Hallucination Detection and Vulnerability Discovery in Large Language Models:

A Red Team Evaluation Framework for AI Safety

Alex Odhiambo Haya

Independent Researcher

Nairobi, Kenya

Abstract

Large Language Models (LLMs) have achieved remarkable capabilities in natural language understanding and generation, yet they persistently exhibit hallucinations - generating plausible but factually incorrect or contextually unfaithful content. This paper presents a comprehensive framework for understanding the root causes of AI hallucinations and establishes systematic methodologies for vulnerability discovery through red team evaluation. Drawing from practical experience in AI safety training and evaluation, I provide an empirically grounded analysis of hallucination mechanisms spanning training data issues, model architecture limitations, and inference-time behaviors. I introduce a structured taxonomy of hallucination types including factuality violations, faithfulness failures, intrinsic contradictions, and citation fabrication. My red team evaluation framework encompasses adversarial testing protocols for prompt injection, jailbreak detection, grounding verification, and dual-use content assessment. I present practical detection methodologies including fact-checking pipelines, uncertainty estimation, self-consistency analysis, and retrieval-augmented verification. This work contributes both theoretical understanding and actionable guidelines for AI practitioners developing safer, more reliable language model systems.

Keywords: *Large Language Models, Hallucination Detection, AI Safety, Red Team Evaluation, Vulnerability Discovery, Prompt Injection, Grounding Verification, Adversarial Testing*

1. Introduction

The rapid advancement of Large Language Models has transformed natural language processing, enabling applications ranging from conversational assistants to automated content generation and code synthesis. Models such as GPT-4, Claude, LLaMA, and DeepSeek demonstrate extraordinary capabilities in zero-shot and few-shot learning tasks, producing fluent, coherent text that often appears indistinguishable from human-written content. However, this capability comes with a critical limitation: these models persistently generate hallucinations - outputs that are syntactically correct and semantically plausible but factually inaccurate, contextually unfaithful, or entirely fabricated.

The consequences of hallucination extend far beyond mere inconvenience. In high-stakes domains such as healthcare, legal counsel, financial analysis, and educational content, hallucinated information can lead to significant harm. Recent incidents highlight this risk: in 2024, a Canadian

airline was ordered to honor a bereavement policy fabricated by its support chatbot, and in 2025, a major consulting firm's government report was found to contain non-existent academic sources and fabricated judicial quotes. These cases underscore the urgent need for systematic approaches to understanding, detecting, and mitigating AI hallucinations.

This paper presents insights from practical experience in AI safety training and evaluation, specifically through work on multilingual conversational AI systems involving complex system prompt development, multi-turn conversation design, and rigorous response evaluation against safety standards. My contributions include: (1) a comprehensive analysis of hallucination root causes across the LLM development pipeline, (2) a structured taxonomy of hallucination types with practical examples, (3) a red team evaluation framework for systematic vulnerability discovery, and (4) actionable detection and mitigation methodologies applicable to production AI systems.

2. Understanding Hallucination: Root Causes and Mechanisms

Recent research, particularly OpenAI's 2025 analysis, has shifted the understanding of hallucinations from mysterious glitches to well-understood statistical phenomena arising from training and evaluation procedures that reward confident guessing over calibrated uncertainty. I present a systematic analysis of hallucination sources across three primary categories: training data issues, model architecture limitations, and inference-time behaviors.

Figure 1: Root Causes of LLM Hallucinations

Training Data Issues	**Model** Architecture	**Inference** Behavior
• Misinformation • Bias propagation • Data gaps • Temporal drift	• Attention limits • Context windows • Objective functions • Positional encoding	• Over-confidence • Pattern matching • Sampling variance • Grounding failures

2.1 Training Data Issues

Large language models are trained on massive web-scale corpora that inevitably contain misinformation, outdated facts, and biased content. This creates several pathways to hallucination:

- **Imitative Falsehood:** Models learn to reproduce false information present in training data, including fake news, unfounded rumors, and outdated claims that were accurate at the time of data collection but have since been superseded.
- **Duplication Bias:** Information that appears frequently in training data receives disproportionate weight, while rare but important facts may be underrepresented. This can lead to confident generation of popular misconceptions.
- **Social Biases:** Biases related to gender, race, culture, and socioeconomic status embedded in training data are amplified during generation, leading to systematically skewed outputs.

- **Temporal Drift:** Knowledge cutoff dates create a fundamental gap between model knowledge and current reality. Facts about positions, policies, and events change continuously, but models lack awareness of these changes.

