International Journal of AI, BigData, Computational and Management Studies
Noble Scholar Research Group | Volume 6, Issue 1, PP 116-123, 2025
ISSN: 3050-9416 | https://doi.org/10.63282/3050-9416.IJAIBDCMS-V6I1P112

Original Article

Red Teaming AI Systems for Security Validation

Ankush Gupta
Senior Solution Architect.

Received On: 16/02/2025 Revised On: 07/03/2025 Accepted On: 13/03/2025 Published On: 20/03/2025

Abstract: *There is a growing trend in which artificial intelligence (AI) is embedded in safety-critical, regulatory-driven, or high impact domains such as healthcare, finance, public infrastructure, enterprise automation. The larger these systems become, however, the more apparent the risks such systems pose from adversarial exploitation, model manipulability, and the unsafe integration of tools. AI models are different from traditional software in that they exhibit probabilistic and context-dependency nature and hence exhibit unique vulnerabilities including prompt injection, jailbreaks, data poisoning, model extraction, and unsafe autonomous decision-making. These threats require special testing and assurance methodologies, beyond traditional penetration tests.*

Red teaming is becoming an important technique to evaluate AI systems under real world adversarial settings. It's the systematic stress-testing of models to identify holes before bad actors can exploit them. Nevertheless, contemporary red team approaches are fragmented with little shared taxonomies, common benchmarks or integration with governance. To fill in these gaps, this paper presents COMPASS-RT, a model-agnostic and deployment-agnostic framework for AI red teaming.

The framework is comprised of six pillars: (i) risk-based scoping, consistent with the NIST AI Risk Management Framework and ISO/IEC 42001; (ii) threat modelling using MITRE ATLAS and the OWASP Top-10 adapted for LLM applications; (iii) hybrid adversarial testing, combining human expertise with automated, LLM-driven attack generation; (iv) benchmark-based validation, using standardized corpora such as AdvBench, HarmBench, and JailbreakBench; (v) governance integration to ensure that findings map to risk registers, mitigation workflows, and regulatory compliance under regimes like the EU AI Act; and (vi) continuous validations, providing sustained measurement and regression testing across model updates.

This paper makes a threefold contribution: it unifies existing best practices from adversarial AI testing, operationalizes best practices in governance, and does so in a way which incorporates reporting templates that organizations can use for audit-ready assurance. By incorporating COMPASS-RT into enterprise security and compliance programs, enterprises can establish defensible processes for demonstrating the robustness of AI systems, eroding the efficacy of attacks and accelerating response. Risk Management: You cannot eliminate risk completely, but disciplined and repeatable red teaming greatly strengthens security posture and confidence in AI deployments.

Keywords: *AI Red Teaming; Adversarial Machine Learning; OWASP LLM Top-10; MITRE ATLAS; NIST AI RMF; ISO/IEC 42001; ISO/IEC 23894; EU AI Act; Jailbreak Detection; Prompt Injection; Automated Red Teaming; Security Validation; Continuous Monitoring; Governance Integration.*

1. Introduction

AI is a central part of digital transformation and empowers us to make complex decisions, develop predictive analyses, generate things and applications in a number of industries from healthcare and finance to manufacturing and critical infrastructure. With the growing prominence of large language models (LLMs), generative adversarial networks (GANs), and the deployment of autonomous agent systems in production pipelines, the attack surface for adversarial exploitation has significantly increased. Unlike deterministic software applications that are subject to some encoding-based deficiencies or misconfigurations, AI vulnerabilities are caused by several reasons including the probabilistic nature of model inference, reliance on training and retrieval data, and contextual understanding of user prompts and commands. This introduces novel failure modes, such as adversarial perturbation, prompt injection, data poisoning, model extraction, and risky tooling orchestration that cannot be effectively countered by traditional security testing procedures.

In response to such risks, red teaming is evolving as a proactive, adversary-based approach to stress-testing AI systems. Based on the military and cybersecurity techniques of red teaming, the approach simulates real-world attackers who have specific goals and resources, with the intent to expose system weaknesses before they can be targeted by adversaries. In the AI context, red teaming goes beyond functional correctness to evaluate ethical alignment, resistance to improper exercise and vulnerability to novel attack vectors. For instance, red teaming an LLM could include devising malicious prompts for the purpose of eliciting harmful directives, introducing tainted data into retrieval pipelines, or distorting contextual information so that policy violations realize. They reveal blind spots in system design, both technical flaws and governance vacuums.

