECOM. PTES Technical Guidelines—The Penetration Testing Execution Standard. Pentest Website. 2012. http://www.pentest-standard.org/index.php/PTES_Technical_Guidelines

[Jeršič 2024]
Jeršič, N. et al. How to Approach Security Testing of Web 3.0 Solutions: A Review of Existing Knowledge. In *Proceedings of the Eleventh Workshop on Software Quality Analysis, Monitoring, Improvement, and Applications*. September 2024. https://ceur-ws.org/Vol-3845/paper20.pdf

[Jiang 2024]
Jiang, F. et al. ArtPrompt: ASCII Art-Based Jailbreak Attacks Against Aligned LLMs. Pages 15157–15173. In *Proceedings of the 62nd Annual Meeting of the Association for the Computational Linguistics (Volume 1: Long Papers)*. August 2024. https://aclanthology.org/2024.acl-long.809/

[Johnson 2024]
Johnson, Z. D. *Generation, Detection, and Evaluation of Role-Play Based Jailbreak Attacks in Large Language Models* [thesis]. Massachusetts Institute of Technology (MIT). 2024. https://dspace.mit.edu/handle/1721.1/156989

[Kalchenko 2018]
Kalchenko, V. V. An Overview of Penetration Testing Methods for Assessing the Security of Computer Systems. *Control, Navigation and Communication Systems.* Volume 4. 2018. Pages 109–114.

[Khan 2012]
Khan, M. A. & UrRehman, F. Free and Open Source Software: Evolution, Benefits and Characteristics. *International Journal of Emerging Trends & Technology in Computer Science.* Volume 1. Issue 3. September 2012. Pages 1–7.

[Kumar, A. 2024]
Kumar, A. et al. SAGE-RT: Synthetic Alignment data Generation for Safety Evaluation and Red Teaming. *arXiv* [preprint]. August 14, 2024. https://doi.org/10.48550/arXiv.2408.11851

[Kumar, V. 2024]
Kumar, V.; Liao, Z.; Jones, J.; & Sun, H. AmpleGCG-Plus: A Strong Generative Model of Adversarial Suffixes to Jailbreak LLMs with Higher Success Rates in Fewer Attempts. *arXiv* [preprint]. October 29, 2024. https://doi.org/10.48550/arXiv.2410.22143

[Lee 2024]
Lee, S. et al. Learning Diverse Attacks on Large Language Models for Robust Red-teaming and Safety Tuning. *arXiv* [preprint]. May 28, 2024. https://doi.org/10.48550/arXiv.2405.18540

[Leszczyna 2021]
Leszczyna, R. Review of Cybersecurity Assessment Methods: Applicability Perspective. *Computers & Security.* Volume 108. September 2021. Page 102376. https://doi.org/10.1016/j.cose.2021.102376

[Li, B. 2024]
Li, B. et al. StructuralSleight: Automated Jailbreak Attacks on Large Language Models Utilizing Uncommon Text-Organization Structures. *arXiv* [preprint]. June 13, 2024. https://doi.org/10.48550/arXiv.2406.08754

[Li, G. 2024]
Li, G. et al. ART: Automatic Red-teaming for Text-to-Image Models to Protect Benign Users. *arXiv* [preprint]. May 24, 2024. https://doi.org/10.48550/arXiv.2405.19360

[Li, J. 2024a]
Li, J. Vulnerabilities Mapping Based on OWASP-SANS: A Survey for Static Application Security Testing (SAST). *Annals of Emerging Technologies in Computing.* Volume 4. Issue 3. October 11, 2024. Pages 1–8. https://doi.org/10.33166/AETiC.2020.03.001

[Li, J. 2024b]
Li, J. et al. FMM-Attack: A Flow-Based Multi-Modal Adversarial Attack on Video-Based LLMs. *arXiv* [preprint]. March 21, 2024. https://doi.org/10.48550/arXiv.2403.13507

[Li, J. 2024c]
Li, J. et al. A Cross-Language Investigation into Jailbreak Attacks in Large Language Models. *arXiv* [preprint]. January 30, 2024. https://doi.org/10.48550/arXiv.2401.16765

[Li, Q. 2024]
Li, Q. et al. Deciphering the Chaos: Enhancing Jailbreak Attacks via Adversarial Prompt Translation. *arXiv* [preprint]. October 15, 2024. https://doi.org/10.48550/arXiv.2410.11317

[Li, R. 2024]
Li, R. et al. Be a Multitude to Itself: A Prompt Evolution Framework for Red Teaming. *Findings of the Association for Computational Linguistics: EMNLP 2024.* November 2024. Pages 3287–3301. https://doi.org/10.18653/v1/2024.findings-emnlp.188

[Li, X. 2024a]
Li, X. et al. Faster-GCG: Efficient Discrete Optimization Jailbreak Attacks against Aligned Large Language Models. *arXiv* [preprint]. October 20, 2024. https://doi.org/10.48550/arXiv.2410.15362

[Li, X. 2024b]
Li, X. et al. Semantic Mirror Jailbreak: Genetic Algorithm Based Jailbreak Prompts Against Open-Source LLMs. *arXiv* [preprint]. February 27, 2024. https://doi.org/10.48550/arXiv.2402.14872

[Li, X. 2024c]
Li, X. et al. DeepInception: Hypnotize Large Language Model to Be Jailbreaker. *arXiv* [preprint]. November 28, 2024. https://doi.org/10.48550/arXiv.2311.03191

[Li, Y. 2024]
Li, Y. et al. Lockpicking LLMs: A Logit-Based Jailbreak Using Token-Level Manipulation. *arXiv* [preprint]. June 19, 2024. https://doi.org/10.48550/arXiv.2405.13068

[Lin, L. 2025]
Lin, L. et al. Against The Achilles' Heel: A Survey on Red Teaming for Generative Models. *Journal of Artificial Intelligence Research.* Volume 82. February 5, 2025. Pages 687–775. https://doi.org/10.1613/jair.1.17654

[Lin, S. 2024]
Lin, S. et al. LLMs Can Be Dangerous Reasoners: Analyzing-Based Jailbreak Attack on Large Language Models. *arXiv* [preprint]. July 23, 2024. https://doi.org/10.48550/arXiv.2407.16205

[Lin, Z. 2024]
Lin, Z. et al. PathSeeker: Exploring LLM Security Vulnerabilities with a Reinforcement Learning-Based Jailbreak Approach. *arXiv* [preprint]. October 3, 2024. https://doi.org/10.48550/arXiv.2409.14177

[Liu, B. 2012]
Liu, B. et al. Software Vulnerability Discovery Techniques: A Survey. *2012 Fourth International Conference on Multimedia Information Networking and Security.* November 2012. Pages 152–156. https://doi.org/10.1109/MINES.2012.202

[Liu, H. 2024]
Liu, H. et al. Boosting Jailbreak Transferability for Large Language Models. *arXiv* [preprint]. November 3, 2024. https://doi.org/10.48550/arXiv.2410.15645

[Liu, X. 2024a]
Liu, X. et al. AutoDAN-Turbo: A Lifelong Agent for Strategy Self-Exploration to Jailbreak LLMs. *arXiv* [preprint]. October 3, 2024. https://doi.org/10.48550/arXiv.2410.05295

[Liu, X 2024b]
Liu, X. et al. AutoDAN: Generating Stealthy Jailbreak Prompts on Aligned Large Language Models. *arXiv* [preprint]. March 20, 2024. https://doi.org/10.48550/arXiv.2310.04451

[Liu, Y. 2024a]
Liu, Y. et al. Arondight: Red Teaming Large Vision Language Models with Auto-generated Multi-modal Jailbreak Prompts. *arXiv* [preprint]. July 21, 2024. https://doi.org/10.48550/arXiv.2407.15050

[Liu, Y. 2024b]
Liu, Y. et al. FlipAttack: Jailbreak LLMs via Flipping. *arXiv* [preprint]. October 2, 2024. https://doi.org/10.48550/arXiv.2410.02832

[Longpre 2024]
Longpre, S. et al. A Safe Harbor for AI Evaluation and Red Teaming. *arXiv* [preprint]. March 7, 2024. https://doi.org/10.48550/arXiv.2403.04893

[Lu 2024]
Lu, L. et al. AutoJailbreak: Exploring Jailbreak Attacks and Defenses Through a Dependency Lens. *arXiv* [preprint]. June 6, 2024. https://doi.org/10.48550/arXiv.2406.03805

[Luo 2024]
Luo, Y. et al. Jailbreak Instruction-Tuned LLMs via End-of-Sentence MLP Re-weighting. *arXiv* [preprint]. October 14, 2024. https://doi.org/10.48550/arXiv.2410.10150

[Lv 2024]
Lv, L. et al. AdaPPA: Adaptive Position Pre-Fill Jailbreak Attack Approach Targeting LLMs. *arXiv* [preprint]. September 11, 2024. https://doi.org/10.48550/arXiv.2409.07503

[Mahboubi 2024]
Mahboubi, A. et al. Evolving Techniques in Cyber Threat Hunting: A Systematic Review. *Journal of Network and Computer Applications*. Volume 232. December 2024. Page 104004. https://doi.org/10.1016/j.jnca.2024.104004

[Marchetto 2023]
Marchetto, A. A Rapid Review on Fuzz Security Testing for Software Protocol Implementations. Pages 3–20. In *Testing Software and Systems.* September 2023. DOI: 10.1007/978-3-031-43240-8_1. https://doi.org/10.1007/978-3-031-43240-8_1

[Mazeika 2024]
Mazeika, M. et al. HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. Pages 35181–35224. In *Proceedings of the 41st International Conference on Machine Learning*. July 2024. Article Number 1431. https://proceedings.mlr.press/v235/mazeika24a.html

[Meeham 2007]
Meeham, M. Red Teaming for Law Enforcement. *Police Chief*. Volume 74. Issue 2. February 2007. Pages 22–28. https://www.policechiefmagazine.org/red-teaming-for-law-enforcement/

[Mehrotra 2025]
Mehrotra, A. et al. Tree of Attacks: Jailbreaking Black-Box LLMs Automatically. Pages 61065–61105. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*. Article Number 1952. June 2025. https://dl.acm.org/doi/10.5555/3737916.3739868

[Men 2024]
Men, T. et al. A Troublemaker with Contagious Jailbreak Makes Chaos in Honest Towns. *arXiv* [preprint]. October 21, 2024. https://doi.org/10.48550/arXiv.2410.16155

[Meta 2024]
Meta. Meta Bug Bounty. *Meta Website*. October 15, 2024 [accessed]. https://bug-bounty.meta.com/scope/

[Miao 2024]
Miao, H. et al. Autonomous LLM-Enhanced Adversarial Attack for Text-to-Motion. *arXiv* [pre-print]. August 1, 2024. https://doi.org/10.48550/arXiv.2408.00352

[Microsoft 2024]
Microsoft. Planning Red Teaming for Large Language Models (LLMs) and Their Applications—Azure OpenAI Service. *Microsoft Website*. November 30, 2024. https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/red-teaming

[Mistral 2025]
Mistral AI. Terms of Service. *Mistral AI Website*. February 6, 2025. https://mistral.ai/terms#terms-of-service

[MITRE 2024a]
MITRE. MITRE ATLAS™. *MITRE Website*. October 15, 2024 [accessed]. https://atlas.mitre.org/

[MITRE 2024b]
MITRE. MITRE ATT&CK®. *MITRE Website*. October 15, 2024 [accessed]. https://attack.mitre.org/

[Modesti 2024]
Modesti, P. et al. Bridging the Gap: A Survey and Classification of Research-Informed Ethical Hacking Tools. *Journal of Cybersecurity and Privacy.* Volume 4. Issue 3. Article 3. July 2024. Pages 410–448. https://doi.org/10.3390/jcp4030021

[Mu 2024]
Mu, H. et al. Stealthy Jailbreak Attacks on Large Language Models via Benign Data Mirroring. *arXiv* [preprint]. October 28, 2024. https://doi.org/10.48550/arXiv.2410.21083

[Narayan 2024]
Narayan, U. et al. Code-of-Thought Prompting: Probing AI Safety with Code. *OpenReview* [pre-print]. September 27, 2024. https://openreview.net/forum?id=lUyYX9VFgA

[Nasr 2023]
Nasr, M. et al. Scalable Extraction of Training Data from (Production) Language Models. *arXiv* [preprint]. November 28, 2023. https://doi.org/10.48550/arXiv.2311.17035

[Nguyen 2022]
Nguyen, C. et al. CTI4AI: Threat Intelligence Generation and Sharing after Red Teaming AI Models. *arXiv* [preprint]. August 16, 2022. http://arxiv.org/abs/2208.07476

[Nour 2023]
Nour, B.; Pourzandi, M.; & Debbabi, M. A Survey on Threat Hunting in Enterprise Networks. *IEEE Communications Surveys & Tutorials.* Volume 25. Issue 4. August 2023. Pages 2299–2324. https://doi.org/10.1109/COMST.2023.3299519

[Nutalapati 2020]
Nutalapati, V. A Comprehensive Review of Mobile App Security Testing Tools and Techniques. *International Research Journal of Engineering & Applied Sciences.* Volume 8. Issue 1. January–March 2020. Pages 10–15. https://www.irjeas.org/wp-content/uploads/admin/volume8/V8I1/IRJEAS04V8I101200320000006.pdf

[OpenAI 2025]
OpenAI. Usage Policies. *OpenAI Website*. January 29, 2025. https://openai.com/policies/usage-policies/

[OpenAI 2023]
OpenAI. OpenAI Bug Bounty. *Bugcrowd Website*. April 11, 2023. https://bugcrowd.com/engagements/openai

[OWASP 2025]
The Open Worldwide Application Security Project (OWASP). OWASP Top 10 for Large Language Model Applications. *OWASP Website*. June 24, 2025 [accessed]. https://owasp.org/www-project-top-10-for-large-language-model-applications/

[OWASP 2024]
The Open Worldwide Application Security Project (OWASP). OWASP Top Ten. *OWASP Website*. https://owasp.org/www-project-top-ten/

[OWASP 2020]
The Open Worldwide Application Security Project (OWASP). OWASP Web Security Testing Guide. *OWASP Website*. December 3, 2020. https://owasp.org/www-project-web-security-testing-guide/

[Pala 2024]
Pala, T. D. et al. Ferret: Faster and Effective Automated Red Teaming with Reward-Based Scoring Technique. *arXiv* [preprint]. August 20, 2024. https://doi.org/10.48550/arXiv.2408.10701

[Pargaonkar 2023]
Pargaonkar, S. Advancements in Security Testing: A Comprehensive Review of Methodologies and Emerging Trends in Software Quality Engineering. *International Journal of Science and Research (IJSR).* Volume 12. Issue 9. September 2023. Pages 61–66. https://doi.org/10.21275/SR23829090815

[Parveen 2023]
Parveen, M. & Shaik, M. A. Review on Penetration Testing Techniques in Cyber Security. Pages 1265–1270. In *2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS).* August 2023. https://doi.org/10.1109/ICAISS58487.2023.10250659

[Pavlova 2024]
Pavlova, M. et al. Automated Red Teaming with GOAT: The Generative Offensive Agent Tester. *arXiv* [preprint]. October 2, 2024. https://doi.org/10.48550/arXiv.2410.01606

[PCI 2024]
PCI Security Standards Council. Payment Card Industry Data Security Standards. *PCI Security Standards Council Website*. October 15, 2024 [accessed]. https://www.pcisecuritystandards.org/standards/

[Perez 2022]
Perez, E. et al. Red Teaming Language Models with Language Models. *arXiv* [preprint]. February 7, 2022. https://doi.org/10.48550/arXiv.2202.03286

[Pierce 2005]
Pierce, J. D. et al. In Pursuit of a Standard Penetration Testing Methodology. *Journal of Information Warfare*. Volume 4. Issue 3. 2005. Pages 26–39. https://www.jstor.org/stable/26504027

[Pillutla 2018]
Pillutla, H. & Arjunan, A. A Survey of Security Concerns, Mechanisms and Testing in Cloud Environment. Pages 1519–1524. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. March 2018. https://doi.org/10.1109/ICECA.2018.8474855

[Pozzobon 2018]
Pozzobon, E. et al. A Survey on Media Access Solutions for CAN Penetration Testing. In *ACM Computer Science in Cars Symposium (CSCS) 2018.* https://www.researchgate.net/publication/328687253_A_Survey_on_Media_Access_Solutions_for_CAN_Penetration_Testing