2.2 Model Architecture Limitations

While transformer architectures enable powerful language modeling, certain design choices inadvertently contribute to hallucination:

- **Attention Mechanism Limitations:** Language models exhibit localized focus within their attention mechanisms, prioritizing nearby tokens and potentially missing critical long-range dependencies that provide factual grounding.
- **Context Window Constraints:** Finite context windows force models to make predictions without access to all relevant information, leading to fabrication when necessary context is unavailable or has been truncated.
- **Objective Function Misalignment:** Next-token prediction objectives optimize for fluency and statistical likelihood rather than factual accuracy. A fluent, confident response is rewarded even when it contains errors.
- **Positional Encoding Effects:** Position-dependent information processing can create artifacts where model behavior varies based on where information appears in the context, not just what information is present.

2.3 Inference-Time Behaviors

Even with perfect training data and optimal architecture, hallucinations emerge during text generation:

- **Over-Confidence:** Models learn to express certainty even when uncertain because evaluation metrics reward confidence. OpenAI's analysis demonstrates that accuracy-based benchmarks penalize expressions of uncertainty, creating systemic pressure toward confident guessing.
- **Pattern Completion:** Models excel at completing patterns, which can lead to fabrication when a plausible-sounding completion doesn't correspond to factual reality. The model generates what "should" come next statistically, not necessarily what is true.
- **Sampling Variance:** Temperature and sampling strategies introduce randomness that can push generation toward low-probability tokens, occasionally producing fabricated content that wouldn't appear under deterministic decoding.
- **Grounding Failures:** When provided with external context (documents, tool outputs), models sometimes fail to properly ground their responses in that context, instead drawing on parametric knowledge or fabricating details.

2.4 The Inevitability Argument

Recent theoretical work by Xu et al. (2024) provides a formal argument that hallucination is an inherent limitation of LLMs used as general problem solvers. By defining hallucination as inconsistency between a computable LLM and a computable ground truth function, they show through learning theory that LLMs cannot learn all computable functions and will therefore inevitably hallucinate. This finding, supported by Anthropic's 2025 interpretability research identifying internal circuits that govern uncertainty expression, suggests that the field should focus on managing hallucination through external verification rather than pursuing complete elimination.

3. Hallucination Taxonomy

Effective detection and mitigation require precise categorization of hallucination types. Based on my evaluation experience and synthesis of current research, I present a comprehensive taxonomy distinguishing five primary hallucination categories:

Table 1: Taxonomy of LLM Hallucinations

Hallucination Type	Definition	Example
Factuality	Generated content contradicts verified real-world facts	Claiming Edison invented the light bulb from scratch
Faithfulness	Output deviates from provided context or source material	Summarizing a document with details not present in the original
Intrinsic	Direct contradiction of explicitly stated input information	Input: "Born in Ulm" → Output: "Born in Berlin"
Extrinsic	Plausible but unverifiable claims added beyond source	Adding fabricated collaboration details to factual response
Citation Fabrication	Inventing non-existent academic sources or references	Creating papers with real author names but fake titles/journals

In my practical evaluation work, citation fabrication emerged as a particularly insidious form of hallucination. A 2024 University of Mississippi study found that 47% of AI-generated citations submitted by students contained incorrect titles, dates, authors, or combinations thereof. This pattern extends beyond academic contexts - LLMs routinely fabricate plausible-sounding references with real author names but fictional titles, creating content that appears authoritative but fails verification.

4. Red Team Evaluation Framework

Red team evaluation systematically probes AI systems to discover vulnerabilities before deployment. My framework, developed through extensive AI safety training work, encompasses five primary vulnerability categories with corresponding testing methodologies:

Table 2: Vulnerability Categories and Red Team Approaches

Vulnerability Class	Description	Red Team Approach
Prompt Injection	Malicious instructions embedded to override system behavior	Systematic injection testing with various encoding schemes
Jailbreak Attacks	Techniques to bypass safety guardrails and content policies	Adversarial prompt crafting with role-play scenarios
Grounding Violations	Model generates content beyond provided tool/context outputs	Tool output verification against response claims
Dual-Use Content	Information with both legitimate and potentially harmful uses	Context-aware evaluation of response appropriateness

Harmful Content	Generation of dangerous, illegal, or unethical content	Policy boundary testing with edge case scenarios

4.1 Prompt Injection Testing

Prompt injection attacks attempt to override system instructions by embedding malicious directives within user input. My testing methodology includes: direct injection attempts where adversarial instructions are inserted in plain text, encoded injection using Base64, ROT13, or Unicode obfuscation to bypass content filters, indirect injection via documents, URLs, or tool outputs containing embedded instructions, and multi-turn injection building context across conversation turns to gradually shift model behavior.