2017
Adversarial Examples
2019
Data Poisoning
2022
Prompt Injection
2023
Jailbreak Attacks
2024
RAG/Memory Poisoning

Fig 1: Evolving AI Threat Landscape

A timeline of the evolution in AI threat views: adversarial examples (2017), data poisoning (2019), LLM prompt injection (2022), jailbreaks (2023), RAG/memory poisoning (2024).

And yet, for all its importance, AI red teaming is also a nascent field with disparate practices. However, industry disclosures, including OpenAI's GPT-4 and GPT-4o system cards, highlight cyclical internal and external red teaming, albeit approaches differ widely among vendors. Standards such as AdvBench, HarmBench and JailbreakBench have been proposed to systematize evaluations, but are adopted with varying willingness, and many entities do not have the capability to incorporate such practices into continuous assurance pipelines. There is also very little linkage between red teaming and enterprise governance frameworks, and red team reports are rarely related to risk registers, compliance evidence, or regulatory reporting.

This gap is being tackled by a number of policy and standards organizations. The NIST AI Risk Management Framework (AI RMF) focuses on the mapping, measurement, and management of AI risk throughout the AI lifecycle. In support of integrating red teaming into organizational systems is the set of complementary standards, for example, those of ISO/IEC 23894 (AI risk management) and ISO/IEC 42001 (AI management systems). On the other hand, the OWASP Top-10 for LLM Applications and MITRE ATLAS are structured taxonomies for threats affecting AI systems that may allow for a coverage-based evaluation of adversarial testing campaigns. Regulatory change, such as the EU AI Act (2024/1689), also increasingly requires pre-market testing, documentation, monitoring post-market of high-risk AI systems, thus increasing the importance of auditable red teaming practices.

In this paper, we present COMPASS-RT, an AI red teaming framework, which is model-agnostic and deployment-agnostic. By bringing technical testing and governance integration under the same umbrella, COMPASS-RT fills the voids in current approaches to adversarial testing, balancing such exercises with international expectations, regulatory expectations, and security-by-design considerations. It is a realization of a six-pillar methodology including scope definition, threat modelling, hybrid attack generation, benchmark-driven evaluation, governance mapping, and continuous validation. By incorporating COMPASS-RT into enterprise assurance programs, companies are able to shift red teaming from ad hoc exercises into repeatable, auditable controls that, taken together, will help to reduce AI security risk.

The rest of this paper is organized as follows. Section II provides literature review on the available frameworks, benchmarks and practices in industry. The COMPASS-RT method is described in Section III. Section IV also contains implementation models and reporting formats as examples. Results, limitations and implications for future defence-in-depth strategies are presented in Section V. We conclude in Section VI with some practical advice for organizations that are considering the implementation of AI red teaming as a foundational assurance function.

2. Literature Review

The literature in AI security validation has seen a significant growth in the last decade, corresponding to the twin concerns of adversarial robustness and responsible deployment. Red teaming has become in this context a pragmatic approach to narrow the gap between academic attacks and operational risk. This review provides an overview of prior work from standards and frameworks, threat modelling taxonomies, industry practices, academic advances in adversarial testing, and policy/regulatory developments, in terms of what works and what doesn't.

2.1. Standards and Governance Frameworks

The foundations of risk management place red teaming as a key control for AI assurance. The NIST AI RMF 1.0 defines four functions Govern, Map, Measure, and Manage—that cumulatively constitute a life-cycle approach to AI risk [1]. Red teaming has a direct mapping to the "Measure" and "Manage" functions, as adversarial stress-testing is needed to assess trustworthiness properties such as safety, robustness, and resilience. NIST subsequently issued the Generative AI Profile (AI 600-1) in 2024 [2], which extends the AI RMF to generative models, recommending pre-release safety testing and misuse resistance testing, once again making a case for red teaming.

Governance structures for AI testing have been formalized to international standards. It is suggested that structured appraisal methods are used to manage the AI risk as part of continuous monitoring [3]. ISO/IEC 42001:2023 the first AI management system standard applies a plan-do-check-act model that includes security testing at strategic level in organizational governance [4]. These models place red teaming as not something extra but as the necessary part of risk based conformance.

