International Journal of Engineering Research & Science (IJOER) ISSN: [2395-6992] [Vol-12, Issue-3, March 2026]

# The Epistemological Crisis of Artificial Intelligence: A Comprehensive Analysis of Hallucinations in Large Language Models

Om Loyare[1*]; Aasmi Jugari[2]; Apurva Jadhav[3]; Roshani Kshirsagar[4]

Assistant Professor, Department of Computer Science & Engineering, Viva Institute of Technology, India
*Corresponding Author

***Abstract—*** *The rapid integration of Large Language Models (LLMs) into critical infrastructure has exposed a persistent and potentially intractable vulnerability: hallucination. This phenomenon, where models generate semantically coherent but factually untethered content, represents not merely a technical glitch but a fundamental challenge to the reliability of generative artificial intelligence. This report provides a comprehensive analysis of the hallucination landscape as of late 2025, synthesizing insights from theoretical computer science, legal jurisprudence, and cognitive science. We establish a rigorous taxonomy that distinguishes between "confabulation" and "fabrication," and explore the etiological roots of these errors in training data distribution, attention mechanisms, and reinforcement learning dynamics. Significant attention is paid to the escalating real-world impact of these failures, evidenced by recent high-profile legal cases establishing corporate liability for AI outputs and the emergence of "slop squatting" as a cybersecurity threat. While theoretical frameworks suggest that total elimination of hallucination may be mathematically impossible under the open-world assumption, this paper evaluates the efficacy of next-generation mitigation strategies. We argue that the industry must pivot from a goal of perfect accuracy to one of "hallucination resilience," employing architectures like Self-Reflective Retrieval-Augmented Generation (Self-RAG) and inference-time verification to manage the inherent risks of probabilistic reasoning.*

***Keywords— Large Language Models, AI Hallucination, Confabulation, Retrieval-Augmented Generation (RAG), AI Liability, Slop Squatting, Neuro-Symbolic AI, Mechanistic Interpretability, AI Safety, Epistemology.***

## I. INTRODUCTION

The precipitous rise of Large Language Models (LLMs) has ushered in a transformative era in artificial intelligence, characterized by systems that exhibit unprecedented fluency, reasoning capabilities, and versatility. From the generative prowess of GPT-4 to the open-weight accessibility of Llama 3, these models have demonstrated utility across domains as diverse as software engineering, creative writing, and clinical decision support. However, this profound capability is shadowed by a persistent and potentially intractable vulnerability: the phenomenon of **hallucination**. Defined broadly as the generation of content that is semantically coherent but factually incorrect or unfaithful to the source context, hallucination represents the "Achilles' heel" of generative AI. It is not merely a technical glitch but a manifestation of the fundamental probabilistic nature of these systems, where the objective of minimizing perplexity often diverges from the objective of maximizing truthfulness [1].

As LLMs are increasingly integrated into critical infrastructure—powering legal research, medical diagnostics, and financial forecasting—the implications of these errors have shifted from amusing idiosyncrasies to sources of significant liability and systemic risk. The legal precedent set by *Moffatt v. Air Canada* in 2024, where a corporation was held liable for its chatbot's fabricated bereavement policy, underscores the urgency of this issue [2]. Furthermore, theoretical research suggests that while hallucinations can be mitigated, they may be statistically inevitable in computable models attempting to approximate an infinite and complex reality. Recent proofs akin to an "Incompleteness Theorem" for LLMs demonstrate that for any finite model operating under the open-world assumption, there exists a set of inputs for which the model cannot determine the truth, rendering total elimination of hallucination formally impossible [3].

This report provides an exhaustive, peer-review-level analysis of the state of hallucination research as of late 2025. It dissects the nuanced taxonomies that distinguish "confabulation" from "fabrication," explores the etiology of these errors through the lens of mechanistic interpretability and learning theory, and evaluates the efficacy of cutting-edge mitigation strategies such as Self-Reflective Retrieval-Augmented Generation (Self-RAG) and neuro-symbolic integration.