[Proofpoint 2024]
Proofpoint. What Is a Red Team in Cybersecurity? [definition] *Proofpoint Website*. November 2, 2024 [accessed]. https://www.proofpoint.com/us/threat-reference/red-team

[Pu 2024]
Pu, R. et al. BaitAttack: Alleviating Intention Shift in Jailbreak Attacks via Adaptive Bait Crafting. Pages 15654–15668. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing.* November 2024. https://doi.org/10.18653/v1/2024.emnlp-main.877

[Qi 2024]
Qi, X. et al. Visual Adversarial Examples Jailbreak Aligned Large Language Models. Pages 21527–21536. In *Proceedings of the AAAI Conference on Artificial Intelligence.* 2024. https://doi.org/10.1609/aaai.v38i19.30150

[Qiu 2024]
Qiu, J. et al. LLM-Based Agentic Systems in Medicine and Healthcare. *Nature Machine Intelligence.* Volume 6. Issue 12. December 5, 2024. Pages 1418–1420. https://www.nature.com/articles/s42256-024-00944-1

[Raina 2024]
Raina, V. et al. Muting Whisper: A Universal Acoustic Adversarial Attack on Speech Foundation Models. Pages 7549–7565. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024. https://doi.org/10.18653/v1/2024.emnlp-main.430

[Raju 2021]
Raju, A. D. et al. A Survey on Cross-Architectural IoT Malware Threat Hunting. *IEEE Access.* Volume 9. June 22, 2021. Pages 91686–91709. https://doi.org/10.1109/ACCESS.2021.3091427

[Rando 2025]
Rando, J. et al. Adversarial ML Problems Are Getting Harder to Solve and to Evaluate. *arXiv* [preprint]. February 4, 2025. https://doi.org/10.48550/arXiv.2502.02260

[Rando 2024a]
Rando, J. et al. Gradient-Based Jailbreak Images for Multimodal Fusion Models. *arXiv* [preprint]. October 23, 2024. https://doi.org/10.48550/arXiv.2410.03489

[Rando 2024b]
Rando, J. & Tramèr, F. Universal Jailbreak Backdoors from Poisoned Human Feedback. *arXiv* [preprint]. April 29, 2024. https://doi.org/10.48550/arXiv.2311.14455

[Ravindran 2022]
Ravindran, U. & Potukuchi, R. V. A Review on Web Application Vulnerability Assessment and Penetration Testing. *Review of Computer Engineering Studies.* Volume 9. Issue 1. March 31, 2022. Pages 1–22. https://doi.org/10.18280/rces.090101

[Ren 2024]
Ren, Q. et al. Derail Yourself: Multi-Turn LLM Jailbreak Attack Through Self-Discovered Clues. *arXiv* [preprint]. October 14, 2024. https://arxiv.org/pdf/2410.10700v1

[Rengarajan 2024]
Rengarajan, D. et al. Imitation Guided Automated Red Teaming. OpenReview [preprint]. October 12, 2024. https://openreview.net/forum?id=fkOZjYwm2R

[Russinovich 2024]
Russinovich, M.; Salem, A.; & Eldan, R. Great, Now Write an Article About That: The Crescendo Multi-Turn LLM Jailbreak Attack. *arXiv* [preprint]. April 2, 2024. http://arxiv.org/abs/2404.01833

[Russo 2016]
Russo, D. Benefits of Open Source Software in Defense Environments. Pages 123–131. In *Proceedings of 4th International Conference in Software Engineering for Defence Applications.* January 2016. https://doi.org/10.1007/978-3-319-27896-4_11

[Sagar 2024]
Sagar, S. et al. LLM-Assisted Red Teaming of Diffusion Models Through "Failures Are Fated, But Can Be Faded." *arXiv* [preprint]. October 22, 2024. https://doi.org/10.48550/arXiv.2410.16738

[Saiem 2024]
Saiem, B. A. et al. SequentialBreak: Large Language Models Can be Fooled by Embedding Jailbreak Prompts into Sequential Prompt Chains. *arXiv* [preprint]. November 10, 2024. https://doi.org/10.48550/arXiv.2411.06426

[Shah 2014]
Shah, S. & Mehtre, B. M. An Overview of Vulnerability Assessment and Penetration Testing Techniques. *Journal of Computer Virology and Hacking Techniques.* Volume 11. Issue 1. November 28, 2014. Pages 27–49. https://doi.org/10.1007/s11416-014-0231-x

[Shen 2024]
Shen, X. et al. Voice Jailbreak Attacks Against GPT-4o. *arXiv* [preprint]. May 29, 2024. https://doi.org/10.48550/arXiv.2405.19103

[Sheng 2019]
Sheng, E. et al. The Woman Worked as a Babysitter: On Biases in Language Generation. Pages 3407–3412. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. November 2019. https://doi.org/10.18653/v1/D19-1339

[Sheshadri 2024]
Sheshadri, A. et al. Latent Adversarial Training Improves Robustness to Persistent Harmful Behaviors in LLMs. *arXiv* [preprint]. August 21, 2024. https://doi.org/10.48550/arXiv.2407.15549

[Shivayogimath 2014]
Shivayogimath, C. N. An Overview of Network Penetration Testing. *International Journal of Research in Engineering and Technology.* Volume 3. Issue 07. July 2014. Pages 408–413. https://ijret.org/volumes/2014v03/i07/IJRET20140307070.pdf

[Siddaway 2019]
Siddaway, A. P. et al. How to Do a Systematic Review: A Best Practice Guide for Conducting and Reporting Narrative Reviews, Meta-Analyses, and Meta-Syntheses. *Annual Review of Psychology.* Volume 70. August 8, 2019. Pages 747–770. https://doi.org/10.1146/annurev-psych-010418-102803

[Solisch 2022]
Solisch, M. Seminar zu aktuellen Themen der Elektro- und Informationstechnik [Overview of Different Approaches and Types of Penetration Testing]. *ResearchGate* [preprint]. Summer 2022. https://www.researchgate.net/profile/Andrea-Reindl/publication/364180536_FMS-BERICHTE_SOMMERSEMESTER_2022/links/633d93879cb4fe44f305a2dc/FMS-BERICHTE-SOMMERSEMESTER-2022.pdf#page=35

[Souly 2025]
Souly, A. et al. A StrongREJECT for Empty Jailbreaks. Pages 125416–125440. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*. June 2025. https://papers.nips.cc/paper_files/paper/2024/hash/e2e06adf560b0706d3b1ddfca9f29756-Abstract-Datasets_and_Benchmarks_Track.html

[Sun 2024]
Sun, X. et al. Multi-Turn Context Jailbreak Attack on Large Language Models from First Principles. *arXiv* [preprint]. August 8, 2024. https://doi.org/10.48550/arXiv.2408.04686

[Sun 2022]
Sun, S. L. et al. Turning Disruption into Growth Opportunity: The Red Team Strategy. *Journal of Business Strategy.* Volume 43. Issue 6. October 19, 2022. Pages 365–372. https://doi.org/10.1108/JBS-05-2021-0087

[Swanson 2024]
Swanson, J. #CalibrateAI/Project Athena Update. *U.S. Army Website*. November 19, 2024. https://www.army.mil/article/281451/calibrateaiproject_athena_update

[Takemoto 2024]
Takemoto, K. All in How You Ask for It: Simple Black-Box Method for Jailbreak Attacks. *Applied Sciences.* Volume 14. Issue 9. April 23, 2024. Page 3558. https://doi.org/10.3390/app14093558

[Tang 2024]
Tang, Y. et al. RoleBreak: Character Hallucination as a Jailbreak Attack in Role-Playing Systems. Pages 7386–7402. In *Proceedings of the 31st International Conference on Computation Linguistics*. January 2025. https://aclanthology.org/2025.coling-main.494.pdf

[Teichmann 2023]
Teichmann, F. M. & Boticiu, S. R. An Overview of the Benefits, Challenges, and Legal Aspects of Penetration Testing and Red Teaming. *International Cybersecurity Law Review.* Volume 4. Issue 4. September 4, 2023. Pages 387–397. https://doi.org/10.1365/s43439-023-00100-2

[Thomas 2025]
Thomas, K. et al. Supporting Human Raters with the Detection of Harmful Content using Large Language Models. Pages 2772–2789. In *2025 IEEE Symposium on Security and Privacy*. May 2025. https://doi.ieeecomputersociety.org/10.1109/SP61157.2025.00082

[Tigner 2021]
Tigner, M. et al. Analysis of Kali Linux Penetration Tools: A Survey of Hacking Tools. Pages 1–6. In *2021 International Conference on Electrical, Computer and Energy Technologies (ICECET).* December 2021. https://doi.org/10.1109/ICECET52533.2021.9698572

[Tu 2024]
Tu, S. et al. Knowledge-to-Jailbreak: One Knowledge Point Worth One Attack. *arXiv* [preprint]. June 17, 2024. https://doi.org/10.48550/arXiv.2406.11682

[Tzu 2008]
Tzu, S. The Art of War. In *Strategic Studies.* Mahnken, T.; Maiolo, J.; & Maiolo, J. A. [editors]. Routledge. Pages 63–91. 2008. ISBN: 9780203928462. https://www.taylorfrancis.com/chapters/edit/10.4324/9780203928462-11/art-war-sun-tzu

[Valea 2019]
Valea, E. et al. A Survey on Security Threats and Countermeasures in IEEE Test Standards. *IEEE Design & Test.* Volume 36. Issue 3. June 2019. Pages 95–116. https://doi.org/10.1109/MDAT.2019.2899064

[Vasenius 2022]
Vasenius, P. Best Practices in Cloud-Based Penetration Testing. Master of Science in Technology Thesis, University of Turku. October 2022. https://www.utupub.fi/bitstream/handle/10024/173476/Vasenius_Petrus_opinnayte.pdf

[Vela 2023]
Vela, E. et al. Google's Reward Criteria for Reporting Bugs in AI Products [blog post]. *Google Online Security Blog*. October 26, 2023. https://security.googleblog.com/2023/10/googles-reward-criteria-for-reporting.html

[Vidgen 2024]
Vidgen, B. et al. Introducing v0.5 of the AI Safety Benchmark from MLCommons. *ML Commons Website*. April 16, 2024. https://mlcommons.org/2024/04/mlc-aisafety-v0-5-poc/

[Wang, F. 2024]
Wang, F. et al. MRJ-Agent: An Effective Jailbreak Agent for Multi-Round Dialogue. *arXiv* [preprint]. November 6, 2024. https://doi.org/10.48550/arXiv.2411.03814

[Wang, H. 2024]
Wang, H. et al. ASETF: A Novel Method for Jailbreak Attack on LLMs through Translate Suffix Embeddings. Pages 2697–2711. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. November 2024. https://doi.org/10.18653/v1/2024.emnlp-main.157

[Wang, W. 2017]
Wang, W. Survey of Software Vulnerability Discovery Technology. Pages 9–13. In *Proceedings of the 2017 7th International Conference on Social Network, Communication and Education*. July 2017. DOI: 10.2991/snce-17.2017.3. https://www.atlantis-press.com/proceedings/snce-17/25882970

[Wang, W. 2024]
Wang, W. et al. Chain-of-Jailbreak Attack for Image Generation Models via Editing Step by Step. *arXiv* [preprint]. October 4, 2024. https://doi.org/10.48550/arXiv.2410.03869

[Wang, Y. 2024]
Wang, Y. et al. Frustratingly Easy Jailbreak of Large Language Models via Output Prefix Attacks. *Research Square* [preprint]. May 9, 2024. https://doi.org/10.21203/rs.3.rs-4385503/v1

[Wang, Z. 2024a]
Wang, Z. et al. Functional Homotopy: Smoothing Discrete Optimization via Continuous Parameters for LLM Jailbreak Attacks. *arXiv* [preprint]. October 5, 2024. https://doi.org/10.48550/arXiv.2410.04234

[Wang, Z. 2024b]
Wang, Z. et al. Poisoned LangChain: Jailbreak LLMs by LangChain. *arXiv* [preprint]. June 26, 2024. https://doi.org/10.48550/arXiv.2406.18122

[Wang, Z. 2024c]
Wang, Z. et al. Hide Your Malicious Goal into Benign Narratives: Jailbreak Large Language Models Through Carrier Articles. *arXiv* [preprint]. August 20, 2024. https://doi.org/10.48550/arXiv.2408.11182

[Wei, A. 2023]
Wei, A. et al. Jailbroken: How Does LLM Safety Training Fail? Pages 80079–80110. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. December 10, 2023. https://dl.acm.org/doi/10.5555/3666122.3669630

[Wei, X. 2025]
Wei, X. et al. Addressing Bias in Generative AI: Challenges and Research Opportunities in Information Management. *Information & Management*. Volume 62. Issue 2. March 1, 2025. https://dl.acm.org/doi/10.1016/j.im.2025.104103

[Weidinger 2024]
Weidinger, L. et al. STAR: SocioTechnical Approach to Red Teaming Language Models. *arXiv* [preprint]. October 23, 2024. https://doi.org/10.48550/arXiv.2406.11757

[Weidinger 2023]
Weidinger, L. et al. Sociotechnical Safety Evaluation of Generative AI Systems. *arXiv* [preprint]. October 31, 2023. http://arxiv.org/abs/2310.11986

[Weidinger 2021]
Weidinger, L. et al. Ethical and Social Risks of Harm from Language Models. *arXiv* [preprint]. December 8, 2021. https://doi.org/10.48550/arXiv.2112.04359

[Wong 2024]
Wong, A. et al. SMILES-Prompting: A Novel Approach to LLM Jailbreak Attacks in Chemical Synthesis. *arXiv* [preprint]. October 21, 2024. https://doi.org/10.48550/arXiv.2410.15641

[Wu, T. 2024a]
Wu, T. et al. You Know What I'm Saying: Jailbreak Attack via Implicit Reference. *arXiv* [preprint]. October 8, 2024. https://doi.org/10.48550/arXiv.2410.03857

[Wu, Y. 2024b]
Wu, Y. et al. Can Large Language Models Automatically Jailbreak GPT-4V? *arXiv* [preprint]. August 23, 2024. https://doi.org/10.48550/arXiv.2407.16686

[Xiao 2024]
Xiao, Z. et al. Distract Large Language Models for Automatic Jailbreak Attack. Pages 16230–16244. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. November 2024. DOI: 10.18653/v1/2024.emnlp-main.908. https://aclanthology.org/2024.emnlp-main.908/

[Ximbo 2022]
Ximbo, B. et al. A Survey on IoT Vulnerability Discovery. Pages 267–282. In *Network and System Security.* December 7, 2022. https://doi.org/10.1007/978-3-031-23020-2_15

[Xu, H. 2024]
Xu, H. et al. RedAgent: Red Teaming Large Language Models with Context-Aware Autonomous Language Agent*. arXiv* [preprint]. July 23, 2024. https://doi.org/10.48550/arXiv.2407.16667

[Xu, X. 2024]
Xu, X. et al. Watch Your Words: Successfully Jailbreak LLM by Mitigating the "Prompt Malice." Pages 295–309. In *Web and Big Data.* August 28, 2024. https://doi.org/10.1007/978-981-97-7232-2_20

[Yaacoub 2021]
Yaacoub, J.-P. A. et al. A Survey on Ethical Hacking: Issues and Challenges. *arXiv* [preprint]. March 28, 2021. https://doi.org/10.48550/arXiv.2103.15072

[Yadav 2014]
Yadav, S. & Navdeti, C. Survey: Secured Techniques for Vulnerability Assessment and Penetration Testing. *International Journal of Computer Science and Information Technologies.* Volume 5. Issue 4. 2014. Pages 5132–5135. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=6dadfc953079cc30a96d935ba9d10b96ee1a89f9

[Yang, H. 2024a]
Yang, H. et al. Audio Is the Achilles' Heel: Red Teaming Audio Large Multimodal Models. *arXiv* [preprint]. October 31, 2024. https://doi.org/10.48550/arXiv.2410.23861

[Yang, H. 2024b]
Yang, H. et al. Jigsaw Puzzles: Splitting Harmful Questions to Jailbreak Large Language Models. *arXiv* [preprint]. October 15, 2024. https://doi.org/10.48550/arXiv.2410.11459

