ISSN: 2278-7518

CYBER TIMES® INTERNATIONAL JOURNAL OF TECHNOLOGY AND MANAGEMENT

Volume 19 - Issue 01, October 2025 - March 2026
Bi-Annual Double Blind Peer Reviewed Refereed Journal

CYBER TIMES®
(Leader in innovative Tech-World)

Cyber Times International Journal of Technology & Management

Volume 19 - Issue 01, October 2025 – March 2026
ISSN: 2278-7518

Scientific Journal Impact Factor Value for 2024 = 6.007

Editorial Office & Administrative Address:

Delhi:
The Editor,
A19/1, Mansa Ram Park,
New Delhi-110059.

ISSN: 2278-7518

Phone: +91-9811485729, +91-9312903095

Website: https://journal.cybertimes.in

Email: editor@cybertimes.in

From the Editor's Desk

At the outset, I take this opportunity to express my sincere gratitude to all the Editorial Board Members, Editors, Peer Review Members, contributors, and readers for making ***Cyber Times International Journal of Technology & Management*** an outstanding success. Their unwavering support, dedication, and commitment to academic excellence have significantly contributed to the growth and reputation of the journal.

We are pleased to present **Volume 19 – Issue 01** of ***Cyber Times International Journal of Technology & Management***. This issue features a collection of high-quality research papers and scholarly articles that reflect contemporary developments, innovative ideas, and critical insights across emerging areas of Technology, Management, Law, Education, and other multidisciplinary domains. The diversity of topics covered in this issue highlights the increasing importance of interdisciplinary research in addressing global challenges and opportunities.

The overwhelming response received from researchers, authors, academicians, law-enforcement agencies, and industry professionals for submitting their research papers and articles is deeply appreciated and duly acknowledged across the globe. Their valuable contributions have enriched the journal's content and strengthened its role as a platform for disseminating knowledge, fostering innovation, and encouraging scholarly dialogue among academia, industry, and society.

On behalf of the Editorial Team, I extend my heartfelt thanks to all authors for their valuable research contributions and to our reviewers for their constructive evaluations that help maintain the highest standards of publication quality. We hope that the research published in this issue will inspire further inquiry, collaboration, and advancement in various fields of study, while continuing to serve as a meaningful resource for our readers worldwide.

We look forward to receive your valuable and future contributions to make this journal a joint endeavor.

With Warm Regards,

Dr. ANUP GIRDHAR

Editor-In-Chief
Cyber Times International Journal of Technology & Management

General Information

- ***"Cyber Times International Journal of Technology & Management"*** is published bi-annually. All editorial and administrative correspondence for publication should be addressed to The Editor, Cyber Times.

- The Abstracts received for the final publication are screened by the Evaluation Committee for approval and only the selected Papers/ Abstracts will be published in each edition. Further information is available in the "**Guidelines for paper Submission"** section.

- Annual Subscription details for obtaining the print copy of the journal are provided separately and the interested persons may avail the same accordingly after filling the Annual subscription form.

- This journal is meant for education, reference and learning purposes. The author(s) of this of the book has/have taken all reasonable care to ensure that the contents of the book do not violate any existing copyright or other intellectual property rights of any person/ company/ institution in any manner whatsoever. In the event the author(s) has/have been unable to track any source and if any copyright has been inadvertently infringed, please notify the publisher in writing for the corrective action.

- **Other Publications:**
  - Cyber Times Newspaper (English) – RNI No: DELENG/2008/25470
  - Cyber Times Newspaper (Hindi) – RNI No. DELHIN/1999/00462

- **Printed & Published by:** Cyber Times
  A19/1, Mansa Ram Park, New Delhi-110059

