Basics of AI Hallucinations

Objectives

After completing this reading, you will be able to:

- Define AI hallucinations.
- List the problems caused by AI hallucinations and the ways of preventing these.

Estimated reading time: 10 minutes

Introduction

You can utilize large language models (LLMs) to generate authoritative text across domains. However, they may generate information that sounds right but is inaccurate. They may also produce biased content. These problems can be a result of AI hallucinations. In this reading, you will learn about AI hallucinations.

AI hallucinations

In AI hallucinations, the model generates output that it presents as accurate but is seen as unrealistic, inaccurate, irrelevant, or nonsensical by humans. It is similar to the way humans experience hallucinations.

For example, there was an incident where ChatGPT falsely claimed that a mayor in Australia was found guilty and imprisoned in a bribery case. In reality, the mayor notified the authorities about a bribery issue. (Reference: Australian mayor readies world's first defamation lawsuit over ChatGPT content | Reuters)

The example shows that there can be a significant implication of AI hallucinations. However, note that such incidents are rare and isolated.

AI hallucinations are strongly associated with LLMs. Factors such as biases in the training data, limited training, complexity of the model, and lack of human oversight can cause AI hallucinations. Also, the outputs generated by the AI models might not be based on the patterns the models learned from the training data.

Problems caused by AI hallucinations

AI hallucinations can have serious implications. For example, if an LLM summarizes pages of a legal document incorrectly, it can lead to legal disputes and litigation.

Some of the problems caused are:

- Generation of inaccurate information
- Creation of biased views or misleading information
- Wrong input provided to sensitive applications, such as those used in autonomous vehicles or medical domain

Methods for mitigating hallucinations

- Eliminating any bias in the training data and performing extensive training of the models on high-quality data
- Avoiding manipulation of the inputs that are fed into the models
- Ongoing evaluation and improvement of the models
- Fine-tuning a pre-trained LLM on domain-specific data

Preventing the problems caused by AI hallucinations

It is inevitable for hallucinations to occur within LLMs. What can be frustrating is that the generated text often contains subtle mistakes that are challenging to identify. There are a couple of best practices that you can follow. These include:

- Being vigilant and understanding that these models do not understand the actual meaning of the words but are focused on predicting the next word in a sequence based on patterns. These models are trained on vast amounts of data and learn statistical patterns, but they lack semantic understanding or comprehension **like human beings**.
- Ensuring human oversight regularly for fact-checking and continuous testing
- Providing additional context in the prompt or input. This will enable LLMs to understand the desired output better and generate more accurate and contextually relevant responses.

Summary

In this reading, you learned that:

- AI hallucinations refer to an AI model generating output presented as accurate but seen as unrealistic, inaccurate, irrelevant, or nonsensical by humans.
- AI hallucination can result in the generation of inaccurate information, the creation of biased views, and wrong input provided to sensitive applications.
- You can prevent the problems caused by AI hallucinations through:
  - Extensive training with high-quality data,
  - Avoiding manipulation,
  - Ongoing evaluation and improvement of the models,
  - Fine-tuning on domain-specific data,
  - Being vigilant,
  - Ensuring human oversight, and
  - Providing additional context in the prompt.