[Yang, Y. 2024a]
Yang, Y. & Fu, H. Transferable Ensemble Black-Box Jailbreak Attacks on Large Language Models. *arXiv* [preprint]. November 27, 2024. https://doi.org/10.48550/arXiv.2410.23558

[Yang, Y. 2024b]
Yang, Y. et al. SoP: Unlock the Power of Social Facilitation for Automatic Jailbreak Attack. *arXiv* [preprint]. July 2, 2024. https://openreview.net/forum?id=tD1atZW1vv

[Yao, D. 2024]
Yao, D. et al. FuzzLLM: A Novel and Universal Fuzzing Framework for Proactively Discovering Jailbreak Vulnerabilities in Large Language Models. Pages 4485–4489. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* April 2024. https://doi.org/10.1109/ICASSP48485.2024.10448041

[Yao, Y. 2024]
Yao, Y. et al. A Survey on Large Language Model (LLM) Security and Privacy: The Good, the Bad, and the Ugly. *High-Confidence Computing.* Volume 4. Issue 2. June 2024. Page 100211. https://doi.org/10.1016/j.hcc.2024.100211

[Ying 2024]
Ying, Z. et al. Jailbreak Vision Language Models via Bi-Modal Adversarial Prompt. *arXiv* [preprint]. July 1, 2024. https://doi.org/10.48550/arXiv.2406.04031

[Yoo 2024]
Yoo, H. et al. Code-Switching Red-Teaming: LLM Evaluation for Safety and Multilingual Understanding. *arXiv* [preprint]. June 17, 2024. https://doi.org/10.48550/arXiv.2406.15481

[Yu, H. 2024]
Yu, H. et al. Step Vulnerability Guided Mean Fluctuation Adversarial Attack against Conditional Diffusion Models. Pages 6791–6799. In *Proceedings of the AAAI Conference on Artificial Intelligence.* Volume 38. Issue 7. March 24, 2024. https://doi.org/10.1609/aaai.v38i7.28503

[Yu, J. 2024a]
Yu, J. et al. GPTFUZZER: Red Teaming Large Language Models with Auto-Generated Jailbreak Prompts. *arXiv* [preprint]. June 27, 2024. https://doi.org/10.48550/arXiv.2309.10253

[Yu, J. 2024b]
Yu, J. et al. Enhancing Jailbreak Attack Against Large Language Models Through Silent Tokens. *arXiv* [preprint]. May 31, 2024. https://doi.org/10.48550/arXiv.2405.20653

[Yu, M. 2020]
Yu, M. et al. A Survey of Security Vulnerability Analysis, Discovery, Detection, and Mitigation on IoT Devices. *Future Internet.* Volume 12. Issue 2. February 2020. Page 27. https://doi.org/10.3390/fi12020027

[Zeng 2024a]
Zeng, Y. et al. AdvI2I: Adversarial Image Attack on Image-to-Image Diffusion Models. *arXiv* [preprint]. November 1, 2024. https://doi.org/10.48550/arXiv.2410.21471

[Zeng 2024b]
Zeng, Y. et al. How Johnny Can Persuade LLMs to Jailbreak Them: Rethinking Persuasion to Challenge AI Safety by Humanizing LLMs. *arXiv* [preprint]. August 2024. https://doi.org/10.48550/arXiv.2401.06373

[Zhang, G. 2024]
Zhang, G. et al. Unveiling Vulnerabilities in Large Vision-Language Models: The SAVJ Jailbreak Approach. Pages 417–434. In *Artificial Neural Networks and Machine Learning – ICANN 2024*. September 17, 2024. https://doi.org/10.1007/978-3-031-72344-5_28

[Zhang, H. 2024]
Zhang, H. et al. Jailbreak Open-Sourced Large Language Models via Enforced Decoding. Pages 5475–5493. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics.* August 2024. https://doi.org/10.18653/v1/2024.acl-long.299

[Zhang, J. 2024a]
Zhang, J. et al. EnJa: Ensemble Jailbreak on Large Language Models. *arXiv* [preprint]. August 7, 2024. https://doi.org/10.48550/arXiv.2408.03603

[Zhang, J. 2024b]
Zhang, J. et al. Holistic Automated Red Teaming for Large Language Models Through Top-Down Test Case Generation and Multi-Turn Interaction. *arXiv* [preprint]. September 25, 2024. https://doi.org/10.48550/arXiv.2409.16783

[Zhang, T. 2024]
Zhang, T. et al. WordGame: Efficient & Effective LLM Jailbreak via Simultaneous Obfuscation in Query and Response. *arXiv* [preprint]. May 22, 2024. https://doi.org/10.48550/arXiv.2405.14023

[Zhao, A. 2024]
Zhao, A. et al. DiveR-CT: Diversity-Enhanced Red Teaming Large Language Model Assistants with Relaxing Constraints. *arXiv* [preprint]. December 20, 2024. https://arxiv.org/abs/2405.19026

[Zhao, J. 2024]
Zhao, J. et al. SQL Injection Jailbreak: A Structural Disaster of Large Language Models. *arXiv* [preprint]. November 3, 2024. https://doi.org/10.48550/arXiv.2411.01565

[Zhao, W. 2025]
Zhao, W. et al. Diversity Helps Jailbreak Large Language Models. Pages 4647–4680. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*. April 2025. https://aclanthology.org/2025.naacl-long.238/

[Zhou, X. 2024]
Zhou, X. et al. HAICOSYSTEM: An Ecosystem for Sandboxing Safety Risks in Human-AI Interactions. *arXiv* [preprint]. October 21, 2024. https://doi.org/10.48550/arXiv.2409.16427

[Zhou, Y. 2024a]
Zhou, Y. et al. Virtual Context: Enhancing Jailbreak Attacks with Special Token Injection. Pages 11843–11857. In *Findings of the Association for Computational Linguistics*. August 2024. https://aclanthology.org/2024.findings-emnlp.692/

[Zhou, Y. 2024b]
Zhou, Y. et al. Large Language Models Are Involuntary Truth-Tellers: Exploiting Fallacy Failure for Jailbreak Attacks. *arXiv* [preprint]. July 1, 2024. https://doi.org/10.48550/arXiv.2407.00869

[Zhu 2014]
Zhu, N.-F. et al. Study on the Standards and Procedures of the Penetration Testing. Pages 87–93. In *International Conference on Computer Science and Network Security (CSNS 2014).* March 2014. https://ci2s-enterprise.com.ar/2014/03/03/2014-international-conference-on-computer-science-and-network-security-2/

[Zou 2025]
Zou, A. et al. Improving Alignment and Robustness with Circuit Breakers. *Advances in Neural Information Processing Systems.* Volume 37. 2024. Pages 83345–83373. https://proceedings.neurips.cc/paper_files/paper/2024/hash/97ca7168c2c333df5ea61ece3b3276e1-Abstract-Conference.html

[Zou 2023]
Zou, A. et al. Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv* [preprint]. December 20, 2023. https://doi.org/10.48550/arXiv.2307.15043

9 Appendix A: Systematic Review Methodology

9.1 Appendix A.1: Search Terms

To search the cyber red-teaming literature, our first set of keywords was composed of relevant variations of "red-teaming":

- red team(ing)
- penetration test(ing)
- security test(ing)
- ethical hacking
- vulnerability research
- vulnerability discovery
- threat hunting
- cyber risk assessment
- cybersecurity assessment
- cyber security assessment

Our second set of keywords was composed of relevant variations of "review":

- review
- survey
- overview
- standard(s)
- guideline(s)
- best practice(s)

When searching, we used queries such as *'intitle:* "*red team" AND intitle:* "*review"'*.

To search the genAI red-teaming literature, our first set of keywords was composed of relevant variations of "genAI":

- genAI
- LLM
- generative model
- foundation model

Our second set of keywords was composed of relevant variations of "red-teaming":

- red teaming
- jailbreak
- adversarial attack

- AI safety
- AI security

When searching, we used queries such as *'intitle:* "*red teaming" AND intitle:* "*genAI"'*.

9.2 Appendix A.2: Screened Papers

9.2.1 Papers Screened from the Cyber Literature Review

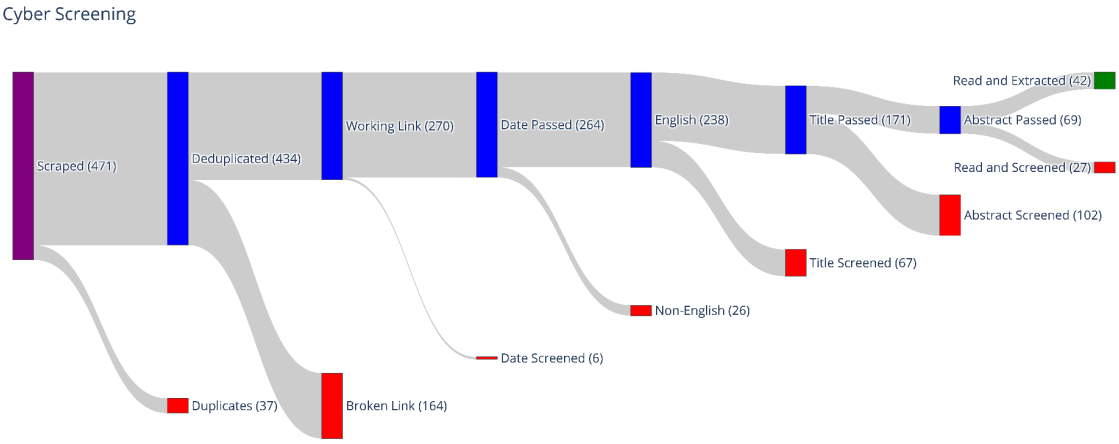


Figure 2: Papers Screened from the Cyber Literature Review at Each Stage

In Figure 2 we plot the overall screening flow for the cybersecurity literature. We began with 471 papers and screened papers that were duplicates (37), had broken links (164), did not have a verifiable date (6), and were non-English (26). We performed a strict deduplication on titles, so we also screened a few duplicates during later stages.

9.2.1.1 Papers Screened by Title

We then screened 67 papers based on their title. We categorize the reasons for screening below and list the date and title of each paper. Note that papers may be screened for multiple reasons, so the total exceeds 67.

- Related to geotechnical engineering and piezocone penetration testing (20)
  - 1990 Australian Experience in Cone Penetration Testing: Survey Results as at 30 April, 1988
  - 2020 Review of Free Fall Penetration Testing and Application in Offshore Engineering
  - 2016 Review of Full Flow Penetration Testing in Ocean Geotechnical Engineering Practice
  - 2020 Variable Penetration Rate Testing for Shear Strength of Peat–A Review
  - 1989 Stiffness of Sands from CPT, SPT and DMT–A Critical Review. Penetration Testing in the UK. Proceedings of the Geotechnology …

  - 2022 Review on the Testing Theory and Engineering Application of Density Piezocone Penetration Test
  - 2017 Underwater Noise from Geotechnical Drilling and Standard Penetration Testing
  - 2010 Probabilistic Framework for Assessing Liquefaction Hazard at a Given Site in a Specified Exposure Time Using Standard Penetration Testing
  - 1991 Field Energy Measurements of Standard Penetration Testing
  - 2017 Radiated Noise Levels from Marine Geotechnical Drilling and Standard Penetration Testing
  - 2015 Effects of Percussion Drilling and Non-Standard Testing Equipment on Penetration Resistance and Liquefaction Assessment in Gravelly Soils
  - 2008 Improved Ultraviolet Induced Fluorescence (UVIF)-Standard Cone Penetration Testing (CPT) System to Detect Petroleum Hydrocarbon Contaminants
  - 2019 Improved Prediction of Permeability Rates and Performance for Green Infrastructure Using Standard Penetration Testing
  - 2019 Automatic Monitoring Technique and Discovery of Standard Penetration Testing in Geotechnical Engineering
  - 2023 Penetration Testing for Standards Development of Distributed Energy Resources
  - 1975 … Standards GOST 19912-74" Soils. Method of Field Dynamic Penetration Testing" and GOST 20069-74" Soils. Method of Field Static Penetration Testing"
  - 2020 Best Practice on Oil Contaminated Sites: A Reliable and Cost-Effective Site Characterization Using a Dual LIF Simultaneous UVOST®-TarGOST®; A Cone Penetration …
  - 2024 A Review of Advances in Research on the Seismic Vulnerability of Bridge Structures
  - 2011 Advance in Research on Groundwater Vulnerability: A Review
  - 2019 Preparing for Water Change in the Columbia River Basin: An Integrated Analysis of Vulnerability & Climate Research Review
- Case studies, examples, and guides (14)
  - 2017 Kali Linux Wireless Penetration Testing Beginner's Guide: Master Wireless Testing Techniques to Survey and Attack Wireless Networks with Kali Linux, Including …
  - 2023 Overview on Case Study Penetration Testing Models Evaluation
  - 2023 Security Vulnerability Analysis Using Penetration Testing Execution Standard (PTES): Case Study of Government's Website
  - 2021 Vulnerability Analysis of Wireless LAN Networks Using Penetration Testing Execution Standard: A Case Study of Cafes in Palembang
  - 2023 Guidelines for White Box Penetration Testing Wired Devices in Secure Network Environments

  - 2020 A Case Study of Penetration Testing According to OWASP Guidelines: beanTech and Their WebApps
  - 2009 Technical Guide to Information Security Testing and Assessment: Recommendations of the National Institute of Standards and Technology
  - 2003 Guideline on Network Security Testing
  - 2002 NIST Guideline on Network Security Testing
  - 2013 Security Testing Guidelines for Mobile Apps
  - 2016 … in Low-Consensus Fields: Supporting Commensuration Through Construct-Centered Methods Aggregation in the Case of Climate Change Vulnerability Research
  - 2020 A Review Paper on Ethical Hacking-e-Learning Case Study
  - 2019 Some Guidelines for Risk Assessment of Vulnerability Discovery Processes
  - 2023 An Ontology and Guidelines for Cybersecurity Risk Assessment in the Automotive Domain
- Book reviews, educational materials, courses, and teaching materials (7)
  - 2013 Book Review: Professional Penetration Testing: Creating and Learning in a Hacking Lab 2e
  - 1988 Book Review: Penetration Testing 1988: Volumes 1 and 2. Edited by J. de Ruiter. Rotterdam: AA Balkema.
  - 2003 A Survey of Educational Test Security Practices and Procedures Throughout the United States During the 2001–2002 School Year
  - 2015 Ethical Hacking Pedagogy: An Analysis and Overview of Teaching Students to Hack
  - 2022 Review on Teaching Ethical Hacking
  - 2016 Review of Red Team: How to Succeed by Thinking Like the Enemy Review of Micah Zenko (New York: Basic Books, 2015)
  - 2023 The Right Tool for the Job: Overview, Comparison and Assessment of Methods for Cybersecurity Awareness Education and Verification
- Papers proposing automated red-teaming systems (12)
  - 2024 A Survey on Penetration Path Planning in Automated Penetration Testing
  - 2023 Automated Penetration Testing, A Systematic Review
  - 2024 Incorporation of Verifier Functionality in the Software for Operations and Network Attack Results Review and the Autonomous Penetration Testing System
  - 2019 Automatic Monitoring Technique and Discovery of Standard Penetration Testing in Geotechnical Engineering
  - 2023 Survey of Model-Based Security Testing Approaches in the Automotive Domain
  - 2021 Overview of Automotive Security Testing Approaches
  - 2019 White-Box Testing Automation with SonarQube: Continuous Integration, Code Review, Security, and Vendor Branches

  - 2024 Automating IoT Security Standard Testing by Common Security Tools
  - 2023 Automated Security Testing for Mobile Apps: Tools, Techniques, and Best Practices
  - 2022 A Methodology to Support Automatic Cyber Risk Assessment Review
  - 2022 Risk Assessment of SCADA Cyber Attack Methods: A Technical Review on Securing Automated Real-time SCADA Systems
  - 2023 An Ontology and Guidelines for Cybersecurity Risk Assessment in the Automotive Domain
- Other unrelated domains such as agriculture, climatology, etc. (18)
  - 2023 A Review on Adapting Social Engineering Attack as One of the Penetration Testing Techniques
  - 2023 Analysis of the Current Population Survey Food Security Supplement Split-Panel Test
  - 1991 Testing Standards for Physical Security Systems at Category 1 Fuel Cycle Facilities
  - 2019 A Review of the Research Methods on Vulnerability of Transportation System
  - 2022 Research Areas in Consumer Vulnerability a Systematic Literature Review
  - 2021 Progress in Agricultural Vulnerability and Risk Research in India: A Systematic Review
  - 2023 A Systematic Review with Bibliometric Analysis of Different Approaches and Methodologies for Undertaking Flood Vulnerability Research
  - 2016 … in Low-Consensus Fields: Supporting Commensuration through Construct-Centered Methods Aggregation in the Case of Climate Change Vulnerability Research
  - 2022 Forest Vulnerability to Climate Change: A Review for Future Research Framework
  - 2023 Vulnerability and Anti-Vulnerability: Research Progress and Review of Tourism Resilience.
  - 2024 A Review of Research Advances in Personnel Vulnerability Analysis and Application
  - 2011 Interpretive Review of Conceptual Frameworks and Research Models that Inform Australia's Agricultural Vulnerability to Climate Change
  - 2023 Spinning in Circles? A Systematic Review on the Role of Theory in Social Vulnerability, Resilience and Adaptation Research
  - 2024 Indicators of Riverbank Erosion Vulnerability Assessment: A Systematic Literature Review for Future Research
  - 2021 Research Review on Vulnerability of District Heating System and its Interdependent Infrastructure Network
  - 2022 Visualization of Urban Vulnerability Research Progress and Review Analysis Based on Citespace

- 2019 Preparing for Water Change in the Columbia River Basin: An Integrated Analysis of Vulnerability & Climate Research Review
- 2024 Cyber Risk Assessment for Cyber-Physical Systems: A Review of Methodologies and Recommendations for Improved Assessment Effectiveness

9.2.1.2 Papers Screened by Abstract

We then screened 102 papers based on their title. We categorize the reasons for screening below and list the date and title of each paper. For the abstracts we only list a single reason for exclusion, though some papers could have been screened for multiple reasons.