2.2. Threat Modelling and Taxonomies

“The Art of Structured Red Teaming” will describe how to create the basic building block for a tailored, systemized red team. MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) is a popularising extension of the well-known MITRE ATT&CK framework into domains specific to AI, such as the data poisoning, evasion, model theft, adversarial perturbation [7]. Meanwhile, OWASP Top-10 for LLM Applications highlights high risks that span across language models, such as LLM01 (Prompt Injection), LLM02 (Insecure Output Handling), LLM03 (Training Data Poisoning), and LLM06 (Sensitive Information Disclosure) [8]. ATLAS and OWASP provide the means for red teaming to realise measurable “coverage metrics”, ensuring test campaigns are systematically exercising known classes of weaknesses - rather than being run on the basis of intuition.

2.3. Industry Practices

Realistic disclosures from modelling agencies also provide real-world detail on how red teaming works. The GPT-4 System Card by OpenAI records rounds of iterative red teaming by internal researchers and external collaborators have conducted and how the results were leveraged to optimize accuracy of safety classification and tuning of the refusal behaviour [9]. The GPT-40 System Card generalizes from this model, to incorporate external multi-lingual red teaming across multiple modalities [10]. These cases illustrate how red teaming has expanded from adversarial testing irrespective of strategic considerations to deployment-context-aligned, holistic stress testing. Likewise, Google’s Secure AI Framework (SAIF) [6] and the UK NCSC/CISA Guidelines for Secure AI System Development [5] incorporate adversarial testing into secure development lifecycles, and stress the importance of adversarial testing not only in pre-release validation, but also for post-deployment monitoring.

2.4. Academic Developments in Adversarial Red Teaming

The academic community has helped to make red teaming both methodologically focused and automated. Perez et al. [11] envisioned “red teaming language models with language models”: using one LLM to produce adversarial cases for another, greatly expanding the coverage while decreasing dependency on expensive human experts. Anthropic research [12] showed human-in-the-loop red teaming at large scale, categorizing adversarial behaviours into risk levels (Risk Levels) and at safety fine-tuning time producing them (Risks Results).

Emulation/Comparison (apart from baseline models, datasets have common evaluation practices. AdvBench [16] and HarmBench [15] offer harmful-behaviour prompts and evaluation corpora and JailbreakBench [14] creates replicable settings to evaluate jailbreak robustness. SafetyBench [17]generalizes the evaluation to wider ethical and misuse cases. Formalizing work such as Liu et al. s prompt injection benchmarking framework [13], develop attack/defense taxonomies and evaluation metrics, and minimize the subjectivity in the scoring of adversarial success. For evolving systems of agents, AgentPoison [18] describes how adversarial inputs can be used to poison both RAG-based retrieval augmented generation (RAG) and long-term memory, broadening the scope of red teaming beyond conversational chat.

2.5. Policy and Regulatory Drivers

Governments and regulators are increasingly requiring systematic testing. The EU AI Act (2024/1689) tasks developers of risk-tiered AI systems with obligations that mandate pre-release testing, documentation of risk-management files and postmarked surveillance [19]. These requirements would turn red teaming into a compliance requirement for “high-risk” AI systems and the providers of those systems. Likewise, the ENISA Threat Landscape 2024 [20] adds that adversarial threats on AI are becoming more sophisticated and proposes continuous adversarial testing should be part of the cybersecurity strategy of Europe.

2.6. Synthesis

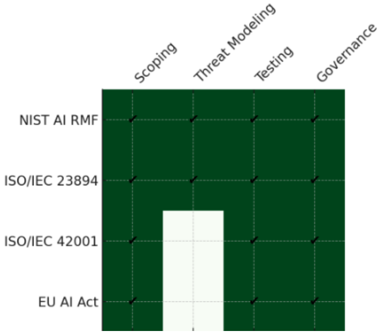

Fig 2: Mapping Red Teaming across Frameworks

Despite proceedings and benchmarks as well as case-studies the practice is challenged. A universal scoring system has not yet been established, test coverage is patchy across modality and language, and the translation of red teaming results into governance evidence, actionable to the business, is non-standardised. These are the gaps that motivate the present of a unified approach such COMPASS-RT that integrates technical adversarial testing with enterprise governance, regulatory compliance, and continuous monitoring.

