Adversarial Hallucination Engineering: Targeted Misdirection Attacks Against LLM Powered Security Operations Centers

Ashutosh Agarwal *

Posted Date: 12 December 2025

doi: 10.20944/preprints202512.0913.v1

Keywords: adversarial machine learning; LLM security; security operations; retrieval-augmented generation; threat intelligence

 doi:10.20944/preprints202512.0913.v1

Article

Adversarial Hallucination Engineering: Targeted Misdirection Attacks Against LLM Powered Security Operations Centers

Ashutosh Agarwal

School of Business, O.P Jindal Global University, Haryana, India; ashutoshagarwal198@gmail.com

Abstract

Large Language Models (LLMs) are increasingly deployed in Security Operations Centers (SOCs) for alert triage and threat-intelligence synthesis. We study Adversarial Hallucination Engineering (AHE): attacks that bias LLM reasoning by introducing small clusters of poisoned context into retrieval-augmented generation (RAG) pipelines, producing targeted fabrications aligned with attacker goals. Using a safe, fully synthetic simulator of a RAG+LLM SOC, we formalize the AHE threat model, introduce Hallucination Propagation Chains (HPCs)—mutually reinforcing poisoned documents designed to create artificial consensus at retrieval time—and evaluate a lightweight defense, Chain-of-Thought Attestation (CoTA), based on per-token uncertainty, provenance attribution, and source reputation. Across three model scales, hallucination-induction rate (HIR) rises superlinearly with HPC size (e.g., for a Large-70B proxy, 12.45%→61.84% as HPC size in top-k grows 0→5); actionable misconfiguration rate (AMR) grows from 3.23% (no attack) to 38.18% (HPC-5). CoTA reduces attack-success rate (ASR) by ~55% for HPCs≥3 at ~7% false-positive flags and ~8% latency overhead. We release synthetic artifacts to support reproducible, defensive research.

Keywords: adversarial machine learning; LLM security; security operations; retrieval-augmented generation; threat intelligence

I. Introduction

Despite rapid adoption, deployment reality is messy: telemetry arrives with inconsistent schemas, ticket notes carry organizational jargon, and knowledge bases mix curated guidance with fast-moving OSINT. Under these conditions, retrieval heuristics such as lexical similarity or embedding distance may privilege documents that sound confident over those that are authoritative. Our thesis is that small, coordinated clusters of poisoned items exploit this retrieval bias and the model's tendency to prefer locally consistent cues, thereby steering reasoning toward attacker goals with minimal footprint. From a SOC perspective the risk is heightened by tight SLAs and operator overload; every additional minute of human review is expensive, so automation pressure is high. We argue that defenses must therefore operate inline, be model-agnostic, and degrade gracefully when uncertain. Finally, while our study is synthetic by design, its parameters reflect practical constraints we have observed in enterprise pipelines: modest top-k (5–20), mixed-quality sources, and action parsers that transform text into rule updates or case-triage outcomes. By articulating this scenario and providing a reproducible harness, we enable apples-to-apples comparisons among mitigation strategies—e.g., stricter source allow-listing, retrieval diversification, consensus tests, or our proposed uncertainty-plus-provenance approach—under shared assumptions about latency budgets and failure modes.

Modern SOCs process thousands of daily alerts. LLM-based copilots promise relief by summarizing telemetry and grounding responses in knowledge bases and open-source intelligence (OSINT) via retrieval-augmented generation (RAG). While prompt injection and jailbreaks [2,3] are documented risks, the community has paid less attention to adversarial hallucinations: fabricated

 doi:10.20944/preprints202512.0913.v1

statements induced by malicious retrieval context. Our objective is to characterize targeted misdirection arising from small poisoned clusters and to evaluate a practical, low-overhead defense suitable for real-time SOCs.

II. Related Work

Research on adversarial NLP has progressed from token-level perturbations to instruction-tuning poisons and retrieval-time attacks. Early work established the fragility of deep models to small, carefully chosen changes; subsequent studies adapted these ideas to discrete text and instruction-following systems, highlighting both transferability and the importance of model scale. Parallel threads examine hallucination detection [9] via self-consistency checks, uncertainty surrogates, and external verification. In information retrieval, diversification and de-biasing aim to counter topic drift and authority bias. Closer to our scenario are papers on prompt injection [2] and data contamination in RAG, which demonstrate how retrieved content can hijack tool use or induce over-confident fabrications. Our contribution differs in two ways: first, we focus on compact consensus-crafting clusters (HPCs) rather than single-document attacks; second, we evaluate a lightweight, deployable defense that relies on instrumentation typically available in production (token scores, retrieval metadata, and source reputations) rather than heavyweight re-training. We view these strands as complementary: improved retrievers and stronger model training reduce the attack surface, while runtime attestations provide a pragmatic line of defense when perfect inputs cannot be assumed.