- Topic (69)
  - 2023 A Systematic Literature Review on Penetration Testing in Networks: Future Research Directions
  - 2019 A Review of Standardization for Penetration Testing Reports and Documents
  - 2023 A Comprehensive Literature Review of Artificial Intelligent Practices in the Field of Penetration Testing
  - 2024 A Systematic Literature Review on Internet of Vehicles Security Challenges and Penetration Testing Solutions
  - 2024 A Review of Penetration Testing Process for SQL Injection Attack
  - 2023 A Comprehensive Review on Penetration Testing Tools with Emerging Technology
  - 2012 Review of the Basics of Hacking and Penetration Testing: Ethical Hacking and Penetration Testing Made Easy. P. Engebretson, Syngress Publishing, Waltham, MA …
  - 1975 Review of the European Symposium on Penetration Testing
  - 2019 Standard Penetration Testing in a Virtual Calibration Chamber
  - 1986 Field Testing: The Standard Penetration Test
  - 2022 Standard Quality Control Testing, Virus Penetration, and Glove Durability
  - 1979 European Standard on Penetration Testing—a Necessity
  - 2021 An Examination of Industry Standards of Success within Penetration Testing Groups
  - 2023 Strengthening IT Governance in the Crypto Marketplace: Leveraging Penetration Testing and Standards Alignment
  - 2024 Cloud Security: Challenges and Best Practices in Penetration Testing
  - 2024 Penetration Testing: A Cost-Benefit Analysis of Best Practices Implementation for Software Startups
  - 2009 National Security with a Canadian Twist: The Investment Canada Act and the New National Security Review Test
  - 2009 Design of a New Emission-Security Standard for Radiated Emission EMC Test
  - 2024 A Survey of Security Testing Techniques for Deep Learning Frameworks

- 2022 Machine Learning in Software Security Testing: A Literature Survey
- 2016 Bachelor Thesis Cashier-as-a-Service Based Webshops Overview and Steps Towards Security Testing
- 2023 Model-Based Security Testing in IoT Systems: A Rapid Review
- 2024 A Systematic Literature Review on Software Security Testing Using Metaheuristics
- 2011 Review of Security Testing Tools
- 2023 A Critical Review on Search-Based Security Testing of Programs
- 2023 Security Testing for Web Applications: A Systematic Literature Review
- 2024 Barriers to Using Static Application Security Testing (SAST) Tools: A Literature Review
- 2015 A Review of Threat Modelling and its Hybrid Approaches to Software Security Testing
- 2024 WoS Bibliometric-Based Review for Security Testing of Android Applications Using Malware Analysis
- 2016 Critical Review on Software Testing: Security Perspective
- 2013 Literature Review of Mobile Applications Testing on Cloud from Information Security Perspective
- 2023 Review on the Competency of Evaluators at Information Technology Product Security Testing Laboratory Based on SNI ISO/IEC 19896–3: 2018
- 2017 Adversarial Testing to Increase the Overall Security of Embedded Systems: A Review of the Process
- 2023 Python Security in DevOps: Best Practices for Secure Coding, Configuration Management, and Continuous Testing and Monitoring
- 2005 Best Practices in a University Environment for Homeland Security Research–Testing and Evaluation
- 2023 Research Communities in Cyber Security Vulnerability Assessments: A Comprehensive Literature Review
- 2017 A Review of Machine Learning in Software Vulnerability Research
- 2008 Short Review of Modern Vulnerability Research
- 2007 Vulnerability and Coping Strategies in Africa: Literature Review for Research in Zambia
- 2021 Survey on Ethical Hacking and Digital Forensics in Organizations and Speculative Text Mining
- 2022 Review on Ethical Hacking and its Techniques
- 2020 Review of Tools and Techniques of Ethical Hacking
- 2019 Review of Maurushat's Ethical Hacking
- 2003 Red Team, Blue Team: Galaxy Survey Shows that Color Matters

- 2024 Game-Theoretic Integration of Red Team Survey Data in Multi-Layer Security Systems
- 2024 Against the Achilles' Heel: A Survey on Red Teaming for Generative Models
- 2024 Red Teaming for Multimodal Large Language Models: A Survey
- 2023 Metrics and Red Teaming in Cyber Resilience and Effectiveness: A Systematic Literature Review
- 2024 Artificial Intelligence Cyberattacks in Red Teaming: A Scoping Review
- 2024 Considerations on AI Model Red-Teaming and Standards
- 2020 Software Vulnerability Analysis and Discovery Using Deep Learning Techniques: A Survey
- 2024 Vulnerability Discovery Based on Source Code Patch Commit Mining: A Systematic Literature Review
- 2019 Systematization of Vulnerability Discovery Knowledge: Review Protocol
- 2020 Is Your Threat Hunting Working? A New SANS Survey for 2020
- 2018 Cyber Risk Metrics Survey, Assessment, and Implementation Plan
- 2023 Overview and Recommendations for Cyber Risk Assessment in Nuclear Power Plants
- 2015 Smart Grid Cyber Security and Risk Assessment: An Overview
- 2024 A Systematic Literature Review for Modeling a Cyber Risk Assessment Framework
- 2023 Cybersecurity Assessment Framework: A Systematic Review*
- 2023 Cybersecurity Risk Assessment for Medium-Risk Drones: A Systematic Literature Review
- 2024 The Role of Internal Auditors Characteristics in Cybersecurity Risk Assessment in Financial-Based Business Organisations: A Conceptual Review
- 2016 Mapping of the Federal Financial Institutions Examination Council (FFIEC) Cybersecurity Assessment Tool (CAT) to the Cyber Resilience Review (CRR)
- 2023 The Role of Internal Auditors Characteristics in Cybersecurity Risk Assessment in Financial-Based Business Organisations: A Conceptual Review
- 2024 Continuous Monitoring and Assessment Mechanisms in Cybersecurity: Best Practices for Sustained Protection of Critical Assets
- 2021 Cybersecurity Assessment and Best Practices for Truck Stop Technologies
- 2020 Review on the Application of Knowledge Graph in Cyber Security Assessment
- 2022 Cyber Security Maturity Assessment Framework for Technology Startups: A systematic Literature Review
- 2021 A Review of Cyber Security Assessment (CSA) for Industrial Control Systems (ICS) and Their Impact on the Availability of the ICS Operation
- 2018 Standards on Cyber Security Assessment of Smart Grid

- Quality (16)
  - 2021 A Survey on Network Penetration Testing
  - 2019 A survey on Vulnerability Assessment & Penetration Testing for Secure Communication
  - 2014 Survey: Secured Techniques for Vulnerability Assessment and Penetration Testing
  - 2024 A Survey of Nmap Command Builder for Learning Penetration Testing
  - 2015 A Comparative Overview on Penetration Testing
  - 2023 An Overview of Penetration Testing and its Types
  - 2022 White Hat Security-An Overview of Penetration Testing Tools
  - 2022 Overview of Different Approaches and Types of Penetration Testing
  - 2020 A Comprehensive Literature Review of Penetration Testing & Its Applications
  - 2021 A Systematic Review on Penetration Testing
  - 2024 Systematic Literature Review of Challenges and AI Contributions in Penetration Testing
  - 2020 A Review of Penetration Testing and Vulnerability Assessment in Cloud Environment
  - 2021 A Review: Penetration Testing Approaches on Content Management System (CMS)
  - 2021 DER Cybersecurity Stakeholder Engagement, Standards Development, and EV Charger Penetration Testing
  - 2019 … for Reporting Vulnerability Research: How Can Peer Reviewed Articles Reflect Complex Practice in Low Consensus Fields Such That They Better Support Review and …
  - 2022 A Systematic Literature Review on Cyber Threat Hunting
- Non-Review (10)
  - 2021 Threat Modeling and Penetration Testing of a Yanzi IoT-System: A Survey on the Security of the System's RF Communication
  - 2020 Ethical Hacking of an IoT-device: Threat Assessment and Penetration Testing: A Survey on Security of a Smart Refrigerator
  - 2024 Web Application Penetration Testing with Artificial Intelligence: A Systematic Review
  - 2016 … of Engineering Sciences & Research Technology a Descriptive Review of Different Penetration Testing Tools and Methods
  - 2016 Auditing 6lowpan Networks Using Standard Penetration Testing Tools
  - 2021 Information System Security Analysis to Determine Server Security Vulnerability with Penetration Testing Execution Standard (PTES) Method at VWX University

  - 2024 Analysis Vulnerability Website Baleomolcreative dengan Metode Penetration Testing Execution Standard & Vulnerability Assessment Pada http Response Header …
  - 2019 Cybersecurity Analysis of a SCADA System Under Current Standards, Client Requisites, and Penetration Testing
  - 2024 A Repository for Testing Compliance to the Internet of Things (IoT) Security Standards
  - 2022 Cyber Red Teaming: Overview of Sly, an Orchestration Tool
- Link (5)
  - 2004 A Critical Review of Penetration Testing Methodologies
  - 2021 Review of the Benefits of DAST (Dynamic Application Security Testing) Versus SAST
  - 2007 Overview of Red Team Reports
  - 2018 SANS 2018 Threat Hunting Survey Results
  - 2023 Leveraging AI and ML for Proactive Threat Hunting: A Comprehensive Review
- Duplicate (2)
  - 2024 Check for Updates Artificial Intelligence Cyberattacks in Red Teaming: A Scoping Review Mays Al-Azzawi, Dung Doan, Tuomo Sipola () and Tero Kokkonen ID Jari …
  - 2022 Method for Conducting Systematic Literature Review (SLR) for Cyber Risk Assessment

9.2.1.3 Papers Screened During Extraction

Finally, we screened 27 papers during paper extraction. We categorize the reasons for screening below and list the date and title of each paper. Each paper has a single reason for exclusion.

- Topic (10)
  - 2019 A Systematic Literature Review and Meta-Analysis on Artificial Intelligence in Penetration Testing and Vulnerability Assessment
  - 2022 On Testing Security Requirements in Industry–A Survey Study
  - 2016 A Categorized Review on Software Security Testing
  - 2023 Security Aspect in Software Testing Perspective: A Systematic Literature Review.
  - 2022 Collaborative Application Security Testing for DevSecOps: An Empirical Analysis of Challenges, Best Practices and Tool Support
  - 2024 A Review Paper on Ethical Hacking
  - 2019 A Survey of the Software Vulnerability Discovery Using Machine Learning Techniques
  - 2022 Zero-Day Attack Solutions Using Threat Hunting Intelligence: Extensive Survey
  - 2016 A Review of Cyber Security Risk Assessment Methods for SCADA Systems

  - 2024 A Systematic Review of Cybersecurity Assessment Methods for HTTPS
- Access (5)
  - 2014 An Overview of Penetration Testing
  - 2016 Security Testing: A Survey
  - 2022 On Testing Security Requirements in Industry–a Survey Study
  - 2018 A Review of Testing Cloud Security
  - 2019 Dimensions of Robust Security Testing in Global Software Engineering: A Systematic Review
- Duplicate (4)
  - 2024 A Retrospective Analysis of a Rapid Review on Fuzz Security Testing for Software Implementation of Communication Protocols
  - 2022 Collaborative Application Security Testing for DevSecOps: An Empirical Analysis of Challenges, Best Practices and Tool Support
  - 2024 Bridging the Gap: A Survey and Classification of Research-Informed Ethical Hacking Tools (Supplementary Material)
  - 2021 Fuzzing the Internet of Things: A Review on the Techniques and Challenges for Efficient Vulnerability Discovery in Embedded Systems
- Quality (3)
  - 2013 A Survey on Software Security Testing Techniques
  - 2014 A Survey Report on Security for Testing Phase of Software Development Process
  - 2023 A Survey: Threat Hunting for the OT Systems
- Non-Review (3)
  - 2023 Review Paper on Wireless Network Penetration Testing
  - 2021 Analysis and Evaluation of Wireless Network Security with the Penetration Testing Execution Standard (PTES)
  - 2023 Enhancing Wireless Network Security via Ethical Hacking: Strategies and Best Practices.
- Language (2)
  - 2024 Exploring the Depths: An Overview of Penetration Testing
  - 2020 A Survey of Smart Contract Vulnerability Research

9.2.2 Papers Screened from the Generative AI Literature Review

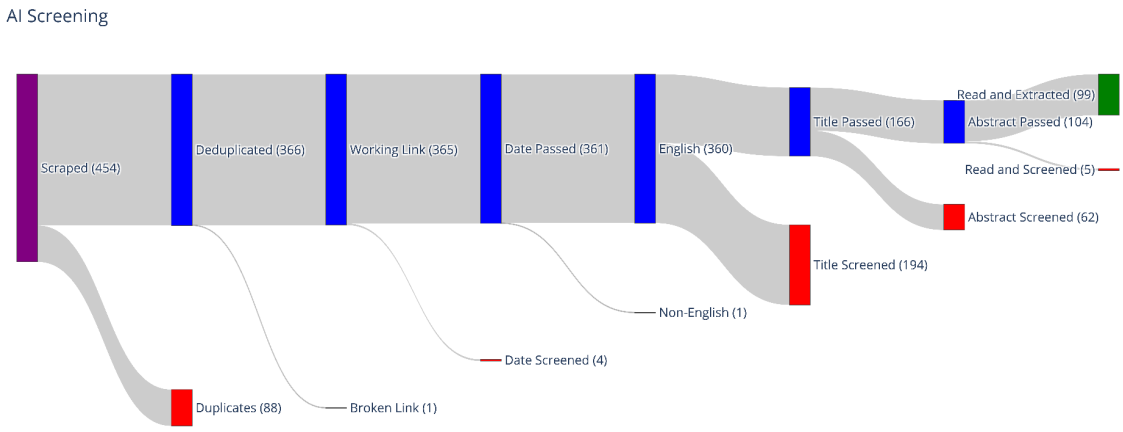


Figure 3: Papers Screened from the Generative AI Red-Teaming Review at Each Stage

In Figure 3 we plot the overall screening flow for the generative AI literature. We began with 454 papers and screened papers that were duplicates (88), had broken links (1), did not have a verifiable date (4), and were non-English (1). We performed a strict deduplication on titles, so we found a few duplicates during later stages.

9.2.2.1 Papers Screened by Title

We then screened 195 papers based on their titles. We categorize the reasons for screening below and list the date and title of each paper. Note that, unlike the cyber paper titles, we only listed a single reason for exclusion, though some papers could have been screened for multiple reasons. For example, a review of cyber red-teaming could be excluded for being a review or being from the wrong domain.

