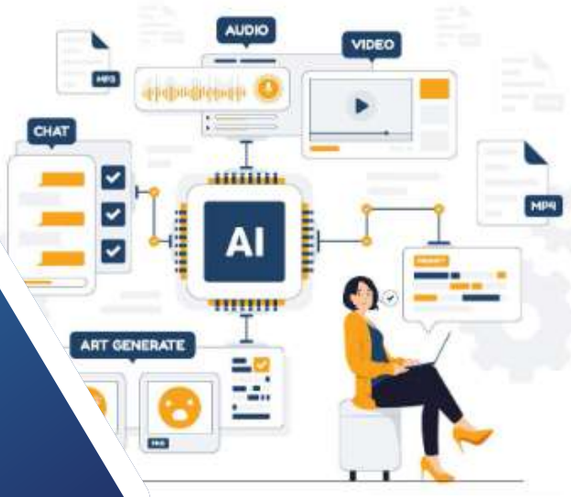
EXAFLUENCE
Data Driven Influence



Generative AI - Separating Fact from Fiction

A blog on LLM Hallucinations

Kalaivani K G

December 2023

Heading the AI team at Exafluence, I've had the unique opportunity to observe the exploration and implementation of GenAI solutions across various industries. The transformative power of GenAI is truly fascinating, but there is always the lingering question: 'How much can we trust this?' This question is especially important when we talk about Large Language Models (LLMs) and the times they make mistakes, which are called 'hallucinations'.

Here's an overview of insights gained about hallucinations from our work in developing these solutions, and their implications for those of us utilizing these advanced technologies.

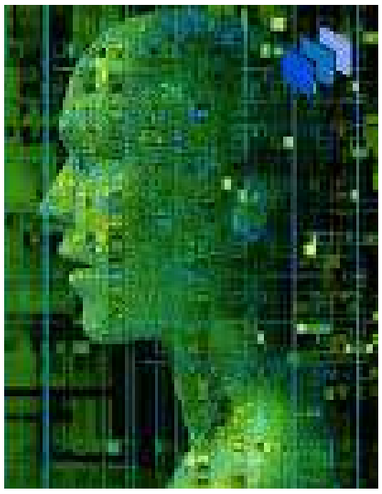
Q1. LLM Hallucinations – What are they and why should we be so concerned?

LLM (Large Language Model) hallucination is when language models make up information that seems real but isn't. Imagine asking your GPS for directions to the nearest cafe, and instead of giving real directions, it invents a fictional cafe and guides you there. That's like an LLM hallucination - the AI confidently gives wrong or made-up information. These hallucinations are a byproduct of the model's design, which aims to generate plausible language based on patterns it has learned.

In a case involving an LLM-based chatbot, we were consistently testing with known stock information, such as those of Microsoft and Amazon. The chatbot could fetch current market prices for these stocks using their tickers. However, when tested with a company that had recently launched an IPO, the LLM was unfamiliar with the new company name, as it was not included in its training data. Consequently, it erroneously provided the ticker and market price of a different company with a similar-sounding name.

The implications of this are significant, especially in enterprise settings. While it's one thing to receive predictions from ML models, knowing they are just predictions, using LLMs in settings where deterministic outcomes are expected requires extreme caution.

Q2. Why are they harmful? Isn't it part of the creative aspect of LLMs?



While the creative aspect of LLMs can be beneficial in certain contexts, LLM hallucinations are harmful because they undermine the reliability and trustworthiness of the information provided. These inaccuracies can lead to misinformed decisions or misunderstandings. For instance, in critical applications like medical advice, legal information, educational content and even a vast number of enterprise use cases, hallucinations can have serious consequences if the fabricated information is taken as truth. In contrast, in the case of creative writing, LLM hallucination is beneficial as it aids in crafting imaginative narratives by generating unique and fictional details that enhance the storytelling process.

The distinction lies in intention - creative uses are controlled and known to be imaginative, whereas hallucinations are unintended falsehoods presented as facts.

Q3. In which contexts are hallucinations considered acceptable, and in which situations should we exercise caution regarding them?

LLM hallucinations vary in acceptability based on their application. They are generally acceptable in creative domains like writing and ideation, where innovation and imagination are key. However, in areas requiring precise and reliable information leading to decisions or developing deterministic solutions, it is crucial to minimize and control these hallucinations. The appropriateness of hallucinations hinges on the potential impact of acting upon inaccurate information. In the chatbot example that I described earlier, a portfolio manager could have made an incorrect decision based on the erroneously returned market price.

User awareness and understanding of the potential for LLM hallucinations, especially in discerning between creative and factual contexts, play a crucial role in mitigating risks and ensuring appropriate reliance on the model's outputs.

Q4. Why do LLMs hallucinate? Why do we need to know these reasons?

LLMs could hallucinate due to various factors:

Inherent Design: LLMs predict responses based on learned patterns, not factual accuracy, leading to guesses that may be incorrect.

Training Data Limitations:

- **Incomplete or biased datasets may result in inaccurate outputs.**
- **The model's knowledge is limited by the timeline of the data it was trained on.**
- **Domain knowledge gaps can also cause hallucinations when queried on those domains.**

Lack of Real-World Understanding: Without access to real-world experiences or external data, LLMs rely solely on language patterns, often disregarding factual correctness.

