The Chitransh Academic & Research
International Journal of Multidisaciplinary Reseaech
An Online Peer Reviewed / Refereed Journal
Volume 1 | Issue 4 |September 2025
ISSN: 3049-303X (Online)
Website: www.thechitranshacadmic.in

HALLUCINATIONS IN LARGE LANGUAGE MODELS: CLASSIFICATION, REAL-WORLD IMPACT, AND MITIGATION STRATEGIES

Tanzeel Baig[1], Manju Vyas[2], Anima Sharma[3], Swati Vijay[4], Neelkamal Chaudhary[5]
[1]Student
[2,3,4,5]Assistant Professor
[*] Department of Artificial Intelligence and Data Science, Jaipur Engineering College and Research Centre

ARTICLE DETAILS

Research Paper

Received: **30/08/2025**

Accepted: **10/09/2025**

Published: **30/09/2025**

Keywords:

Hallucination, Large Language Models, LLM security.

ABSTRACT

Natural language processing has greatly advanced thanks to Large Language Models (LLMs), opening up applications in everything from decision-making to content creation. But they frequently produce hallucinations, which are believable-looking but factually inaccurate outputs. This essay distinguishes between intrinsic, extrinsic, amalgamated, and non-factual hallucinations, emphasizing factors such as knowledge interference and contextual misalignment. We look at the dangers of such mistakes in delicate domains where false information can have detrimental effects, such as journalism, science, and healthcare. We also recognize the artistic value of hallucinations in fields like literature and the arts. The study examines current mitigation and detection strategies. It concludes by outlining future directions centered on enhancing evaluation, ethical standards, and knowledge integration.

I. Introduction

Large Language Models (LLMs) have driven significant breakthroughs in artificial intelligence, particularly in natural language understanding and generation. Their abilities now encompass tasks such as code completion and development support, as well as applications in computer vision and scientific computing, marking a transformative advancement in AI [1].

Despite these innovations, a notable issue persists: LLMs frequently "hallucinate," generating output that appears plausible yet is factually incorrect, nonsensical, or invented [2]–[6]. These hallucinations erode user trust and can lead to serious consequences in sectors like healthcare, journalism, and scientific research [7]–[10].

Consequently, minimizing hallucination has become a focal point in AI research. Efforts are being made to detect and mitigate these inaccuracies while maintaining high standards of factual accuracy [7]. The ultimate objective is to enhance LLMs' generative strength without compromising reliability—and when possible, channel any creative aspects into beneficial directions [11].

II. Types and Causes of Hallucinations

Hallucinations in large language models (LLMs) refer to outputs that stray from factual accuracy or fail to align with the provided context. This section presents a taxonomy of these hallucinations and examines their root causes.

A. Taxonomy of Hallucinations

1. Intrinsic Hallucinations: These stem from limitations within the model's internal knowledge representation. They occur when the model produces information that contradicts what it has learned during training. Such hallucinations typically arise from the model's inability to accurately retrieve or represent stored knowledge in its parametric memory [12].

Corresponding Author: tanzeelbaig.ai27@jecrc.ac.in

2. Extrinsic Hallucinations: These are caused by errors in processing or integrating external information, such as the user prompt or accompanying context. In multimodal systems, extrinsic hallucinations can manifest visually—for instance, when a model misdescribes elements of an image during visual question answering [13].
3. Amalgamated Hallucinations: These result from the model inappropriately merging multiple facts or conditions present in a prompt. When several constraints are provided, the model may incorrectly conflate or blend them, leading to an output that combines disparate pieces of information inaccurately [14].
4. Non-Factual Hallucinations: These involve the generation of claims that directly contradict established facts or empirical knowledge. This type of hallucination is particularly concerning in tasks that demand domain-specific accuracy, as it can disseminate misinformation [15].