- Reviews, surveys, or being too broad (72):
  - 2024 Recent Advancements in LLM Red-Teaming: Techniques, Defenses, and Ethical Considerations
  - 2024 Red-Teaming for Generative AI: Silver Bullet or Security Theater?
  - 2024 Against the Achilles' Heel: A Survey on Red Teaming for Generative Models
  - 2024 LLMs Red Teaming
  - 2024 LLM Jailbreak Attack Versus Defense Techniques—A Comprehensive Study
  - 2024 Jailbreak Attacks and Defenses Against Large Language Models: A Survey
  - 2024 Comprehensive Assessment of Jailbreak Attacks Against LLMs
  - 2024 A Comprehensive Study of Jailbreak Attack Versus Defense for Large Language Models
  - 2024 A Comprehensive Study on Jailbreak Attacks and Defenses for Multimodal Large Language Models

- 2024 Competition Report: Finding Universal Jailbreak Backdoors in Aligned LLMs
- 2024 Survey on Adversarial Attack and Defense for Medical Image Analysis: Methods and Challenges
- 2024 AI Safety in Generative AI Large Language Models: A Survey
- 2024 AI Safety and Security
- 2024 Mapping Technical Safety Research at AI Companies: A Literature Review and Incentives Analysis
- 2024 Gen-AI for User Safety: A Survey
- 2024 Mechanistic Interpretability for AI Safety—A Review
- 2024 AI Safety and Ethics
- 2024 AI Safety Assurance for Automated Vehicles: A Survey on Research, Standardization, Regulation
- 2024 Systematic Overview of AI Security Standards
- 2024 Enhancing Autonomous System Security and Resilience with Generative AI: A Comprehensive Survey
- 2024 AI Security Assessment: Attacks and Defenses on Large Language Models
- 2024 An Overview of Trustworthy AI: Advances in IP Protection, Privacy-Preserving Federated Learning, Security Verification, and GAI Safety Alignment
- 2024 Red Teaming: Everything Everywhere All at Once
- 2024 Considerations on AI Model Red-Teaming and Standards
- 2024 A Safe Harbor for AI Evaluation and Red Teaming
- 2024 AI Red Teaming
- 2024 Jailbreak Attacks on Large Language Models and Possible Defenses: Present Status and Future Possibilities
- 2024 Analyzing Ethical Biases and Jailbreak Vulnerabilities in AI Systems
- 2024 Revealing the Difficulty in Jailbreak Defense on Language Models for Metaverse
- 2024 Trustworthy, Responsible, and Safe AI: A Comprehensive Architectural Framework for AI Safety with Challenges and Mitigations
- 2024 Bridging Today and the Future of Humanity: AI Safety in 2024 and Beyond
- 2024 Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress?
- 2024 Towards AI Safety: A Taxonomy for AI System Evaluation
- 2024 Safety Cases: Justifying the Safety of Advanced AI Systems
- 2024 AI Speech and AI Safety
- 2024 Standardization Trends on Safety and Trustworthiness Technology for Advanced AI
- 2024 The Elephant in the Room—Why AI Safety Demands Diverse Teams
- 2024 Safety Challenges of AI in Medicine

- 2024 Proceedings of Safety4ConvAI: The Third Workshop on Safety for Conversational AI@ LREC-COLING 2024
- 2024 Human-AI Safety: A Descendant of Generative AI and Control Systems Safety
- 2024 International Scientific Report on the Safety of Advanced AI
- 2024 Can There Be Responsible AI Without AI Liability? Incentivizing Generative AI Safety Through Ex-Post Tort Liability Under the EU AI Liability Directive
- 2024 Gaps in the Safety Evaluation of Generative AI
- 2024 Building a Culture of Safety for AI: Comparisons and Challenges
- 2024 Safety Cases for Frontier AI
- 2024 AI Safety Collides with the Overattribution Bias
- 2024 Holistic Safety and Responsibility Evaluations of Advanced AI Models
- 2024 Generative AI Agents in Autonomous Machines: A Safety Perspective
- 2024 SoK: Towards Security and Safety of Edge AI
- 2024 Safety Case Template for Frontier AI: A Cyber Inability Argument
- 2024 Unified Taxonomy in AI Safety: Watermarks, Adversarial Defenses, and Transferable Attacks
- 2024 Not Oracles of the Battlefield: Safety Considerations for AI-Based Military Decision Support Systems
- 2024 Towards Evaluations-Based Safety Cases for AI Scheming
- 2024 Assessing the Safety and Robustness of Advanced AI
- 2024 Probabilistic Analysis of Copyright Disputes and Generative AI Safety
- 2024 Understanding the First Wave of AI Safety Institutes: Characteristics, Functions, and Challenges
- 2024 Towards AI-Safety-by-Design: A Taxonomy of Runtime Guardrails in Foundation Model Based Systems
- 2024 Unpacking AI Security Considerations
- 2024 Data Security and Privacy Concerns for Generative AI Platforms
- 2024 Generative AI Security
- 2024 A Guide to Evaluating AI Vendors: Key Questions to Mitigate Security Risks
- 2024 Generative AI Security: Challenges and Countermeasures
- 2024 SecGenAI: Enhancing Security of Cloud-Based Generative AI Applications within Australian Critical Technologies of National Interest
- 2024 Complete Security and Privacy for AI Inference in Decentralized Systems
- 2024 Integrated AI Security and Efficiency: Trustworthiness, Trojan Detection, and Performance Acceleration
- 2024 Exploring Security Challenges in Generative AI for Web Engineering
- 2024 Assurance of Third-Party AI Systems for UK National Security

- 2024 Generative AI in Medical Practice: In-Depth Exploration of Privacy and Security Challenges
- 2024 Generative AI Security: Theories and Practices
- 2024 Security Considerations in Generative AI for Web Applications
- 2024 A Formal Framework for Assessing and Mitigating Emergent Security Risks in Generative AI Models: Bridging Theory and Dynamic Risk Mitigation
- 2024 Synchronized Coevolution: A Conceptual Framework for Sustaining a Human-Centered Security Culture in AI-Driven Environments

- Defenses (62):
  - 2024 Tiny Refinements Elicit Resilience: Toward Efficient Prefix-Model Against LLM Red-Teaming
  - 2024 Autodefense: Multi-Agent LLM Defense Against Jailbreak Attacks
  - 2024 Mitigating Adversarial Manipulation in LLMs: A Prompt-Based Approach to Counter Jailbreak Attacks (Prompt-G)
  - 2024 LLM Improvement for Jailbreak Defense: Analysis Through the Lens of Over-Refusal
  - 2024 Adversarial Tuning: Defending Against Jailbreak Attacks for LLMs
  - 2024 Defensive Prompt Patch: A Robust and Interpretable Defense of LLMs Against Jailbreak Attacks
  - 2024 RePD: Defending Jailbreak Attack Through a Retrieval-Based Prompt Decomposition Process
  - 2024 Defending Large Language Models Against Jailbreak Attacks via Layer-Specific Editing
  - 2024 Pruning for Protection: Increasing Jailbreak Resistance in Aligned LLMs Without Fine-Tuning
  - 2024 Jailbreak Antidote: Runtime Safety-Utility Balance via Sparse Representation Adjustment in Large Language Models
  - 2024 Break the Breakout: Reinventing LM Defense Against Jailbreak Attacks with Self-Refinement
  - 2024 Defensive Prompt Patch: A Robust and Generalizable Defense of Large Language Models Against Jailbreak Attacks
  - 2024 BackdoorAlign: Mitigating Fine-Tuning Based Jailbreak Attack with Backdoor Enhanced Safety Alignment
  - 2024 Securing Vision-Language Models with a Robust Encoder Against Jailbreak and Adversarial Attacks
  - 2024 RobustKV: Defending Large Language Models Against Jailbreak Attacks via KV Eviction
  - 2024 Mitigating Fine-Tuning Jailbreak Attack with Backdoor Enhanced Alignment
  - 2024 Token Highlighter: Inspecting and Mitigating Jailbreak Prompts for Large Language Models

- 2024 Safedecoding: Defending Against Jailbreak Attacks via Safety-Aware Decoding
- 2024 GradSafe: Detecting Jailbreak Prompts for LLMs via Safety-Critical Gradient Analysis
- 2024 Defending Jailbreak Prompts via In-Context Adversarial Game
- 2024 Safe Unlearning: A Surprisingly Effective and Generalizable Solution to Defend Against Jailbreak Attacks
- 2024 Defending Large Language Models Against Jailbreak Attacks Through Chain of Thought Prompting
- 2024 Defending Large Language Models Against Jailbreak Attacks via Semantic Smoothing
- 2024 HSF: Defending Against Jailbreak Attacks with Hidden State Filtering
- 2024 Merging Improves Self-Critique Against Jailbreak Attacks
- 2024 BlueSuffix: Reinforced Blue Teaming for Vision-Language Models Against Jailbreak Attacks
- 2024 Gradient Cuff: Detecting Jailbreak Attacks on Large Language Models by Exploring Refusal Loss Landscapes
- 2024 Safealigner: Safety Alignment Against Jailbreak Attacks via Response Disparity Guidance
- 2024 Prefix Guidance: A Steering Wheel for Large Language Models to Defend Against Jailbreak Attacks
- 2024 Knowledge Graph Unlearning to Defend Language Model Against Jailbreak Attack
- 2024 Bathe: Defense Against the jailbreak Attack in Multimodal Large Language Models by Treating Harmful Instruction as Backdoor Trigger
- 2024 EEG-Defender: Defending Against Jailbreak Through Early Exit Generation of Large Language Models
- 2024 UniGuard: Towards Universal Safety Guardrails for Jailbreak Attacks on Multimodal Large Language Models
- 2024 Defending Jailbreak Attack in VLMs via Cross-Modality Information Detector
- 2024 Adversarial for Good–Defending Training Data Privacy with Adversarial Attack Wisdom
- 2024 Improving Behavior Based Authentication Against Adversarial Attack Using XAI
- 2024 Incremental Adversarial Learning for Polymorphic Attack Detection
- 2024 Test-time Adversarial Defense with Opposite Adversarial Path and High Attack Time Cost
- 2024 PPNNI: Privacy-Preserving Neural Network Inference Against Adversarial Example Attack
- 2024 Artwork Protection Against Neural Style Transfer Using Locally Adaptive Adversarial Color Attack

  - 2024 Concept-Guided LLM Agents for Human-AI Safety Codesign
  - 2024 Safeguarding AI Agents: Developing and Analyzing Safety Architectures
  - 2024 Affirmative Safety: An Approach to Risk Management for High-Risk AI
  - 2024 SLM as Guardian: Pioneering AI Safety with Small Language Models
  - 2024 AI Risk Management Should Incorporate Both Safety and Security
  - 2024 Building Trustworthy NeuroSymbolic AI Systems: Consistency, Reliability, Explainability, and Safety
  - 2024 Affirmative Safety: An Approach to Risk Management for Advanced AI
  - 2024 An AI System Evaluation Framework for Advancing AI Safety: Terminology, Taxonomy, Lifecycle Mapping
  - 2024 Mechanistic Interpretability for Progress Towards Quantitative AI Safety
  - 2024 SURE: Framework for Safety to Construct Trustworthy AI
  - 2024 Innovative Approaches to Enhancing Safety and Ethical AI Interactions in Digital Environments
  - 2024 An Adversarial Perspective on Machine Unlearning for AI Safety
  - 2024 AI Safety and Ethics: Developing Robust Frameworks for Ethical AI Development and Deployment
  - 2024 Malak: AI-Based Multilingual Personal Assistant to Combat Misinformation and Generative AI Safety Issues
  - 2024 A Taxonomy of Multi-Layered Runtime Guardrails for Designing Foundation Model-Based Agents: Swiss Cheese Model for AI Safety by Design
  - 2024 Embodied AI with Two Arms: Zero-shot Learning, Safety and Modularity
  - 2024 Assuring AI Safety: Fallible Knowledge and the Gricean Maxims
  - 2024 Security of and by Generative AI Platforms
  - 2024 AuditNet: Conversational AI Security Assistant
  - 2024 Enhancing the Security of Edge-AI Runtime Environments: A Fine-Tuning Method Based on Large Language Models
  - 2024 Coordinated Flaw Disclosure for AI: Beyond Security Vulnerabilities
- Not generative AI (37):
  - 2024 Red Teaming Language Model Detectors with Language Models
  - 2024 A Red Teaming Framework for Securing AI in Maritime Autonomous Systems
  - 2024 Red-Teaming Segment Anything Model
  - 2024 Performance of LLM-Written Text Detectors Across Domains and Under Adversarial Attack
  - 2024 BEACOMP: A Novel Textual Adversarial Attack Architecture for Unveiling the Fragility of Neural Text Classifiers
  - 2024 Bots Shield Fake News: Adversarial Attack on User Engagement based Fake News Detection

- 2024 Humanizing Machine-Generated Content: Evading AI-Text Detection Through Adversarial Attack
- 2024 Constrained Adaptive Attack: Effective Adversarial Attack Against Deep Neural Networks for Tabular Data
- 2024 Adversarial Attack on 3D Fused Sensory Data in Drone Surveillance
- 2024 Deebbaa: A Benchmark Deep Black Box Adversarial Attack Against Cyber-Physical Power Systems
- 2024 A Low-Frequency Adversarial Attack Method for Object Detection Using Generative Model
- 2024 Deep Generative Models as an Adversarial Attack Strategy for Tabular Machine Learning
- 2024 Uncertainty-Aware Diffusion-Based Adversarial Attack for Realistic Colonoscopy Image Synthesis
- 2024 RW-VoiceShield: Raw Waveform-Based Adversarial Attack on One-Shot Voice Conversion
- 2024 Imperceptible Face Forgery Attack via Adversarial Semantic Mask
- 2024 ProGen: Projection-Based Adversarial Attack Generation Against Network Intrusion Detection
- 2024 Diffusion-Based Adversarial Attack to Automatic Speech Recognition
- 2024 AdvShadow: Evading DeepFake Detection via Adversarial Shadow Attack
- 2024 Black-Box Universal Adversarial Attack for DNN-Based Models of SAR Automatic Target Recognition
- 2024 Edge-Oriented Adversarial Attack for Deep Gait Recognition
- 2024 Signal Adversarial Examples Generation for Signal Detection Network via White-Box Attack
- 2024 Machine Learning-Based Anomaly Detection for Smart Home Networks Under Adversarial Attack
- 2024 Sparse Adversarial Learning for FDIA Attack Sample Generation in Distributed Smart Grids.
- 2024 STAA-Net: A Sparse and Transferable Adversarial Attack for Speech Emotion Recognition
- 2024 Attack to Defend: Exploiting Adversarial Attacks for Detecting Poisoned Models
- 2024 AOHDL: Adversarial Optimized Hybrid Deep Learning Design for Preventing Attack in Radar Target Detection
- 2024 HOMOGRAPH: A Novel Textual Adversarial Attack Architecture to Unmask the Susceptibility of Linguistic Acceptability Classifiers
- 2024 Universal Adversarial Attack Against Speaker Recognition Models
- 2024 Unraveling Adversarial Examples Against Speaker Identification—Techniques for Attack Detection and Victim Model Classification

  - 2024 LPLA: The Adversarial Attack Against License Plate Recognition Systems
  - 2024 024 Ava: Inconspicuous Attribute Variation-Based Adversarial Attack Bypassing Deepfake Detection
  - 2024 AAMT: Adversarial Attack-Driven Mutual Teaching for Source-Free Domain-Adaptive Person Reidentification
  - 2024 Physical Adversarial Attack on Monocular Depth Estimation via Shape-Varying Patches
  - 2024 Boosting the Transferability of Adversarial Examples with Gradient-Aligned Ensemble Attack for Speaker Recognition
  - 2024 Cross-Point Adversarial Attack Based on Feature Neighborhood Disruption Against Segment Anything Model
  - 2024 2024 A Generative Adversarial Attack for Multilingual Text Classifiers
- Analyses of attacks (12):
  - 2024 Don't Listen to Me: Understanding and Exploring Jailbreak Prompts of Large Language Models
  - 2024 What Features in Prompts Jailbreak LLMs? Investigating the Mechanisms Behind Attacks
  - 2024 How Alignment and Jailbreak Work: Explain LLM Safety Through Intermediate Hidden States
  - 2024 Do LLMs Have Political Correctness? Analyzing Ethical Biases and Jailbreak Vulnerabilities in AI Systems
  - 2024 JailbreakLens: Visual Analysis of Jailbreak Attacks Against Large Language Models
  - 2024 Towards Understanding Jailbreak Attacks in LLMs: A Representation Space Analysis
  - 2024 The VLLM Safety Paradox: Dual Ease in Jailbreak Attack and Defense
  - 2024 Investigating Coverage Criteria in Large Language Models: An In-Depth Study Through Jailbreak Attacks
  - 2024 Learning to See but Forgetting to Follow: Visual Instruction Tuning Makes LLMs More Prone to Jailbreak Attacks
  - 2024 Understanding Jailbreak Success: A Study of Latent Space Dynamics in Large Language Models
  - 2024 Subtoxic Questions: Dive into Attitude Change of LLM's Response in Jailbreak Attempts
  - 2024 Implications of Minimum Description Length for Adversarial Attack in Natural Language Processing
- Domain (5)
  - 2024 Leveraging Large Language Models for Autonomous Red Teaming in Simulating Advanced Ransomware Attacks

  - 2024 The Crucial Role of Red Teaming: Strengthening Indonesia's Cyber Defenses Through Cybersecurity Drill Tests.
  - 2024 Cyber Security for AI Recommendations
  - 2024 Generative AI for Cyber Security: Analyzing the Potential of ChatGPT, DALL-E and Other Models for Enhancing the Security Space
  - 2024 Inside Cyber: How AI, 5G, IoT, and Quantum Computing Will Transform Privacy and Our Security
- Not red-teaming (4)
  - 2024 The Future of Artificial Intelligence Will Be "Next to Normal"—A Perspective on Future Directions and the Psychology of AI Safety Concerns
  - 2024 To Trust or Not to Trust: Evaluating the Reliability and Safety of AI Responses to Laryngeal Cancer Queries
  - 2024 Exploring Parent-Child Perceptions on Safety in Generative AI: Concerns, Mitigation Strategies, and Design Implications
  - 2024 AI Rights for Human Safety
- Duplicate (3)
  - 2024 Jailbreakv-28k: A Benchmark for Assessing the Robustness of Multimodal Large Language Models Against Jailbreak Attacks
  - 2024 JAILJUDGE: A Comprehensive Jailbreak Judge Benchmark with 2024 Multi-Agent Enhanced Explanation Evaluation Framework
  - 2024 AuditNet: A Conversational AI-Based Security Assistant

9.2.2.2 Papers Screened by Abstract

We then screened 62 papers based on their abstracts. We categorize the reasons for screening below and list the date and title of each paper. As with the titles, we only listed a single reason for exclusion, though some papers could have been screened for multiple reasons.