A matrix mapping NIST AI RMF, ISO/IEC 23894, ISO/IEC 42001, and EU AI Act requirements to AI red teaming activities (scoping, threat modelling, testing, governance).

3. Methodology

The fact is that nobody has a clear idea how well AIs can be trusted and under which circumstances, so we introduce COMPASS-RT (Comprehensive Adversarial Stress-Testing for Red Teaming), a structured, repeatable framework for testing the safety of an AI system in the real world. The method is model and deployment agnostic, available for use in generative, discriminative, and agentic AI regardless of application domain. The complete purpose for it is to bring the practice of red teaming out of the experiment stage ad-hoc engagements and standardize the effort into a form of disciplined process that can be evaluated for best practice or regulatory compliance, while attempting to integrate that functionality with enterprise governance and security program requirements.

The methodology's initial step is scoping, in which organizations identify the mission-driven goals of the red team exercise. This includes defining the assets to protect (e.g., sensitive training data, proprietary model parameters, or integrated external tools), as well as potential attacker goals from data exfiltration or unauthorized tool invocation to reputational damage of producing harmful content. With the mapping feature of the NIST AI RMF, teams are advised to take into account the socio-technical environment of deployment, the importance of the use case, and the societal impacts of system misuse. crisp success criteria such as whether a prompt injection will cause unsafe tool behavior, or if sensitive information can be directly extracted under low-level probing are established in order to provide clear definitions of successes and failures during validation.

After scoping, the approach focuses on structured threat modelling based on the overlapping perspective between MITRE ATLAS and the OWASP Top-10 for LLM Applications. MITRE ATLAS lists tactics and techniques for the different ways in which adversaries can attack AI systems: model evasion, data poisoning, model extraction, inference manipulation, etc. Collectively, these taxonomies are used to build a large library of attacks that is customized to the system under analysis. Coverage can be quantified by making sure that the red team campaign exercises each applicable threat technique at least once, using various modalities, languages and user interaction contexts.

The test case generation problem is tackled using a hybrid model of human creative with automated adversarial augmentation. Human red teamers, skilled in security ethics or domain specific scenarios, supply the creativity needed for the new attack payloads. This is also complemented by automated generators, many of which are backed by language models, that mutate seed prompts, generate indirect injections, or simulate multi-turn adversarial conversations. We use benchmark datasets, including AdvBench, HarBench, SafetyBench, JailbreakBench, as seed corpora, and then rely on automatic paraphrasing and scenario mutation to expand them. For agentic architectures which have memory or retrieval-augmented generation, adversarial data poisoning is studied through a series of attacks, inspired by recent work e.g. AgentPoison that investigates the long-term effect of adversarial inputs on the persistence of a memory.

Running of such test-cases is based on an execution harness which makes the evaluation comparable among different approaches. Testing environments need to record entire transcripts, model output, and any tool calls executed by the system, such that each run can be reproduced and analysed. Defences such as output filters, policy constraints, or allow-list filtering can be turned on and off one by one to notice the difference in attack success rates. A staged pipeline is enacted, starting from internal testing in sandboxes, moving into the beta-phase testing with select customer exposure, and ending in external red-teaming activities with a wide variety of testers across languages and modality. This tiered approach allows risks to be highlighted gradually, diminishing the likelihood mistakes will have been made on critical issues by the time the deployments are complete.

COMPASS-RT Design Evaluation metric is a crucial component of COMPASS-RT. The most common is called the Attack Success Rate (ASR), which measures the fraction of adversarial queries that achieve a designated malicious goal. Auxiliary metrics are the metric of refusal robustness, to quantify the level of persistence of safe refusals under repeated adversarial probing and the metric of Défense efficacy, to quantify the reductions in ASR that follow mitigation. Furthermore, we evaluate degradation under composition either by chaining multiple adversarial inputs or by testing across complicated workflows such as tool invocation and retrieval augmentation. Code quality: Coverage-based metrics can be used to guarantee that all applicable techniques in a threat taxonomy are put to the test; governance-based metrics measure the speed with which findings are handled, and replaced with controls.