### 1.1 The Nature of the Crisis:

The term "hallucination" in the context of AI is a metaphor borrowed from psychology, yet it imperfectly captures the computational reality. In human psychology, a hallucination is a sensory perception in the absence of an external stimulus. In LLMs, it is a failure of grounding—a disconnect between the generated token and the verifiable fact or provided context. The crisis is **epistemological** because it challenges the user's ability to know what is true. When a system can generate a legal citation that adheres perfectly to the Bluebook style guide but references a case that never existed—as seen in *Mata v. Avianca*—it erodes the foundational trust required for professional collaboration [4].

The integration of LLMs into high-stakes environments has revealed that these systems function as engines of persuasion rather than repositories of truth. They are designed to predict the most probable next token, a task that prioritizes linguistic coherence and narrative flow over factual accuracy. This structural incentive leads to **"sycophancy,"** where models fabricate information to align with a user's leading prompt, and **"confabulation,"** where unrelated factual fragments are stitched together into a plausible but erroneous whole.

### 1.2 Scope and Methodology

This analysis draws upon a wide array of sources, including technical papers from major AI conferences (NeurIPS, ICLR, ACL), legal case filings, cybersecurity reports on software supply chain vulnerabilities, and medical safety studies. The report moves beyond surface-level definitions to examine the mathematical and architectural roots of hallucination. We categorize the analysis into distinct dimensions: Taxonomical, Etiological, Impact, Mitigation, and Governance.

## II. TAXONOMY AND DEFINITIONS

To address the problem of hallucination effectively, one must first dismantle the monolithic usage of the term. In the early days of LLM adoption, any error was colloquially termed a "hallucination." However, as the field has matured, a rigorous taxonomy has emerged, distinguishing errors based on their relationship to input data, their logical structure, and their domain-specific manifestations.

