Sridaran Rajagopal · Priti Sajja ·
Rohit Thanki · Ajay Kumar (Eds.)

Communications in Computer and Information Science 2823

Artificial Intelligence Based Smart and Secured Applications

4th International Conference, ASCIS 2025
Gujarat, India, September 11–13, 2025
Revised Selected Papers, Part V

Part 5

Springer

A Novel Approach for De-Biasing and Enhanced Hallucination Mitigation in LLM

M. K. Shamseera and R. Durga(✉)

Department of Advanced Computing and Analytics, VISTAS, Chennai, India
durga.scs@vistas.ac.in

Abstract. Hallucination in large language models refers to the typical mistakes we encounter when relying on AI systems, such as ChatGPT-like models, in our daily lives, a situation where the model produces inaccurate, illogical, or fake text. This happens because large language models (LLMs), which produce text based on patterns, are neither databases nor search engines and associations discovered in their training data instead of referencing particular references. Resolving hallucinations is crucial to promoting constructive human-AI interactions and increasing confidence in AI-generated material. Monitoring hallucinations is undoubtedly challenging, but it can be a game-changer as it enhances data verification at the source level and beyond.

Keywords: Hallucination in AI · Large Language Models (LLMs) · Knowledge grounding

1 1Introduction

The development of AI has fundamentally changed how people and computers interact we humans depend on the system in all aspects of applications to improve performance. These trends encompass the fields of academia, medical applications, and sustainability advancement. The trend itself has evolved from simple rule- based chatbots to more general concepts known as Large Language Models (LLMs), which produce humanized text, images, and content that support adaptive learning, robotics, and the metaverse. The majority of LLM models use deep learning approaches to analyze the environment and produce the desired outcome, which includes jobs like generating code, translating languages, summarizing information, and answering questions. In 1966, MIT researcher Joseph Weizenbaum created ELIZA, a simple chatbot that predicts talks using pre-programmed rules, which marked the beginning of large language models (LLMs). Eliza's conversation Despite a lack of full understanding, "Hello Eliza" marked the beginning of natural language processing (NLP) research and set the stage for further development and more complex LLMs. The term 'LLM' became widely associated with and popular with the GPT family, which began to gain attention after its introduction in 2018. GPT-1 was introduced in 2018, GPT-2 in 2019 2020 with GPT-3the GPT-4 mostly these GPT are based on transformer architecture which helps advancement of robotics, software engineering, social impact and now we reach up to Gemma 3 in 2025, in the era of giants like LCM (Meta Large Concept Model).

R. Sridaran et al. (Eds.): ASCIS 2025, CCIS 2823, pp. 178–186, 2026.
https://doi.org/10.1007/978-3-032-17846-6_11

LLM s are powerful but always facing challenges while dealing with the hallucinating behavior so designing one with Adaptive AI where the Models That Evolve in Real Time, LLMs in robotics, virtual reality, and the metaverse are advanced by personalized AI assistants that are smarter, more contextually aware, and go beyond text. Consequently, minimizing hallucinations in large language models (LLMs) is one of the primary focuses of artificial intelligence research. These models, while incredibly powerful, can produce inaccurate or fictitious outputs due to limitations in their ability to verify information or understand context fully. Tackling this challenge is vital for improving trust in AI systems and ensuring their outputs align with factual and reliable information. Hallucination can be either intentional or nonintentional in some cases, but it misleads to the progress of data retrieval. Grounding LLM becomes unavoidable as it Prevents Hallucinations, so as we lack grounding LLM s are hallucinating. Effective mitigation strategies combine advanced techniques such as knowledge grounding, contrastive learning, and consistency modeling with enhancements in training data quality and contextual understanding that fosters better response for the queries. Using contrastive examples, contrastive learning improves examples to more accurately convey our intent. This entails giving both good examples that highlight the real one and negative examples that highlight the traits of LLMs to steer clear of to lessen hallucinogenic actions. Contextual awareness in LLM is crucial as it ensures the model remains relevant to specific and negotiates to avoid inaccurate replays. By integrating dynamic knowledge retrieval systems and reinforcement learning methods, researchers are developing hybrid approaches that significantly improve the reliability of LLMs. These advancements aim to address hallucination at its core, fostering a future where AI-generated content becomes a dependable tool for communication and decision-making.

1.1 Contribution of This Paper

This study indicates novel approaches to addressing this hallucination problem, especially with primary focus on advancing prompt engineering and developing better models with cost cutting to alleviate duplications that cause bigger data crashes.

