International Journal of
Science, Architecture, Technology and Environment

Mitigating Hallucinations in Large Language Models

Dr. Vikas Joshi[1*], Dr. Pallavi Krishna Purohit[2]

[1,2]*Assistant Professor, Sangam University, Bhilwara, Rajasthan, India*

[*]*Corresponding author, vikas.joshi@sangamuniversity.ac.in*

DOI: https://doi.org/10.63680/ijsate0426112.080

Abstract

Large language models have become central to modern natural language processing because they can generate fluent, context-aware, and useful text across a wide range of tasks. However, their tendency to produce hallucinated content remains one of the biggest barriers to safe and trustworthy deployment. Hallucination can take the form of fabricated facts, unsupported claims, or confident statements that are not grounded in evidence. This paper reviews the causes of hallucination in large language models and examines the most widely used mitigation strategies, including prompt engineering, retrieval-augmented generation, verification pipelines, reasoning-based refinement, and model alignment. The discussion shows that no single technique fully eliminates hallucination; instead, systems that combine retrieval, reasoning, and post-generation checking tend to perform more reliably. The paper also highlights important implementation challenges such as latency, retrieval noise, evaluation difficulty, and the trade-off between factuality and usability. Finally, future research directions are outlined for building more transparent, reliable, and trustworthy language models.

Keywords: Large language models, hallucination, factuality, retrieval-augmented generation, verification, prompt engineering, alignment

I. Introduction

Large language models are now used in education, software development, search, summarization, and decision support. Their popularity comes from their ability to generate fluent and often persuasive text. Yet that fluency can also hide factual errors, and in many cases the model may sound more confident than the evidence allows. This mismatch between style and truth is what makes hallucination such a serious research problem.

Hallucination is not a single defect with one cause. It can arise from weak grounding, noisy or incomplete training data, ambiguous prompts, or generation strategies that favor plausible completion over evidence-based response. In practical terms, this means that hallucination reduction must be treated as a system design problem rather than only a model training problem.

This paper focuses on methods that can be applied in real systems. The goal is not merely to describe hallucination, but to organize the main mitigation ideas into a practical framework. The discussion is structured around four layers: prompt control, retrieval grounding, verification, and training-time alignment.

II. Hallucination Sources

Hallucinations often begin when the model lacks enough evidence to support a response. Because large language models are trained to predict likely text, they may produce an answer that is grammatically strong but factually weak. This is especially risky in technical, medical, legal, and academic domains where precision matters more than fluency.

Prompts can also trigger hallucination. If the instruction is vague, the model may infer missing details and present them as facts. If the prompt is overly broad, the model may compress multiple possibilities into a single confident statement. Recent surveys show that many hallucinations are not random errors but predictable outcomes of weak conditioning and incomplete context.

Another source is decoding behavior. A model that is tuned for creativity may introduce more variation, but that variation can also increase factual drift. For that reason, hallucination reduction must consider both what the model knows and how it chooses the final wording.

III. Detection Methods

Hallucination detection tries to identify unsupported content before the final answer is accepted. One common strategy is span-level detection, where the system highlights phrases that do not match the source evidence. This is useful for summarization and structured generation tasks because it can isolate the problematic portion instead of rejecting the whole output.

Another approach is self-consistency checking. If a model produces different answers under slightly different prompts, that inconsistency can be used as a signal of uncertainty. Research has shown that models may often recall real references accurately while becoming inconsistent when they invent them, which makes internal checking a useful clue for hallucination detection.

A more practical option is external verification. Here, the system compares generated claims with retrieved documents or a knowledge base. This is especially relevant in retrieval-augmented systems, where the model is expected to answer from evidence rather than from memory alone.

IV. Mitigation Methods

Prompt engineering is the simplest mitigation method. Clear instructions such as “answer only from the provided context” or “say when evidence is insufficient” can reduce unsupported output. This works best when the task is narrow and the prompt is well designed, but it is not enough for deeply uncertain or knowledge-heavy queries.

Retrieval-augmented generation is one of the most effective and widely used strategies. In this setup, the model retrieves documents from an external source and uses them to guide generation. This makes the system more grounded and especially useful for current information, domain knowledge, and enterprise search tasks. Recent surveys continue to identify RAG as a leading method for reducing hallucination, although its quality depends heavily on the retrieval stage.

