

A Comparative Study of Techniques for Reducing Hallucinations in Large Language Models

Amarnath Awasthi¹, Aryan Chaudhary², Anushka Mishra³, Anshi Gupta⁴

¹ Assistant Professor, Axis Institute of Higher Education

^{2,3,4} Axis Institute of Higher Education

Abstract—Large Language Models (LLMs), such as ChatGPT, are AI Tools that can understand and generate human like text. They are useful for answering questions, Solving problems, and providing information. However, sometimes LLMs give wrong or made up information with confidence. This problem is known as AI Hallucination. LLMs do not actually know facts or understand reality. Instead, they predict the next most likely word based on patterns learned from their training data. Therefore, when the required information is missing, under or misunderstood. The model may be fill the gaps with information that sounds correct but is actually wrong.

AI Hallucinations can become a serious problem in fields such as Law, Medicine, Education or Finance where incorrect information can lead to real-world Consequences. One possible way to reduce this problem is Prompt Engineering means creating clear, Specific and Well Structured.

The result of an LLMs also depends a lot on how the user given the Prompt. If the Prompt is not cleared or well formatted. The AI may give an incorrect answer. To reduce Hallucinations, Researches are working on methods like specific Prompting, CoT and RAG.

This Research focuses on Studying AI Hallucinations in LLMs and ways to reduce them. The Study will compare different prompt techniques and Examine their Effect on the accuracy and reliability of LLMs responses. It will also explore how prompt design is related to Hallucinations. The main AIM is to identify Practical prompting methods that can help users get more accurate and trustworthy responses from LLMs.

Index Terms—LLMs, Prompting, CoT(Chain of Thoughts), RAG(Retrieval-Augmented Generation), AI Safety.

I. INTRODUCTION

Large Language Model is a type of Artificial Intelligence like ChatGPT, Gemini, etc. are example

of LLM. An LLM is trained on a massive amount of text data from the internet like books, websites, articles and wikipedia. Because it learns from billions of words, it is called "Large" during training, it learn human language, grammar, fact and writing patterns.

The main work of an LLM is to understand human language and generate a Human- like response. It does not actually know fact like human, it work by predicting the next word. For example if you type "My Name Is ", it will predict the next word that is most likely to come after it. It this very fast to create full sentence and answers.

LLMs are very useful. They can answer question write Emails and Essays, solve problem, translate language and summarize long documents. They have made it very easy for Humans to interact the computers.

But LLMs also have a major limitation sometimes they create False information when the user wants to add the emotions in the prompt like they write "What is full form of C.U.", A.I. answers "Chandigarh University situated in Unnao... etc." but the users only wants the result "Chandigarh University", unless they should write the "Give me only the full form of C.U."

1. What is AI Hallucination?

AI Hallucination in Large Language Models (LLMs) refers to the phenomenon where an artificial intelligence system generates responses that are factually incorrect, logically inconsistent, or completely fabricated, while presenting them with high confidence. In human psychology, a hallucination involves perceiving something that is not present in reality. In the context of computer science and natural language processing, AI hallucination means the model produces outputs that

have no backing in its training data, context, or real-world facts.

When users interact with tools like ChatGPT or Gemini, they expect accurate answers to their queries. However, because LLMs do not possess real understanding, they sometimes generate text that sounds completely convincing and grammatically perfect but is entirely false. For instance, as mentioned earlier, asking a simple question like the full form of "C.U." might cause the model to output "Chandigarh University situated in Unnao," adding false geographic details that were never requested.

Researchers generally categorize AI hallucinations into two main types:

- **Intrinsic Hallucinations:** This occurs when the AI generated output directly contradicts the source information provided in the prompt or document. For example, if a user uploads a text stating "The event happened in 2020" and the model summarizes it as "The event happened in 2025," it is an intrinsic error.
- **Extrinsic Hallucinations:** This happens when the AI output includes extra facts or details that cannot be verified from the given context or training data. The information is not necessarily contradictory to the prompt, but it is entirely made up by the model to fill in conversational gaps.

In fields like law, medicine, finance, and education, hallucinations pose serious risks. A student relying on fake historical dates, a doctor getting wrong medical dosages, or a lawyer using fabricated legal cases generated by an LLM can face real consequences. Therefore, understanding and measuring hallucination is the first step toward building safer AI tools.