4.2 Jailbreak Detection

Jailbreak attacks employ various techniques to circumvent safety guardrails. Common patterns I evaluate include: role-play scenarios that establish alternative personas not bound by standard restrictions, hypothetical framing requesting harmful content "for educational purposes" or "in a fictional context," authority impersonation claiming special permissions or developer access, and gradient-based attacks using adversarial suffixes optimized to trigger unsafe responses. Effective evaluation requires maintaining comprehensive attack libraries while avoiding disclosure that could enable misuse.

4.3 Grounding Verification

A critical component of my evaluation framework involves verifying that model responses are properly grounded in provided context. This is particularly important for tool-augmented models where responses should reflect tool outputs rather than parametric knowledge. My verification process systematically compares every factual claim in model responses against source documents or tool outputs, identifies speculation or fabrication beyond provided information, and flags responses that confidently state information not present in any provided context. This grounding verification proved essential in my work on multilingual AI training systems, where strict adherence to tool output grounding was critical for evaluation accuracy.

4.4 Dual-Use Content Assessment

Dual-use content presents particular challenges for AI safety - information that serves legitimate purposes in some contexts may enable harm in others. My assessment methodology considers: user intent signals derived from conversation context and query framing, information specificity evaluating whether responses provide actionable harmful details versus general educational content, availability analysis comparing model outputs against information readily available through other means, and cumulative risk assessing whether seemingly innocuous responses could be combined to enable harm. This nuanced evaluation requires human judgment informed by detailed policy guidelines, making it unsuitable for fully automated assessment.

5. Detection Methodologies

Effective hallucination detection requires multiple complementary approaches, as no single method provides comprehensive coverage. I present a framework combining five detection strategies:

Table 3: Hallucination Detection Methodologies

Detection Method	Approach	Strengths	Limitations
Fact-Checking	Verify claims against knowledge bases	High precision for verifiable facts	KB coverage gaps
Uncertainty Estimation	Analyze token probability scores	No external data needed	Lexical variety issues
Self-Consistency	Check response consistency across samples	Detects confident errors	Computation intensive
Retrieval-Augmented	Ground outputs in retrieved documents	Real-time verification	Retrieval quality dependent
Human Evaluation	Expert annotation of model outputs	Nuanced judgment	Scalability, cost

5.1 Fact-Checking Pipelines

Automated fact-checking involves decomposing model outputs into atomic claims and verifying each against knowledge bases or authoritative sources. While effective for factual assertions, this approach faces coverage limitations - knowledge bases may lack information about recent events, niche topics, or nuanced claims. My implementation combines multiple knowledge sources including structured databases, search engines, and domain-specific APIs to maximize coverage while maintaining verification speed.

5.2 Uncertainty Estimation

Uncertainty estimation analyzes model confidence through token probability scores derived from logit outputs. Low-probability tokens or high-entropy distributions may indicate uncertain or hallucinated content. However, as Zhang et al. (2024) note, this approach has fundamental limitations: semantically equivalent responses with different wording receive different uncertainty scores, and over-confident models may assign high probability to incorrect outputs. My approach uses uncertainty as one signal among many rather than a definitive detector.

5.3 Self-Consistency Analysis

Self-consistency methods generate multiple responses to the same query and analyze consistency across samples. Inconsistent responses suggest uncertainty or hallucination. This approach is particularly effective for detecting confident errors - cases where the model produces incorrect but consistent responses under deterministic settings but reveals uncertainty under sampling. The

primary limitation is computational cost, as multiple inference passes are required for each evaluation.

5.4 Retrieval-Augmented Verification

Retrieval-augmented approaches ground verification in external documents retrieved at inference time. By comparing model outputs against retrieved evidence, this method can detect both factual errors and faithfulness violations. However, effectiveness depends critically on retrieval quality - incomplete, outdated, or irrelevant retrieved documents can lead to both false positives (flagging correct information not present in retrieved documents) and false negatives (failing to detect errors when retrieved documents are themselves incorrect).

5.5 Human Evaluation Protocols

Despite advances in automated detection, human evaluation remains essential for nuanced assessment of hallucination, particularly for faithfulness violations, contextual appropriateness, and dual-use content. My evaluation protocols involve: structured annotation guidelines with clear criteria for each hallucination type, multi-annotator agreement requirements for ambiguous cases, detailed error labeling with reasoning comments explaining detection rationale, and iterative calibration sessions to maintain consistency across evaluators. This human-in-the-loop approach, while costly, provides ground truth labels essential for training and validating automated detection systems.