Verification pipelines add a second layer of protection. After the model generates a draft, a checker compares the claims with evidence and either approves, revises, or rejects the response. This architecture is useful because it can be attached to an existing model without redesigning the full system. Its main weakness is added latency and the possibility of false rejection when the checker is too strict.

Alignment-based methods try to shape model behavior during training. Instruction tuning, preference optimization, and human feedback can make the model more careful and less likely to invent details. These methods are helpful, but the literature increasingly suggests that they work best when combined with retrieval and verification rather than used alone.

V. Comparative View

Method	Main Advantage	Main Limitation
Prompt engineering	Fast, low-cost, easy to deploy	Limited against deeper factual errors
Retrieval-augmented generation	Grounds answers in external evidence	Sensitive to retrieval quality
Verification layer	Detects unsupported claims	Adds latency and complexity
Alignment tuning	Improves general response behavior	Harder to evaluate and retrain

In practice, hybrid systems are more robust than single-method solutions. A common architecture is to retrieve evidence, generate a response, and then verify it before output. This layered design reduces the chance that one weakness will dominate the final result.

VI. Practical Issues

Hallucination mitigation is only useful if it can be deployed reliably. Retrieval can fail when the search layer returns irrelevant or incomplete documents. Verification can also fail if the checker lacks context or if the evidence is ambiguous. These limitations mean that each application needs its own tuning strategy rather than a universal recipe.

Latency is another important issue. A system with retrieval and checking stages will usually respond more slowly than a simple generation-only model. In some applications, such as enterprise search or academic support, this cost is acceptable because factual reliability matters more than speed. In conversational

systems, however, the trade-off must be carefully managed.

Evaluation is still a difficult research problem. Hallucination is not always fully false; sometimes it is partly correct, incomplete, or misleading in context. Because of this, simple accuracy metrics are not enough. Better evaluation often requires benchmark data, human judgment, and task-specific factuality measures.

VII. Future Direction

Future work will likely focus on adaptive systems that decide when to retrieve, when to answer directly, and when to abstain. This is important because not all prompts need the same level of caution. A model that can recognize uncertainty would be more useful than one that always answers or always refuses.

There is also a need for better benchmarks. Current evaluation sets do not fully represent real-world use, especially long-context reasoning, domain-specific terminology, and multi-step tasks. More realistic benchmarks would make it easier to compare mitigation methods fairly.

Another promising direction is combining knowledge graphs, retrieval, and reasoning in a single pipeline. Such systems may improve both factuality and explainability. That combination is especially valuable in academic, medical, and enterprise settings where users need not only a response, but also a justified response.

VIII. Conclusion

Hallucination reduction in large language models is best understood as a layered engineering challenge. Prompt control, retrieval, verification, and alignment each address part of the problem, but none of them is sufficient on its own. The most reliable systems are those that combine these methods according to task needs and risk level.

For an IEEE paper, the strongest writing style is careful, analytical, and specific. It should explain why a method works, where it fails, and how it fits into a larger system. The draft above follows that approach and can be expanded into a polished submission-ready manuscript. owl.

Declaration of Conflicting Interests

The authors declare no potential conflicts of interest with respect to the research, authorship and publication of this article.

Funding

The author received no financial support for the research, authorship and publication of this article.

References

A. Survey on Hallucination in Large Language Models, arXiv:2311.05232, 2023, revised 2024.

B. Jiang et al., "On Large Language Models' Hallucination with Regard to Known Facts," in Proc. NAACL, 2024.

C. A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models, arXiv:2401.01313, 2024.

D. Large Language Models Hallucination: A Comprehensive Survey, arXiv:2510.06265, 2025.aws.

E. Survey and analysis of hallucinations in large language models, Frontiers in Artificial Intelligence, 2025.

F. A hallucination detection and mitigation framework for faithful text summarization using LLMs, Scientific Reports, 2025.

G. Hallucinations in AI Models, IEEE Computer Society, 2025.YouTubecomputer

H. Mitigating Hallucination in Retrieval-Augmented and Agentic Systems, 2025 survey.

I. Do Language Models Know When They're Hallucinating?, EACL Findings, 2024.

J. FENJI at SemEval-2025 Task 3: Retrieval-Augmented Hallucination Span Detection, 2025.