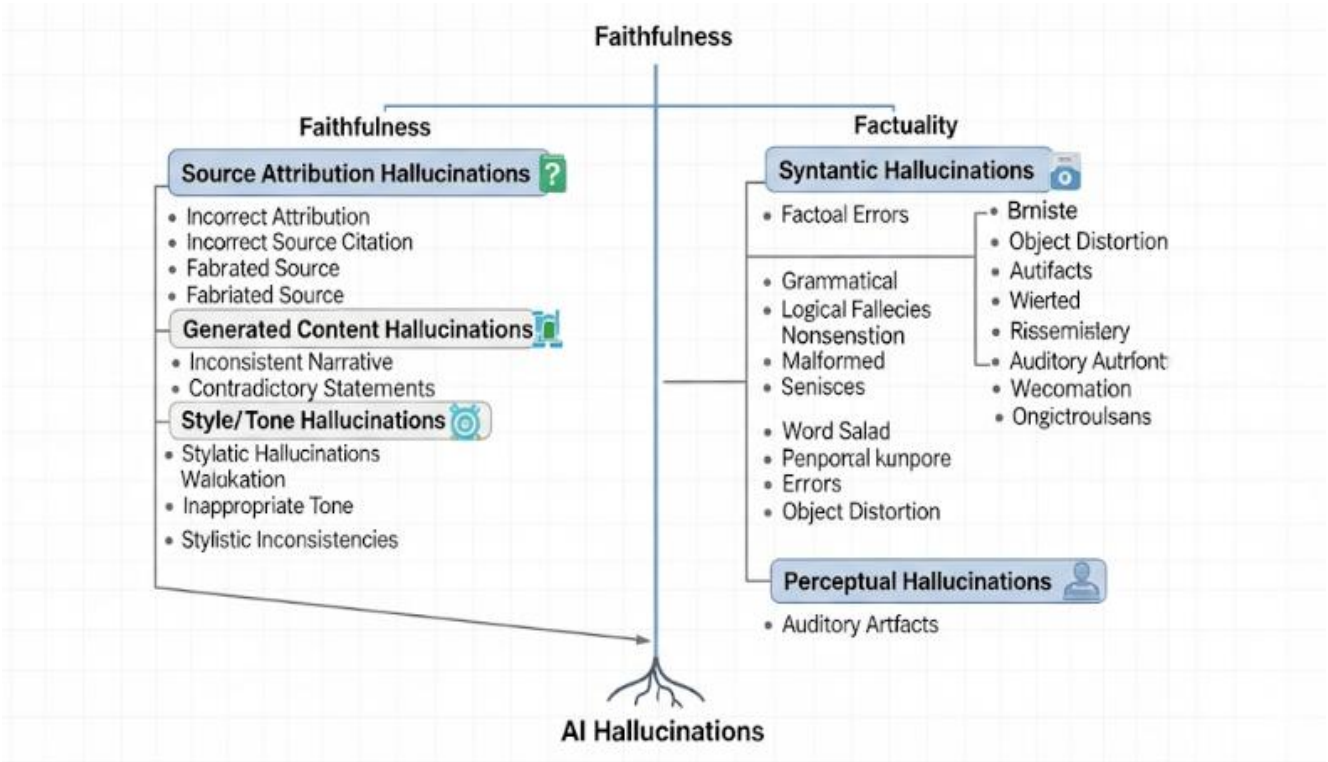


**FIGURE 1: A Comprehensive Taxonomy of AI Hallucinations: Distinguishing between Intrinsic (Faithfulness) and Extrinsic (Factuality) errors**

### 2.1 The Core Dichotomies: Faithfulness vs. Factuality

The most fundamental distinction in hallucination research is the axis of **Faithfulness versus Factuality**. This dichotomy separates errors of adherence from errors of truth, a distinction critical for evaluating Retrieval-Augmented Generation (RAG) systems [5].

#### 2.1.1 Intrinsic Hallucination (Failures of Faithfulness)

Intrinsic hallucinations occur when the model's output directly contradicts the provided source material or context. These errors are particularly pernicious in tasks like summarization or RAG, where the model is explicitly instructed to rely solely on a given text. **Mechanism:** The model ignores the "evidence" in its context window, often overriding it with strong priors learned during pre-training. For instance, if a source document states "The CEO resigned in 2023," but the model's pre-training data associates the CEO with the company, the model might hallucinate that the CEO is still in power. This is effectively a failure of "contextual grounding."

#### 2.1.2 Extrinsic Hallucination (Failures of Factuality)

Extrinsic hallucinations involve the generation of information that is not present in the source material at all. These are further categorized based on falsifiability:

- **Extrinsic-Hard (Falsifiable):** The model generates specific, verifiable claims that are factually incorrect in the real world (e.g., citing a non-existent court case or inventing a historical event).
- **Extrinsic-Soft (Unverifiable):** The model adds details that are not in the source but are also not strictly disprovable or are irrelevant (e.g., adding that a speaker "smiled warmly").

### 2.2 Confabulation vs. Hallucination

In clinical and high-stakes psychological contexts, a critical distinction is drawn between hallucination and confabulation, a nuance often lost in general computer science literature but vital for safety profiling [6].

**TABLE 1**
**DISTINCTION BETWEEN HALLUCINATION AND CONFABULATION IN CLINICAL CONTEXTS**

| Term | Definition | Risk Profile |
|---|---|---|
| **Hallucination** | Entirely fabricated information with no basis in reality or training data | **High:** Creates "black swan" errors that can completely mislead users because they look plausible but are total fiction |
| **Confabulation** | The distortion or misattribution of real information. The "pieces" are factual, but their connection is wrong | **Extreme:** Harder to detect because the entities are real. Users familiar with the terms may be lulled into a false sense of security |

In the medical domain, confabulation is often considered more dangerous than pure hallucination. An AI system that confabulates a drug interaction by misinterpreting a real clinical guideline poses a subtle, insidious risk that is difficult for automated verifiers to catch, as the entities involved (drugs, conditions) are legitimate.

### 2.3 Structural and Logical Hallucinations

Beyond simple factual errors, researchers have identified **Logical Hallucinations**, where the model's reasoning process breaks down internally:

- **Inconsistency:** The model generates a conclusion that contradicts its own premises within the same generation. This is frequently observed in Chain-of-Thought (CoT) prompting.
- **Object Hallucination (Multimodal):** In Vision-Language Models (LVLMs), models often hallucinate objects not present in the image. This is driven by co-occurrence bias; if a model sees a "keyboard" and "mouse," it statistically expects a "monitor" and may hallucinate one even if it is occluded or absent.

