IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁓ Peer-reviewed & Refereed journal ⁓ Vol. 15, Issue 2, February 2026

DOI: 10.17148/IJARCCE.2026.15203

Abstractive Summarization Via Contrastive Prompt Constructed By LLMS Hallucination

Kruthi S[1], Preksha M P[2]

Department of CSE, SJB Institute of Technology, Bengaluru, India[1,2]

Abstract: Recent progress in Large Language Models (LLMs) has significantly advanced natural language processing tasks such as summarization, translation, and text generation. Despite their impressive capabilities, these models frequently generate hallucinations—responses that appear fluent and convincing but lack factual correctness or logical grounding. Such behavior raises serious concerns regarding the dependability and ethical deployment of LLMs in real-world scenarios. This paper reviews and analyzes existing research on hallucinations in LLMs, focusing on their underlying causes, practical consequences, and mitigation strategies. Studies by Reddy et al. (2024) and Perković et al. (2024) investigate both internal model limitations and external influencing factors, including biased datasets, inadequate contextual understanding, and poorly structured prompts. Their findings highlight ethical and operational risks across multiple application domains. Research presented at ICALT 2024 emphasizes the dangers of hallucinated content in educational environments and proposes comparative and cross-verification techniques to preserve factual integrity. Furthermore, Sun et al. (2025) introduce a Markov Chain–based multi-agent debate framework that enhances post- generation verification through structured evidence retrieval and claim validation.

I. INTRODUCTION

Large Language Models such as GPT, LaMDA, and LLaMA have reshaped the field of artificial intelligence by enabling machines to generate coherent and contextually relevant human-like text. Built upon transformer architectures and trained on vast collections of textual data, these models exhibit strong performance across tasks including reasoning, summarization, question answering, and creative writing. However, alongside these achievements, LLMs exhibit a critical weakness: hallucinations. Hallucinations occur when a model generates information that is syntactically fluent and contextually plausible but factually incorrect, fabricated, or unsupported by reliable evidence. This phenomenon significantly undermines trust in AI-generated outputs, particularly in sensitive domains such as healthcare, law, and education. The emergence of hallucinations can be attributed to several interacting factors, including incomplete or biased training data, ambiguous or underspecified prompts, limited reasoning capabilities, and overfitting to surface-level patterns. Prior research, including the works of Reddy et al. (2024) and Perković et al. (2024), categorizes hallucinations into intrinsic types—where outputs conflict with provided input—and extrinsic types—where outputs introduce unverifiable or external information. These studies also emphasize the broader ethical and societal implications of unchecked hallucinations. Addressing hallucinations is therefore a fundamental requirement for ensuring the transparency, reliability, and ethical deployment of LLMs. A combination of improved dataset quality, enhanced prompt engineering, retrieval-based grounding, and automated verification systems is necessary to reduce hallucination frequency and impact.

II. METHODOLOGY

The methodologies examined across the reviewed studies span conceptual analysis, architectural evaluation, comparative validation, and computational verification. Together, they provide a comprehensive framework for understanding and mitigating hallucinations in LLMs.

1. Analytical and Conceptual Framework:

This approach employs descriptive and analytical methods to examine the origins, characteristics, and consequences of hallucinations in LLMs. Researchers begin by analyzing transformer architectures, focusing on attention mechanisms, encoder–decoder layers, and positional encoding. While these components enable flexible language generation, their complexity can also contribute to unintended hallucinated outputs.

Hallucinations are categorized into intrinsic and extrinsic forms. Intrinsic hallucinations arise from inconsistencies between the input and generated output, whereas extrinsic hallucinations introduce information that cannot be validated against the input or known facts.

This framework establishes a cause-and-effect relationship between hallucinations and contributing factors such as dataset bias, overfitting, and limited contextual reasoning. A multidimensional mitigation strategy is proposed to address hallucinations from multiple angles rather than relying on a single solution.

Dataset curation: is emphasized to reduce bias, remove low-quality data, and improve the overall representativeness of training corporation

Prompt engineering: techniques are applied to provide clearer contextual instructions, helping models generate more accurate and relevant outputs. **Human-in-the-loop verification:** is incorporated during both training and post-deployment phases to monitor model behavior and correct errors.

Regularization and filtering techniques: Regularization and data filtering techniques play an essential role in limiting overfitting and preventing incorrect memorization of training data within large language models. By constraining model complexity and removing noisy or low-quality data, these techniques help improve generalization and reduce the likelihood of hallucinated outputs.

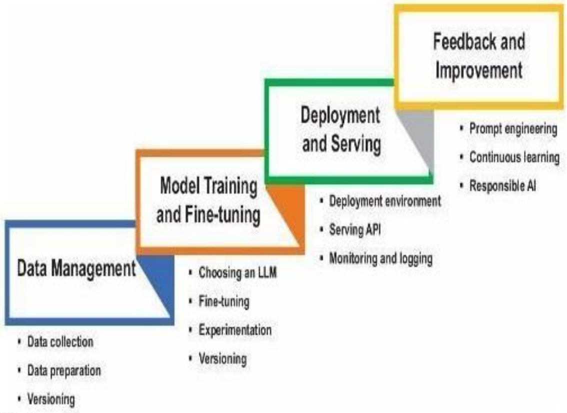


Fig. 2. LLMOps lifecycle.