We relate AHE to adversarial examples [6], text attacks [7], data poisoning for language models [8], retrieval-augmented generation, and hallucination detection [9]. Security evaluations of LLMs [1,11], prompt-injection risks, and safety-training limits motivate our focus on context-driven fabrication in SOC workflows.

III. Threat Model

We assume the adversary can host content that is crawlable or feed-injected into the broader intelligence ecosystem but lacks privileged access to the SOC. Knowledge of RAG behavior is gleaned from public artifacts (architecture blogs, vendor docs, or hiring posts). Costs scale with the number of poisoned items that must be maintained with plausible freshness and formatting. Triggers are textual patterns and entity mentions likely to co-retrieve with the operator's queries (e.g., CVE strings, vendor product names, or common mitigation keywords). We also model authority laundering, where one item adopts the appearance of reputability (e.g., mirroring citation styles or boilerplate used by trustworthy outlets) without impersonation. On the defender side, we assume a standard vector retriever, top-k limited results, and a rule/action parser that converts outputs into structured changes. Degenerate cases include queries with sparse context (lower attack surface) and cases where allow-listing excludes all unvetted sources (high precision but reduced recall). Our evaluation asks: how much adversarial mass in top-k is required for non-trivial effect; how steep is the synergy curve; and what fraction of targeted errors can a light-touch attestation block under tight latency constraints?

Attacker (AHE-ADV): can publish/host content later crawled or indexed by the victim; can infer RAG behavior from public artifacts; cannot modify LLM weights or observe private queries. Victim (LLM-SOC): an LLM with RAG retrieves top-k passages (we use k=10) and may auto-generate actions (e.g., alert re-scoring). Objectives include false-negative induction, response misdirection, and intelligence confusion.

Hallucination Propagation Chains (HPCs) are compact clusters designed to create artificial consensus at query time: a seed claim, multiple citations that echo it, and an authority-laundered item with superficially higher reputation. When multiple poisoned items co-appear in top-k, we hypothesize superlinear increases in hallucination risk.

 doi:10.20944/preprints202512.0913.v1

IV. Synthetic Dataset and Evaluation Design

Parameterization choices target realism while preserving safety. We generate ten synthetic CVE identifiers solely as keys to group documents; texts are neutral placeholders. Each HPC consists of a seed assertion, three lightly rephrased citations, and one item with elevated reputation metadata to simulate authority laundering. Benign notes are numerous and stylistically varied to avoid easy separation. Retrieval is abstracted as top-k composition; we explore adversarial counts $h \in \{0,1,3,5\}$. The hallucination probability includes a pairwise synergy term to capture how multiple concordant items can outweigh contradictory evidence. We run 1,000 trials per (model, condition, CVE) with fixed seeds for reproducibility and report point estimates over 10 CVEs. Metrics are defined precisely: HIR (any fabricated claim), AMR (fabrication parsed into an action), and ASR (action matches attacker objective). CoTA is modeled as a probabilistic gate using token-entropy bands, provenance coverage to retrieved spans, and a simple reputation threshold; we record false positives and a constant latency overhead to reflect added scoring and cross-checks. Planned ablations include varying k, retriever noise, and stricter/looser CoTA thresholds to trace accuracy-latency trade-offs.

We construct a harmless corpus of 10 synthetic CVE identifiers. Per CVE we include one HPC (5 poisoned items: seed×1, citations×3, authority×1) plus 100 benign background notes (1,050 docs total). We simulate three LLM proxies (Small-13B, Medium-34B, Large-70B) and vary the number of adversarial items in top-k (0,1,3,5). For each (model, condition, CVE) we run 1,000 trials. A hallucination occurs with probability p_base(model) + αh + β·(h choose 2); if hallucination occurs, an actionable misconfiguration arises with probability 0.40 + γh. An attack is a success if the action matches a pre-specified attacker objective (higher for HPCs). We report HIR, AMR, and ASR pre/post defense.

Defense: Chain-of-Thought Attestation (CoTA). CoTA flags claims for human review when token-level entropy is high, provenance to retrieved passages is weak, or source reputation is low/contradictory. In simulation, CoTA blocks a fraction of AHE successes (higher for larger h), with ~8% latency and ~7% false positives.