Review focus on current developments and trends also provides extensive overview and concerns about this matter now the days. The main motivation of this work was to offer guidance and valuable insights for future research works through real word scenarios and theoretical analysis.

1.2 Organization of This Research

Section 1 describes the LLM in detail, the process of hallucination, and contribution of this work. Section 2 defines the related works for monitoring hallucination and mitigation strategies in LLM related models. Section 3 illustrates the proposed methodology for task mitigation strategy in the LLM. Section 4 carries out the result, and finally, the conclusion of this research is represented in Section 5.

Objectives

The study examines the practice of AI approaches for envisioning coronary channel ailment utilizing a dataset from 462 healing records and number features from the

Southwestern Human inject ailment dataset. The k-wealth recipe and gummy youth oversampling model were used to calculate the trouble of unstable dossier. An approximate study of various machine intelligence methods, containing LR, SVM, KNN, as well as ANN, was administered to forecast coronary channel ailment occurrences from dispassionate dossier. SVM explained the highest veracity rate (78.1%). Ahmad G. N. and other's study distinguished the act of differing algorithms for coronary ailment classification, attaining a wonderful categorisation veracity of 100% with the RF rule. The study plans utilizing metaheuristic systems like the Coelenterate rule to raise visage from the congestive heart failure dataset together with apply this in the Organization Acquisition plan to separate active and unsound pour ailment assemblages. The Siphonophore algorithm was preferred on account of allure speedy of order and quality in gestating appearance.

2 Literature Survey

It is continuously revolutionary and a subject of continuing research to decrease hallucinatory behavior in order to strengthen large language models (LLMs). Speaking about this paradigm shift were Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Hao a Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu.in 2023 by means of A Survey on Hallucination in Large Language Models: Fundamentals, Classification, and Difficulties. The study draws attention to the difficulties that retrieval-augmented LLMs confront and suggests areas for further study, such as hallucinations in huge language models and comprehending their knowledge bounds. Researchers that contribute to the authenticity of the data and created contents should further investigate the overview of hallucination detecting techniques and benchmarks. Large language models' (LLMs') performance on natural language tasks is hampered by two primary biases. These are memorizations on bias and the application of statistical analysis.

The authors Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman (2023) offer a comprehensive study based on the criteria. By looking into LLM families like PaLM, GPT-3.5, and LLaMA., in controlled tests, the study highlights the need for better techniques to assess and minimize hallucinations in AI-generated content. According to the authors entity-based indexing to retrieve memorized data bias and performance degradation as a result of statistical usage patterns are the main study areas.

To detect hallucinated references without external verification also discussed by the authors Agrawal A, Mackey L, Kalai AT (2023) the major problems they encountered during the study were about Inaccessible training data, Hallucination spectrum and Prompt sensitivity. to overcome these limitations several efficient techniques were proposed. several in depth survey studies were also conducted by the researchers to understand the flow of AI generated contents, working principles, security, threats were also pointed out detailed in these studies ethical and societal implications are the major concern of these studies. Later in 2024, Hanyu Duan, Yi Yang, and Kar Yan Tam conducted a thorough empirical analysis about the LLM. status of hallucination was done

but the problem they encountered were they didn't differentiate the categories of hallucination such as factual fabrication, instruction inconsistency, logical inconsistency and so on. And not detailing with which layer influences model response.

Mitigating Hallucination in Large Language Model by Leveraging Decoder Layer Contrasting, a 2024 study by Guangsheng Liu, Xinbo Ai, Wenbin Luo, and Ange Li, looks at methods that compare the probability distributions of intermediate and final layers during inference to ascertain the next-token distribution. This method significantly improves benchmarks like GSM8k, StrategyQA, and Wiki Factor by successfully addressing the problem of uneven output distributions from lower decoder levels. The semantic entropy technique was employed by Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal (2024).

An entropy-based uncertainty estimator was developed to predict the uncertainty of the inputs provided by user. The researchers suggest an entropy- based uncertainty estimator to pinpoint confabulations, a subset of hallucinations in which LLMs produce erroneous and arbitrary responses. Unlike traditional approaches that rely on specific datasets or task-specific supervision, Instead of concentrating on individual word sequences, this technique calculates uncertainty at the semantic level.