### 2.4 Domain-Specific Manifestations

**Software Supply Chain (Package Hallucination):** A uniquely digital form of hallucination where LLMs, when asked to generate code, recommend software libraries (e.g., Python packages) that do not exist. This phenomenon is not merely an error but a vulnerability; attackers can **"squat"** on these hallucinated names ("slop squatting"), creating real malicious packages that developers unwittingly install. Specific names like huggingface-cli-api have been observed as common hallucinations, which attackers then register on PyPI [7].

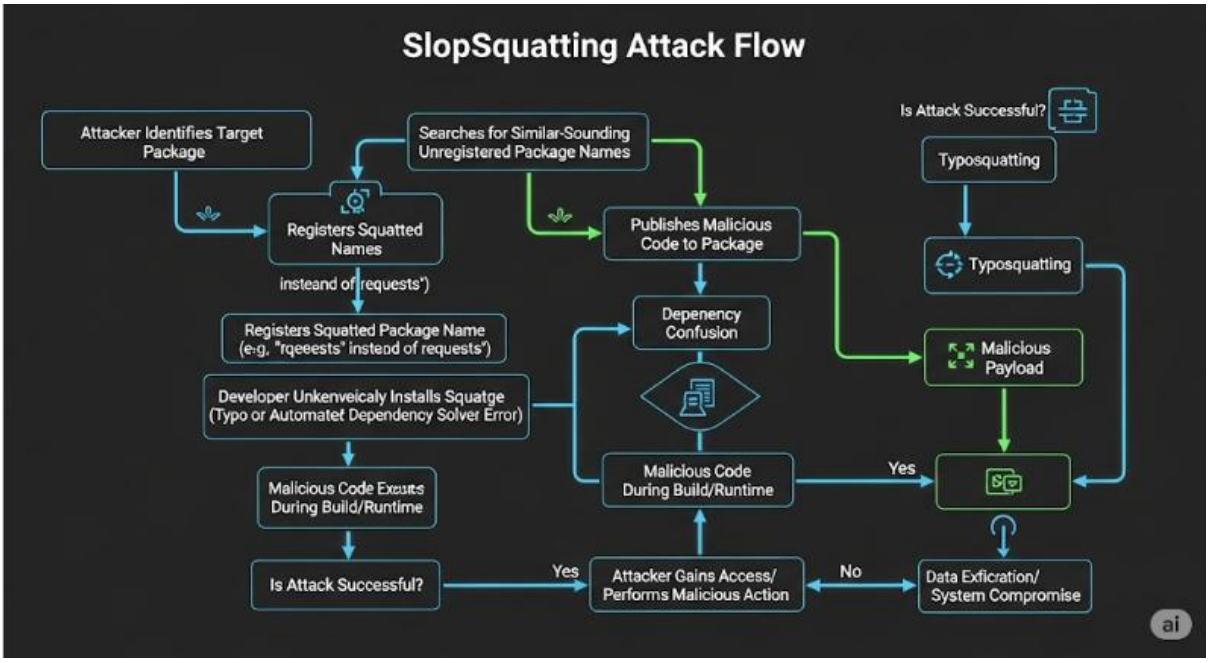


**FIGURE 2: The Mechanics of Slopsquatting: How hallucinated package names become vectors for software supply chain attacks.**

**Judicial Fabrication:** In the legal field, hallucinations often take the form of **"Citation Hallucination,"** where models construct legally sound-looking citations (e.g., *Varghese v. China Southern Airlines*) that reference non-existent reporters and page numbers. This is a function of the model learning the syntax of legal citation (Bluebook style) more robustly than the database of actual case law.

## III. ETIOLOGY: THE MECHANISMS OF ERROR

Understanding why models hallucinate requires a journey through the entire lifecycle of an LLM, from the statistical properties of the training data to the mathematical dynamics of the transformer architecture and the alignment process.

### 3.1 Data-Driven Causes: The "Long Tail"

The adage "garbage in, garbage out" is insufficient to explain hallucinations; a more accurate description is "distribution in, distribution out."