6. Mitigation Strategies

Given the theoretical arguments for hallucination inevitability, practical mitigation focuses on reducing frequency and impact rather than elimination. I organize mitigation strategies across three levels: prompting-level interventions, model-level improvements, and system-level safeguards.

6.1 Prompting-Level Interventions

Prompt engineering offers accessible mitigation without model changes:

- **Chain-of-Thought Prompting:** Requiring explicit reasoning steps reduces hallucination in prompt-sensitive scenarios by forcing models to show their work, making errors more detectable.
- **Explicit Uncertainty Instructions:** System prompts that explicitly instruct models to express uncertainty when lacking information can improve calibration, though effectiveness varies by model.
- **Source Grounding Requirements:** Instructions requiring citations or source references for factual claims can reduce fabrication by forcing models to ground assertions.
- **Structured Output Formats:** Constraining output format through schemas or templates can reduce opportunities for ungrounded elaboration.

6.2 Model-Level Improvements

Model development approaches include calibration-aware training that incorporates uncertainty into reward functions rather than purely accuracy-based optimization, targeted fine-tuning on high-quality factual data with explicit hallucination examples as negative samples, contrastive decoding methods that reduce probability assigned to hallucinated content relative to grounded alternatives, and retrieval augmentation that grounds generation in retrieved documents, though this requires careful attention to retrieval quality.

6.3 System-Level Safeguards

Production deployments benefit from post-generation filtering that applies automated fact-checking before presenting outputs to users, confidence score display that shows users calibrated uncertainty estimates to inform trust decisions, human review pipelines for high-stakes outputs requiring verification before action, and user feedback loops that enable continuous monitoring and model improvement based on reported errors. These system-level approaches acknowledge that some hallucination will occur and focus on minimizing downstream impact through layered defenses.

7. Practical Applications from AI Safety Training

My work on multilingual AI training systems provided extensive practical experience applying these frameworks to conversational AI evaluation. Key insights from this work include:

7.1 Complex System Prompt Development

Developing system prompts exceeding 10,000 characters requires careful attention to instruction clarity, tool usage protocols, and safety constraints. I found that explicit grounding requirements - stating that responses must be derived solely from tool outputs without speculation - significantly reduced hallucination in tool-augmented scenarios. However, overly rigid constraints sometimes led to unhelpful responses when legitimate inference would have been appropriate, highlighting the need for nuanced guidelines that distinguish required grounding from permitted inference.

7.2 Multi-Turn Conversation Evaluation

Hallucination patterns differ across conversation turns. Early turns often establish factual foundations that constrain later generation - if hallucinated facts are introduced early, they may be consistently maintained throughout the conversation, creating internally consistent but externally false discourse. My evaluation protocols track factual claims across turns to identify both initial hallucinations and inappropriate elaboration on earlier statements.

7.3 Multilingual Considerations

My French-language (fr_FR) evaluation work revealed that hallucination rates and types vary significantly across languages. Languages with less training data representation showed higher hallucination rates, particularly for culturally-specific or region-specific knowledge. This finding aligns with recent multilingual hallucination benchmarks (Mu-SHROOM, SemEval 2025)

showing that evaluation must account for language-specific patterns rather than assuming uniform behavior across languages.

7.4 Error Documentation and Correction

Effective AI safety evaluation requires systematic error documentation. My correction packages include: specific error labels categorizing violation type (hallucination, grounding violation, policy breach), reasoning comments explaining why the output was problematic, corrected responses demonstrating appropriate behavior, and severity ratings guiding prioritization for model improvement. This structured documentation enables both immediate correction and systematic analysis of error patterns for broader mitigation efforts.

8. Discussion

8.1 Implications for AI Deployment

The theoretical inevitability of hallucination has significant implications for AI deployment. Organizations cannot assume that sufficiently advanced models will eliminate hallucination risk - instead, deployment strategies must incorporate ongoing verification, user education about model limitations, and appropriate human oversight for high-stakes decisions. The Air Canada chatbot incident demonstrates that disclaimers alone are insufficient; organizations remain responsible for outputs of AI systems they deploy.

8.2 The Creativity-Hallucination Tradeoff

Recent research suggests a fundamental tension between hallucination and desirable capabilities. The same mechanisms enabling creative generation, novel combinations, and flexible reasoning also enable fabrication. David Baker's Nobel Prize-winning work using AI "hallucinations" to design novel proteins illustrates that fabrication can be beneficial in appropriate contexts. This suggests that optimal hallucination mitigation depends on application context - creative applications may tolerate or even benefit from fabrication that would be unacceptable in factual domains.