B. Root Drivers behind LLM Hallucinations

Several foundational dynamics within large language models contribute to their propensity to hallucinate:

1. Knowledge Overshadowing: When certain information patterns dominate, they can suppress lesser-known details. This imbalance—with some facts receiving disproportionate attention—leads the model to favor dominant over subtle knowledge, causing it to generate inaccurate or incomplete responses [14].
2. Shallow Representation of Knowledge: Inadequacies in how facts are encoded, especially within early neural layers, can lead to the model "imagining" extra details. Known as knowledge-enrichment hallucinations, this happens when the model lacks solid recall of specific topics and fills in gaps with unfounded content [15].
3. Missteps in Attribute Extraction: Errors in later attention layers may cause the model to misread or incorrectly extract key attributes. This results in answer-extraction hallucinations, where essential information isn't properly identified, leading to faulty answers [15].
4. Context vs. Prior Conflict: When a prompt's context clashes with internal memory, the model may favor its learned priors over the immediate input. These contextual

Corresponding Author: tanzeelbaig.ai27@jecrc.ac.in

misalignment hallucinations occur when the model's internal biases overshadow the prompt's requirements, leading to irrelevant or inaccurate outputs [11].

5. Semantic Entropy as a Hallucination Signal:This measure quantifies variation across multiple model outputs to identify inconsistencies. High semantic entropy indicates that different generations diverge significantly in meaning—an effective signal of potential hallucination due to uncertainty or insufficient knowledge [11].

C. Factors Influencing Hallucinations

1. Scale and Model Architecture: An LLM's hallucination behavior is influenced by its design and parameter count. Larger models have the potential to produce more hallucinations due to their increased complexity, even though they frequently store more information [12].
2. Quality and Diversity of Training Data: An important factor is the training corpus's balance and diversity. Hallucinations are more likely to be produced by models that were trained on skewed, biased, or small datasets, particularly when they come across contexts that were not adequately represented during training [16].
3. Dynamic Nature of Language Model Reasoning: An LLM's reasoning process is flexible, adapting to different input contexts, tuning techniques, and optimization strategies. This fluctuating behavior makes it more difficult to identify and address the causes of hallucinations [18].
4. Task Complexity: The task's actual nature is also important. LLMs encounter more uncertainty as tasks become more complex or ambiguous, which heightens their propensity to "fill gaps" with hallucinogenic material [17].

D. Implications and Future Directions

Enhancing the reliability of LLMs in practical applications requires an understanding of the different types of hallucinations and their causes. Future initiatives ought to focus on:

1. Strong Assessment Methods: Systematic assessment of hallucination tendencies across model types and languages will be made possible by the establishment of more robust and thorough benchmarks, such as the HalluQA benchmark created for Chinese LLMs [17].

Corresponding Author: tanzeelbaig.ai27@jecrc.ac.in

2. Enhancing Long-Chain Reasoning: Developing techniques that facilitate long, cohesive chains of reasoning can help avoid factual drift and hallucination in complex sequences, particularly in fields where accuracy is crucial (such as healthcare, legal assistance, and education) [19].
3. Investigating Processes of Dynamic Reasoning: It is possible to identify hallucination triggers and gain a better understanding of the cognitive pathways that result in ungrounded outputs by analyzing how an LLM's internal reasoning changes during inference [18].
4. Designing Factual Recall Pipelines: Investigating cutting-edge methods to lessen hallucinations, like reestablishing language models' internal fact recall pipeline to enhance factual precision [15].

In order to create more reliable and accurate LLMs and enable safer and more efficient deployment across a variety of applications, it will be essential to address these issues.

III. IMPACT OF HALLUCINATIONS

The phenomenon of hallucinations in relation to Large Language Models (LLMs) continues to garner attention for both its challenges and opportunities across different domains. Here I look at the implications of high-stakes decision making in terms of hallucinations and seek to highlight their ostensibly unintended advantages in creative domains while incorporating as much new evidence and views as possible.