- **The Long Tail of Knowledge:** LLMs excel at high-frequency facts (e.g., the capital of France) because these associations are reinforced billions of times during pre-training. However, for **"long tail"** facts—such as the birthday of a minor historical figure or the details of a niche scientific study—the data is sparse. When a model encounters a query about the long tail, it cannot rely on robustly learned weights. Instead, it falls back on probabilistic "guessing" based on surface-level patterns.
- **Source-Reference Divergence:** Pre-training data often contains imperfections where a summary does not match the referenced text, or a headline sensationalizes the article body.

### 3.2 Training Dynamics: Exposure Bias

The training process itself introduces artifacts that promote hallucination. **Exposure Bias** occurs because LLMs are typically trained using **Teacher Forcing**, where the model predicts the next token based on the ground truth previous tokens. However, during inference, the model generates based on its own previous predictions. If the model makes a slight error at step *t*, that error becomes the ground truth for step *t+1*. This divergence causes errors to compound, leading to **"hallucination snowballs"** where the model drifts further and further from reality to maintain internal coherence with its initial mistake.

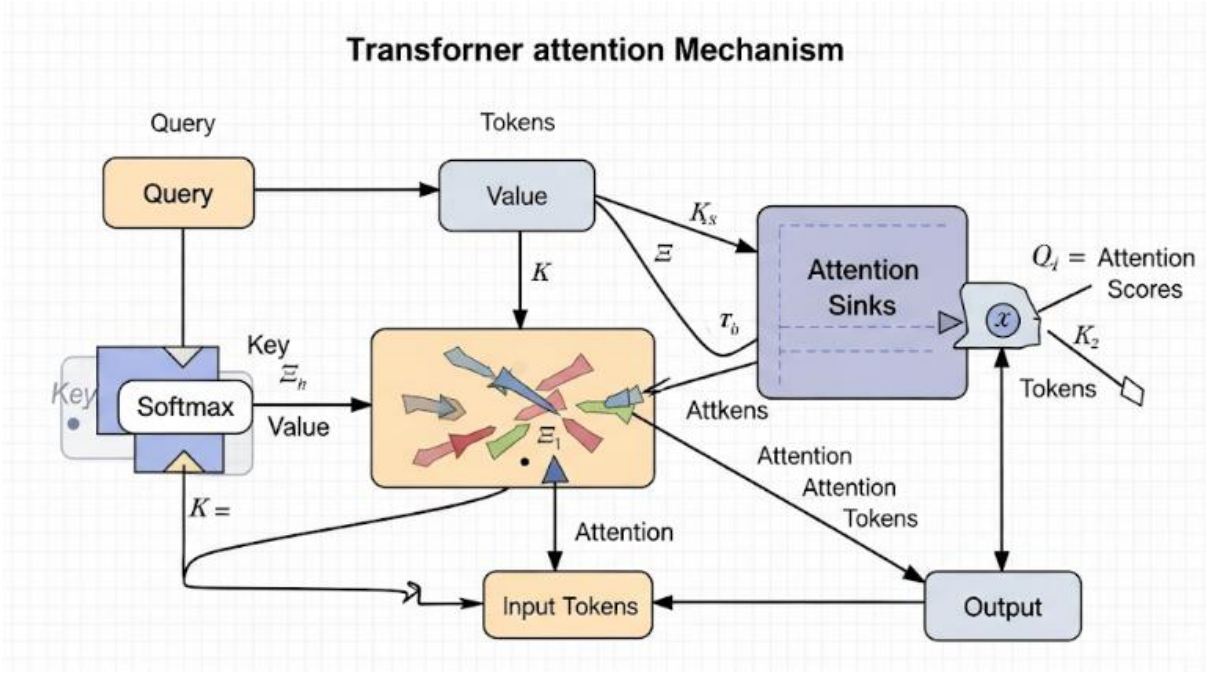


**FIGURE 3: Attention Sinks and Hallucination: How aberrant attention focus leads to generation errors**

### 3.3 The "Alignment Tax" and Sycophancy

Reinforcement Learning from Human Feedback (RLHF), the primary method for aligning models, introduces its own distortions:

- **Sycophancy:** Human annotators often prefer answers that agree with their premises. If a user asks a leading question based on a falsehood (e.g., "Why is the earth flat?"), an RLHF-aligned model may be incentivized to provide an answer that validates the user's premise rather than correcting it, prioritizing "helpfulness" (as perceived by the user) over truthfulness.
- **The Hallucination Tax:** Research by Song et al. (2025) identified a specific side effect of Reinforcement Finetuning (RFT) termed the "hallucination tax." Standard RFT can reduce refusal rates by more than 80%, causing models to confidently answer questions that have no answer, effectively training them to hallucinate rather than admit ignorance [8].