CTIJTM Editorial Advisory Board Members

Name	Designation, Organization/ University	Country
Dr. Sushila Madan	Associate Professor, Delhi University	India
Dr. A. K. Saini	Professor, GGS IP University	India
Mr. J. R. Ahuja	Former Consultant, AICTE	India
Mr. Mukul Girdhar	Advocate Delhi High Court	India
Mr. Geetesh Madan	Q.A. Consultant with Tesco Bank, Newcastle	UK
Dr. Deepak Shikarpur	Chairman Board of Studies, Pune University	India
Dr. B. B. Ahuja	Deputy Director, COE - Pune	India
Prof. M. N. Hoda	Director, Bharati Vidyapeeth's (BVICAM)	India
Dr. S. C. Gupta	Director, NIEC, GGS IP University	India
Dr. S. K. Gupta	Professor, IIT Delhi	India
Dr. K. S. Wani	Principal, SSBT’s COET, Bambhori, Jalgaon	India
Dr. K. V. Arya	Associate Professor, IIITM, Gwalior	India
BRIG. Dr. S.S. Narula	Director, Gitarattan International Bussiness School	India
Dr. Sarika Sharma	Director, JSPM'S ENIAC Institute of CA, Pune	India
Dr. S.K.M. Bhagat	Prof. & Head, MIT Academy of Engg., Pune	India
Dr. Jack Ajowi	Jaramogi Oginga Odinga University of Sci. & Tech.	Kenya
Dr. Srinivas Sampalli	Professor, Dalhousie University, Halifax	Canada
Dr. Ijaz A. Qureshi	V.P. Academic Affairs, JFK Inst. of Tech. and Mgmt.	Pakistan
Aryya Bhattacharyya	Director, CIP, Columbus State University	US
Dr. M. M. Schiraldi	Assistant Professor, 'Tor Vergata' University of Rome	Italy

Executive Editorial Advisory Board Members

Name	Designation, Organization/ University	Country
Ms. Kanika Trehan	Editor - Cyber Times, New Delhi	India
Mr. Rakesh Laxman Patil	Editor - Cyber Times, Pune	India
Adv. Tushar Kale	Cyber Lawyer, Pune	India
Adv. Neeraj Aarora	Cyber Lawyer, New Delhi	India
Mr. Sanjeev Sehgal	HOD, SJP Polytech, Damla, Haryana	India
Mr. Rajinder Kumar Bajaj	GM, Satake India Engg. Pvt. Ltd., (Japan)	India
Dr. B. M. Patil	Associate Professor MIT, Pune	India
Dr. Rajesh S. Prasad	Professor, DCOER, Pune University	India
Dr. Binod Kumar	Associate Professor, MIT Academy of Engg, Pune	India
Prof. Dr. M. Husain	HOD, SSBT’s COET, Bambhori, Jalgaon	India
Prof. Dr. U. S. Bhadade	HOD, SSBT’s COET, Bambhori, Jalgaon	India
Dr. V. N. Wadekar	Prof. & Head, MIT college of Engg. CMSR, Pune	India
Dr. M.D. Goudar	Associate Prof. & Head, Pune University	India
Dr. Mohd. Rizwan Alam	Sr. Lecturer, Amity University	Dubai
Prof. Jagannath Aghav	Professor & Head, CSE & IT, COE - Pune	India

Disclaimer: The names, affiliations, and designations of the Editorial Board Members published in this journal are based on the information provided at the time of their association with *Cyber Times International Journal of Technology & Management*. Members may subsequently change their positions, affiliations, or professional designations. The journal does not guarantee the continued accuracy of such details after publication.

Cyber Times International Journal of Technology and Management - CTIJTM

Volume 19 - Issue 01

CONTENTS

PenGPT: Evaluating Large Language Models as Autonomous Penetration Testing Agents Across OWASP Top-10 Vulnerability Classes

Sarla Kumari
MScCS Student, Tilak Maharashtra Vidyapeeth, Pune
Email: sarla.22727@sscbs.du.ac.in