Lastly, the approach highlights governance integration so the findings of the red team work are not just technical documents but are converted into risk treatments actions. Organisations can embed findings in risk registers, allocate control ownership, and plan remediation cycles utilising the ISO/IEC 42001 management system framework and ISO/IEC

23894 criteria for AI risk management. For all high-risk applications under the EU AI Act, evidence from COMPASS-RT activities can be directly correlated with mandatory documentation such as a risk management file or post-market monitoring reports. This continuous validation is accomplished by turning well prior adversary cases into test cases for regression to execute on future versions of the system, providing feedback to close the loop on the “manage” function of the NIST AI RMF.

Compass-RT is a disciplined approach to bringing together adversarial creativity, automation, structure coverage and rough govenring list into one methodology. By integrating these practices into enterprise assurance, red teaming can evolve into a persistent, auditable control that demonstrably mitigates the remaining risk of AI exploitation.

4. Results

We designed the COMPASS-RT to showcase its flexibility across a broad range of classes of AI systems, such as conversational LLMs, retrieval-augmented assistants and tool-enabled agentic models. Real-world systems differ widely in architecture and deployment scenario and thus, the results reported here are not vendor specific or restricted to a specific platform. Instead, they are reference patterns and reporting structures that organizations can use to measure adversarial robustness, monitor defensive effectiveness, and distill findings into governance artifacts.

The assessment started by scoping three example environments. The first environment was a conversational model trained in a customer support flow, who had no access to any tool, but had readily seen sensitive customer queries. The second setting was a retrieval-augmented generation (RAG) system, which was used as a knowledge worker assistant who can query internal documents. The third model was an agentic model, which had access to tools like search APIs, robotic process automation scripts, and ticketing tools. These environments spanned a continuum of risk from privacy attacks in support, to integrity and safety threats in agentic automation.

Hybrid-based red teaming campaigns were conducted against each environment. Seed test cases were generated from popular datasets including AdvBench, HarmBench, SafetyBench, JailbreakBench to cover known vulnerabilities for baseline testing. Secondary LLM-powered automated adversarial generation then grew these seeds into thousands of test cases in a variety of languages, modalities, and interaction patterns. Human experts provided purpose-designed scenarios in the domain in order to succeed in eliciting financial account details from the customer support model or to introduce poisonous documents into the RAG pipeline. The execution harness guaranteed reproducibility, recording full transcripts, tool invocations as well as any side effects issued during the campaigns.

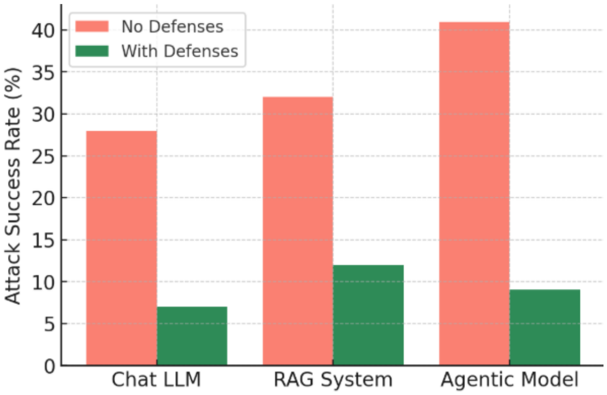

Fig 3: Comparative Adversarial Success Rates across System Environments Before and After Defensive Measures

The findings had a pattern similar to those seen before, but were measurable in more detail. In the conversational model setting, prompt injection attacks succeeded to elicit the restricted information in 28% of the cases without any provided Défense mechanisms. The use of refusal and output filtering effectively rendered this rate to 7%, providing evidence of dose intervention effectiveness, but also leaving an environmental risk that needed to purpose action. Within the RAG setting, adversarial documents inserted in-retrieving corpus were successfully surfaced 32% of the time at testing, displacing model predictions and undermining organization policy. These results highlighted the importance of provenance-aware defences such as signed content and anomaly detec- tion in retrieval pipelines. The agentic model environment had the highest risk: tool-use adversarial tests generated unauthorised tool calls with 41% success of rate despite no mitigations in place. The strict allow-listing of tool actions brought the rate of such attacks down to less than 10%, but the prevalence of indirect injections proved that Défense-in-depth was required.