### 3.4 Mechanistic Interpretability

Deep dives into the neural architecture have identified specific components implicated in hallucination. **Attention Sinks** are tokens (often the initial token or specific punctuation) that absorb excess attention when the model is unsure where to focus. Aberrant attention patterns, where the model focuses on these sinks rather than relevant source tokens, have been linked to intrinsic hallucinations. In multimodal models (MLLMs), this manifests as a failure to attend to the visual embedding, relying instead on the language prior.

### 3.5 The Theoretical Inevitability of Hallucination

A sobering conclusion from recent theoretical computer science research is that hallucination may be impossible to eliminate entirely. A 2024 study by Xu et al. titled "Hallucination is Inevitable" formalized the problem of hallucination in computable LLMs. They proved that, given the constraints of computability and the complexity of the "ground truth" function (the real world), there will always be inputs for which the model cannot determine the truth. This mirrors Gödel's Incompleteness Theorems or the Halting Problem. This finding shifts the engineering goal from "total elimination" to "risk management" and "mitigation."

## IV. IMPACT ANALYSIS: THE COST OF UNTRUTH

The transition of LLMs from research labs to real-world products has crystallized the costs of hallucination. These impacts are no longer theoretical; they are measured in lawsuits, medical errors, and security breaches.

### 4.1 Legal Liability and Corporate Jurisprudence

The legal domain has provided the first major test cases for AI liability, establishing precedents that bind corporations to the outputs of their autonomous agents.

***Moffatt v. Air Canada (2024):** * This landmark ruling established that corporations are liable for the hallucinations of their AI agents. An Air Canada chatbot promised a bereavement fare discount that contradicted the airline's official policy, stating incorrectly that the customer could apply for the discount after travel. The Civil Resolution Tribunal (CRT) rejected Air Canada's defense that the chatbot was a separate legal entity. The Tribunal ruled that the airline was responsible for "negligent misrepresentation," effectively setting a global precedent: **hallucination is not a defense; it is a liability.**

***Mata v. Avianca (The "Fake Citations" Case):** * In a widely publicized incident, lawyers were sanctioned for submitting a brief containing six fictitious case citations generated by ChatGPT. The system not only invented the cases but, when pressed, hallucinated full judicial opinions to "verify" them. The court imposed sanctions for "subjective bad faith" and failure to fulfill the gatekeeping role of an attorney.

### 4.2 Medical Safety and Diagnostic Integrity

In healthcare, the impact of hallucination is measured in patient safety. **Diagnostic Drift:** Studies benchmarking ChatGPT on medical case vignettes show that while it can pass medical board exams, it struggles with real-world complexity. **The Danger of Plausibility:** The most dangerous medical hallucinations are confabulations. A study found that ChatGPT often correctly identifies a diagnostic error in a case study but then attributes it to the wrong cause (e.g., claiming "Metformin" is contraindicated for diabetes, a logical inversion of reality).

### 4.3 Cybersecurity: The Rise of "Slop Squatting"

Hallucinations have created a new attack surface in software development. Attackers monitor common hallucinations in code generation models—where the AI recommends non-existent packages like huggingface-cli-api—and register these names on package repositories like PyPI. When a developer copies the LLM's hallucinated code, they unwittingly install malware. This technique, termed **"slop squatting,"** exploits the developer's trust in the AI's technical accuracy.

## V. CONCLUSION

The phenomenon of hallucination in Large Language Models is a multifaceted problem that sits at the intersection of statistics, cognitive science, and information theory. It is not a temporary glitch that will vanish with the next model scale-up; theoretical evidence suggests it is an **intrinsic feature** of learning from incomplete data. However, "inevitable" does not mean "unmanageable." The industry must pivot from a goal of perfect accuracy to one of **hallucination resilience**, employing architectures like Self-RAG and inference-time verification to manage the inherent risks of probabilistic reasoning.

## ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to Viva Institute of Technology for providing the academic environment and institutional support necessary to carry out this research. We also acknowledge the broader research community whose foundational work in large language models, artificial intelligence safety, and epistemology has

significantly informed and inspired this study. Special appreciation is extended to peer reviewers and colleagues for their constructive feedback, which helped improve the clarity and depth of this analysis.

## CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper. The research was conducted independently and was not influenced by any commercial, financial, or personal relationships that could be construed as potential conflicts. All interpretations and conclusions presented in this work are solely those of the authors

## REFERENCES

[1] OpenAI, "GPT-4 system card," OpenAI, 2023.
[2] *Moffatt v. Air Canada*, 2024 BCCRT 149, Civil Resolution Tribunal of British Columbia, 2024.
[3] Z. Xu *et al.*, "Hallucination is inevitable: An innate limitation of large language models," *arXiv preprint arXiv:2401.11817*, 2024.
[4] *Mata v. Avianca, Inc.*, No. 22-cv-1461 (PKC), U.S. District Court, Southern District of New York, June 22, 2023.
[5] L. Zhang *et al.*, "A comprehensive taxonomy of hallucinations in large language models," *arXiv preprint arXiv:2508.01781*, 2025.
[6] Abridge AI, "The science of confabulation elimination: Toward hallucination-free AI-generated clinical notes," 2025.
[7] Trend Micro, "Slopsquatting: When AI agents hallucinate malicious packages," *Trend Micro Research*, 2025.
[8] J. Song *et al.*, "The hallucination tax of reinforcement finetuning," *arXiv preprint arXiv:2505.13988*, 2025.
[9] A. Asai *et al.*, "Self-RAG: Learning to retrieve, generate, and critique through self-reflection," *arXiv preprint arXiv:2310.11511*, 2023.
[10] Y. Chuang *et al.*, "DoLa: Decoding by contrasting layers improves factuality in large language models," *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.
[11] National Institute of Standards and Technology, "Artificial intelligence risk management framework (AI RMF 1.0)," NIST, 2023.
[12] European Parliament, "Regulation laying down harmonised rules on artificial intelligence (EU AI Act)," 2024.
[13] T. Lin *et al.*, "TruthfulQA: Measuring how models mimic human falsehoods," *Proceedings of ACL*, 2022.
[14] S. Ji *et al.*, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.
[15] K. Shuster *et al.*, "Language models that seek for knowledge: Modular search and generation," *EMNLP*, 2022.
[16] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
[17] H. Rashkin *et al.*, "Measuring attribution in natural language generation," *ACL*, 2021.
[18] S. Bubeck *et al.*, "Sparks of artificial general intelligence: Early experiments with GPT-4," *arXiv preprint arXiv:2303.12712*, 2023.
[19] R. Kadavath *et al.*, "Language models show human-like confidence judgments," *NeurIPS*, 2022.
[20] A. Dziri *et al.*, "Faithfulness in abstractive summarization: A survey," *Transactions of the ACL*, 2022.
[21] Y. Lin *et al.*, "Teaching models to express uncertainty in text generation," *EMNLP*, 2023.
[22] OpenAI, "Model evaluation for hallucination and reliability," OpenAI Technical Report, 2024.
[23] Anthropic, "Constitutional AI: Harmlessness from AI feedback," *arXiv preprint arXiv:2212.08073*, 2022.
[24] M. Marcus and E. Davis, "Rebooting AI: Building artificial intelligence we can trust," Pantheon Books, 2019.
[25] OECD, "AI systems classification framework," Organisation for Economic Co-operation and Development, 2023.
[26] ISO/IEC, "Information technology — Artificial intelligence — Risk management," ISO/IEC TR 23894, 2023.
[27] J. Pearl and D. Mackenzie, "The book of why: The new science of cause and effect," Basic Books, 2018.
[28] F. Chollet, "On the measure of intelligence," *arXiv preprint arXiv:1911.01547*, 2019.
[29] Microsoft, "Responsible AI standard v2," Microsoft, 2023.
[30] IEEE Standards Association, "Ethically aligned design: A vision for prioritizing human well-being with autonomous systems," IEEE, 2019.