Hallucinations at the Firewall

Woo Jon Hou Ainsley

Singapore University of Technology and Design

Abstract

Generative AI shows strong capabilities in language, reasoning, and code but remains prone to hallucinations—outputs that are fluent yet incorrect. In cybersecurity, such errors pose serious risks, from misleading analysts to potential adversarial exploitation. This project investigates hallucinations in three directions: (1) creating benchmarks and interpretability tools to characterize them in security contexts; (2) developing mitigation strategies such as retrieval-augmented generation, symbolic-neural hybrids, and uncertainty-aware decoding; and (3) integrating these methods into real-world workflows like vulnerability assessment, malware analysis, and penetration testing, while exploring how attackers might exploit hallucinations. Evaluation will combine accuracy metrics, human-in-the-loop studies, and red-team simulations. By bridging theory and applied system design, the work aims to advance understanding of hallucinations and improve the reliability of AI in cybersecurity, with broader implications for other high-stakes areas such as healthcare and law.

Code — https://github.com/Kantosaurus/AAAI-2026

Introduction

Large language models (LLMs) are rapidly becoming part of cybersecurity workflows, assisting with vulnerability triage, malware interpretation, penetration testing, and secure configuration analysis. However, these systems remain prone to hallucinations—confident but incorrect outputs such as fabricated CVE identifiers, fictional exploit paths, or misleading reverse-engineering summaries. Unlike creative applications, hallucinations in cybersecurity can cause real harm: analysts may deploy incorrect patches, misconfigure systems, or waste time investigating nonexistent threats. Worse, attackers may intentionally trigger hallucinations through crafted inputs, turning an AI reliability issue into a security vulnerability. This project aims to systematically understand, mitigate, and operationalize defenses against hallucinations, ensuring LLM-powered systems remain trustworthy in adversarial environments.

Background

Research on AI Hallucinations

Hallucinations are common in LLMs due to mismatches between training and inference contexts. While mitigation methods like retrieval-augmented generation and uncertainty estimation exist for general-purpose settings, cybersecurity remains underexplored—despite hallucinations being particularly dangerous in security contexts where confidently false information can mislead analysts and introduce vulnerabilities.

AI in Cybersecurity

AI aids in threat detection, malware classification, and vulnerability discovery, but hallucinations—fabricated CVEs, invented best practices, or flawed code—can mislead analysts into overlooking threats or introducing vulnerabilities. Organizations adopt AI to address expertise shortages, yet verifying outputs requires that same expertise. In cybersecurity's adversarial context, where defenders must be correct while attackers need only one weakness, hallucinations demand both technical mitigations and impact assessment frameworks.

Approach

Our investigation proceeds in three stages: characterization of hallucinations in cybersecurity contexts, development and evaluation of mitigation techniques, and integration into realistic security workflows with adversarial testing.

Phase 1: Characterizing Hallucinations in Security Contexts In Stage 1, we characterize hallucinations across cybersecurity tasks by evaluating closed-source models (GPT-4.1, Claude 3.7, Gemini 2.0) and open-source alternatives (Llama-3, Qwen2.5-Instruct, Mistral-Large) on vulnerability assessment and malware analysis. Vulnerability assessment uses CVE/NVD, MITRE ATT&CK, ExploitDB, Metasploit metadata, and CIS benchmarks; malware analysis employs EMBER, SOREL-20M, AndroZoo, and Cuckoo sandbox logs. We develop a taxonomy of hallucinations—fabricated CVEs, incorrect severity ratings, fictional malware behaviors, and inconsistent attack chains—and employ interpretability techniques including causal tracing, activation steering, and contrastive decoding to understand their mechanisms.

Phase 2: Preventing and Mitigating Hallucinations Stage 2 evaluates mitigation strategies using cybersecurity-specific evidence stores built from NVD/CVE, NIST 800-53, CIS benchmarks, MITRE ATTCK, and malware reports from VirusTotal and SOREL. We test three approaches: retrieval-augmented generation with security-domain indices, symbolic-neural hybrids incorporating rule-based CVE validation and exploit chain logic constraints, and uncertainty-aware generation using conformal prediction, calibration models, abstention mechanisms, and hallucination scoring via entropy and self-consistency metrics.

Phase 3: Cybersecurity-Specific Integration In Stage 3, we assess real-world impact by integrating hallucination-aware models into cybersecurity workflows. For vulnerability assessment, mock pipelines evaluate whether hallucinations produce incorrect patch advice, fabricated CVEs, or inaccurate risk scores. For malware analysis, we test hallucination effects on family classification, behavior summarization, and sandbox trace explanation. Finally, adversarial exploitation testing uses prompt-injection suites, adversarial inputs, and red-team-generated exploit paths to assess whether attackers can deliberately trigger hallucinations to compromise security decisions.

Evaluation

We evaluate models along four dimensions: factual accuracy against trusted cybersecurity sources; hallucination reduction between baseline and mitigated models; analyst trust calibration, measuring how well confidence reflects true accuracy and helps detect hallucinations; and task completion impact, assessing time, decision accuracy, and critical errors in AI-assisted vulnerability and malware analysis.

Discussion

This research identifies systematic causes of cybersecurity hallucinations from data gaps and inference misalignment. It evaluates mitigation strategies that reduce errors but introduce efficieency and usability trade-offs and delivers practical frameworks for deploying reliable AI in high-stakes settings. Beyond cybersecurity, the findings generalize to critical sectors like healthcare, finance, medicine, and law, where grounding in verified knowledge, logical constraints, and calibrated uncertainty are essential to reducing hallucination risks and supporting global stability.

Conclusion

This project presents a systematic investigation into AI hallucinations in cybersecurity, examining their characterization, mitigation, and real-world impact. By integrating theoretical inquiry with practical system design, we develop methods that enhance reliability in contexts where errors can compromise security. The approach is grounded in established research on hallucinations and cybersecurity, leverages existing benchmarks and datasets for feasibility, and addresses a critical gap in AI safety for security applications. Ultimately, this work represents a step toward AI systems that function as trustworthy collaborators in defending society's most essential digital infrastructure systems that acknowledge their limitations, ground their outputs in evidence, and support rather than undermine human expertise

References

Huang, W.; Xia, F.; Shah, D.; Driess, D.; Zeng, A.; Lu, Y.; Florence, P.; Mordatch, I.; Levine, S.; Hausman, K.; and Ichter, B. 2023. Grounded Decoding: Guiding Text Generation with Grounded Models for Embodied Agents. In *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*.

Joren, H.; Zhang, J.; Ferng, C.-S.; Juan, D.-C.; Taly, A.; and Rashtchian, C. 2025. Sufficient Context: A New Lens on Retrieval Augmented Generation Systems. arXiv:2411.06037.

Kalai, A. T.; Nachum, O.; Vempala, S. S.; and Zhang, E. 2025. Why Language Models Hallucinate. arXiv:2509.04664.

Mendes, P.; Romano, P.; and Garlan, D. 2025. CLUE: Neural Networks Calibration via Learning Uncertainty-Error Alignment. arXiv:2505.22803.