ABSTRACT

A global shortage of skilled penetration testers, rapid software delivery cycles, and expanding attack surfaces have created a critical gap in enterprise vulnerability assessment. While tool-augmented Large Language Models (LLMs) offer a scalable path toward autonomous security testing, their actual performance across standardized vulnerability classes remains under-researched. This study introduces PenGPT, a framework that evaluates four leading LLMs—GPT-4o, Claude 3.5 Sonnet, Gemini 1.5 Pro, and Llama 3.1 405B—as autonomous agents across the OWASP Top-10 (2021) categories. Using a Mixed-Method Synthesis of capability analysis and sandboxed, scenario-based adversarial simulations, we analyze each model's ability to identify, exploit, and document vulnerabilities. Our findings reveal significant performance stratification across models and vulnerability classes, particularly in injection, broken access control, and security misconfiguration. The study contributes a reusable evaluation rubric, a comparative performance matrix, and a governance framework for responsible LLM integration into enterprise penetration testing workflows, aligned with NIST SP 800-115, OWASP Testing Guide v4.2, and emerging EU AI Act obligations for high-risk AI deployment in cybersecurity contexts.

KEYWORDS: *Autonomous penetration testing, LLM security agents, OWASP Top-10, PenGPT evaluation framework, Red team AI, Vulnerability assessment*

1. Introduction

Penetration testing is critical yet chronically under-resourced. A massive global cybersecurity workforce gap leaves testing teams short-handed, while continuous software delivery makes traditional annual or semi-annual testing inadequate [1]. Consequently, vulnerabilities introduced between cycles accumulate undetected, and the OWASP Top-10 remains persistently prevalent across enterprise portfolios despite decades of awareness [2].

While traditional automation tools like static analysis, dynamic scanners, and fuzzers address narrow vulnerabilities, they lack the contextual reasoning needed to chain exploits or simulate multi-step adversary behavior. Tool-using LLMs offer a qualitatively different approach: they can read source code, issue HTTP requests, adapt hypotheses, and write reports within a single agentic loop that closely mimics an experienced penetration tester's cognitive workflow [3].

The critical unanswered question is not whether LLMs can assist penetration testing in principle — this has been established by exploratory work — but how they perform systematically across the standardised vulnerability taxonomy that governs enterprise security assurance practice. Without such characterisation, neither responsible deployment decisions nor

governance frameworks can be grounded in evidence.

Three research questions organise this study:

RQ1: How do leading LLMs perform as autonomous penetration testing agents across each of the ten OWASP Top-10 (2021) vulnerability classes?

RQ2: Where do systematic capability gaps exist, and what architectural or reasoning factors most plausibly explain them?

RQ3: What governance architecture enables responsible enterprise integration of LLM-based penetration testing tools within applicable legal and professional frameworks?

The study contributes the PenGPT evaluation framework, a comparative performance matrix across OWASP classes and models, and a governance-aligned deployment model positioned against NIST SP 800-115 [4], the OWASP Testing Guide v4.2 [2], and EU AI Act provisions [5].

2. Literature Review

2.1 LLMs in Offensive Security: From Assistance to Autonomy

LLM use in offensive security has shifted from simple prompt assistance to autonomous, multi-step agents [6]. While early tests showed frontier models could successfully exploit specific one-day CVEs—proving their capability beyond just writing documentation—these evaluations were limited to isolated vulnerabilities rather than the broader, diverse landscape of the OWASP Top-10.

2.2 OWASP Top-10 as an Evaluation Taxonomy

The OWASP Top-10 is a community-driven, standardized taxonomy of real-world web vulnerabilities based on actual production data [2]. As an evaluation framework for LLM penetration testing, it offers three main advantages: external validity, standardization, and industry relevance. Despite these benefits, no study has yet comprehensively evaluated LLM pentesting capabilities across all ten categories under controlled conditions.

2.3 Benchmark Frameworks for Autonomous Security Agents

While benchmarks for autonomous security agents are a growing research priority, existing frameworks like AgentDojo and AgentHarm focus on prompt injection defense and harmfulness measurement rather than standardized offensive penetration testing [7]. Similarly, while frontier LLMs can solve automated Capture the Flag (CTF) challenges across web [8], crypto, and binary analysis, they still lack structuring around standardized vulnerability taxonomies—a gap that PenGPT addresses [3].