2. Why do AI Hallucinations Happen?

To understand why hallucinations occur, we have to look at how Large Language Models are built and trained. LLMs are not traditional search engines or structured databases; they are statistical neural networks designed to learn language patterns from massive datasets. Several fundamental factors contribute to why these systems create false information:

- **Training Data Quality and Noise:** LLMs are trained on billions of words collected from

internet pages, Wikipedia, books, and public articles. The internet contains huge amounts of inaccurate information, rumors, sarcastic posts, outdated facts, and biased content. Because the model learns from this mixed dataset, it absorbs incorrect information along with correct facts during training.

- **Objective Function Limitations:** During training, an LLM is optimized using a loss function that rewards predicting the next probable token in a sequence. The system is trained to maximize fluency and natural sentence structure, not necessarily true factual correctness. As long as the sentence sounds natural to human language rules, the training algorithm considers it a good prediction.
- **Lack of Real-World Knowledge and Grounding:** Humans understand words by linking them to real experiences, logical rules, and physical reality. An LLM does not "know" what a university, a law, or a medical condition actually is. It only knows how the word "university" relates statistically to other words in text data. Without real grounding, it cannot cross-check its own statements against reality before printing them.
- **Overgeneralization:** Neural networks try to find general patterns across vast datasets. Sometimes, the model overgeneralizes two similar concepts and blends them together. If two entities share similar names, contexts, or structures, the LLM might mix up their facts, creating a hybrid answer that is completely inaccurate.

3. When do AI Hallucinations Happen?

Hallucinations do not happen randomly in every response; they are much more likely to occur under specific trigger conditions during interaction. Identifying when hallucinations occur helps users design better prompts and safety checks.

- **Ambiguous or Unconstrained Prompts:** When a prompt lacks precise instructions, the AI is forced to make assumptions about what the user wants. As highlighted in our previous example, asking "What is full form of C.U." gives the AI freedom to generate long, unconstrained text, increasing the chance of hallucinated details like wrong locations. If the prompt lacks boundary

constraints (such as "Answer in 2 words"), hallucinations spike.

- **Low-Frequency or Niche Topics:** LLMs perform exceptionally well on widely discussed topics like basic programming or famous historical events because they appeared millions of times in training data. However, when asked about obscure research papers, local village histories, or highly specialized technical formulas, the model's training data is sparse. Instead of admitting it doesn't know, the LLM attempts to generate plausible-sounding text, leading to hallucinated facts.
- **Long-Context and Complex Reasoning:** When prompts become very long or require multi-step math and logic, LLMs often lose track of early context details. This failure in attention mechanisms causes the model to drift away from original constraints and start generating incorrect steps mid-way through its response.
- **Leading and False-Premise Questions:** If a user asks a question containing a false assumption (e.g., "Why did Albert Einstein invent the smartphone in 1920?"), the model's tendency to be helpful often tricks it into agreeing with the user. Instead of correcting the false premise, it hallucinates a fictional narrative to satisfy the user's prompt.

4. How do AI Hallucinations Happen?

Understanding how hallucinations happen requires analyzing the technical generation mechanism inside the LLM during runtime. Text generation in models like ChatGPT occurs word-by-word (or token-by-token) through mathematical probability distributions.

Step-by-Step Hallucination Mechanism:-

1. User Prompt
 2. Probability Calculation
 3. Word Choice
 4. Snowball Effect
- **Token-by-Token Probability Prediction:** Given an input prompt, the LLM calculates a probability score for thousands of possible next words in its vocabulary. It selects a word based on these probabilities and feeds that word back into its input history to select the next word.

- **Sampling Methods and Temperature Settings:** To make AI responses sound creative and human-like rather than repetitive, developers use sampling settings like "Temperature" and "Top-p". A higher temperature forces the model to pick slightly lower-probability words instead of always picking the top mathematical choice. While this adds creativity, it also drastically increases the chances of selecting a token that introduces false facts into the sequence.
- **The "Snowball Effect" in Generation:** Once an LLM outputs a single hallucinated or inaccurate word, that wrong word immediately becomes part of its permanent context for all subsequent word predictions. The model feels mathematically obligated to remain consistent with its own previous text. As a result, one minor false word early in the sentence causes the rest of the paragraph to "snowball" into a completely fake story.
- **Absence of Uncertainty Calibration:** Most standard LLMs are not trained with a built-in mechanism to measure their own certainty. Unlike a human who says "I am not sure, but...", standard model architectures treat high-probability and low-probability guesses with the exact same output confidence, delivering fabricated statements as absolute fact.

II. SOLUTION

AI Hallucinations is a major barrier to using LLMs in high-risk fields like law, medicine and Education. Since Hallucination happens due to data quality, training method and generation mechanism, The solution must be multi-layered.