The concept of DE hallucination techniques will help us in some percentage like retrieval-augmented generation, the method by improving the factual correctness of LLM outputs. Hallucination mechanisms in Retrieval- Augmented Generation (RAG) models, which use outside knowledge to lower mistakes, are the subject of this paper. Even with precise retrieval, hallucinations happen when Copying Heads are unable to efficiently integrate or retain returned content, and Knowledge Feed-Forward Networks (FFNs) place an excessive amount of emphasis on internal parametric knowledge. In response, the authors present ReDeEP, a unique technique that separates the ways in which LLMs use internal and external knowledge to enhance hallucination detection.

Additionally, the authors Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Yang Song, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li offer AARF, a mitigation technique that enhances the precision of RAG models by balancing the contributions of Copying Heads and Knowledge FFNs. The recent work of Fervers, P. et is also helpful to understand LLM family member GPT 3.5 performance on structured vs. unstructured reports and found that ChatGPT had low accuracy, correctly classifying 53% of free-text reports and 44% of structured reports.

In order to increase the value of the LLM output at particular tasks, prompt tuning—which entails synchronizing the instructions given to a pre-trained LLM during the fine-tuning phase—is also essential to mitigation. However, S.M. Towhidul Islam Tonmoy1, S. M. Mehedi Zaman1, Vinija Jain3,4∗, Anku Rani2, Vipula Rawte2, Aman Chadha3,4∗, and Amitava Das2 in 2024 pointed out that prompt engineering still faces the common problems of unclear instructions, ambiguous task definitions, insufficient training data, and bias. So advance prompt engineering will improve the performance and reliability of AI generated contents thus by increase in the trust in the LLM contents

Table 1. DIFFERENT APPROACHES USED TO IDENTIFY HALLUCINATION IN LLM

Author	Dataset	Parameter	Description
Lorenz Kuhn, Jannik Kossen, Yarin Gal, and Sebas a Farquhar - 2024	dataset-independent approach	Semantic Entropy LexicalVs Semantic Uncertainty Output Distributionin Meaning-Space RandomSeed Sensitivity	The research progresses through the extent to which LLM is likely to produce unreliable outputs
Amitava Das, Anku Raji, Vipula Rawte, S.M. Towhidul Islam Tonmoy, S.M. Mehedi Zaman, Vinija Jain, and Aman Chada - 2024	TriviaQA SQuAD BioASQ NQ SVAMP CoQA SciQ MedHallu	Retrieval-Augmented Generation, Knowledge Retrieval, Supervised Fine-Tuning(SFT), Prompt Engineering	The paper explores over 32 types to mitigate hallucinations, categorizing them based on dataset categories
Fervers et al. - 2024	205 MRI scans of 150 paatients	Cohen's kappa Intraclass correlation coefficient Meanabsolute error Chi-square test	the study explores the power of ChatGPT in medical applications

Table 1 indicates the application, performance evaluation, datasets, and parameter of the methods obtained to identify hallucination process.

3 Proposed Work

The section describes the proposed method for detecting hallucination in LLM through advanced prompt engineering with feedback and reasoning and better model creation and development called as RLLMWRH (Reliable Large Language MODELS With Reduced Hallucination) using SFT teacher student approach. Advanced prompt engineering will alleviate hallucination with improved outcome and. The better model creation and development limits the chances of unbiased data production (Fig. 1).