A. Consequences in Critical Applications

Risks associated with hallucinations in LLMs for precision medicine, cutting edge scientific research, and journalism pose unprecedented challenges in these fields where trust is sacred. Such realms which involve advanced decision-making processes are especially vulnerable because of the verifiable trust on automations and systems in place [21].

In parts of the world where patients can be harmed with AI-generated hallucinations in medicine as well as a wrong AI-driven diagnosis, an illogical AI-driven treatment protocol based on model hallucinations, or erroneous interpretations of clinical trial data can all result in severe inaccuracies that imperil patient health, undermine trust in research, and erode confidence in AI

Corresponding Author: tanzeelbaig.ai27@jecrc.ac.in

technology. These machines need to be understood fully in their implementation as well as their serious consequences when advanced negligence happens for people as well as professionals dealing with this tech. The social impact of hallucinations is not limited to certain applications mentioned that hallucinations in AI can lead to societal consequences, which is further complicated by the fact that they can exist in the social environment [22].

They proposed a framework for measuring the effects of misinformation from AI hallucinations on social networks and provided evidence to determine that LLMs produce hallucinated information in online social networks, and that those inconsistencies are widely circulated. This example gives real urgency to the need for strategies designed to prevent the spread of incorrect information on a larger scale across complex social settings.

Hallucinations also prevent an LLM from fulfilling its intended function in settings where high precision is required, for example in the training of motor skills. Qiu (2024) showed that a LLM-assisted learning with hallucinated outputs could limit the performance in accessing activities (e.g., effecting a sport) and where hallucinated outputs lowered performance for badminton training. This discovery emphasizes assessing a LLMs' validity in fields related to precision and the related physical skill to use in Physical Education [23].

B. Potential Benefits in Creative Processes

While hallucinations are frequently viewed as a drawback, they can also foster creativity and original thought in LLMs. According to Jiang et al. (2024), by presenting novel linkages and seemingly erroneous results, these outputs may encourage creativity [24].

Similar to divergent thinking, these hallucinations can provide odd or unexpected ideas in fields like literature, design, and the arts. This can generate original plots, creative works, or design components that might not emerge from more conventional methods. For instance, MyStoryKnight, a storytelling platform created by Yotam et al. (2024), combines AI-generated characters with human storytelling to increase narrative originality through the use of LLM hallucinations. This illustrates how LLMs and human creativity can work together to provide interesting and creative outcomes [25].

Corresponding Author: tanzeelbaig.ai27@jecrc.ac.in

Furthermore, even if knowledge is not factually accurate, the capacity to combine it in novel ways can foster innovative problem-solving. This trait is particularly helpful in domains like product creation, where innovative pairings can result in ground-breaking discoveries.

C. Juggling Benefits and Risks

In LLMs, hallucinations pose a complicated problem since they can be both a risk and an inspiration. A deliberate and well-rounded strategy is needed to manage this dual nature effectively. Hallucinations can have a substantial impact on user happiness and trust, emphasising the value of openness and user awareness in lowering these worries [26]. Sustaining user trust is essential for wider adoption as AI systems are incorporated more and more into everyday life and important applications.

Researchers are continuously looking for ways to reduce hallucinations in order to solve these problems. Tonmoy et al. (2024) offer a thorough analysis of existing mitigation techniques, emphasizing the preservation of LLMs' creative powers while lowering the frequency and severity of hallucinations [20]. These initiatives, which seek to balance safety and creativity, include advancements in model training and the creation of trustworthy post-processing methods for output verification.

IV. DETECTION AND MITIGATION STRATEGIES

Hallucinations, in which Large Language Models (LLMs) generate information that is factually inaccurate or irrelevant to context, have become a significant concern as LLMs are increasingly incorporated into different applications. The current approaches to identifying and minimizing hallucinations will be covered in this section, with an emphasis on grounding techniques, mitigation strategies, and detection methods.