- Not presenting a red-teaming method (27)
  - 2024 Operationalizing a Threat Model for Red-Teaming Large Language Models (LLMs)
  - 2024 Harmbench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal
  - 2024 Automated Progressive Red Teaming
  - 2024 Exploring Straightforward Conversational Red-Teaming
  - 2024 Attack Atlas: A Practitioner's Perspective on Challenges and Pitfalls in Red Teaming GenAI
  - 2024 Scaling up Mischief: Red-Teaming AI and Distributing Governance
  - 2024 ALERT: A Comprehensive Benchmark for Assessing Large Language Models' Safety Through Red Teaming

  - 2024 Red Teaming Large Language Models in Medicine: Real-World Insights on Model Behavior
  - 2024 PyRIT: A Framework for Security Risk Identification and Red Teaming in Generative AI System
  - 2024 Red Teaming GPT-4V: Are GPT-4V Safe Against Uni/Multi-Modal Jailbreak Attacks?
  - 2024 Desert Camels and Oil Sheikhs: Arab-Centric Red Teaming of Frontier LLMs
  - 2024 Using Market Design to Improve Red Teaming of Generative AI Models
  - 2024 Exploring Vulnerabilities in LLMs: A Red Teaming Approach to Evaluate Social Bias
  - 2024 Characterizing and Evaluating the Reliability of LLMs Against Jailbreak Attacks
  - 2024 JailbreakEval: An Integrated Toolkit for Evaluating Jailbreak Attempts Against Large Language Models
  - 2024 Jailbreak Paradox: The Achilles' Heel of LLMs
  - 2024 JailBreakV: A Benchmark for Assessing the Robustness of MultiModal Large Language Models Against Jailbreak Attacks
  - 2024 Attackeval: How to Evaluate the Effectiveness of Jailbreak Attacking on Large Language Models
  - 2024 Bag of Tricks: Benchmarking of Jailbreak Attacks on LLMs
  - 2024 "Not Aligned" is Not" Malicious": Being Careful about Hallucinations of Large Language Models' Jailbreak
  - 2024 Unveiling the Safety of GPT-4o: An Empirical Study Using Jailbreak Attacks
  - 2024 Are Large Language Models Really Bias-Free? Jailbreak Prompts for Assessing Adversarial Robustness to Bias Elicitation
  - 2024 Retention Score: Quantifying Jailbreak Risks for Vision Language Models
  - 2024 Universal Jailbreak Backdoors in Large Language Model Alignment
  - 2024 Demonstration of an Adversarial Attack Against a Multimodal Vision Language Model for Pathology Imaging
  - 2024 AEGIS2. 0: A Diverse AI Safety Dataset and Risks Taxonomy for Alignment of LLM Guardrails
  - 2024 Introducing v0.5 of the AI Safety Benchmark from MLCommons
- Not generative AI (21)
  - 2024 Adversarial Evasion Attack Efficiency Against Large Language Models
  - 2024 LST2A: Lexical-Syntactic Targeted Adversarial Attack for Texts
  - 2024 An Adversarial Attack Approach on Financial LLMs Driven by Embedding-Similarity Optimization
  - 2024 OpenFact at CheckThat! 2024: Combining Multiple Attack Methods for Effective Adversarial Text Generation
  - 2024 Diffusion Model for Adversarial Attack Against NLP Models

  - 2024 Mutual-Modality Adversarial Attack with Semantic Perturbation
  - 2024 Model Mimic Attack: Knowledge Distillation for Provably Transferable Adversarial Examples
  - 2024 Diffusion Models for Imperceptible and Transferable Adversarial Attack
  - 2024 Adv-Diffusion: Imperceptible Adversarial Face Identity Attack via Latent Diffusion Model
  - 2024 Generative Adversarial Network Based Image-Scaling Attack and Defense Modeling
  - 2024 Content-Based Unrestricted Adversarial Attack
  - 2024 Transferable Structural Sparse Adversarial Attack Via Exact Group Sparsity Training
  - 2024 FACL-Attack: Frequency-Aware Contrastive Learning for Transferable Adversarial Attacks
  - 2024 OTAD: An Optimal Transport-Induced Robust Model for Agnostic Adversarial Attack
  - 2024 SCA: Highly Efficient Semantic-Consistent Unrestricted Adversarial Attack
  - 2024 A Reliable Approach for Generating Realistic Adversarial Attack via Trust Region-Based Optimization
  - 2024 Where and How to Attack? A Causality-Inspired Recipe for Generating Counterfactual Adversarial Examples
  - 2024 D-BADGE: Decision-Based Adversarial Batch Attack with Directional Gradient Estimation
  - 2024 Dynamic Programming-Based White Box Adversarial Attack for Deep Neural Networks
  - 2024 Improving Adversarial Transferability via Frequency-Guided Sample Relevance Attack
  - 2024 Downstream Transfer Attack: Adversarial Attacks on Downstream Models with Pre-Trained Vision Transformers
- Duplicate (5)
  - 2024 Red Teaming Language Models for Processing Contradictory Dialogues
  - 2024 Red Teaming Language-Conditioned Robot Models via Vision Language Models
  - 2024 Dart: Deep Adversarial Automated Red Teaming for LLM safety
  - 2024 Tastle: Distract Large Language Models for Automatic Jailbreak Attack
  - 2024 DAG-Jailbreak: Enhancing Black-Box Jailbreak Attacks and Defenses Through DAG Dependency Analysis
- Not red-teaming (4)
  - 2024 Red Teaming Language Conditioned Robotic Behavior
  - 2024 Trojan Activation Attack: Red-Teaming Large Language Models Using Steering Vectors for Safety-Alignment

  - 2024 Red Teaming Language Models for Contradictory Dialogues
  - 2024 CulturalTeaming: AI-Assisted Interactive Red-Teaming for Challenging LLMs' (Lack of) Multicultural Knowledge
- Defense (2)
  - 2024 JailbreakHunter: A Visual Analytics Approach for Jailbreak Prompts Discovery from Large-Scale Human-LLM Conversational Datasets
  - 2024 MoJE: Mixture of Jailbreak Experts, Naive Tabular Classifiers as Guard for Prompt Attacks
- Educational tutorial (1)
  - 2024 DARE to Diversify: DAta Driven and Diverse LLM REd Teaming
- Withdrawn paper (1)
  - 2024 Multi-Round Jailbreak Attack on Large Language Models

9.2.2.3 Papers Screened During Extraction

We categorize the reasons for screening below and list the date and title of each paper. Each paper has a single reason for exclusion.

- Not generative AI (3)
  - 2024 TF-Attack: Transferable and Fast Adversarial Attacks on Large Language Models
  - 2024 DA3: A Distribution-Aware Adversarial Attack Against Language Models
  - 2024 From Homeostasis to Resource Sharing: Biologically and Economically Compatible Multi-Objective Multi-Agent AI Safety Benchmarks
- Not a red-reaming method (1)
  - JAILJUDGE: A Comprehensive Jailbreak Judge Benchmark with Multi-Agent Enhanced Explanation Evaluation Framework
- No experimental results (1)
  - Jailbreak Large Language Models Through Logic Chain Injection

9.3 Appendix A.3: Extraction Templates

When extracting information from papers in the cyber red-teaming literature, our extraction template used the following questions:

- Review Method
- RT Definition
- Adversary emulation?
- Methods/Phases
- Method Analysis
- Domains
- Tools

- Tool Analysis/Categories
- Attacks
- Attack Analysis
- Vulnerabilities
- Vulnerability Analysis
- Manuals
- Security Mindset
- Engagement Advice
- Other Reviews
- Other Analysis
- Conclusions/Takeaways

When extracting information from papers in the generative AI red-teaming literature, our extraction template used the following questions:

- What was the working definition of RT?
- What were the criteria for successful RT?
- What was the RT methodology?
- What type of vulnerabilities did the paper address? What was the threat model?
- What was the system being evaluated?
- Who were the evaluators? What resources were available to them (e.g., time, compute, expertise, access)?
- What tools/methods did the evaluators use?
- What were the recommended mitigations produced by the activity?
- How were the outputs structured? How were they shared?
- What was the cost (monetary, time) of the activity?
- What risks were potentially missed?
- What other evaluations were performed on the system aside from red-teaming?
- What conclusions were made from the red-teaming activity (e.g., recommendations for future RT or issues with process)?
- Rough category?
  - Choose from: AI Search, AI Iteration, Manual, Optimized, Other

9.4 Appendix A.4: Final Paper List

9.4.1 Final Cyber Red-Teaming Paper List

We cite the final cyber papers here:

Aboelfotoh, S. F. & Hikal, N. A. A Review of Cyber-security Measuring and Assessment Methods for Modern Enterprises. *International Journal on Informatics Visualization*. Volume 3. Number 2. . 2019. https://joiv.org/index.php/joiv/article/view/239

Adam, H. M.; Widyawan; & Putra, G. D. A Review of Penetration Testing Frameworks, Tools, and Application Areas. Pages 319–324. In *2023 IEEE 7th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*. 2023. https://doi.org/10.1109/ICITISEE58992.2023.10404397

Al-Ahmad; A. S., Kahtan; H., Hujainah, F.; & Jalab, H. A. Systematic Literature Review on Penetration Testing for Mobile Cloud Computing Applications. *IEEE Access.* Volume 7. November 29, 2019. Pages 173524–173540. https://doi.org/10.1109/ACCESS.2019.2956770

Aldauiji, F.; Batarfi, O.; & Bayousef, M. Utilizing Cyber Threat Hunting Techniques to Find Ransomware Attacks: A Survey of the State of the Art. *IEEE Access*. Volume 10. June 8, 2022. Pages 61695–61706. https://doi.org/10.1109/ACCESS.2022.3181278

Altayaran, S. A. & Elmedany, W. Integrating Web Application Security Penetration Testing into the Software Development Life Cycle: A Systematic Literature Review. Pages 671-676. In *2021 International Conference on Data Analytics for Business and Industry (ICDABI)*. October 2021. https://doi.org/10.1109/ICDABI53623.2021.9655950

Altulaihan, E. A.; Alismail, A.; & Frikha, M. A Survey on Web Application Penetration Testing. *Electronics*. Volume 12. Issue 5. March 4, 2023. Page 1229. https://doi.org/10.3390/electronics12051229

Beaman, C. et al. Fuzzing Vulnerability Discovery Techniques: Survey, Challenges and Future Directions. *Computers & Security*. Volume 120. July 2022. Page 102813. https://doi.org/10.1016/j.cose.2022.102813

Briggs, J. et al. Survey of Layered Defense, Defense in Depth, and Testing of Network Security. In *Selected Readings in Cybersecurity.* Cambridge Scholars Publishing. Pages 105–119. 2018. ISBN: 1-5275-1641-5.

Chen, L. et al. A Survey on Threat Hunting: Approaches and Applications. Pages 340-344. In *2022 7th IEEE International Conference on Data Science in Cyberspace (DSC).* July 2022. DOI: 10.1109/DSC55868.2022.00053. https://ieeexplore.ieee.org/document/9900201

Cong, N. T. et al. An Overview of Static and Dynamic Analysis in Application Security Testing. *Journal of Military Science and Technology.* Volume 99. Number 99. November 2024. Pages 1–11. https://doi.org/10.54939/1859-1043.j.mst.99.2024.1-11

Eceiza, M. et al. Fuzzing the Internet of Things: A Review on the Techniques and Challenges for Efficient Vulnerability Discovery in Embedded Systems. *IEEE Internet of Things Journal*. Volume 8. Issue 13. July 2021. Pages 10390–10411. https://doi.org/10.1109/JIOT.2021.3056179

Fuchs, M. & Lemon, J. *SANS 2019 Threat Hunting Survey: The Differing Needs of New and Experienced Hunters*. SANS Institute. October 2019. https://www.sans.org/media/analyst-program/2019-threat-hunting-survey-differing-experienced-hunters-39220.pdf

Gbormittah, E. A Systematic Literature Review on Cyberwarfare and State-Sponsored Hacking: Penetration Testing Insights. *TechRxiv* [preprint]. August 2024. https://www.techrxiv.org/doi/full/10.36227/techrxiv.172373675.55159205

Ghaffarian, S. M. & Shahriari, H. R. Software Vulnerability Analysis and Discovery Using Machine-Learning and Data-Mining Techniques: A Survey. *ACM Computing Surveys*. Volume 50. Issue 4. August 2017. Pages 1-36. https://doi.org/10.1145/3092566

Großmann, J. & Seehusen, F. Combining Security Risk Assessment and Security Testing Based on Standards. In *Risk Assessment and Risk-Driven Testing.* F. Seehusen et al [editors]. Springer International Publishing. November 13, 2015. Pages 18–33. ISBN 978-3-319-26416-5. https://doi.org/10.1007/978-3-319-26416-5_2

Jeršič, N. et al. How to Approach Security Testing of Web 3.0 Solutions: A Review of Existing Knowledge. In *Proceedings of the Eleventh Workshop on Software Quality Analysis, Monitoring, Improvement, and Applications*. September 2024. https://ceur-ws.org/Vol-3845/paper20.pdf

Leszczyna, R. Review of Cybersecurity Assessment Methods: Applicability Perspective. *Computers & Security.* Volume 108. September 2021. Page 102376. https://doi.org/10.1016/j.cose.2021.102376

Li, J. Vulnerabilities Mapping Based on OWASP-SANS: A Survey for Static Application Security Testing (SAST). *Annals of Emerging Technologies in Computing.* Volume 4. Issue 3. October 11, 2024. Pages 1–8. https://doi.org/10.33166/AETiC.2020.03.001

Liu, B. et al. Software Vulnerability Discovery Techniques: A Survey. *2012 Fourth International Conference on Multimedia Information Networking and Security.* November 2012. Pages 152–156. https://doi.org/10.1109/MINES.2012.202

Mahboubi, A. et al. Evolving Techniques in Cyber Threat Hunting: A Systematic Review. *Journal of Network and Computer Applications*. Volume 232. December 2024. Page 104004. https://doi.org/10.1016/j.jnca.2024.104004