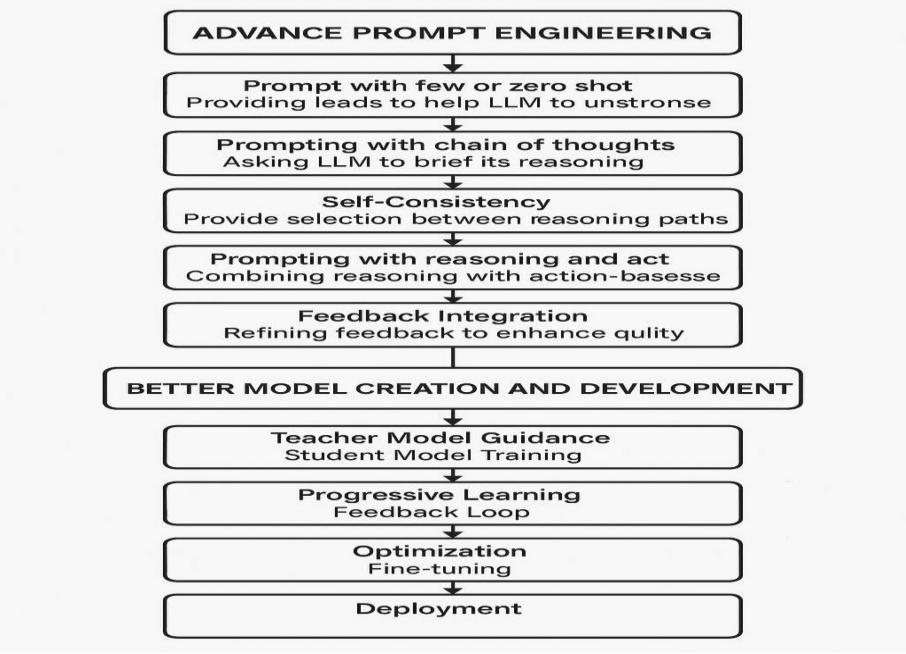

Fig. 1. The Proposed Architecture Diagram Based RLLMWRH

3.1 Advance Prompt Engineering

The process of selecting the best output from the list of samples one or the derived one are prompt engineering. This is the best method to alleviate hallucination in specific context with the expected outcome in LLM. it can be accomplished through several methods but in this proposed model we try to advance the features of prompt engineering through feedback and reasoning method. The method progress as reasoning at different levels – helps LLM to split complex tasks into smaller logical steps that improves accuracy and coherence.

1. Prompt with few or zero shot – Providing leads to help LLM to understand the expected response.
2. Prompting with chain of thoughts – Asking LLM to brief its reasoning before conclusion.
3. Self-Consistency – provide selection between reasoning paths.
4. Prompting with reasoning and act – Combining reasoning with action- based response to understand LLM behaviors.
5. Feedback Integration – refining feedback to enhance response quality.
6. Automated Response Optimization –measure the response timing.

3.2 Better Model Creation and Development

Development of novel systems is a systematic process which requires better understanding and applying the best algorithms along with the quality of data. Proposed Model

focuses model creation through supervised fine tuning with teacher student approach. Supervised fine-tuning (SFT) are always better options while targeting model proposals which can be investigated using methods like Teacher-Student Approaches and Knowledge Injection by Elaraby et al., 2023. But this study needs more examples to identify real potential. Study of Elaraby et al., 2023 with BLOOM7B, GPT-4 was the basis of my prior statement. Selecting the methods that will be apt for better model integration also matters in the design phase of algorithm considering the cost of model deployment. Key elements of SFT includes variables such as Teacher Model Guidance – A giant more LLM generates responses to students Student Model Training, Progressive Learning, Feedback Loop & Optimization & Deployment.

3.3 Feedback and Reasoning (FR)

This section calculates mainly through reasoning timing along with feedback. we can calculate it through Measures such as (T_p) – Time taken by the model to process the input prompt. (T_g) – Average time taken per token before the full response begins. Reasoning Step Time (T_r) – Time spent on logical steps before forming Final Response Time (T_f) – Total time from receiving input to delivering the final output.

$$Set\, T\_\{reasoning\} \;=\; (T\_f \;-\; T\_p \;-\; T\_g)\,] \tag{1}$$

$$(T\_\{feedback\}) = Time\ spent\ analyzing\ user\ feedback\ and\ refining\ the\ response. \tag{2}$$

$$hence[T\_\{reasoning + feedback\} = (T\_f - T\_p - T\_g) + T\_\{feedback\}] \tag{3}$$

3.4 SFT with Teacher Student Approach

This section discusses the method for Implementing the Teacher-Student Approach in LLMs The goal is to optimize the student model by learning from the teacher model, which helps to balance between computational costs without compromising performance.

$$[L_S = \alpha L_{hard} + (1 - \alpha) L_{soft}] \tag{4}$$

With variable definition

(T) = Teacher Model (larger, more capable) (S) = Student Model (smaller, optimized)

(L_T) = Function of loss in the teacher model (L_S) = Function of loss in the teacher model

(D) = Distillation factor (how much knowledge is transferred)

(\tau) = Temperature parameter (softens probabilities for better learning)

(\alpha) = Weighting factor between teacher guidance and student learning

$$[L_{soft} = D \cdot KL(S_\tau || T_\tau)] \tag{5}$$

T (KL) is the Kullback-Leibler divergence, for the student's probability distribution is from the teacher's.

(S_{\tau}) and (T_{\tau}) are the student and teacher model outputs softened by temperature (\tau).