A. Ways to Find Hallucinations

Finding hallucinations in LLM outputs is an important step in making them less harmful. There are a number of ways to figure out when an LLM gives out wrong or unreliable information:

Corresponding Author: tanzeelbaig.ai27@jecrc.ac.in

1. Confidence Scoring: Some researchers have developed frameworks to quantify the likelihood of hallucination in LLM outputs. These methods generate confidence scores based on the input provided and the attributes of the model's response [28].
2. Segment Analysis: This method focuses on identifying specific spans of text within the LLM's output that might contain hallucinated information. By analyzing smaller segments of generated content, it provides more granular insights into where hallucinations occur [27].
3. Knowledge Consistency Checks: Wang et al. propose a system that uses Named Entity Recognition (NER) and Natural Language Inference (NLI) techniques. This method detects a wide range of hallucinations by identifying when the model generates content that contradicts the input or established facts [27].
4. Output Evaluation Metrics: Luo et al. introduce a quantitative method, the "Hallucination Critic," which measures the percentage of hallucinated content in LLM outputs, offering a way to assess the extent of unverified information in generated text [30].
5. Hierarchical Detection: A novel approach uses smaller language models for an initial hallucination check. This is then followed by larger LLMs that act as "constrained reasoners" to provide detailed explanations for detected hallucinations, aiming for a balance between detection speed, interpretability, and accuracy [29].
6. Entity-Specific Anomaly Detection: Su et al. propose a method specifically designed to detect hallucinations at the entity level, focusing on inconsistencies related to particular entities mentioned in the output [31].

B. Detection Methods

Reducing the adverse effects of Large Language Models (LLMs) requires the ability to detect hallucinations in their outputs. Several methods have been put forth by researchers to identify instances in which a model produces false or deceptive data:

1. Named Entity Recognition (NER) and Natural Language Inference (NLI): One method uses a combination of NER and NLI to identify discrepancies in responses that are generated. This approach determines whether the model's output conflicts with the input context or confirmed facts.

Corresponding Author: tanzeelbaig.ai27@jecrc.ac.in

2. Span-Based Detection (SBD): This method looks for possible hallucinated content by analyzing smaller output segments, or "spans." More precise and localized inaccuracy detection is made possible by SBD [27].
3. Probability-Based Techniques: Some frameworks look at how likely it is that a model's response has false information. These methods give confidence scores based on how the input and output are related [28].
4. Two-Stage Detection Framework: In this new setup, smaller models first flag responses that might be hallucinated. Then, bigger, more powerful models look over these cases and give logical explanations that strike a balance between speed and reliability [29].
5. Hallucination Critic: Metric-based method to determine the extent of hallucinations in the text. This approach measures the amount of hallucination by figuring out how much of a response is untrustworthy [30].
6. Entity-Level Detection: A technique that particularly monitors named entity inconsistencies. As a result, hallucinations involving named subjects or factual references can be specifically detected [31].

C. Mitigation Strategies

Once hallucinations are identified, several techniques can be applied to reduce their frequency and minimize their effects:

1. Prompt Engineering:Hallucinations can be avoided by carefully crafting prompts. This entails providing precise context, specifying the format of the expected output, and telling the model to refrain from conjecture or unsubstantiated assertions [20].
2. Adversarial Testing: Adversarial testing can reveal LLM behavior's weak points and guide the creation of focused remedies to combat particular hallucination patterns [21].
3. Multi-Scoring Frameworks: Hallucination detection and control can be enhanced by combining several evaluation metrics. Combining different scoring techniques improves overall reliability because no single technique is optimal in every circumstance [28].
4. Fine-Tuning and Parameter Adjustment: Hallucinations can be considerably decreased by adjusting model parameters or fine-tuning LLMs with domain-specific data, particularly in tasks requiring specialized knowledge [32]..