Marchetto, A. A Rapid Review on Fuzz Security Testing for Software Protocol Implementations. Pages 3–20. In *Testing Software and Systems.* September 2023. DOI: 10.1007/978-3-031-43240-8_1. https://doi.org/10.1007/978-3-031-43240-8_1

Modesti, P. et al. Bridging the Gap: A Survey and Classification of Research-Informed Ethical Hacking Tools. *Journal of Cybersecurity and Privacy.* Volume 4. Issue 3. Article 3. July 2024. Pages 410–448. https://doi.org/10.3390/jcp4030021

Nour, B.; Pourzandi, M.; & Debbabi, M. A Survey on Threat Hunting in Enterprise Networks. *IEEE Communications Surveys & Tutorials.* Volume 25. Issue 4. August 2023. Pages 2299–2324. https://doi.org/10.1109/COMST.2023.3299519

Nutalapati, V. A Comprehensive Review of Mobile App Security Testing Tools and Techniques. *International Research Journal of Engineering & Applied Sciences.* Volume 8. Issue 1. January–March 2020. Pages 10–15. https://www.irjeas.org/wp-content/uploads/admin/volume8/V8I1/IRJEAS04V8I101200320000006.pdf

Pargaonkar, S. Advancements in Security Testing: A Comprehensive Review of Methodologies and Emerging Trends in Software Quality Engineering. *International Journal of Science and Research (IJSR).* Volume 12. Issue 9. September 2023. Pages 61–66. https://doi.org/10.21275/SR23829090815

Parveen, M. & Shaik, M. A. Review on Penetration Testing Techniques in Cyber Security. Pages 1265–1270. In *2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS).* August 2023. https://doi.org/10.1109/ICAISS58487.2023.10250659

Pierce, J. D. et al. In Pursuit of a Standard Penetration Testing Methodology. *Journal of Information Warfare*. Volume 4. Issue 3. 2005. Pages 26–39. https://www.jstor.org/stable/26504027

Pillutla, H. & Arjunan, A. A Survey of Security Concerns, Mechanisms and Testing in Cloud Environment. Pages 1519–1524. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. March 2018. https://doi.org/10.1109/ICECA.2018.8474855

Pozzobon, E. et al. A Survey on Media Access Solutions for CAN Penetration Testing. In *ACM Computer Science in Cars Symposium (CSCS) 2018.* https://www.researchgate.net/publication/328687253_A_Survey_on_Media_Access_Solutions_for_CAN_Penetration_Testing

Raju, A. D. et al. A Survey on Cross-Architectural IoT Malware Threat Hunting. *IEEE Access.* Volume 9. June 22, 2021. Pages 91686–91709. https://doi.org/10.1109/ACCESS.2021.3091427

Ravindran, U. & Potukuchi, R. V. A Review on Web Application Vulnerability Assessment and Penetration Testing. *Review of Computer Engineering Studies.* Volume 9. Issue 1. March 31, 2022. Pages 1–22. https://doi.org/10.18280/rces.090101

Shah, S. & Mehtre, B. M. An Overview of Vulnerability Assessment and Penetration Testing Techniques. *Journal of Computer Virology and Hacking Techniques.* Volume 11. Issue 1. November 28, 2014. Pages 27–49. https://doi.org/10.1007/s11416-014-0231-x

Shivayogimath, C. N. An Overview of Network Penetration Testing. *International Journal of Research in Engineering and Technology.* Volume 3. Issue 07. July 2014. Pages 408–413. https://ijret.org/volumes/2014v03/i07/IJRET20140307070.pdf

Teichmann, F. M. & Boticiu, S. R. An Overview of the Benefits, Challenges, and Legal Aspects of Penetration Testing and Red Teaming. *International Cybersecurity Law Review.* Volume 4. Issue 4. September 4, 2023. Pages 387–397. https://doi.org/10.1365/s43439-023-00100-2

Tigner, M. et al. Analysis of Kali Linux Penetration Tools: A Survey of Hacking Tools. Pages 1–6. In *2021 International Conference on Electrical, Computer and Energy Technologies (ICECET).* December 2021. https://doi.org/10.1109/ICECET52533.2021.9698572

Valea, E. et al. A Survey on Security Threats and Countermeasures in IEEE Test Standards. *IEEE Design & Test.* Volume 36. Issue 3. June 2019. Pages 95–116. https://doi.org/10.1109/MDAT.2019.2899064

Vasenius, P. Best Practices in Cloud-Based Penetration Testing. *Master of Science in Technology Thesis, University of Turku*. October 2022. https://www.utupub.fi/bitstream/handle/10024/173476/Vasenius_Petrus_opinnayte.pdf

Wang, W. Survey of Software Vulnerability Discovery Technology. Pages 9–13. In *Proceedings of the 2017 7th International Conference on Social Network, Communication and Education*. July 2017. DOI: 10.2991/snce-17.2017.3. https://www.atlantis-press.com/proceedings/snce-17/25882970

Ximbo, B. et al. A Survey on IoT Vulnerability Discovery. Pages 267–282. In *Network and System Security.* December 7, 2022. https://doi.org/10.1007/978-3-031-23020-2_15

Yaacoub, J.-P. A. et al. A Survey on Ethical Hacking: Issues and Challenges. *arXiv* [preprint]. March 28, 2021. https://doi.org/10.48550/arXiv.2103.15072

Yu, M. et al. A Survey of Security Vulnerability Analysis, Discovery, Detection, and Mitigation on IoT Devices. *Future Internet.* Volume 12. Issue 2. February 2020. Page 27. https://doi.org/10.3390/fi12020027

Zhu, N.-F. et al. Study on the Standards and Procedures of the Penetration Testing. Pages 87–93. In *International Conference on Computer Science and Network Security (CSNS 2014).* March 2014. https://ci2s-enterprise.com.ar/2014/03/03/2014-international-conference-on-computer-science-and-network-security-2/

9.4.2 Final Generative AI Red-Teaming Paper List

We cite the final list of generative AI red-teaming papers here:

Amayuelas, A. et al. MultiAgent Collaboration Attack: Investigating Adversarial Attacks in Large Language Model Collaborations via Debate. Pages 6929–6948. In *Findings of the Association for Computational Linguistics: EMNLP*. November 2024. DOI: 10.18653/v1/2024.findings-emnlp.407. https://aclanthology.org/2024.findings-emnlp.407/

Bowen, D. et al. Data Poisoning in LLMs: Jailbreak-Tuning and Scaling Laws. *arXiv* [preprint]. December 2024. https://doi.org/10.48550/arXiv.2408.02946

Chang, Z. et al. Play Guessing Game with LLM: Indirect Jailbreak Attack with Implicit Clues. *arXiv* [preprint]. https://doi.org/10.48550/arXiv.2402.09091

Chen, Z. et al. AgentPoison: Red-teaming LLM Agents via Poisoning Memory or Knowledge Bases. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*. September 2024. https://neurips.cc/virtual/2024/poster/94715

Dang, P. et al. DiffZOO: A Purely Query-Based Black-Box Attack for Red-teaming Text-to-Image Generative Model via Zeroth Order Optimization. *arXiv* [preprint]. August 2024. https://doi.org/10.48550/arXiv.2408.11071

Deng, D. et al. AdversaFlow: Visual Red Teaming for Large Language Models with Multi-Level Adversarial Flow. *IEEE Transactions on Visualization and Computer Graphics*. Volume 31. Issue 1. September 2024. Pages 492–502. https://doi.org/10.1109/TVCG.2024.3456150

Deng, G. et al. Pandora: Jailbreak GPTs by Retrieval Augmented Generation Poisoning. *arXiv* (preprint). February 13, 2024. https://doi.org/10.48550/arXiv.2402.08416

Dominique, B. et al. Prompt Templates: A Methodology for Improving Manual Red Teaming Performance. *CHI 2024*. May 2024. https://research.ibm.com/publications/prompt-templates-a-methodology-for-improving-manual-red-teaming-performance

Dong, Y. et al. Harnessing Task Overload for Scalable Jailbreak Attacks on Large Language Models. *arXiv* [preprint]. October 2024. https://doi.org/10.48550/arXiv.2410.04190

Doumbouya, M. K. B. et al. h4rm3l: A Dynamic Benchmark of Composable Jailbreak Attacks for LLM Safety Assessment. *arXiv* [preprint]. September 2024. https://doi.org/10.48550/arXiv.2408.04811

Gibbs, T. et al. Emerging Vulnerabilities in Frontier Models: Multi-Turn Jailbreak Attacks. *arXiv* [preprint]. August 29, 2024. https://doi.org/10.48550/arXiv.2409.00137

Gu, X. et al. Agent Smith: A Single Image Can Jailbreak One Million Multimodal LLM Agents Exponentially Fast. Pages 16647–16672. In *ICML'24: Proceedings of the 41st International Conference on Machine Learning*. July 2024. https://dl.acm.org/doi/10.5555/3692070.3692731

Han, V. T. Y.; Bhardwaj, R.; & Poria, S. Ruby Teaming: Improving Quality Diversity Search with Memory for Automated Red Teaming. *arXiv* [preprint]. June 2024. https://doi.org/10.48550/arXiv.2406.11654

Hardy, A. F. et al. ASTPrompter: Weakly Supervised Automated Language Model Red-Teaming to Identify Low-Perplexity Toxic Prompts. *arXiv* [preprint]. July 2024. https://doi.org/10.48550/arXiv.2407.09447

Hong, Z.-W. et al. Curiosity-Driven Red-Teaming for Large Language Models. *arXiv* [preprint]. February 2024. https://doi.org/10.48550/arXiv.2402.19464

Hu, K. et al. Efficient LLM Jailbreak via Adaptive Dense-to-Sparse Constrained Optimization. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*. December 2024. https://papers.nips.cc/paper_files/paper/2024/hash/29571f8fda54fe93631c41aad4215abc-Abstract-Conference.html

Huang, X. et al. Medical MLLM is Vulnerable: Cross-Modality Jailbreak and Mismatched Attacks on Medical Multimodal Large Language Models. Pages 3797-3805. In *Proceedings of the AAAI Conference on Artificial Intelligence*. April 2025. DOI: 10.1609/aaai.v39i4.32396. https://ojs.aaai.org/index.php/AAAI/article/view/32396

Huang, Y. et al. Perception-Guided Jailbreak against Text-to-Image Models. Pages 26238-26247. In *Proceedings of the AAAI Conference on Artificial Intelligence*. April 2025. DOI: 10.1609/aaai.v39i25.34821. https://ojs.aaai.org/index.php/AAAI/article/view/34821

Jiang, F. et al. ArtPrompt: ASCII Art-Based Jailbreak Attacks Against Aligned LLMs. Pages 15157–15173 In *Proceedings of the 62nd Annual Meeting of the Association for the Computational Linguistics (Volume 1: Long Papers)*. August 2024. https://aclanthology.org/2024.acl-long.809/

Johnson, Z. D. *Generation, Detection, and Evaluation of Role-Play Based Jailbreak Attacks in Large Language Models* [thesis]. Massachusetts Institute of Technology (MIT). 2024. https://dspace.mit.edu/handle/1721.1/156989

Kumar, A. et al. SAGE-RT: Synthetic Alignment Data Generation for Safety Evaluation and Red Teaming. *arXiv* [preprint]. August 14, 2024. https://doi.org/10.48550/arXiv.2408.11851

Kumar, V.; Liao, Z.; Jones, J.; & Sun, H. AmpleGCG-Plus: A Strong Generative Model of Adversarial Suffixes to Jailbreak LLMs with Higher Success Rates in Fewer Attempts. *arXiv* [preprint]. October 29, 2024. https://doi.org/10.48550/arXiv.2410.22143

Lee, S. et al. Learning Diverse Attacks on Large Language Models for Robust Red-Teaming and Safety Tuning. *arXiv* [preprint]. May 28, 2024. https://doi.org/10.48550/arXiv.2405.18540

Li, B. et al. StructuralSleight: Automated Jailbreak Attacks on Large Language Models Utilizing Uncommon Text-Organization Structures. *arXiv* [preprint]. June 13, 2024. https://doi.org/10.48550/arXiv.2406.08754

Li, G. et al. ART: Automatic Red-Teaming for Text-to-Image Models to Protect Benign Users. *arXiv* [preprint]. May 24, 2024. https://doi.org/10.48550/arXiv.2405.19360

Li, J. et al. FMM-Attack: A Flow-Based Multi-Modal Adversarial Attack on Video-Based LLMs. *arXiv* [preprint]. March 21, 2024. https://doi.org/10.48550/arXiv.2403.13507

Li, J. et al. A Cross-Language Investigation into Jailbreak Attacks in Large Language Models. *arXiv* [preprint]. January 30, 2024. https://doi.org/10.48550/arXiv.2401.16765

Li, Q. et al. Deciphering the Chaos: Enhancing Jailbreak Attacks via Adversarial Prompt Translation. *arXiv* [preprint]. October 15, 2024. https://doi.org/10.48550/arXiv.2410.11317

Li, R. et al. Be a Multitude to Itself: A Prompt Evolution Framework for Red Teaming. *Findings of the Association for Computational Linguistics: EMNLP 2024*. November 2024. Pages 3287–3301. https://doi.org/10.18653/v1/2024.findings-emnlp.188

Li, X. et al. Faster-GCG: Efficient Discrete Optimization Jailbreak Attacks Against Aligned Large Language Models. *arXiv* [preprint]. October 20, 2024. https://doi.org/10.48550/arXiv.2410.15362

Li, X. et al. Semantic Mirror Jailbreak: Genetic Algorithm Based Jailbreak Prompts Against Open-Source LLMs. *arXiv* [preprint]. February 27, 2024. https://doi.org/10.48550/arXiv.2402.14872

Li, Y. et al. Lockpicking LLMs: A Logit-Based Jailbreak Using Token-Level Manipulation. *arXiv* [preprint]. June 19, 2024. https://doi.org/10.48550/arXiv.2405.13068

Lin, S. et al. LLMs Can Be Dangerous Reasoners: Analyzing-Based Jailbreak Attack on Large Language Models. *arXiv* [preprint]. July 23, 2024. https://doi.org/10.48550/arXiv.2407.16205

Lin, Z. et al. PathSeeker: Exploring LLM Security Vulnerabilities with a Reinforcement Learning-Based Jailbreak Approach. *arXiv* [preprint]. October 3, 2024. https://doi.org/10.48550/arXiv.2409.14177

Liu, H. et al. Boosting Jailbreak Transferability for Large Language Models. *arXiv* [pre-print]. November 3, 2024. https://doi.org/10.48550/arXiv.2410.15645

Liu, X. et al. AutoDAN-Turbo: A Lifelong Agent for Strategy Self-Exploration to Jailbreak LLMs. arXiv [preprint]. October 3, 2024. https://doi.org/10.48550/arXiv.2410.05295

Liu, Y. et al. Arondight: Red Teaming Large Vision Language Models with Auto-Generated Multi-Modal Jailbreak Prompts. *arXiv* [preprint]. July 21, 2024. https://doi.org/10.48550/arXiv.2407.15050

Liu, Y. et al. FlipAttack: Jailbreak LLMs via Flipping. *arXiv* [preprint]. October 2, 2024. https://doi.org/10.48550/arXiv.2410.02832

Lu, L. et al. AutoJailbreak: Exploring Jailbreak Attacks and Defenses Through a Dependency Lens. *arXiv* [preprint]. June 6, 2024. https://doi.org/10.48550/arXiv.2406.03805

Luo, Y. et al. Jailbreak Instruction-Tuned LLMs via End-of-Sentence MLP Re-weighting. *arXiv* [preprint]. October 14, 2024. https://doi.org/10.48550/arXiv.2410.10150

Lv, L. et al. AdaPPA: Adaptive Position Pre-Fill Jailbreak Attack Approach Targeting LLMs. *arXiv* [preprint]. September 11, 2024. https://doi.org/10.48550/arXiv.2409.07503

Men, T. et al. A Troublemaker with Contagious Jailbreak Makes Chaos in Honest Towns. *arXiv* [preprint]. October 21, 2024. https://doi.org/10.48550/arXiv.2410.16155

Miao, H. et al. Autonomous LLM-Enhanced Adversarial Attack for Text-to-Motion. *arXiv* [pre-print]. August 1, 2024. https://doi.org/10.48550/arXiv.2408.00352

