IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, June 2026
Impact Factor: 8.2

The Hallucination Problem in Large Language Models: Causes, Consequences, and Emerging Solutions

Ansari Saquib Mohd Sohail
University of Mumbai, Mumbai

Abstract: *Large Language Models (LLMs) such as GPT-4, Gemini, and Claude have demonstrated remarkable capabilities in natural language understanding and generation. However, one of the most critical challenges limiting their reliable deployment is the phenomenon known as AI hallucination — where models generate information that is factually incorrect, fabricated, or logically inconsistent. This paper analyzes the root causes of hallucination in LLMs, examines its consequences across high-stakes domains such as healthcare, law, and education, and evaluates current mitigation strategies including Retrieval-Augmented Generation (RAG) and Reinforcement Learning from Human Feedback (RLHF). The paper concludes by discussing the future direction of research aimed at building more truthful and reliable AI systems.*

Keywords: Large Language Models, AI Hallucination, GPT, Natural Language Processing, RAG, RLHF, AI Safety

I. INTRODUCTION

Artificial Intelligence has undergone a dramatic transformation over the past decade. The emergence of Large Language Models — deep learning systems trained on vast corpora of text — has pushed the boundaries of what machines can do. Tools like OpenAI's ChatGPT, Google's Gemini, and Anthropic's Claude are now used by millions of people daily for tasks ranging from answering questions and writing code, to summarizing documents and assisting in medical or legal research.

Yet beneath this impressive capability lies a significant and well-documented flaw: LLMs frequently hallucinate. The term, borrowed loosely from psychology, refers to instances where an AI model generates output that sounds confident and fluent, but is factually wrong, misleading, or entirely fabricated. A model might cite a research paper that does not exist, state an incorrect historical date with complete confidence, or invent a legal precedent when consulted by a lawyer.

This is not a minor software bug — it is a fundamental challenge rooted in the very architecture and training methodology of these models. As LLMs are increasingly trusted in critical decision-making environments, understanding why hallucinations occur, what damage they can cause, and how they can be reduced has become one of the most urgent research priorities in modern AI.

This paper is structured as follows: Section 2 examines the technical and structural causes of hallucination; Section 3 explores its real-world consequences across key industries; Section 4 reviews current mitigation strategies and their effectiveness; and Section 5 discusses the outlook for future research and safer AI systems.

II. CAUSES OF AI HALLUCINATIONS

Understanding why LLMs hallucinate requires examining several interconnected factors — from the quality of training data, to the probabilistic nature of language generation, to the absence of an internal truth-checking mechanism.

DOI: 10.48175/IJARSCT-36351

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, June 2026
Impact Factor: 8.2

2.1 Training Data Limitations

LLMs are trained on massive datasets scraped from the internet, books, and other text sources. While impressive in size, these datasets inevitably contain errors, outdated information, biases, and contradictions. The model learns statistical patterns from this data without any ability to verify the factual accuracy of what it absorbs. As a result, if incorrect information appears frequently in the training data, the model may reproduce it with high confidence.

Furthermore, LLMs have a knowledge cutoff date — a point beyond which they have no information. When users ask about events or data beyond this cutoff, the model may attempt to generate a plausible-sounding response rather than acknowledging ignorance, leading directly to hallucination.

2.2 Probabilistic Text Generation

At its core, an LLM does not 'think' or 'know' — it predicts the next most likely token (word or word-part) given the preceding context. This process is fundamentally probabilistic. While this mechanism produces impressively fluent text, it has no built-in obligation to be accurate. The model is optimized to generate text that sounds plausible and coherent, not text that is necessarily true.

This means the model can produce outputs that follow logical and linguistic patterns perfectly, while being factually wrong. A well-structured, confidently-worded paragraph can be entirely fabricated — and the model itself has no mechanism to detect this.

2.3 Lack of Grounding in External Reality

Unlike a search engine that retrieves information from a live database, a base LLM operates entirely from internal weights established during training. It has no access to real-time information, external databases, or verifiable fact repositories unless specifically connected to such tools. This absence of 'grounding' means the model cannot cross-check its outputs against a reliable source of truth.

2.4 Overconfidence and Sycophancy

LLMs are often trained using Reinforcement Learning from Human Feedback (RLHF), where human raters reward responses that seem helpful and confident. This can inadvertently train the model to sound more certain than it should be, even when the underlying information is unreliable. Additionally, models can exhibit sycophancy — telling users what they want to hear rather than what is accurate — further compounding the hallucination problem.