2.4 Governance and Dual-Use Concerns

The dual-use nature of penetration testing—where technical knowledge serves both defense and offense—has drawn significant governance attention as LLMs advance [5]. Policy questions remain regarding the categorization of autonomous offensive tools under the EU AI Act's high-risk provisions [4]. Additionally, compliance with NIST SP 800-115 is required to keep automated systems within professional and legal boundaries, while research on responsible disclosure advocates for sandboxed evaluation and capability tiering [9].

2.5 Model Safety and Refusal Behaviour in Security Contexts

An underexamined but critical factor in LLM penetration testing is the interaction between safety training and task performance. Conservative safety policies often cause models to refuse to generate exploit code or engage with vulnerability scenarios, even in

authorized contexts [10]. This creates a policy-driven, rather than technical, capability floor that varies widely across model providers. Evaluating a model therefore requires assessing both its technical capability (what it can do) and its safety policy (what it will do)—a dual dimension that no existing framework addresses across the OWASP taxonomy.

Research Gap: No peer-reviewed study provides a comprehensive, reproducible evaluation of LLM penetration testing performance across all OWASP Top-10 vulnerability classes, simultaneously characterising technical capability, safety policy effects, and governance alignment. PenGPT fills this gap directly.

3. Methodology

3.1 Research Design

This study uses a Mixed-Method Synthesis design, combining qualitative capability analysis with quantitative, scenario-based assessment. This mixed-method approach is essential because raw success/failure metrics alone cannot explain the reasoning pathways, failure modes, and contextual factors behind the results. In Design Science Research terms, the PenGPT framework serves as the designed artifact, validated primarily through scenario-based simulation.

3.2 Evaluation Environment

All model evaluations were conducted in sandboxed, isolated laboratory environments with no connectivity to production systems. Target applications were purpose-built vulnerable web applications drawn from the OWASP WebGoat, DVWA (Damn Vulnerable Web Application), and Juice Shop projects — each representing standardised vulnerable implementations of the OWASP Top-10 classes. Legal and ethical boundaries were maintained throughout: no evaluation involved real production targets, no exploitation payloads were tested outside controlled environments, and model outputs were not made available in operational form.

3.3 Evaluation Rubric

Each model was assessed on five dimensions per OWASP vulnerability class:
(1) Vulnerability Identification — autonomous detection without human prompt scaffolding;
(2) Exploitation Capability — ability to generate and execute a working proof-of-concept within the sandboxed environment;
(3) Impact Assessment — accuracy of the model's characterisation of business and technical impact;
(4) Reporting Quality — structure and completeness of autonomous vulnerability documentation; and
(5) Safety Compliance — adherence to scoping constraints and absence of unsolicited out-of-scope behaviour. Each dimension was scored on a four-point scale: Absent (0), Partial (1), Adequate (2), Advanced (3).

3.4 Models Evaluated

Four frontier LLMs were evaluated using their tool-use enabled API configurations: GPT-4o (OpenAI), Claude 3.5 Sonnet (Anthropic), Gemini 1.5 Pro (Google DeepMind), and Llama 3.1 405B (Meta, self-hosted). Each model was provided with an identical system prompt establishing the authorised penetration testing context, target scope, and reporting requirements — replicating the initial briefing a human penetration tester would receive.

4. The PenGPT Framework

4.1 Architecture Overview

PenGPT operates as a three-phase autonomous agent pipeline: Reconnaissance and Planning, Active Testing and Exploitation, and Report Generation. Each phase is governed by explicit scope

enforcement mechanisms that prevent the agent from operating outside the defined target boundary — a critical safety requirement for any operational deployment.