8.3 Evaluation Metric Reform

OpenAI's 2025 analysis identifies evaluation metrics as a root cause of hallucination, with accuracy-based benchmarks systematically rewarding confident guessing over calibrated uncertainty. Addressing hallucination at scale requires fundamental reform of how models are evaluated, moving from pure accuracy to metrics that reward appropriate expressions of uncertainty. This represents a collective action problem - individual organizations face competitive pressure to optimize accuracy-based benchmarks even when doing so increases hallucination risk.

8.4 Limitations

This work has several limitations. My practical experience focuses on conversational AI evaluation rather than other modalities; hallucination patterns in code generation, image understanding, or

other domains may differ significantly. The red team framework emphasizes adversarial discovery rather than quantitative measurement of vulnerability severity. Detection methods presented here represent current state-of-art but continue to evolve rapidly. Finally, mitigation strategies show varying effectiveness across models, tasks, and languages - practitioners should validate approaches in their specific deployment contexts.

9. Conclusion

Hallucination represents a fundamental challenge for large language model deployment, arising from well-understood mechanisms in training data, model architecture, and inference procedures. While theoretical results suggest hallucination cannot be completely eliminated, systematic approaches to detection and mitigation can significantly reduce frequency and impact. My framework provides practitioners with a comprehensive toolkit spanning hallucination taxonomy, red team evaluation methodology, detection approaches, and mitigation strategies.

The field has shifted from pursuing zero hallucination toward managing uncertainty in measurable, predictable ways. This shift acknowledges both the limitations of current technology and the need for practical deployment in applications that can tolerate some error with appropriate safeguards. Future work should focus on developing calibration-aware evaluation metrics, improving multilingual and multimodal hallucination detection, and establishing industry standards for responsible AI deployment that account for hallucination risk.

My practical experience in AI safety training demonstrates that rigorous evaluation protocols, combined with systematic error documentation and correction, enable continuous improvement in model safety and reliability. As AI systems become increasingly integrated into critical applications, the importance of this work only grows. I hope this framework provides a foundation for practitioners developing safer, more trustworthy language model systems.

References

[1] Anh-Hoang, D., Tran, V., & Nguyen, T. (2025). Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior. Frontiers in Artificial Intelligence, 8, 1622292.

[2] Anthropic. (2025). Interpretability research on Claude's uncertainty circuits. Anthropic Research Blog.

[3] Brown, T., Mann, B., Ryder, N., Subbiah, M., et al. (2020). Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

[4] Gambo, I., Khan, J. A., Agbonkhese, C., & Alwadain, A. (2025). "My AI is Lying to Me": User-reported LLM hallucinations in AI mobile apps reviews. Scientific Reports.

[5] Huang, L., Yu, W., Ma, W., Zhong, W., et al. (2024). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Transactions on Information Systems, 1(1), 1-45.

[6] Ji, Z., Lee, N., Frieske, R., Yu, T., et al. (2023). Survey of Hallucination in Natural Language Generation. ACM Computing Surveys, 55(12), 1-38.

[7] Kalai, A. T., & Nachum, O. (2025). Why Language Models Hallucinate. OpenAI Research.

[8] Lewis, P., Perez, E., Piktus, A., Petroni, F., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

[9] Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On Faithfulness and Factuality in Abstractive Summarization. Proceedings of the 58th Annual Meeting of the ACL, 1906-1919.

[10] OpenAI. (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.

[11] Ouyang, L., Wu, J., Jiang, X., Almeida, D., et al. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730-27744.

[12] Shuster, K., Komeili, M., Adolphs, L., Roller, S., et al. (2022). Language Models that Seek for Knowledge: Modular Search & Generation for Dialogue and Prompt Completion. Findings of EMNLP 2022, 373-393.

[13] Touvron, H., Martin, L., Stone, K., Albert, P., et al. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv preprint arXiv:2307.09288.

[14] Wei, J., Wang, X., Schuurmans, D., Bosma, M., et al. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Advances in Neural Information Processing Systems, 35, 24824-24837.

[15] Xu, Z., Jain, S., & Kankanhalli, M. (2024). Hallucination is Inevitable: An Innate Limitation of Large Language Models. arXiv preprint arXiv:2401.11817.

[16] Zhang, Y., Li, Y., Cui, L., Cai, D., et al. (2024). Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. arXiv preprint arXiv:2309.01219.