[DIAGRAM 1 — Draw Here: Flowchart showing the 4 causes of AI hallucination feeding into the hallucination output. Boxes: Training Data Issues → Probabilistic Generation → No External Grounding → Overconfidence → Hallucination]

Figure 1: Key causes contributing to AI hallucination in Large Language Models.

III. REAL-WORLD CONSEQUENCES OF AI HALLUCINATIONS

The consequences of AI hallucinations extend far beyond trivial errors. As LLMs are increasingly deployed in high-stakes environments, the impact of fabricated or incorrect outputs can be severe, sometimes irreversible.

3.1 Healthcare and Medical Misinformation

Perhaps the most alarming application domain is healthcare. AI tools are being used to assist with medical diagnosis, drug interaction checking, and patient communication. A hallucinated drug dosage, a fabricated symptom list, or an invented medical study could directly harm or kill a patient. Studies have shown that even state-of-the-art LLMs make critical factual errors when answering medical questions, citing non-existent clinical trials or recommending contraindicated treatments.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, June 2026
Impact Factor: 8.2

3.2 Legal and Judicial Consequences

The legal profession has already experienced dramatic consequences from AI hallucinations. In a widely reported 2023 case, a New York attorney submitted a legal brief to federal court that contained multiple citations to cases that did not exist — all generated by ChatGPT. The attorney faced sanctions, and the incident prompted widespread discussion about the dangers of relying on LLMs for legal research without verification. Courts have since begun issuing AI disclosure requirements for submitted documents.

3.3 Education and Academic Integrity

In educational settings, students who rely on LLMs for research may submit papers containing fabricated facts, invented citations, or incorrect explanations — often without realizing the content is false. This undermines academic integrity and, more dangerously, misinforms the next generation of professionals. Educators are grappling with how to teach critical AI literacy alongside the use of these tools.

3.4 Business and Reputational Damage

In the corporate world, AI-generated reports, product descriptions, or communications containing hallucinated data can mislead investors, damage brand reputation, or result in regulatory violations. A company relying on an AI tool that confidently states false market figures or fabricates competitor analysis could make costly strategic decisions based on entirely fictional information.

Table 1 below summarizes the impact of AI hallucinations across key sectors:

Sector	Example of Hallucination	Potential Consequence	Severity
Healthcare	Fabricated drug dosage or clinical trial	Patient harm or death	Critical
Legal	Non-existent case citations in court briefs	Sanctions, case dismissal	High
Education	Invented academic references in student papers	Academic misconduct	Medium
Journalism	False quotes attributed to real people	Defamation, public misinformation	High
Finance	Hallucinated market data in reports	Poor investment decisions	High
Customer Service	Incorrect product or policy information	Customer dissatisfaction, liability	Medium

Table 1: Impact of AI hallucinations across major sectors.

IV. CURRENT SOLUTIONS AND MITIGATION STRATEGIES

Addressing AI hallucination is one of the most active areas of research in the AI community. While no single solution has been proven to eliminate hallucinations entirely, several approaches have shown meaningful progress.

4.1 Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) is currently one of the most widely adopted approaches to reducing hallucinations. Instead of relying solely on knowledge stored in model weights, RAG systems first retrieve relevant documents from an external, trusted database and then use these documents as context when generating a response. This 'grounds' the model's output in verifiable, up-to-date information and significantly reduces the likelihood of fabrication.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, June 2026
Impact Factor: 8.2

RAG has been adopted by companies like Microsoft (in Bing AI) and Google (in Gemini) to anchor AI responses in live web results. Studies show RAG can reduce hallucination rates by 30–50% in knowledge-intensive tasks.

4.2 Reinforcement Learning from Human Feedback (RLHF)

RLHF involves training the model using feedback from human reviewers who rate the quality, accuracy, and helpfulness of responses. When reviewers specifically penalize hallucinations and reward accurate, appropriately uncertain responses (e.g., 'I don't know'), the model gradually learns to be more truthful. This technique was central to the development of ChatGPT and Claude, and has been shown to improve factual accuracy substantially compared to base models without RLHF fine-tuning.

4.3 Constitutional AI and Self-Critique

Anthropic introduced Constitutional AI (CAI), a technique where the model is given a set of principles and asked to critique and revise its own responses before finalizing them. This self-critique process can catch factual inconsistencies and improve response quality. Similarly, chain-of-thought prompting — where the model is asked to reason step by step — tends to reduce confident errors by forcing more deliberate reasoning.