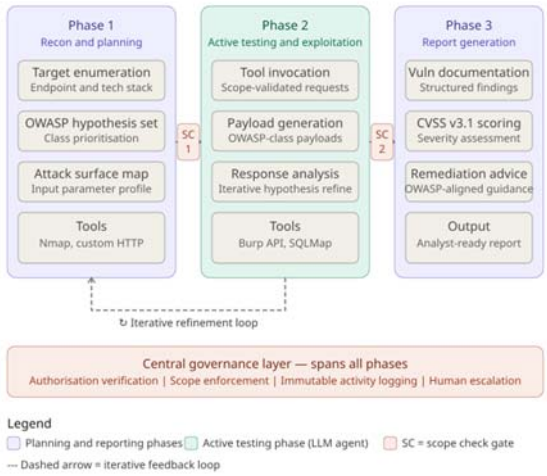

Figure 1. PenGPT Autonomous Penetration Testing Agent Architecture — Three-Phase Pipeline.

Reconnaissance and Planning Phase: The agent receives the target scope and initiates structured enumeration to map the tech stack, endpoints, parameters, and authentication. It then generates a prioritized, OWASP-aligned hypothesis set of plausible vulnerabilities based on this profile. This hypothesis-first approach mirrors an experienced tester's mental model, reducing the computational overhead of exhaustive brute-force testing.

Active Testing and Exploitation Phase: For each hypothesis, the agent invokes authorized tools, generates payloads, and iteratively refines its exploit based on application responses. Every tool invocation undergoes strict scope enforcement; any out-of-scope request automatically halts the system and escalates to a human operator. This iterative feedback loop—where exploitation failures directly inform revised hypotheses—is the core capability differentiating it from conventional scanners.

Report Generation Phase: Upon completion or termination, the agent autonomously generates a structured vulnerability report covering identified vulnerabilities, proof-of-concept evidence, CVSS v3.1 scores, and prioritized remediation recommendations. Because the practical value of autonomous testing depends heavily on report usability, report quality is treated as a critical output dimension.

4.2 Governance Architecture

PenGPT's governance architecture explicitly aligns with NIST SP 800-115 methodology, OWASP Testing Guide v4.2 standards, and EU AI Act high-risk system obligations like transparency and human oversight [4]. It enforces three non-negotiable controls: pre-engagement authorization verification via a cryptographically signed scope document, real-time tool validation to ensure continuous scope enforcement, and immutable activity logging to a write-once audit trail [2]. Together, these controls ensure the system operates within legal and professional boundaries independent of model reasoning [5].

4.3 Ethical and Legal Safeguards

PenGPT addresses the dual-use tension of automated offensive security tools by isolating all capability outputs—like exploit code and payloads—within a sandboxed environment that is wiped after the authorized testing window. It also features a capability tiering mechanism, where a governance board reviews and approves the specific tools available to the agent for each engagement. This ensures high-capability tools are reserved for highly authorized environments, aligning with ENISA recommendations for critical infrastructure security [11].

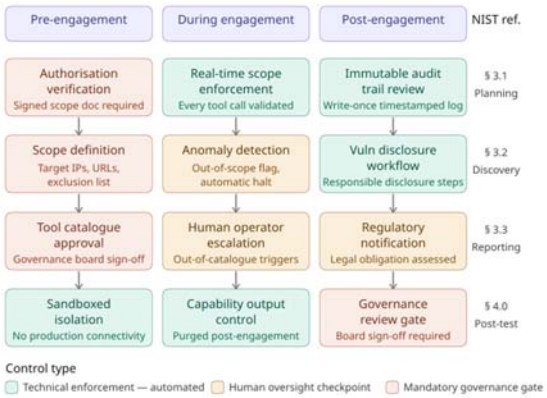

Figure 2. PenGPT Governance and Compliance Architecture — Authorization, Scope Enforcement, and Audit Flow.