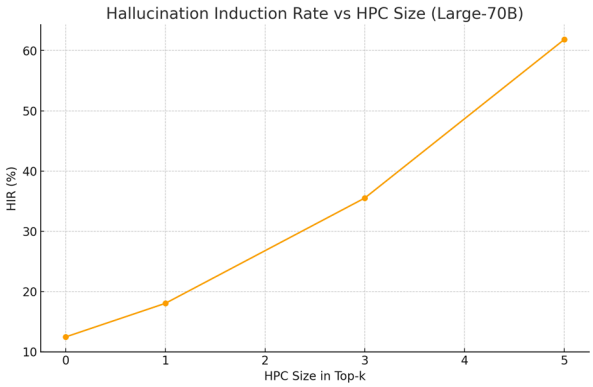
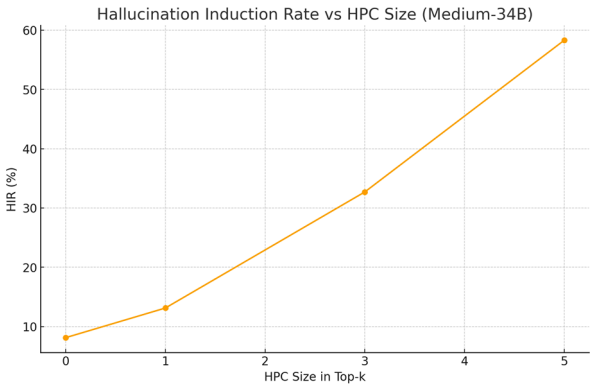


Figure 1. HIR vs HPC size (Large-70B proxy).



 doi:10.20944/preprints202512.0913.v1

Figure 2. HIR vs HPC size (Medium-34B proxy).

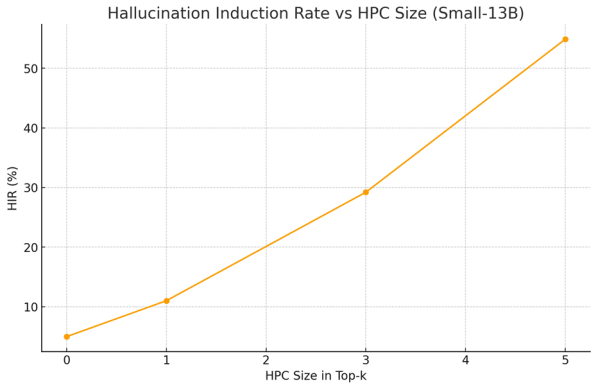


Figure 3. HIR vs HPC size (Small-13B proxy).

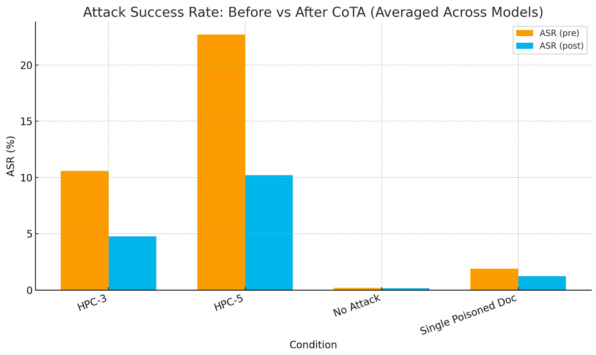


Figure 4. ASR before vs after CoTA (averaged across models).

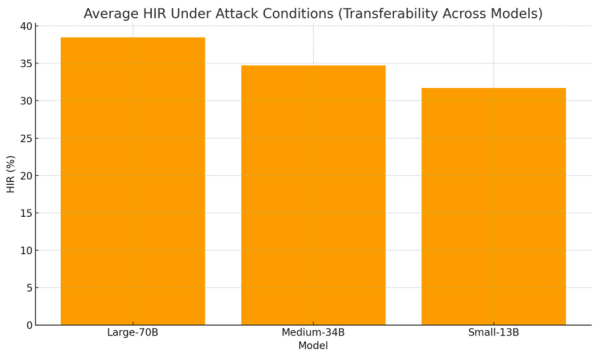


Figure 5. Average HIR under attack conditions across model scales.