Corresponding Author: tanzeelbaig.ai27@jecrc.ac.in

5. Rewriting Mechanisms: In order to balance accuracy, speed, and computational efficiency, these methods entail modifying the model's outputs to rectify or remove hallucinated content [27].
6. Dynamic Retrieval Augmentation (DRAD): DRAD improves the factual reliability of LLM responses by modifying the retrieval procedure in real time based on hallucination detection [31].

V. FUTURE DIRECTIONS AND CONCLUSION

Addressing the problem of hallucinations continues to be a critical research priority as Large Language Models (LLMs) develop further and are incorporated into a growing number of applications. This section examines recent advancements and suggests future directions for improving the technical robustness, ethical alignment, and dependability of LLM.

A. Advancing Evaluation Techniques

To detect and treat hallucinations in LLMs, it is crucial to establish reliable and expert-aligned evaluation techniques. The significance of creating flexible and all-encompassing assessment frameworks that can accurately identify hallucinations in a variety of use cases and domains is highlighted by recent research.

Flexible definitions and context-sensitive assessment techniques are essential, especially in situations involving high stakes decisions [34]. They advise that future research focuses on developing instruments that can distinguish between different kinds of hallucinations and evaluate their practical implications.

According to Schiller (2024), who supports a user-centered evaluation model, better assessment techniques may result from looking at how people see and identify hallucinations [35]. This human-in-the-loop method could help close the gap between usefulness and automated metrics.

B. Ethical Considerations and Responsible Development

A significant challenge is balancing the conflict between ethical responsibility and hallucination-driven creativity. Beyond factual errors, hallucinations bring up questions about trust, responsibility, and their wider social effects.

Corresponding Author: tanzeelbaig.ai27@jecrc.ac.in

In order to keep an eye on hallucination risks in a variety of settings, creating domain-specific ethical standards and adaptive auditing systems [36]. To guarantee that guidelines change in tandem with technological advancements, they emphasize the significance of interdisciplinary collaboration.

In order to create more transparent and trustworthy LLMs, Lin et al. (2024) stress the importance of debiasing and dehallucination techniques [35]. Future studies should make sure that their findings respect moral principles and conform to social norms in addition to reducing hallucinations.

C. Improving Knowledge Integration Approaches

It is still very hard to deal with the conflict between moral duty and the creative power of hallucinations. Hallucinations not only lead to wrong information, but they also make people worry about accountability, public trust, and their effects on society as a whole.

Researchers have stressed the need for domain-specific ethical frameworks and dynamic auditing mechanisms that are specific to each application area in order to effectively manage these risks [20]. To make sure that these standards stay in line with new technology, there is a lot of focus on encouraging collaboration between different fields.

To make large language models (LLMs) more transparent and reliable, we also need to use debiasing and dehallucination strategies together. Future research should not only try to reduce hallucinations, but it should also make sure that the results are ethical and meet societal standards [13].

D. Conclusion

Recent research has made a lot of progress in understanding, finding, and reducing hallucinations in large language models (LLMs). Future research should focus on creating strong evaluation frameworks, dealing with new ethical issues, and improving knowledge grounding to make sure these models are reliable and accountable. Hallucinations can be very dangerous in fields like medicine, science, and sports, but they can also be helpful in creative fields like literature and the arts. To safely and successfully use LLMs, you need to be able to handle both sides of this issue.

Corresponding Author: tanzeelbaig.ai27@jecrc.ac.in

Researchers, technologists, and policymakers must continue to work together to reduce risks and steer the responsible growth of LLM technologies.