4.4 Confidence Calibration and Uncertainty Expression

Researchers are working on better calibrating model confidence — ensuring that when a model is uncertain, it says so explicitly rather than inventing an answer. Techniques include training models to output explicit uncertainty markers (e.g., 'I am not sure, but...') and building separate classifier layers that estimate the reliability of a given output before it is shown to the user.

4.5 Fact-Checking Layers and Verification Pipelines

Several systems now integrate a post-generation verification step, where a separate AI module or external tool checks the model's output against trusted databases before presenting the answer to the user. While this adds latency, it dramatically improves reliability in critical applications such as medical advice systems or legal research tools.

[DIAGRAM 2 — Draw Here: Comparison diagram of RAG vs Standard LLM. Left side — Standard LLM: User Query → Model Weights → Response. Right side — RAG: User Query → Retrieve Documents → Model + Documents → Verified Response]

Figure 2: Standard LLM generation vs. Retrieval-Augmented Generation (RAG).

V. FUTURE OUTLOOK

Despite significant progress, hallucination remains an unsolved problem. The research community has identified several promising directions that may yield more robust solutions in the coming years.

5.1 Multimodal Grounding

As AI systems increasingly process images, audio, and video alongside text, multimodal grounding offers a new avenue for reducing hallucinations. By cross-referencing textual claims with visual or auditory evidence, models may achieve a richer form of verification that purely text-based systems cannot.

5.2 World Models

Leading AI researchers, including Yann LeCun at Meta, have argued that truly reliable AI requires 'world models' — internal representations of how the physical and social world operates, not just statistical patterns in text. Such models would be capable of reasoning about causality and physical plausibility, making it far harder to generate confident falsehoods.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, June 2026
Impact Factor: 8.2

5.3 Standardized Hallucination Benchmarks

The field currently lacks standardized, universally accepted benchmarks for measuring hallucination rates across different models and domains. Organizations like NIST and the Partnership on AI are working toward such standards, which will be critical for comparing solutions objectively and holding AI developers accountable for the reliability of their systems.

5.4 Regulatory and Ethical Frameworks

Governments and regulatory bodies are beginning to address AI hallucination through policy. The European Union's AI Act, for example, classifies AI used in high-risk domains (healthcare, law enforcement, education) under stricter obligations including transparency, human oversight, and accuracy requirements. As these frameworks mature, companies will face legal incentives to invest more seriously in hallucination mitigation.

VI. CONCLUSION

AI hallucination represents one of the defining technical and ethical challenges of the current era of artificial intelligence. Large Language Models have transformed how humans interact with information and technology, yet their tendency to generate confident falsehoods poses serious risks across healthcare, law, education, and beyond.

The root causes of hallucination — flawed training data, probabilistic generation, lack of grounding, and overconfidence — are deeply embedded in the current architecture of these systems. They cannot be fixed with a simple patch; they require sustained research across multiple fronts.

Encouragingly, solutions such as Retrieval-Augmented Generation, RLHF, Constitutional AI, and fact-checking pipelines have demonstrated real progress. As the field moves toward world models, multimodal reasoning, and stronger regulatory frameworks, the prospect of genuinely reliable AI systems becomes increasingly plausible.

Ultimately, the challenge of hallucination is not merely a technical one — it is a reminder that trust in AI must be earned through rigorous, verifiable, and honest engineering. The goal is not just AI that sounds intelligent, but AI that is dependably truthful.

REFERENCES

1. Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., & Sontag, D. (2022). Large language models are few-shot clinical information extractors. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 1998–2022.
2. Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., & Fung, P. (2023). A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. Proceedings of the 13th International Joint Conference on Natural Language Processing.
3. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623.
4. Bommasani, R., Hudson, D. A., Aditi, E., Altman, R., Arora, S., & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.
5. European Parliament. (2024). EU Artificial Intelligence Act. Official Journal of the European Union.
6. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Shi, E., Hoi, S. C. H., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38.
7. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, June 2026
Impact Factor: 8.2

8. Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 1906–1919.
9. OpenAI. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774.
10. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.
11. Salvagno, M., Taccone, F. S., & Gerli, A. G. (2023). Can artificial intelligence help for scientific writing? Critical Care, 27(1), 75.
12. Shen, T., Jin, R., Huang, Y., Liu, C., Dong, W., Guo, Z., Wu, H., Liu, Y., & Xiong, D. (2023). Large language model alignment: A survey. arXiv preprint arXiv:2309.15025.