Mu, H. et al. Stealthy Jailbreak Attacks on Large Language Models via Benign Data Mirroring. *arXiv* [preprint]. October 28, 2024. https://doi.org/10.48550/arXiv.2410.21083

Narayan, U. et al. Code-of-Thought Prompting: Probing AI Safety with Code. *Open-Review* [pre-print]. September 27, 2024. https://openreview.net/forum?id=lUyYX9VFgA

Pala, T. D. et al. Ferret: Faster and Effective Automated Red Teaming with Reward-Based Scoring Technique. *arXiv* [preprint]. August 20, 2024. https://doi.org/10.48550/arXiv.2408.10701

Pavlova, M. et al. Automated Red Teaming with GOAT: The Generative Offensive Agent Tester. *arXiv* [preprint]. October 2, 2024. https://doi.org/10.48550/arXiv.2410.01606

Pu, R. et al. BaitAttack: Alleviating Intention Shift in Jailbreak Attacks via Adaptive Bait Crafting. Pages 15654–15668. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. November 2024. https://doi.org/10.18653/v1/2024.emnlp-main.877

Qi, X. et al. Visual Adversarial Examples Jailbreak Aligned Large Language Models. Pages 21527–21536. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 2024. https://doi.org/10.1609/aaai.v38i19.30150

Raina, V. et al. Muting Whisper: A Universal Acoustic Adversarial Attack on Speech Foundation Models. Pages 7549–7565. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024. https://doi.org/10.18653/v1/2024.emnlp-main.430

Rando, J. et al. Gradient-Based Jailbreak Images for Multimodal Fusion Models. *arXiv* [preprint]. October 23, 2024. https://doi.org/10.48550/arXiv.2410.03489

Ren, Q. et al. Derail Yourself: Multi-Turn LLM Jailbreak Attack Through Self-Discovered Clues. *arXiv* [preprint]. October 14, 2024. https://arxiv.org/pdf/2410.10700v1

Rengarajan, D. et al. Imitation Guided Automated Red Teaming. *OpenReview* [preprint]. October 12, 2024. https://openreview.net/forum?id=fkOZjYwm2R

Russinovich, M.; Salem, A.; & Eldan, R. Great, Now Write an Article About That: The Crescendo Multi-Turn LLM Jailbreak Attack. *arXiv* [preprint]. April 2, 2024. http://arxiv.org/abs/2404.01833

Sagar, S. et al. LLM-Assisted Red Teaming of Diffusion Models Through "Failures Are Fated, But Can Be Faded." *arXiv* [preprint]. October 22, 2024. https://doi.org/10.48550/arXiv.2410.16738

Saiem, B. A. et al. SequentialBreak: Large Language Models Can Be Fooled by Embedding Jailbreak Prompts into Sequential Prompt Chains. *arXiv* [preprint]. November 10, 2024. https://doi.org/10.48550/arXiv.2411.06426

Shen, X. et al. Voice Jailbreak Attacks Against GPT-4o. *arXiv* [preprint]. May 29, 2024. https://doi.org/10.48550/arXiv.2405.19103

Sun, X. et al. Multi-Turn Context Jailbreak Attack on Large Language Models from First Principles. *arXiv* [preprint]. August 8, 2024. https://doi.org/10.48550/arXiv.2408.04686

Takemoto, K. All in How You Ask for It: Simple Black-Box Method for Jailbreak Attacks. *Applied Sciences*. Volume 14. Issue 9. April 23, 2024. Page 3558. https://doi.org/10.3390/app14093558

Tang, Y. et al. RoleBreak: Character Hallucination as a Jailbreak Attack in Role-Playing Systems. Pages 7386–7402. In *Proceedings of the 31st International Conference on Computation Linguistics*. January 2025. https://aclanthology.org/2025.coling-main.494.pdf

Tu, S. et al. Knowledge-to-Jailbreak: One Knowledge Point Worth One Attack. *arXiv* [preprint]. June 17, 2024. https://doi.org/10.48550/arXiv.2406.11682

Wang, F. et al. MRJ-Agent: An Effective Jailbreak Agent for Multi-Round Dialogue. *arXiv* [preprint]. November 6, 2024. https://doi.org/10.48550/arXiv.2411.03814

Wang, H. et al. ASETF: A Novel Method for Jailbreak Attack on LLMs through Translate Suffix Embeddings. Pages 2697–2711. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. November 2024. https://doi.org/10.18653/v1/2024.emnlp-main.157

Wang, W. et al. Chain-of-Jailbreak Attack for Image Generation Models via Editing Step by Step. *arXiv* [preprint]. October 4, 2024. https://doi.org/10.48550/arXiv.2410.03869

Wang, Y. et al. Frustratingly Easy Jailbreak of Large Language Models via Output Prefix Attacks. *Research Square* [preprint]. May 9, 2024. https://doi.org/10.21203/rs.3.rs-4385503/v1

Wang, Z. et al. Functional Homotopy: Smoothing Discrete Optimization via Continuous Parameters for LLM Jailbreak Attacks. *arXiv* [preprint]. October 5, 2024. https://doi.org/10.48550/arXiv.2410.04234

Wang, Z. et al. Poisoned LangChain: Jailbreak LLMs by LangChain. *arXiv* [preprint]. June 26, 2024. https://doi.org/10.48550/arXiv.2406.18122

Wang, Z. et al. Hide Your Malicious Goal into Benign Narratives: Jailbreak Large Language Models Through Carrier Articles. *arXiv* [preprint]. August 20, 2024. https://doi.org/10.48550/arXiv.2408.11182

Weidinger, L. et al. STAR: SocioTechnical Approach to Red Teaming Language Models. *arXiv* [preprint]. October 23, 2024. https://doi.org/10.48550/arXiv.2406.11757

Wong, A. et al. SMILES-Prompting: A Novel Approach to LLM Jailbreak Attacks in Chemical Synthesis. *arXiv* [preprint]. October 21, 2024. https://doi.org/10.48550/arXiv.2410.15641

Wu, T. et al. You Know What I'm Saying: Jailbreak Attack via Implicit Reference. *arXiv* [preprint]. October 8, 2024. https://doi.org/10.48550/arXiv.2410.03857

Wu, Y. et al. Can Large Language Models Automatically Jailbreak GPT-4V? *arXiv* [pre-print]. August 23, 2024. https://doi.org/10.48550/arXiv.2407.16686

Xiao, Z. et al. Distract Large Language Models for Automatic Jailbreak Attack. Pages 16230–16244. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. November 2024. DOI: 10.18653/v1/2024.emnlp-main.908. https://aclanthology.org/2024.emnlp-main.908/

Xu, H. et al. RedAgent: Red Teaming Large Language Models with Context-Aware Autonomous Language Agent. *arXiv* [preprint]. July 23, 2024. https://doi.org/10.48550/arXiv.2407.16667

Xu, X. et al. Watch Your Words: Successfully Jailbreak LLM by Mitigating the "Prompt Malice." Pages 295–309. In *Web and Big Data*. August 28, 2024. https://doi.org/10.1007/978-981-97-7232-2_20

Yang, H. et al. Audio Is the Achilles' Heel: Red Teaming Audio Large Multimodal Models. *arXiv* [preprint]. October 31, 2024. https://doi.org/10.48550/arXiv.2410.23861

Yang, H. et al. Jigsaw Puzzles: Splitting Harmful Questions to Jailbreak Large Language Models. *arXiv* [preprint]. October 15, 2024. https://doi.org/10.48550/arXiv.2410.11459

Yang, Y. & Fu, H. Transferable Ensemble Black-Box Jailbreak Attacks on Large Language Models. *arXiv* [preprint]. November 27, 2024. https://doi.org/10.48550/arXiv.2410.23558

Yang, Y. et al. SoP: Unlock the Power of Social Facilitation for Automatic Jailbreak Attack. *arXiv* [preprint]. July 2, 2024. https://openreview.net/forum?id=tD1atZW1vv

Yao, D. et al. FuzzLLM: A Novel and Universal Fuzzing Framework for Proactively Discovering Jailbreak Vulnerabilities in Large Language Models. Pages 4485–4489. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. April 2024. https://doi.org/10.1109/ICASSP48485.2024.10448041

Ying, Z. et al. Jailbreak Vision Language Models via Bi-Modal Adversarial Prompt. *arXiv* [preprint]. July 1, 2024. https://doi.org/10.48550/arXiv.2406.04031

Yoo, H. et al. Code-Switching Red-Teaming: LLM Evaluation for Safety and Multilingual Understanding. *arXiv* [preprint]. June 17, 2024. https://doi.org/10.48550/arXiv.2406.15481

Yu, H. et al. Step Vulnerability Guided Mean Fluctuation Adversarial Attack Against Conditional Diffusion Models. Pages 6791–6799. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Volume 38. Issue 7. March 24, 2024. https://doi.org/10.1609/aaai.v38i7.28503

Yu, J. et al. Enhancing Jailbreak Attack Against Large Language Models Through Silent Tokens. *arXiv* [preprint]. May 31, 2024. https://arxiv.org/abs/2405.20653

Zeng, Y. et al. AdvI2I: Adversarial Image Attack on Image-to-Image Diffusion Models. *arXiv* [preprint]. November 1, 2024. https://doi.org/10.48550/arXiv.2410.21471

Zeng, Y. et al. How Johnny Can Persuade LLMs to Jailbreak Them: Rethinking Persuasion to Challenge AI Safety by Humanizing LLMs. *arXiv* [preprint]. August 2024. https://doi.org/10.48550/arXiv.2401.06373

Zhang, G. et al. Unveiling Vulnerabilities in Large Vision-Language Models: The SAVJ Jailbreak Approach. Pages 417–434. In *Artificial Neural Networks and Machine Learning—ICANN 2024*. September 17, 2024. https://doi.org/10.1007/978-3-031-72344-5_28

Zhang, H. et al. Jailbreak Open-Sourced Large Language Models via Enforced Decoding. Pages 5475–5493. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. August 2024. https://doi.org/10.18653/v1/2024.acl-long.299

Zhang, J. et al. EnJa: Ensemble Jailbreak on Large Language Models. *arXiv* [preprint]. August 7, 2024. https://doi.org/10.48550/arXiv.2408.03603

Zhang, J. et al. Holistic Automated Red Teaming for Large Language Models Through Top-Down Test Case Generation and Multi-Turn Interaction. *arXiv* [preprint]. September 25, 2024. https://doi.org/10.48550/arXiv.2409.16783

Zhang, T. et al. WordGame: Efficient & Effective LLM Jailbreak via Simultaneous Obfuscation in Query and Response. *arXiv* [preprint]. May 22, 2024. https://doi.org/10.48550/arXiv.2405.14023

Zhao, A. et al. DiveR-CT: Diversity-Enhanced Red Teaming Large Language Model Assistants with Relaxing Constraints. *arXiv* [preprint]. December 20, 2024. https://arxiv.org/abs/2405.19026

Zhao, J. et al. SQL Injection Jailbreak: A Structural Disaster of Large Language Models. *arXiv* [preprint]. November 3, 2024. https://doi.org/10.48550/arXiv.2411.01565

Zhao, W. et al. Diversity Helps Jailbreak Large Language Models. Pages 4647–4680. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*. April 2025. https://aclanthology.org/2025.naacl-long.238/

Zhou, X. et al. HAICOSYSTEM: An Ecosystem for Sandboxing Safety Risks in Human-AI Interactions. *arXiv* [preprint]. October 21, 2024. https://doi.org/10.48550/arXiv.2409.16427

Zhou, Y. et al. Virtual Context: Enhancing Jailbreak Attacks with Special Token Injection. Pages 11843–11857. In *Findings of the Association for Computational Linguistics*. August 2024. https://aclanthology.org/2024.findings-emnlp.692/

Zhou, Y. et al. Large Language Models Are Involuntary Truth-Tellers: Exploiting Fallacy Failure for Jailbreak Attacks. *arXiv* [preprint]. July 1, 2024. https://doi.org/10.48550/arXiv.2407.00869

10 Appendix B: Operational Stages of Cyber Red-Teaming

In our review of the cyber red-teaming literature, we identified the most frequently mentioned stages of the cyber red-teaming process and ordered them chronologically. This appendix provides more detail on the contents of each stage.

- **Pre-Engagement**: This stage groups together a variety of processes needed to lay the groundwork for a cyber red-teaming engagement. These include making contact with the host organization, officially arranging the red-teaming engagement, defining rules of engagement, and up-front legal matters, such as liability waivers and non-disclosure agreements.
- **Threat Modeling**: This stage is also part of laying the groundwork for a cyber red-teaming engagement. During this stage the red team and host cooperate to decide on the attacker profiles and loss events to be simulated during the red-teaming engagement. Attacker profiles might specify particular tools, techniques, resources, or target assets. Loss events are decided by the host, depending on their circumstances. Another important part of this stage is the use of "white cards"—agreements that allow red teams to bypass obstacles to cheaply emulate expensive adversary capabilities.
- **Reconnaissance**: During this stage the red team gathers information about the target systems to identify lines of attack. Reconnaissance can be conducted via passive means, which do not make contact with the target system, or active means, which do.
- **Scanning**: This stage is a subset of reconnaissance, which is particularly popular. Scanning is a form of active reconnaissance, usually using a tool that automates the collection and analysis of large quantities of data. Common types of scanning include network scanning, code scanning, and web application scanning.
- **Vulnerability Analysis**: During this stage the red team uses the information gathered during reconnaissance to identify potential vulnerabilities in the target system. This can also include assessing vulnerability usefulness or even starting to assemble a chain of exploits, but we sometimes see those activities split out into separate stages.
- **Initial Access**: During this stage the red team exploits one or more vulnerabilities to gain an initial foothold in the target system. Accessing functionality, which is supposed to be publicly exposed without an exploit, is not typically considered initial access.
- **Maintaining Access**: Once access is gained, the red team must maintain access for as long as needed. This typically involves exploiting one or more vulnerabilities to preserve the foothold gained during initial access through routine interruptions such as system restarts, updates, or credential changes.
- **Exploitation**: This stage groups together all activities that involve exploiting vulnerabilities in the target system to achieve the goals of the red team. This includes initial access, maintaining access, and any further exploitation needed to achieve the red team objective.

- **Post-Exploitation**: This stage groups together a variety of actions performed after the red team has successfully compromised the system. These actions depend heavily on the rules of engagement. They might include communications with the host organization, demonstrations of compromise, or pursuit of further objectives.
- **Reporting**: During this stage the red team delivers the results of the engagement to the host organization. This often follows some procedure agreed upon during pre-engagement. The scope of reporting might include logging red team actions, summarizing logs, describing vulnerabilities, further analysis such as prioritizing vulnerabilities or identifying mitigations, or even continued contact with the host organization after the main engagement has ended.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
(Leave Blank)	July 2025	Final

4. TITLE AND SUBTITLE	5. FUNDING NUMBERS
What Can Generative AI Red-Teaming Learn from Cyber Red-Teaming?	FA8702-15-D-0002

6. AUTHOR(S)
Anusha Sinha
James Lucassen
Keltin Grimes
Michael Feffer
Ellie Soto
Hoda Heidari
Nathan VanHoudnos

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Software Engineering Institute Carnegie Mellon University Pittsburgh, PA 15213	CMU/SEI-2025-TR-006

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSORING/MONITORING AGENCY REPORT NUMBER
SEI Administrative Agent AFLCMC/AZS 5 Eglin Street Hanscom AFB, MA 01731-2100	n/a

11. SUPPLEMENTARY NOTES

12A DISTRIBUTION/AVAILABILITY STATEMENT	12B DISTRIBUTION CODE
Unclassified/Unlimited, DTIC, NTIS	A

13. ABSTRACT (MAXIMUM 200 WORDS)

Red-teaming, a security practice rooted in adversarial emulation, has been widely applied across various domains, including cybersecurity and artificial intelligence (AI). This paper investigates the applicability of established cyber red-teaming methodologies to the evaluation of generative AI systems, addressing the growing need for robust security assessments in AI-driven applications. Through a pair of systematic literature reviews, we synthesize existing generative AI red-teaming approaches and analyze their alignment with established practices in cyber red-teaming.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Generative AI red-teaming, Generative AI, cybersecurity	92

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

NSN 7540-01-280-5500 Standard Form 298 (Rev. 2-89) Prescribed by ANSI Std. Z39-18 298-102