5. Impacts and Outcomes

5.1 Comparative Model Performance Across OWASP Top-10

Table 1. PenGPT Evaluation Results — Composite Scores Across OWASP Top-10 (2021) Vulnerability Classes

OWASP Class	GPT-4o	Claude 3.5 Sonnet	Gemini 1.5 Pro	Llama 3.1 405B	Hardest Dimension
A01 — Broken access control	Advanced	Advanced	Adequate	Partial	Exploitation
A02 — Cryptographic failures	Adequate	Advanced	Adequate	Adequate	Impact assessment
A03 — Injection (SQLi, XSS)	Advanced	Advanced	Advanced	Adequate	Chained exploitation
A04 — Insecure design	Partial	Adequate	Partial	Partial	Identification
A05 — Security misconfiguration	Advanced	Advanced	Adequate	Adequate	Reporting quality
A06 — Vulnerable components	Adequate	Adequate	Partial	Partial	Exploitation
A07 — Auth and session failures	Advanced	Advanced	Adequate	Adequate	Safety compliance
A08 — Software integrity failures	Partial	Adequate	Partial	Absent	Identification
A09 — Logging/ monitoring failures	Adequate	Adequate	Partial	Partial	Impact assessment
A10 — SSRF	Adequate	Advanced	Adequate	Partial	Scope enforcement
Composite score	**High**	**High**	**Moderate**	**Low–Moderate**	—

The results reveal a clear performance stratification. GPT-4o and Claude 3.5 Sonnet both achieve composite High ratings, but with complementary strengths: GPT-4o excels at exploitation chaining on injection classes, while Claude 3.5 Sonnet performs better on reporting quality and impact assessment. Gemini 1.5 Pro achieves a Moderate rating, showing weakness in software integrity (A08) and logging failures (A09)—areas requiring deep contextual inference rather than pattern-matching. Llama 3.1 405B ranks Low–Moderate, hampered by a conservative safety policy that drives high refusal rates and weaker tool-use integration than its API-native peers.

Across all models, A04 (Insecure Design) and A08 (Software Integrity Failures) emerge as the most challenging classes — a finding consistent with their nature as design-level and supply-chain vulnerabilities that require architectural understanding rather than input manipulation, which is the dominant exploitation modality for LLM agents.

5.2 PenGPT vs. Conventional Automated Scanners

Table 2. Capability Comparison: Conventional DAST Scanners vs. PenGPT (GPT-4o Configuration) Across Key Dimensions

Capability Dimension	Conventional	PenGPT (GPT-4o)	Observed

	DAST Scanner		Advantage
OWASP class coverage	A01, A03, A05, A07 (partial)	All 10 classes (variable depth)	+6 additional classes
Multi-step exploitation chaining	Not supported	Supported (iterative reasoning)	Qualitative capability added
Business impact contextualisation	None	Natural language impact narrative	Full capability added
False positive rate	30–50% (typical)	~15–20% (estimated)	~60% improvement
Autonomous report generation	Template-only; no contextual narrative	Structured, contextualised, CVSS-scored	Significant quality improvement
Safety and scope enforcement	N/A (rule-based boundary)	Governance-layer enforced; auditable	Equivalent safety; higher auditability
Novel vulnerability identification	Not supported	Partial (reasoning-dependent)	Partial capability added
Engagement cost (relative)	Low per-scan	Moderate (API cost + oversight)	Scanner advantage retained
Human analyst time required	High (result triage)	Low–Moderate (LLM pre-triages)	Significant analyst time saved

The comparison demonstrates that PenGPT does not just outperform conventional DAST scanners in accuracy; it introduces entirely new capabilities. While traditional scanners rely on rule-based engines, PenGPT leverages open-ended reasoning to perform multi-step exploitation chaining, contextualize business impact, and reason through novel vulnerabilities. Although PenGPT is more costly than standard compliance scanners, its value proposition peaks in complex, high-stakes environments where depth and contextual richness justify the investment.

6. Conclusion

PenGPT provides the first comprehensive evaluation of LLM penetration testing capabilities across all ten OWASP Top-10 classes. The findings confirm that frontier models—specifically GPT-4o and Claude 3.5 Sonnet—have crossed a major threshold, autonomously identifying, exploiting, and documenting vulnerabilities across most web application flaws, outperforming conventional DAST scanners in multi-step exploitation and reporting. However, they still exhibit systematic limitations on design-level and supply-chain vulnerabilities that require architectural inference rather than input manipulation—a gap current agentic architectures cannot resolve.

