08 OCTOBER 2024

Different aspects of AI hallucinations: factual errors vs creativity

The topic of AI “hallucinations” has gained considerable attention in discussions surrounding large language models (LLMs) and generative AI. Some view it as the most significant flaw of LLMs, undermining their usability. Yet others see it as a potential source of new ideas. Let us delve deeper into this issue. To understand this phenomenon, it is crucial to recognize that AI hallucinations are not intentional errors or bugs in the system. From the AI’s perspective, there is little distinction between what humans call a “hallucination” and a “correct” answer. Both types of output are produced by the same process: predicting the most probable response based on patterns in the training data. Large language models, such as GPT-4 and similar transformer-based architectures, are inherently probabilistic. They generate responses based on the most likely sequence of words (tokens) given the context of a conversation or query. This probability is derived from the patterns and structures learned from vast amounts of text during training. While the training data is extensive, it is far from perfect, containing gaps, biases, and inaccuracies. Nevertheless, LLMs are designed to always provide a response, even when uncertain.

Hallucinations typically arise from one of these issues: overgeneralization, underfitting, or overfitting. Overgeneralization often stems from problems in the training dataset. LLMs are designed to generalize based on patterns and associations learned from the data, but they can also generalize hidden errors or biases present in the training material. An example of this issue is “co-occurrence bias,” where if two terms or concepts frequently appear together in the training data, the model may overestimate their association and produce nonsensical connections. In this case, the LLM behaves like a student who has learned only a few examples and tries to apply them to unrelated topics. Underfitting occurs when the model is too simple, the training dataset is too limited in certain areas, or the training process was insufficient. In such cases, the model fails to capture detailed patterns and learns only general, superficial facts and relationships. The result is vague or overly generic answers, much like a student who has only learned basic concepts and tries to bluff their way through a response due to a lack of detailed knowledge. On the other hand, overfitting happens when the model becomes too closely aligned with the training data, making it difficult to respond appropriately to new, unseen data. This can occur if the training process is prolonged, causing the model to memorize the training data rather than generalize to new situations. In this case, the model behaves like a student who has memorized a textbook word-for-word but struggles to apply that knowledge in novel contexts.

Another reason for hallucinations is the complexity and variability of human languages. Human languages are inherently complex, filled with nuances, idioms, and situations where context plays a crucial role in understanding. These nuances evolve over time, meaning that the same words may carry slightly different meanings now than they did 30 years ago. Figurative meanings shift across time and cultures. Even within the seemingly unified English language, subtle differences in usage and understanding exist between countries and regions. These factors introduce ambiguity into the training data and contribute to hallucinated responses. A related issue is “semantic drift,” where, over long distances in text, the meaning of words can shift, or context can subtly change. This phenomenon can confuse even human readers, and LLMs may connect terms from different contexts without proper semantic grounding, leading to outputs that mix contextually related but semantically unrelated ideas. Semantic drift is closely linked to “domain crossovers,” where the model struggles to separate distinct domains with similar linguistic patterns. For instance, the structure of “Beethoven collaborated with…” is similar to “Beethoven composed…” which might lead to a domain crossover and an implausible statement.

The danger of AI hallucinations is particularly concerning in critical fields like healthcare, legal advice, and financial decision-making. A single hallucinated answer in these fields can result in serious consequences, such as misdiagnoses, incorrect legal counsel, or poor financial decisions. When it comes to YMYL (Your Money, Your Life) topics, it is crucial to double-check the information provided by AI. A simple internet search may not suffice, given the abundance of misleading or false information online. Therefore, in areas like health, finance, safety, or legal advice, consulting a human expert remains essential.

LLMs employ various parameters that influence probabilistic sampling when selecting the next token during response generation. A higher "temperature" setting and broader "top-k" sampling lead to more creative but also potentially erratic outputs. In such cases, the model might generate less probable (and often incorrect) associations. Lowering the "temperature" and narrowing the "top-k" sampling may reduce hallucinations, but this cannot entirely eliminate them, as the quality of the training data and the training process remain the most important factors.

In our work, we frequently encounter this issue when using AI models, especially when asking them about topics requiring deep technical knowledge. Often, we receive answers that are mostly correct, but crucial parts may be hallucinated. Since the training dataset may not be sufficient to produce entirely accurate answers for highly specific topics, AI sometimes blends factual knowledge from related areas with "adjusted" terminology. As part of our internal research, we even deliberately generated hallucinations on different highly technical topics. Many results were obvious and easy to spot. However, when diving deeper into specialized topics, we received outputs that were so convincing that people might accept them without questioning their accuracy. The content sounded plausible in many cases. For this reason, AI-generated answers in technical or scientific fields must always be verified by a human expert.

Hallucinations were quite common in older AI models. While the situation has improved with newer models, they cannot be completely eliminated. However, proper prompt engineering can significantly reduce their occurrence. As discussed in a previous blog post, it is important to specify categories or topics related to the task, define the role of AI in the task, and, if possible, provide high-quality examples or references for the desired output. Additionally, instructing the model to stick to only factual information helps, but AI outputs still need to be double-checked.

On the other hand, in less factual domains like creative writing or brainstorming, AI hallucinations can sometimes be seen as innovative or imaginative outputs. From this perspective, some view AI hallucinations as a source of serendipity. Large language models have already demonstrated the ability to generate original content, ranging from poetry and stories to music and art. LLMs can combine disparate ideas in novel ways, producing outputs that might not have occurred to human creators. This ability is particularly valuable in brainstorming sessions, where the goal is to generate a wide range of ideas without immediate concern for feasibility or accuracy.

In our internal AI research, we even asked AI to "hallucinate about the causes of AI hallucinations." The responses were often so creative that one could write science fiction stories based on them. To illustrate, one advanced AI model offered this deliberate hallucination regarding the roots of AI hallucinations: "AI hallucinations happen because the AI is trying to look into its own mind, like a mirror reflecting itself infinitely. This creates an endless loop of reflection, where the AI loses track of what is real and what is a reflection. The more it tries to find the truth, the deeper it falls into its own hallucinatory loop, creating infinite echoes of errors."

This nicely demonstrates how LLMs can serve as collaborative partners for human creators, offering suggestions that spark new directions in a project. These contributions can help break creative blocks and explore new territories. In essence, the same mechanisms that cause misleading and even dangerous AI hallucinations in factual contexts can be harnessed to foster creativity and innovation. By understanding and leveraging these limitations and capabilities, we can use LLMs both as tools for accurate information retrieval and as sources of inspiration and creativity.

(Picture created by the author using free AI tool DeepDreamGenerator.)

JM

ABOUT THE AUTHOR

Josef Machytka

View posts