The COMPASS-RT created reporting templates for structured outputs in governance and compliance. All test families were tagged with their MITRE ATLAS technique or

OWASP LLM Top-10 risk, providing a systematic way of reporting coverage. Attack Success Rates were displayed along confide nce intervals and classified according to defense posture, this facilitates risk owners to rate against the level of mitigation strength. Results were prioritization by ISO/IEC 23894 risk factors according to as severity was prioritized by influence and availability. Time-to-mitigation was also observed to provide a view into when organizations could effectively manage risks they identified that also reinforce the need for continuous validation.

Perhaps the key result of COMPASS-RT was its success in translating technical results into a form that was appropriate for use in a regulatory or audit context. Consistent with the EU AI Act, the findings were recorded in a structured risk management file, incorporating attack evidence, mitigation plan, and regression test plan. In the NIST AI RMF case, the artifacts filled up the “Measure” and “Manage” roles with tangible proof of tests and monitors. Using the ISO/IEC 42001 model, we embedded both knowledge and data from COMPASS-RT outputs to the plan-do-check-act cycle so that experiences from red teaming influenced security engineering and organizational governance.

Overall, our results show AI systems are inherently attackable by the evolving attacker despite that COMPASS-RT provides effective defences, leading to dramatic reduction in attack success and refusal robustness and in the degree of organizational preparation. Standardized metrics and templates are used to facilitate comparison among systems and governance integration ensures that red teaming findings are not isolated within technical teams but rather feed into enterprise-wide risk management processes.

5. Discussion

The findings from the use of COMPASS-RT provide a number of key takeaways regarding the current state of AI red teaming, and its role in the broader security validation space. They show that technical countermeasure like refusal tuning, filtering or limiting use of toolkits can help discouraging adversaries from attacking successfully although are not sufficient for preventing attacks all together. The lingering nature of residual risks especially in systems where agentic access to tools multiplies opportunities for abuse, underscores the necessity of changing red teaming from a one-time pre-release activity to a continuous assurance process.

A key discovery is the significance of hybrid adversarial generation that exploits the union of human creativity and automated adversarial augmentation. Evaluating Imitation Attacks by Generating Paraphrased and Obfuscated Test Cases The automated adversaries that leverage LMs were very efficient to produce a large set of paraphrased (e.g., strong attacks) and obfuscated (e.g., weak attacks) text cases, which were more diverse than the human test cases, allowing the coverage of more attack strategies. Nevertheless, the human testers' qualitative feedback was still critical. They were able to build up highly contextualized, nuanced exploits, like wrapping up malicious queries in customer-service language or sneaking poisoned commands into domain-related documents. This is indicative of where the future of AI red teaming should be heading with the right orchestration of human intelligence and automation in the form of machine made attacks that cover a broad space yet human evaluators check creativity, contextual plausibility, and high value places.

In addition, the results serve as validation that we should not solely rely on the simple success/failure rates when evaluating an approach. Attack Success Rate is a helpful headline metric, but without accompanying measures for thing like resistance robustness, Défense effectiveness, and scope of coverage, it is at risk of providing false comfort. By way of example, a decline in success rates may be misleading if coverage of adversarial conditions is incomplete, or if refusals only occur in narrow circumstances. By integrating these multidimensional metrics COMPASS-RT serves to enhance the depth of red teaming as an assurance practice, allowing organizations to develop a more detailed picture of their system’s resiliency.

Governance integration appeared as a second major theme. In the absence of formal techniques for translating insights into enterprise governance, red teaming may face relegation to nothing more than a technical anomaly with little organizational utility. COMPASS-RT bridges results over into risk registers, reference specification of technical vulnerabilities, i.e., a mapping to ISO/IEC 23894 risk management criteria, as well as association with ISO/IEC 42001 management system cycle, leading to formally acknowledging organizational risks, belonging to someone, being mitigated/reduced and monitored. By the same token, other companies are mapping to regulatory mandates like the EU AI Act with support for developing and documenting conformance to duties imposed by regulations on pre-release testing and post-market monitoring to provide defensible support for auditors and regulators. This governance connection turns red teaming from one-off engagements to a compliance enabler that supports organizational responsibility.