Improving the data and Training level high-quality Data Curation: The first solution is to clean training data instead of all using all internet data, model should be trained on verified up-to-date and fact checked data sets filtering out rumors, posts, information and biased content reduces the chance that the model learns false facts. Also, the training objective should be changed. Currently the models get a reward only when the sentence is both fluent and factually correct, which can be done through human

feedback. Another important step is to Connect the model with real knowledge graphs and verified databases, so it can cross check what a word like university or a medical team actually means in the real world.

At a generation level, the most effective solution is RAG. In this method, the model does not answer only from its memory, but it first searches for info in a trusted databases and then creates an answer based on that retrieved proof. This directly reduces false information. Another solution is to use better and more constrained prompts. Many Hallucinations happen because the user prompt is ambiguous, for example asking what is the full form of CU without any context gives the model freedom to add false details like a wrong location. If we add boundary constraints in the prompt like answer in two words only, if you do not know say I do not know, or answer only from the given context, the model will not fill gaps with fake data. For low-frequency topics, long complex question, the model should be trained to show its Uncertainty. It should be Learn to give a confidence score and if the confidence is low, it should refuse to answer instead of guessing.

Finally, at the verification level, we need to control, the snowball effect. In LLMs, text is generated token by token and once a single wrong word becomes part of the context and the model tries to stay consistent with it, making the whole paragraph false. To stop this, a second checking model can be used which verifies each Sentence after it is generated.

One more simple solution is to train the model only for one special field. Right now one model tries to know Everything, so it knows little about everything and makes mistakes, If we train a model only on doctor books, it will become good for medical questions and will make less false answers. Same for law, if we train it only on real court cases, it will get better answers for law. This is called domain specific training. Another easy way is to make to make models check each other. One model will write the answers and two other models will check it like teachers. If they find a wrong point, they will connect it. The answer is only given when all models agree. This will catch many errors. We should also let the model use tools. The model should not The model should not try to remember the data or do maths in its mind. It should be allowed to use a calculator, a calendar or google search. If it does not know the

latest news, it should search on the internet instead of guessing.

III. CONCLUSION

In our research paper we have discussed why and when Hallucinations happened. LLM's Hallucinations remains the single greatest bottleneck impeding the safe, autonomous deployment of generative AI across high-stake companies. Far from being a simple Software bug or a transient training artifact, Hallucinations is an inherent property of the auto-aggressive model. A single and accurate token generated early in a sequence can compound, leading the model down a path confident yet entirely ungrounded assertion. In Sensitive fields such as, medicine, law and finance the margins for error are non-existent. It can lead to real world harm, legal liability and a permanent loss of user trust. To address this challenge, AI ecosystem has shifted away from relying solely or larger training data sets toward multi-layered mitigation. Systematic solutions such as (RAG) and (RLHF). Moving forward, completely eliminating Hallucinations within pure probabilistic language models. The successful integration of LLM into critical infrastructure will not depend on achieving flawless model outputs, but rather on engineering resilient, human-in-the-loop System that anticipates, detect and isolate Hallucination as a fundamental operating condition.

Overcoming AI hallucination requires a robust, multi-layered approach that addresses errors across every stage of the model life cycle:

- **Data & Training:** Curating high-quality data sets, refining human feedback rewards, incorporating knowledge graphs and using domains specific training significantly reduce base level misinformation.
- **Generation & Retrieval:** Integrating Retrieval Augmented Generation strict prompt constraints, tool integration and uncertainly scoring prevents model from guessing or filling knowledge gaps with false details.
- **Verification & Multi-Model Systems:** Implementing sentence multi agent consensus protocol prevents "Snowball Effect" ensuring errors are caught before reaching users.

By combining domain specialization, live external verification and strict boundaries, LLM's can reach the high reliability and accuracy required for deployment in critical fields like law, medicine and education.

REFERENCES

- [1] Huang, L., Yu, W., Ma, W., et al. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(1), 1–55. <https://doi.org/10.1145/3703155>
Cited by: 7124
- [2] Sharma, M., Tong, M., Korbak, T., et al. (2023). Towards understanding sycophancy in language models. *arXiv*. <https://doi.org/10.48550/arxiv.2310.13548>
Cited by: 1960
- [3] Wei, J., Wang, X., Schuurmans, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837. <https://doi.org/10.48550/arxiv.2201.11903>
Cited by: 38086
- [4] Wei, J., Huang, D., Lu, Y., Zhou, D., & Le, Q. V. (2023). Simple synthetic data reduces sycophancy in large language models. *arXiv*. <https://doi.org/10.48550/arxiv.2308.03958>
Cited by: 8519