Bane and Boon of Hallucinations in the Context of Generative AI

S. M. Nazmuz Sakib
https://orcid.org/0000-0001-9310-3014
School of Business and Trade, International MBA Institute, Dhaka International University, Bangladesh

EXECUTIVE SUMMARY

The phenomenon of hallucinations takes place when generative artificial intelligence systems, such as large language models (LLMs) like ChatGPT, generate outputs that are illogical, factually incorrect, or otherwise unreal. In generative artificial intelligence, hallucinations have the ability to unlock creative potential, but they also create challenges for producing accurate and trustworthy AI outputs. Both concerns will be covered in this abstract. Artificial intelligence hallucinations can be caused by a variety of factors. There is a possibility that the model will show an inaccurate response to novel situations or edge cases if the training data is insufficient, incomplete, or biassed. It is common for generative artificial intelligence to generate content in response to cues, regardless of the model's "understanding" or the quality of its output.

INTRODUCTION

AI has progressed leaps and bounds. It can do a plethora of tasks which includes texts, images, sounds and code. This is generally termed as generative AI. The content these software's produce is indistinguishable from the content that humans can produce. This is possible due to the fact there is a huge amount of training data behind which has been fed to the system, it uses these datasets to create detect patterns and synthesize data in a creative manner. Generative AI uses machine learning models like deep neural networks. The data generated by an AI model is not necessarily accurate, this is where it gets interested because it might occasionally cause problems creating hallucinations. There are different cases in hallucinations, sometimes the information is false and hypothetical and other times the information can be plausible but not genuine. AI has integrated itself into our daily lives and people have begun relying

on it for many tasks making it an extremely important topic. Hallucinations can cause the spread of misinformation and harm to trust.

Artificial intelligence hallucinations can be extremely harmful in fact-based areas such as the media, academia, and the legal system. Artificial intelligence has been employed in court filings to generate information containing fake judicial opinions and legal citations, with real-world consequences for the parties concerned. AI hallucinations is a problem which crossed beyond the point of dissemination of incorrect information to ethical problems. There are certain risks attached to the usage of AI and they have negative impacts which can result in spreading of stereotypes and maligning reputations. The research will focus on understanding the causes of AI hallucinations and how can they be mitigated, because it has become deeply ingrained within our lives and its use only widens by the passing moment. The overall goal is to determine the flaws to improve the design and the outputs which AI provides us with so the systems in place can become more dependable and ethical.

Purpose of the Study

The purpose of this study is to examine AI models and determine their causes of hallucinations, the study would go on to determine the underlying causes by examining the people who are working on AI and automation. Data would be collected from 200 individuals which would include students, start-ups, and managers to evaluate creative techniques to mitigate the use of AI.

Background of Study

When AI began back in the 20th century it was simpler, but it has developed to the complex form that can be seen today, since work on AI began hallucinations also began from a simpler spectrum to a more complex one(Natale & Ballatore, 2020). These nascent AI models were produced from basic logic and minimal datasets which caused them to become susceptible to Hallucinations. Equipped with an insufficient training data these models were prone to mistakes. Due to the simplicity of AI models back in the day, singling out hallucinations was easier thus opening ways to their resolution and detection in an easier manner (Challen et al., 2019). The dynamics of AI changed in the late 20th century when machine learning came to be. **Neuronal Networks** were considered to be a giant, as from the name it can be deduced that these models were based on the model of the human brain which went on to provide flexibility and intelligence in a better way as compared to the era before this innovation. Since, AI came to be on this model, it began resulting in even more complex hallucinations.

To further this after the advent of **Deep Learning**, utilizing deep neural networks with layers after layers which could learn from enormous quantities of information (Sze et al., 2017). Producing data of unprecedented complexity, the **convolutional neural network (CNN) and Recurrent Neural Network (RNN)** gained a lot of popularity(Hoffmann et al., 2017). CNN forte was in analysing visible interpretation and analysis while RNN specialized in analysing text-based data making it better at sequential data processing. Diving further into the problem of adequate training data for AI models, there seemed to be not a lack of data butt the lack of proper information due to which it began to create hallucinations. To further enhance the interplay between huge datasets and model architectures a **transformer-based models** were form, this was the most detailed model which processed massive amounts of information (Gillioz et al., 2020), now these hallucinations became even harder to detect and treat. There has been