Fig 1.1 LLM lifecycle

2. Mechanistic Process and Training-Based Examination:

This approach adopts a mechanistic and process-oriented perspective to explain how hallucinations are produced during text generation in LLMs. The analysis follows the complete training lifecycle, beginning with large-scale pretraining and extending through fine-tuning and reinforcement learning from human feedback (RLHF). It demonstrates that the reliance on probabilistic token prediction, when not anchored to verified factual information, can result in confident yet inaccurate responses.

This uses a cause-driven analytical model by examining:

Unreliable training data: can introduce hallucinations when datasets contain fictional material, outdated information, or unverified sources.
Limited logical reasoning: capabilities contribute to hallucinations, as models rely on probabilistic token prediction without built-in mechanisms for factual verification.
Vague or incomplete prompts: combined with overfitting, often lead to incorrect interpretations and unsupported responses.
Prompt refinement and segmentation: involve breaking complex instructions into smaller, clearer components to reduce ambiguity and improve response accuracy.
Multi-model verification: Responses generated through different LLMs are being compared for factual inconsistencies.

The Educational Cross-Verification and Comparative Approach:

This study presents a comparative and cross-validation framework specifically designed for the use of large language models in educational contexts. Given the high sensitivity of learning environments to mis-information, the proposed approach emphasizes systematic verification of AI-generated content through structured comparative reasoning.

Key methodological elements include:
Multi-model verification: Outputs generated by different large language models are compared to identify factual errors and inconsistencies.
Iterative validation loop: ach model output undergoes multiple stages of review to ensure factual accuracy and logical coherence.
Comparative evaluation metrics: Consistency among model responses is assessed using statistical comparison methods and correlation analysis.
Comparative reasoning and collective intelligence: The framework relies on multiple models and verification layers to improve reliability in educational applications and reduce the occurrence of hallucinated content.

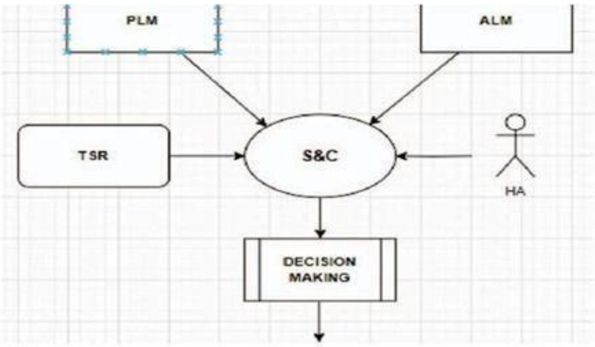


Fig 3.1 Comparative and Cross-Verification

1. Experimental and Computational Verification:

This work introduces a quantitative and experimentally driven framework for hallucination detection in large language models. The proposed method integrates Markov Chain–based modeling with multi-agent systems, where autonomous content.

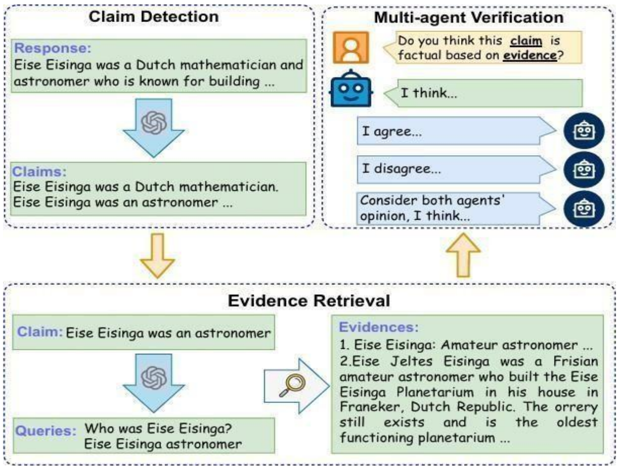
Claim Detection: The process begins by identifying factual statements from LLM-generated outputs, such as responses produced by ChatGPT or Factool.

Evidence Retrieval: Relevant supporting or contradicting evidence is collected using web search tools, APIs like Google Search, or local and vector-based databases.

Multi-Agent Verification The framework employs three agents—Trust, Skeptic, and Leader—to verify claims through probabilistic debate and consensus. It is validated on Factool and HaluEval datasets, achieving better accuracy and F1 scores than Self-Check and The ICALT approach further enhances educational AI by cross-checking multiple models to identify inconsistencies and improve factual reliability.

Responses generated by different systems are compared to detect inconsistencies or hallucinated content.