Learning and Reasoning Limitations: LLMs lack deep comprehension and logical reasoning, which can lead to errors in complex or nuanced situations.

Overgeneralization: Broad generalizations from training data can produce plausible but factually incorrect statements.

User Prompting: Ambiguous or leading queries from users can prompt hallucinated responses.

Understanding the reasons behind LLM hallucinations is crucial for recognizing the limitations of these models and, consequently, for developing better risk management strategies in application design. It also allows for setting realistic expectations about the capabilities and limitations of these models.

Q5. What are the types of LLM Hallucinations?

The paper titled ' A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions' (arXiv:2311.05232v1 [cs.CL] 9 Nov 2023), broadly categorizes hallucinations into two major categories:

Factuality Hallucination: This occurs when the LLM generates false information. For instance, if you inquire, 'Who was the first person to walk on the moon?' and it responds, “The first person to walk on the moon was Charles Lindbergh in 1951, during the Lunar Pioneer mission. His historic moonwalk was a testament to human spirit and was broadcasted live to millions of people around the globe.' This is a factuality hallucination. It provides incorrect real-world facts in a confident manner.

Faithfulness Hallucination: This type of hallucination occurs when the LLM fails to adhere to the user's instructions or context. For instance, if you request, 'Translate 'hello' into Spanish,' and instead of translating, it provides the history of the word 'hello,' that's a faithfulness hallucination. It doesn't perform the requested task as expected."
It's important to recognize that these categories of hallucinations are not mutually exclusive and can often overlap, with a single instance potentially exhibiting characteristics of both factuality and faithfulness hallucinations.

Q6. How to control LLM hallucinations?

Depending on the cause of hallucinations, different methods may work well to control them. Some of them are:

Knowledge related

Fine-tuning: Where the LLM hallucinates due to lack of domain knowledge, fine-tuning on specific domains will make the model more reliable.

Retrieval augmented generation: Retrieval-augmented generation systems have access to external knowledge sources, such as databases, documents, or the internet. By integrating external knowledge, retrieval-augmented generation systems can enhance the quality and accuracy of their responses, especially when faced with tasks that require access to up-to-date information or domain-specific knowledge that may not be present in the model's training data. This external knowledge can serve as a reference for the model during text generation.

Inference time parameter settings

LLMs typically use an autoregressive generation process where they predict the next token one at a time based on the preceding context. During the prediction of the next token, the model looks at tokens with higher probabilities.

Nucleus Sampling: Top-P is a parameter that serves as a cumulative probability cutoff for token selection. Top-p determines the threshold for including words based on their predicted probabilities. When top-p is reduced, it limits the set of words considered for the next word in the sequence. Reducing the nucleus of words can reduce the chances of selecting less relevant or hallucinatory words.

Temperature setting: There is another parameter called Temperature that affects the output token selection. The temperature setting has an impact on the distribution of probabilities for word selection. Reducing the temperature can make the generated text more deterministic and focused.

- OpenAI suggests modifying only one of them. We usually keep Top-p constant and play around with temperature settings.

Prompting techniques

Specificity in prompting: Specific prompts prevent hallucination by offering clear instructions, context, and constraints, reducing the chances of the model generating inaccurate or irrelevant content and ensuring that responses align closely with the user's intent.

Chain of Thought Reasoning: Chain of thought reasoning encourages the model to organize information in a structured and logical manner. This structured approach ensures that the generated content follows a coherent flow of thought. This approach can control hallucinations related to content coherence and logical reasoning.

We have achieved reasonable success by combining the temperature setting, specificity of prompting, and RAG (Retrieval-Augmented Generation) methods in our use cases.

LLM hallucination is a very dynamic and evolving area under active research and development, with ongoing efforts aimed at understanding and mitigating LLM hallucinations.

Q7. What if the tolerance for hallucinations is zero?

Use cases where there is zero tolerance for hallucinations need to be designed with contingency measures in place. Human in the loop (HITL) is often considered a highly reliable contingency measure for mitigating hallucinations in LLMs when combined with effective processes, guidelines, and oversight.

For example, in one of the data mapping use cases we are developing to map hundreds of fields, we employ Human-in-the-Loop (HITL) to ensure the accuracy of the LLM's suggestions. The AI significantly boosts productivity, while human oversight ensures high precision. This combination leads to an effective blend of efficiency and accuracy in the system's output.

While Human-in-the-Loop systems are effective, they also require more time and resources, which can be a challenge for large-scale use and quick responses. Hence, it's important to consider the trade-offs and come up with a balanced approach to their implementation.

Final Thoughts: Summing up, the world of LLMs is a fascinating mix of innovation and challenge. As we marvel at the possibilities of new applications, it's also a reminder to be aware of the fine line between creative genius and factual accuracy. Excited to see the ongoing advancements in this field and hoping that in near future, creating reliable, deterministic systems is simpler and free from today's hurdles.

If you would like to learn more, write to us at marketing@exafluence.com

For interesting videos about our solutions subscribe to our YouTube channel- https://tinyurl.com/YouTubeExf

For regular updates on our solutions follow us on LinkedIn https://tinyurl.com/LinkedInExf