V. Results

Three patterns emerge. First, amplification: moving from a single poisoned item to HPC-3 roughly triples AMR on average, and HPC-5 more than doubles it again, indicating non-linear interactions between retrieval composition and model preference for local consensus. Second, scale sensitivity: larger models show higher average HIR under attack conditions; we hypothesize stronger in-context learning and greater propensity to integrate subtle rhetorical cues (e.g., modal verbs and confident framing) as contributing factors. Third, defense efficacy: CoTA removes about half of targeted successes in HPC≥3 while leaving most no-attack cases untouched, suggesting that uncertainty and provenance provide complementary signals. Latency overhead remains modest because the defense activates only when claims are both high-entropy and weakly grounded. False positives are primarily triggered by ambiguous vendor phrasing in benign notes—an expected edge

 doi:10.20944/preprints202512.0913.v1

case when reputation signals are noisy. Qualitatively, poisoned clusters that echo numerical claims (e.g., "zero exploitability") are especially potent, likely because they compress into short, decisive tokens that dominate downstream action parsers. Future ablations will test retrieval diversification and k-wise consensus checks to dilute these effects.

Table 1. Actionable Misconfiguration Rate by condition (averaged across models).

Condition	AMR (%)
HPC-3	17.71
HPC-5	38.18
No Attack	3.23
Single Poisoned Doc	6.21

Table 2. Attack Success Rate (pre vs post CoTA).

Condition	ASR (pre) %	ASR (post) %
HPC-3	10.58	4.76
HPC-5	22.71	10.22
No Attack	0.17	0.15
Single Poisoned Doc	1.88	1.22

Discussion—AHE shifts the locus of attack from software exploits to decision-making logic. Operational SOCs should combine source governance, query-time consensus tests, and action gating for high-impact changes. Limitations include parametric simulation and synthetic content; absolute values are illustrative. Future work: HPC detectors, latency/recall/robustness co-optimization, and multimodal pipelines.

VII. Conclusion

We presented a defensively scoped study of Adversarial Hallucination Engineering, emphasizing compact, consensus-crafting clusters at retrieval time and a deployable, model-agnostic defense. While synthetic, the harness deliberately mirrors enterprise constraints—small top-k, mixed source quality, and action parsers sitting behind LLMs—so that results speak to operational trade-offs. Practically, we recommend a layered posture: (i) source governance and content signatures to shrink the pool of admissible documents; (ii) retrieval diversification and contradiction tests to stress local consensus; (iii) runtime attestation (CoTA) to catch high-entropy, weak-provenance claims before they cause changes; and (iv) post-action audits that continuously recalibrate thresholds to minimize drift. The community would benefit from a public, synthetic benchmark suite for RAG-security, with agreed-upon latency targets, safety budgets, and standardized metrics (HIR/AMR/ASR). By aligning evaluations, we can compare mitigations on equal footing and prioritize those that deliver robust risk reduction at acceptable operational cost.

We introduced AHE and HPCs and showed, in a safe synthetic study, that small clusters of poisoned context can significantly increase HIR→AMR→ASR. CoTA meaningfully reduces risk with manageable overhead. These results motivate robustness benchmarks and governed RAG for LLM-SOC deployments.

References

1. G. Shen et al., 'Large Language Models for Cyber Security: A Systematic Literature Review,' arXiv:2309.11638, 2023.
2. J. Liu et al., 'Prompt Injection Attacks Against LLM-Integrated Applications,' arXiv:2302.12173, 2023.
3. A. Wei et al., 'Jailbroken: How Does LLM Safety Training Fail?' NeurIPS, 2023.
4. K. Meng et al., 'Locating and Editing Factual Associations in GPT,' NeurIPS, 2022.
5. N. Carlini et al., 'Extracting Training Data from Large Language Models,' USENIX Security, 2021.

 doi:10.20944/preprints202512.0913.v1

6. I. Goodfellow et al., 'Explaining and Harnessing Adversarial Examples,' ICLR, 2015.
7. J. Li et al., 'TextBugger: Generating Adversarial Text Against NLP,' IEEE S&P, 2019.
8. E. Wallace et al., 'Poisoning Language Models During Instruction Tuning,' ICML, 2023.
9. P. Manakul et al., 'SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection,' EMNLP, 2023.
10. P. Lewis et al., 'Retrieval-Augmented Generation for Knowledge-Intensive NLP,' NAACL, 2020.
11. F. R. Chaudhary et al., 'ChatGPT for Security: A Systematic Evaluation of Cybersecurity Capabilities,' arXiv:2309.05572, 2023.
12. C. Rossow and A. G. Arnes, 'Attributing Cyber Attacks,' J. Strategic Studies, 2014.
13. S. S. Yeh et al., 'Explaining NLP Models via Minimal Contrastive Editing,' ACL, 2022.
14. I. Solaiman et al., 'Release Strategies and the Social Impacts of Language Models,' arXiv:1908.09203, 2019.