This cross-verification strategy strengthens factual reliability and promotes trustworthy educational AI systems.



Response: Eise Eisinga was a Dutch mathematician and astronomer who is known for building ...

Claims: Eise Eisinga was a Dutch mathematician. Eise Eisinga was an astronomer ...

Do you think this claim is factual based on evidence?

I think...

I agree...

I disagree...

Consider both agents' opinion, I think...

Claim: Eise Eisinga was an astronomer

Queries: Who was Eise Eisinga? Eise Eisinga astronomer

Evidences:
1. Eise Eisinga: Amateur astronomer ...
2.Eise Jeltes Eisinga was a Frisian amateur astronomer who built the Eise Eisinga Planetarium in his house in Franeker, Dutch Republic. The orrery still exists and is the oldest functioning planetarium ...

Fig 4.1 Overview of 3 stages

DOI: 10.17148/IJARCCE.2026.15203

III. CONCLUSION AND DISCUSSION

Discussion

Consequences and Ethical Implications

Misinformation and trust issues: Incorrect or fabricated AI outputs can mislead users, spread false information, and gradually reduce confidence in AI- based systems.

Legal, ethical, and reputational risks: Errors in sensitive domains such as healthcare, law, or education may lead to legal consequences, ethical concerns, and damage to an organization's public image.

Legal and social risks: Incorrect outputs in fields like healthcare, law or education can cause ethical and legal complications.

Reputational harm: Organizations using hallucination prone AI will face public skepticism.

Mitigation by Data and Model Improvements: Improved data quality: Curating reliable training datasets through cleaning, filtering, and balancing helps minimize bias and factual errors.

Effective prompt design: Crafting clear, context-rich prompts guides the model toward more accurate and relevant outputs.

Model optimization: Regularization and fine-tuning techniques reduce overfitting and improve the model's ability to generalize facts.

Human-guided learning: Reinforcement Learning

from Human Feedback (RLHF) aligns model behavior with human judgment and factual correctness. sets and responsible deployment procedures to create safer and more reliable AI systems.

Why It's Difficult to Get Rid of Hallucinations

Errors are always present in large datasets, therefore they cannot be cleaned. Models cannot distinguish fiction and non-fiction without explicit training. LLMs do non internally. Verify sentence before creating one. LLMs prioritize fluency over factual accuracy, which Leads to confident but incrorrect. This cross-verification strategy strengthens factual reliability and promotes trustworthy educational AI systems.

Conclusion

Hallucinations remain one of the most persistent challenges in large language models. Despite significant advancements, these models still struggle to consistently distinguish accurate information from incorrect content due to limitations in training have shown significant gains in reducing these errors. By adding layers of expert over- sight, external grounding, and evidence-checking, these tactics together increase the dependability of model output.

In conclusion, it is possible to greatly lessen but not completely eliminate hallucinations. In summary, while hallucinations cannot be entirely removed, they can be significantly minimized. Achieving safer and more dependable AI systems requires continuous research, stronger verification mechanisms, improved datasets, and responsible deployment practices. Ongoing attention to accuracy and transparency will be essential for the safe use of LLMs across diverse application domains.

IV. ACKNOWLEDGMENT

We express our heartfelt gratitude to the Department of Computer Science and Engineering at SJB Institute of Technology for their unwavering support and resources through- out this research. We also extend our thanks to the Kaggle community for providing open-source datasets critical to this project. The encouragement and guidance from our peers and mentors have been instrumental in shaping this work, and we are deeply appreciative of their contributions. Special thanks to out guide **Mrs. Srinidhi KS.**

Who gave us lot of support in building the project and making everything come in a organized way

REFERENCES

[1]. K. P. Reddy, S. Kautish, and A. S. Sidhu, "Hallucinations in Large Language Models (LLMs)," 2024. Available in: *Hallucinations_in_Large_Language_Models_LLMs.pdf.*

[2]. T. Perković, K. Grgić, and S. Kezić, "Hallucinations in LLMs: Understanding and Addressing Challenges," 2024. *Hallucinations_in_LLMs_Understanding_and_Addressi ng_Challenges.pdf.*

[3]. V. Saini and T. Kaur, "Mitigation of Hallucinations in Language Models in Education: A New Approach of Comparative and Cross-Verification," in *Proc. 2024 IEEE Int. Conf. Advanced Learning Technologies (ICALT)*, pp. 207– 209, 2024.

[4]. C. Sun, Y. Zhang, J. Guo, J. Liu, Z. Xu, and C. Huang, "Towards Detecting LLM Hallucination via Markov Chain- based Multi-agent Debate Framework," 2024. Available in: *Towards_Detecting_LLMs_Hallucination_via_Markov_Chain-based_Multi-agent_Debate_Framework.pdf.*

[5]. *Lin, S., Hilton, J., & Evans, O. (2021). Rethinking evaluation of language models: Beyond accuracy.*