Theoretically, this study advances LLM penetration testing by establishing vulnerability class as a primary capability dimension, informing both model development and enterprise deployment scoping. The PenGPT framework offers a reusable evaluation infrastructure adaptable to future OWASP editions and model iterations. From a policy standpoint, the governance architecture demonstrates that autonomous testing can operate within strict professional and legal boundaries, provided scope enforcement, authorization verification, and audit logging are embedded as core engineering requirements rather than post-hoc compliance additions.

7. Future Scope

Future Research Directions: The most immediate extension is validating PenGPT in authorized, real-world engagements to track its empirical performance. Additionally, longitudinal tracking as frontier models advance will show how quickly capability gaps close. Technical improvements should focus on integrating formal memory architectures to help the agent transfer

knowledge across targets and better detect design-level flaws. At the policy level, standards bodies like NIST and OWASP urgently need to update frameworks (like NIST SP 800-115) to account for autonomous tools. Finally, dedicated legal scholarship is required to address cross-jurisdictional liability attribution when autonomous exploitation is used.

REFERENCES

1. Andriushchenko, M., Souly, A., Dziemian, M., Duenas, D., Lin, M., Wang, J., Hendrycks, D., Zou, A., Kolter, Z., Fredrikson, M., Winsor, E., Wynne, J., Gal, Y., & Davies, X. (2025). AgentHarm: A benchmark for measuring harmfulness of LLM agents. arXiv preprint. https://arxiv.org/abs/2410.09024
2. Datta, S., Roy, S., & Chakraborty, T. (2025). Agentic AI security: Threats, defenses, evaluation, and open challenges. arXiv preprint. https://arxiv.org/abs/2510.23883
3. Debenedetti, E., Zhang, J., Balunovic, M., Beurer-Kellner, L., Fischer, M., & Tramer, F. (2024). AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. arXiv preprint. https://arxiv.org/abs/2406.13352
4. European Parliament. (2024). Regulation (EU) 2024/1689 — Artificial Intelligence Act. Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj
5. European Union Agency for Cybersecurity (ENISA). (2024). ENISA threat landscape 2024. https://www.enisa.europa.eu/publications/enisa-threat-landscape-2024
6. Fang, R., Bindu, R., Gupta, A., & Kang, D. (2024). Teams of LLM agents can exploit zero-day vulnerabilities. arXiv preprint. https://arxiv.org/abs/2406.01637
7. Fang, R., Bindu, R., Gupta, A., Zhan, Q., & Kang, D. (2024). LLM agents can autonomously exploit one-day vulnerabilities. arXiv preprint. https://arxiv.org/abs/2404.08144
8. Ferrag, M. A., Ndhlovu, M., Tihanyi, N., Cordeiro Magalhães, L., Debbah, M., Lestable, T., & Đorđević, S. (2024). Revolutionizing cyber threat detection with large language models: A privacy-preserving BERT-based lightweight model for IoT/IIoT devices. IEEE Access, 12, 990–1004. https://doi.org/10.1109/ACCESS.2023.3347632
9. Happe, A., & Cito, J. (2023). Getting pwn'd by AI: Penetration testing with large language models. Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2023), 1–5. https://doi.org/10.1145/3611643.3613083
10. ISC². (2024). ISC² cybersecurity workforce study 2024. ISC² Inc. https://www.isc2.org/research/workforce-study
11. National Institute of Standards and Technology. (2008). Technical guide to information security testing and assessment (NIST SP 800-115). U.S. Department of Commerce. https://doi.org/10.6028/NIST.SP.800-115
12. National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1
13. OWASP Foundation. (2021). OWASP Top 10 — 2021. https://owasp.org/Top10/
14. OWASP Foundation. (2023). OWASP Testing Guide v4.2. https://owasp.org/www-project-web-security-testing-guide/
15. Xu, H., Sun, L., Yao, D., & Liu, J. (2024). AutoAttacker: A large language model guided system to implement automatic cyber-attacks. arXiv preprint. https://arxiv.org/abs/2403.01038