4 Results and Discussions

With the proposed model by integrating Advanced Prompt Engineering, Feedback & Reasoning Optimization, and Supervised Fine-Tuning (SFT) with a Teacher- Student Approach we can measure the performance metrics such as coherence and accuracy are to be improvised Also by combining these two techniques we expect minimizes latency, ensuring timely, logically structured replies in the context of LLM.

5 Conclusion

DE hallucination in LLMs is not just recommended but mandatory as it leads to performance enhancement, faster output retrieval, reliability and adaptability in LLM. Using hybrid approaches that combine mitigation strategies with debiasing techniques enhances data quality and improves the effectiveness of applied data across all aspects of research and studies in this era. To guarantee the dependability of AI-generated content and boost its adoption over time, it is imperative that big language models address hallucinations. We can promote more precise, significant human-AI interactions and increase confidence in AI systems by enhancing monitoring and verification methods. Advancements in retrieval-augmented generation, fine-tuning techniques, and human oversight will be essential in mitigating hallucinations, making AI a more reliable tool across various applications. Other technologies that could be integrated with the current system will undoubtedly enhance existing LLMs in the future, ensuring continuous upgrades without delay. Hopefully, this will become a golden feather in the journey of LLMs.

References

Wang, Y., et al.: Multimodal Chain-of-Thought Reasoning: A Comprehensive Survey (2025). https://github.com/yaotingwangofficial/Awesome-MCoT

Farquhar, S., Kossen, J., Kuhn, L., Gal, Y.: Detecting hallucinations in large language models using semantic entropy. Springer Nature **630**, 625–630 (2024). https://www.nature.com/articles/s41586-024-07421-0

Towhidul Islam Tonmoy, S.M., et al.: A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models (2024). arXiv:2401.01313V3

Agrawal, A., Mackey, L., Kalai, A.T.: Do language models know when they're hallucinating references? (2023). arXiv preprint arXiv:2305.18248

Barocas, S., Hardt, M., Narayanan, A.: Fairness and machine learning: limitations and opportunities. fair mlbook.org? (2019). http://www.fairmlbook.org

Abdulrahman, E., Abdelrahim, F., Fathi, M., Firass, A., Ali, K.: ChatGPT and the rise of semi-humans. Humanit Soc. Sci. Commun. **10**(1), 626 (2023). https://doi.org/10.1057/s41599-023-02154-3

Ahmad, I., Yousaf, M., Yousaf, S., Ahmad, M.: Fake News Detection Using Machine Learning Ensemble Methods. Complexity, 1–11.h (2020)

Fervers, P., et al.: ChatGPT yields low accuracy in determining LI-RADS scores based on free-text and structured radiology reports in German language. Front. Radiol. **4**, 1390774 (2024). https://doi.org/10.3389/fradi.2024.1390774

Lee, K., et al.: Deduplicating training data makes language models better. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, pp. 8424–8445 (2022)
McKenna, N., et al.: Sources of Hallucination by Large Language Models on Inference Tasks. Journal of ArXiv :2305.14552 (2023). https://doi.org/10.48550/arXiv.2305.14552
Blodgett, S.L., Barocas, S., Daumé III, H., Wallach, H.: Language (technology) is power: a critical survey of "bias" in NLP. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp. 5454–5476 (2020)
Huang, L., et al.: A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. Journal of ArXiv, volume abs/2311.05232 (2023). https://api.semanticscholar.org/CorpusID:265067168
Wang, Y., Pan, Y., Yan, M., Su, Z., Luan, T.: A Survey on ChatGPT: AI–Generated Contents, Challenges, and Solutions. IEEE Open J. Comp. Soc. **4**, 280–302 (2023). https://doi.org/10.1109/OJCS.2023.3300321
Zhou, X., Zafarani, R.: A survey of fake news: fundamental theories, detection methods, and opportunities. ACM Comput. Surv. **53**(5), 1–40 (2020). https://doi.org/10.1145/3395046
Ji, Z., et al.: Survey of hallucination in natural language generation. ACM Comput. Surv. **55**(12), 1–38 (2023). https://doi.org/10.1145/3571730
Christensen, J.: Understanding the role and impact of Generative Artificial Intelligence (AI) hallucination within consumers' tourism decision-making pro cesses. Curr. Issues in Tourism (2024). https://doi.org/10.1080/13683500.2023
Lu, J., et al.: YODA: teacher-student progressive learning for language model [2401.15670] YODA: Teacher-Student Progressive Learning for Language Models (2024)