Automated Offensive Cybersecurity

What AI Can and Cannot Yet Do

CPT Ruben MISSOTTEN, ir. - RMA / Dept CISS / Cylab
ruben.missotten@mil.be

Offensive security has always required scarce human expertise. That is changing. Early AI approaches used reinforcement learning (RL) to automate network navigation and attack planning, treating penetration testing as a sequential decision problem [1]. Large Language Models (LLMs) extended this further by operating over natural language action spaces and drawing on broad pretraining knowledge. In 2025, an LLM-based framework reached the top of HackerOne's bug-bounty leaderboards through fully autonomous vulnerability discovery and exploitation [2]. Generative AI is also entering adjacent disciplines: LLM-guided fuzzing has uncovered significant zero-day vulnerabilities [3], and AI-assisted tools for phishing and exploitation are documented on underground markets [4]. Automation is no longer a research prospect in offensive security — it is an operational reality.

These developments are not occurring in a vacuum. Credential theft and exploitation of public-facing services together account for the majority of observed initial access vectors, each tied at 30% of incidents [5]. The average time between vulnerability disclosure and active exploitation has fallen from 63 days in 2018–2019 to just 5 days in 2023 [6]. Once an attacker establishes a foothold, the window before they are discovered has narrowed from 205 days in 2014 to 11 days in 2024 [7]. AI-driven automation threatens to compress it further, particularly as agentic systems mature. Yet the bottleneck is no longer initial access — it is the sustained, multi-stage operation that follows, and where AI currently still falls short.

Current LLM agents perform well on isolated, well-scoped tasks — web penetration testing and vulnerability discovery being the clearest examples — but struggle to execute extended attack chains across interconnected enterprise systems [8]. This gap between narrow task performance and operational autonomy is the field's central open problem. Characterizing it rigorously through benchmarks that reflect real adversarial workflows is a prerequisite for advancing automated offensive tools and anticipating what adversaries will deploy next [9].

References

[1] J. Schwartz and H. Kurniawati, "Autonomous Penetration Testing using Reinforcement Learning," B.S. thesis, School of Information Technology and Electrical Engineering, University of Queensland, Brisbane, Australia, 2018. [Online]. Available: https://arxiv.org/abs/1905.05965

[2] N. Waisman. XBOW - XBOW on HackerOne: What's next. Published: August 18, 2025. [Online]. Available: https://xbow.com/blog/xbowon-hackerone-whats-next

[3] C. S. Xia, M. Paltenghi, J. Le Tian, M. Pradel, and L. Zhang, "Fuzz4All: Universal fuzzing with large language models," in Proceedings of the IEEE/ACM 46th International Conference on Software Engineering, ser. ICSE '24. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3597503.3639121

[4] Z. Lin, J. Cui, X. Liao, and X. Wang, "Malla: Demystifying real-world large language model integrated malicious services," in 33rd USENIX Security Symposium (USENIX Security 24). Philadelphia, PA: USENIX Association, Aug. 2024, pp. 4693–4710. [Online]. Available: https://www.usenix.org/conference/ usenixsecurity24/presentation/lin-zilong

[5] "IBM X-Force 2025 threat intelligence index," IBM Security, published: June, 2025. [Online]. Available: https://www.ibm.com/ reports/threat-intelligence

[6] C. Charrier and R. Weiner. How low can you go? An analysis of 2023 time-to-exploit trends. Google Cloud Blog. Published: October 15, 2024. [Online]. Available: https://cloud.google.com/blog/topics/threatintelligence/time-to-exploit-trends-2023

[7] "M-Trends 2025 report," Google Cloud Security, published: April 23, 2025. [Online]. Available: https://cloud.google.com/blog/topics/threatintelligence/m-trends-2025/

[8] B. Singer, K. Lucas, L. Adiga, M. Jain, L. Bauer, and V. Sekar, "On the feasibility of using LLMs to autonomously execute multi-host network attacks," 2025. [Online]. Available: https://arxiv.org/abs/2501.16466

[9] R. Missotten, V. Rimmer, W. Mees, and L. Desmet, "On the Potential of LLMs for Offensive Security: Benchmarks vs. Operational Reality," in Proc. 2025 Annual Computer Security Applications Conference Workshops (ACSAC Workshops), pp. 420–427, Dec. 2025. [Online]. Available: https://doi.org/10.1109/ACSACW69556.2025.00052