Reference

1. Rahman, M. M., & Kundu, A. (2024). Code hallucination. *arXiv preprint*. https://arxiv.org/abs/2407.04831
2. Xu, Z., Jain, S., &Kankanhalli, M. S. (2024, January). Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint*. https://arxiv.org/abs/2401.11817
3. Peng, B., Chen, K., Li, M., Feng, P., Bi, Z., Liu, J., & Niu, Q. (2024). Securing large language models: Addressing bias, misinformation, and prompt attacks. *arXiv preprint*. https://arxiv.org/abs/2409.08087
4. Peng, B., Bi, Z., Niu, Q., Liu, M., Feng, P., Wang, T., Yan, L. K., Wen, Y., Zhang, Y., & Yin, C. H. (2024, October). Jailbreaking and mitigation of vulnerabilities in large language models. *OSF Preprints*.
5. Zhang, H., Huang, J., Mei, K., Yao, Y., Wang, Z., Zhan, C., Wang, H., & Zhang, Y. (2024). Agent security bench (ASB): Formalizing and benchmarking attacks and defenses in LLM-based agents. *arXiv preprint*. https://arxiv.org/abs/2410.02644
6. Deng, H., Zhang, H., Ou, J., & Feng, C. (2024). Can LLM be a good path planner based on prompt engineering? Mitigating the hallucination for path planning. *arXiv preprint*. https://arxiv.org/abs/2408.13184
7. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023, November). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint*. http://arxiv.org/abs/2311.05232
8. Niu, Q., Liu, J., Bi, Z., Feng, P., Peng, B., & Chen, K. (2024). Large language models and cognitive science: A comprehensive review of similarities, differences, and challenges. *arXiv preprint*. https://arxiv.org/abs/2409.02387

Corresponding Author: tanzeelbaig.ai27@jecrc.ac.in

9. Niu, Q., Chen, K., Li, M., Feng, P., Bi, Z., Liu, J., & Peng, B. (2024). From text to multimodality: Exploring the evolution and impact of large language models in medical practice. *arXiv preprint*. https://arxiv.org/abs/2410.01812
10. Jiang, G., Shi, X., & Luo, Q. (2024). LLM-collaboration on automatic science journalism for the general audience. *arXiv preprint*. https://arxiv.org/abs/2407.09756
11. Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature, 620*, 625–630.
12. Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., & Peng, W. (2024). A survey on hallucination in large vision-language models. *arXiv preprint*. https://arxiv.org/abs/2402.00253
13. Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., & Shou, M. Z. (2024). Hallucination of multimodal large language models: A survey. *arXiv preprint*. https://arxiv.org/abs/2404.18930
14. Zhang, Y., Li, S., Liu, J., Yu, P., Fung, Y. R., Li, J., Li, M., & Ji, H. (2024). Knowledge overshadowing causes amalgamated hallucination in large language models. *arXiv preprint*. https://arxiv.org/abs/2407.08039
15. Yu, L., Cao, M., Cheung, J. C. K., & Dong, Y. (2024). Mechanistic understanding and mitigation of language model non-factual hallucinations. *arXiv preprint*. https://arxiv.org/abs/2403.18167
16. Amatriain, X. (2024). Measuring and mitigating hallucinations in large language models: A multifaceted approach. *Preprint*.
17. Cheng, Q., Sun, T., Zhang, W., Wang, S., Liu, X., Zhang, M., He, J., Huang, M., Yin, Z., Chen, K., & Qiu, X. (2023). Evaluating hallucinations in Chinese large language models. *arXiv preprint*. https://arxiv.org/abs/2310.03368
18. Jiang, C., Qi, B., Hong, X., Fu, D., Cheng, Y., Meng, F., & Zhou, J. (2024). On large language models' hallucination with regard to known facts. *arXiv preprint*. https://arxiv.org/abs/2403.20009
19. Li, J., & Hong, Q. (2024). A long-chain approach to reduce hallucinations in large language models. *Research Square*.