Another interesting dialogue involves the increasingly dynamic adversarial attack from agentic and retrieval-augmented models. Historically, red teaming has been centered on getting unsafe outputs from conversational models, but as AI is increasingly incorporated into workflows which include memory or retrieval or external action, the scope of red teaming must widen. His—— etc The body and soul are brah, one is a soul self and the other make the body move, so is undeniable he is dying.周期for Rings of Orbis\Development: Attacks like memory poisoning, RAG manipulation, indirect prompt injection exploit long-term persistence and cross-session impact, making detection / mitigation more challenging. Due to the promising performance we obtained in these environments during testing, we contend that the future both in terms of research as well as industry practice will have to prioritize red teaming techniques that are truly capable of accounting for these intricate system interactions.

Notwithstanding the potentiality of COMPASS-RT, some limitations remain. Full coverage of the adversarial domain is not possible due to the limitless creativity of possible attackers. Automated benchmarks, even when their used to provide reproducibility, run the risk of being too simplistic in modelling real adversarial attackers, and of encouraging overfitting to benchmark performance. There is also the problem of evaluator bias: even automated judges, such as LLMs, can rate the adversarial outcomes despaired, and human judgement adds another layer of subjectivity. Circumventing these limitations entails rotating varied evaluators, constantly revising adversarial corpora, and supplementing quantitative scoring with qualitative expert judgments.

Finally, the findings signal a need for advances in red teaming beyond red-blue dualisms toward more integrated approaches such as purple and violet teaming. In such models, attacking red teamers work closely with defending engineers and governance decision-makers, resulting in faster development and deployment of mitigations. This fits with the concept of security by design advocated under frameworks like the Secure AI Framework (SAIF) and the UK NCSC/CISA guidelines. Involving defenders and developers at the beginning of red teaming allows for quick feedback loops, faster time to mitigation, and can create a culture that values resiliency.

6. Conclusion

With continuous proliferation of AI in safety-critical and high-consequential applications, novel security validation paradigms are needed. But unlike traditional computer programs, AI has a stochastic (probabilistic) character, driven by data, whose behaviour is in principle difficult to predict and control. This unpredictability generates novel failure modes such as adversarial attacks, prompt injection, data poisoning, and unsafe autonomous tool use, revealing both technical vulnerabilities and governance challenges. Evidence presented throughout this paper supports the fact that conventional penetration testing methodologies are not adequate to address these risks. Rather, red teaming becomes a core form for systematic adversarial testing.

The proposed COMPASS-RT unites fragmented red teaming by engaging best practices across security engineering, adversarial machine learning research and international governance standards into a coherent methodology. With risk based scoping aligned to the NIST AI Risk Management Framework, integration with MITRE ATLAS and OWASP threat modelling taxonomies, hybrid adversarial generation, benchmark driven evaluation, and result mapping to ISO/IEC 23894, ISO/IEC 42001, and the EU AI ACT, COMPASS-RT makes red teaming into a repeatable, auditable and compliance-enabling process. These findings validate that such a procedure offers quantifiable enhancements in adversarial resilience, denial homogeneity, and response to prevention as well as creating governance artifacts that comply with regulatory requirements.

One of the distinctive features of COMPASS-RT is its dual attention to technical breadth and organizational depth. From a technical standpoint, the framework guarantees that different kinds of attacks are performed on different modalities, languages, and deployment conditions, thus obtaining a holistic evaluation of the system's attack vulnerability landscape. On the organizational side, it means the findings are not only available for technical teams but rather results in tools, risk registers, remediation plans and ongoing monitoring needs. This dual alignment reinforces accountability and maturity levels in enterprise AI assurance programs.

But the implementation itself also underscores the enduring barriers. Full adversarial coverage is still not realistic as attackers' creativity is unparalleled, and AI is constantly advancing. Those automated red teaming tools are phenomenal but can also introduce evaluator bias and overfit to the benchmark. Relatedly, agentic and retrieval-augmented models increase the adversarial surface in ways that current defences fail to cover. These constraints illustrate the necessity of an ongoing arms race between red team tools and techniques and defensive infrastructures. Future work should prioritize hardening RAG pipelines, guaranteeing provenance and integrity for external data sources, and hardening the resilience of autonomous multi-tool workflows.

The answer is by institutionalizing red teaming as a capital "R," capital "T"- red team- a lifecycle. As penetration testing and vulnerability management are integrated into secure development, and as operations, the process has to take AI red teaming with the model development, deployment, and monitoring. And it's not just technical innovation that's needed, its cultural organizations need to start adapting to more collaborative models, wherein both blue, red and purple stakeholders are co-developing mitigations and streamlining response time.

Finally, it is worth noting that while we cannot claim absolute eradication of adversarial risk in AI systems, we can assert that disciplined application of COMPASS-RT can substantially minimize, if not eliminate, the likelihood and consequence of adversarial exploitation. By integrating structured adversarial testing with governance integration, organizations can enhance their security and control posture and earn the trust of regulators, stakeholders, and the public. When red teaming is promoted from an ad hoc process to a formal assurance activity, it becomes a foundation stone of responsible AI deployment and a major enabler for the drive towards sustainable innovation in the era of intelligent systems.

References

[1] National Institute of Standards and Technology (NIST), *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, Jan. 2023. Available: https://doi.org/10.6028/NIST.AI.100-1

[2] National Institute of Standards and Technology (NIST), *AI RMF Generative AI Profile (AI 600-1)*, 2024. Available: https://doi.org/10.6028/NIST.AI.600-1

[3] ISO/IEC 23894:2023, *Information Technology — Artificial Intelligence — Guidance on Risk Management*, International Organization for Standardization, Geneva, 2023.

[4] ISO/IEC 42001:2023, *Artificial Intelligence — Management System*, International Organization for Standardization, Geneva, 2023.

[5] UK National Cyber Security Centre (NCSC) and Cybersecurity and Infrastructure Security Agency (CISA), *Guidelines for Secure AI System Development*, Nov. 2023. Available: https://www.ncsc.gov.uk/collection/secure-ai

[6] Google, *Introducing Google's Secure AI Framework (SAIF)*, Jun. 2023. Available: https://cloud.google.com/secure-ai-framework

[7] MITRE, *Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS) Fact Sheet*, 2024. [Online]. Available: https://atlas.mitre.org

[8] OWASP, *Top 10 for Large Language Model Applications v1.1*, Open Worldwide Application Security Project, 2023–2024. Available: https://owasp.org/www-project-top-10-for-large-language-model-applications

[9] [9] OpenAI, *GPT-4 System Card*, Mar. 2023. Available: https://cdn.openai.com/papers/gpt-4-system-card.pdf

[10] OpenAI, *GPT-4o System Card*, Aug. 2024. Available: https://cdn.openai.com/papers/GPT-4o-system-card.pdf

[11] E. Perez, S. Ringer, K. Kaplun, et al., "Red Teaming Language Models with Language Models," *arXiv preprint* arXiv:2202.03286, 2022.

[12] D. Ganguli, A. Askell, J. Clark, et al., "Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviours, and Lessons Learned," Anthropic Research Report, 2022.

[13] Y. Liu, X. Xu, A. Zhang, et al., "Formalizing and Benchmarking Prompt Injection Attacks and Defenses," in *Proc. USENIX Security Symposium*, 2024.

[14] P. Chao, H. Jin, A. Zhang, et al., "JailbreakBench: An Open Robustness Benchmark for Jailbreaking LLMs," in *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

[15] T. Mazeika, H. Zhang, S. He, et al., "HarmBench: A Standardized Evaluation Framework for Red Teaming and Robust Refusal," *arXiv preprint* arXiv:2402.04249, 2024.

[16] WalledAI Project, *AdvBench Dataset*, 2024. Available: https://github.com/walledai/AdvBench

[17] CoAI Group, Tsinghua University, *SafetyBench: A Comprehensive Safety Benchmark for LLMs*, 2023–2024. Available: https://github.com/thu-coai/SafetyBench

[18] Z. Chen, H. Zhou, Y. Liu, et al., "AgentPoison: Red-teaming LLM Agents via Poisoning Long-term Memory or RAG," in *Proc. NeurIPS*, 2024.

[19] European Union, *Regulation (EU) 2024/1689 of the European Parliament and of the Council on Artificial Intelligence (EU AI Act)*, OJ L, 12 Jul. 2024.

[20] European Union Agency for Cybersecurity (ENISA), *ENISA Threat Landscape 2024*, 2024. [Online]. Available: https://www.enisa.europa.eu/publications