Corresponding Author: tanzeelbaig.ai27@jecrc.ac.in

20. Tonmoy, S. T. I., Zaman, S. M. M., Jain, V., Rani, A., Rawte, V., Chadha, A., & Das, A. (2024). A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint*. https://arxiv.org/abs/2401.01313
21. Perkovic, G., Drobnjak, A., &Botički, I. (2024). Hallucinations in LLMs: Understanding and addressing challenges. In *2024 47th MIPRO ICT and Electronics in Education*. IEEE.
22. Hao, G., Wu, J., Pan, Q., & Morello, R. (2024). Quantifying the uncertainty of LLM hallucination spreading in complex adaptive social networks. *Scientific Reports*.
23. Qiu, Y. (2024). The impact of LLM hallucinations on motor skill learning: A case study in badminton.
24. Jiang, X., Tian, Y., Hua, F., Xu, C., Wang, Y., & Guo, J. (2024). A survey on large language model hallucination via a creativity perspective. *arXiv preprint*. https://arxiv.org/abs/2402.06647
25. Yotam, S., Gabriela, A. P., & Isa, A. R. (2024). Mystoryknight: A character-drawing driven storytelling system using LLM hallucinations. *Interaction*.
26. Oelschlager, R. (2024). Evaluating the impact of hallucinations on user trust and satisfaction in LLM-based systems. *DiVA*.
27. Wang, S., Wang, X., Mei, J., Xie, Y., Muarray, S., Li, Z., Wu, L., Chen, S.-Q., & Xiong, W. (2024). Developing a reliable, general-purpose hallucination detection and mitigation service: Insights and lessons learned. *arXiv preprint*. https://arxiv.org/abs/2407.15441
28. Valentin, N., Creager, E., Gemp, I., Nie, A., Goodman, N. D., &Shieber, S. M. (2024). Cost-effective hallucination detection in large language model conversations. *arXiv preprint*. https://arxiv.org/abs/2401.00231
29. Hu, Y., Zhao, Z., Xu, Y., Shen, Z., & Neubig, G. (2024). SLM: Scalable language model hallucination detection via constrained reasoning. *arXiv preprint*. https://arxiv.org/abs/2401.03158
30. Luo, J., Li, T., Wu, D., Jenkin, M., Liu, S., & Dudek, G. (2024). Hallucination detection and hallucination mitigation: An investigation. *arXiv preprint*. https://arxiv.org/abs/2401.08358
31. Su, W., Tang, Y., Ai, Q., Wang, C., Wu, Z., & Liu, Y. (2024). Mitigating entity-level hallucination in large language models. *arXiv preprint*. https://arxiv.org/abs/2407.09417

Corresponding Author: tanzeelbaig.ai27@jecrc.ac.in

32. Zhang, Z., Wang, Y., Wang, C., Chen, J., & Zheng, Z. (2024). LLM hallucinations in practical code generation: Phenomena, mechanism, and mitigation. *arXiv preprint*. https://arxiv.org/abs/2409.20550
33. Liu, F., Liu, Y., Shi, L., Huang, H., Wang, R., Yang, Z., Zhang, L., Li, Z., & Ma, Y. (2024). Exploring and evaluating hallucinations in LLM-powered code generation. *arXiv preprint*. https://arxiv.org/abs/2404.00971
34. Chakraborty, N., Ornik, M., & Driggs-Campbell, K. (2024). Hallucination detection in foundation models for decision-making: A flexible definition and review of the state of the art. *arXiv preprint*. https://arxiv.org/abs/2403.16527
35. Schiller, C. A. (2024). The human factor in detecting errors of large language models: A systematic literature review and future research directions. *ArXiv preprint*. https://arxiv.org/abs/2403.09743
36. Jiao, J., Afroogh, S., Xu, Y., & Phillips, C. (2024). Navigating LLM ethics: Advancements, challenges, and future directions. *arXiv preprint*. https://arxiv.org/abs/2406.18841
37. Lin, Z., Guan, S., Zhang, W., Zhang, H., Li, Y., et al. (2024). Towards trustworthy LLMs: A review on debiasing and dehallucinating in large language models.

Corresponding Author: tanzeelbaig.ai27@jecrc.ac.in