Special Issue for 3rd International Conference on Information Literacy
Issue 73, December 2024
DOI: 10.70000/cj.2024.73.622

Cybrarians Journal
E-ISSN 1687-2215

Understanding the impact of AI Hallucinations on the university community

Research – Literature review

Hend Kamel
Librarian, New Giza University, Egypt
Hend.kamel@ngu.edu.eg

Abstract

Since we live in the era of the information revolution, finding trusted and accurate information takestime and effort made students and researchers aim to find an easier way. Generative AI (Artificial Intelligence) tools offer an easy solution for accessing the required information easily and accessible;however, these tools rely on vast datasets to predict statistically probable outputs, not guaranteed accuracy. This can lead to misinformation, factual errors, biases, and fabricated content, which is termed "hallucinations." The research problem focuses on the challenges of detecting these AI hallucinations,the main issue for all users of AI technologies. The main objective of the study is to raise awareness about AI hallucinations and promote the ethical and effective use of AI tools among New Giza University students, faculty, and staff. This involves the approach to understanding the biases and errors associated with AI outputs. Methodologically, the study will employ a mixed-methods approach, combining quantitative analyses of AI tool accuracy with collecting qualitative data via survey of users across a range of fields to gather insights on the impact of AI hallucinations. The expected results of this research are to reveal the pitfalls that researchers might run into when relying on AI technology for their work. Additionally, the findings will contribute significantly to information literacy programs, by advocating for the including of AI tool

assessments within the broader information literacy curriculum and equipping users with the skills tocritically evaluate AI-generated content.

Keywords
Artificial Intelligence (AI), Generative AI, Natural language processing (NLP) – AI Hallucinations, Information literacy, AI literacy

Literature Review

AI tools spread caused a huge effect on students, researchers and even citizens' attitude and interaction with information. It became a main part for all of us in our daily life. serving as a primary guide in everything from simple tasks to complex decision-making.

Artificial intelligence is one of the most important reasons for this transformative field in computer science. It produces systems that make our daily life easier as it can do human intelligence duties such as learning, reasoning, and even creativity. machine learning which made computers able to learn from various data and adapt its responses according to it (Saudi Data and Artificial Intelligence Authority, 2023).
As AI technologies have evolved, so have tools. One of the most significant advancements has been the development of **Generative AI**, which can understand and respond to natural language inputs after they trained on large datasets and algorithms led it to generate human-like text, by natural language processing (NLP) or understanding and responding the human input (Gold, n.d.). Generative AI (GAI) uses machine learning and neural networks to automatically generate fresh and original content, such as images, text, and videos. This advancement represents a major step forward in AI capabilities. However, the term artificial intelligence (AI) encompasses a wider range of applications, with generative AI being a notable example. (Al-Khalifa, 2023).

Large Language Models (LLMs), such as GPT-3 "Generative Pre-trained Transformer 3", generate text that is both coherent and contextually appropriate. by being trained on massive datasets, allowing them to produce relevant text based on a given prompt or input even if it was simple (Najjar, 2023).

This capability in LLMs (Large Language Model) made them invaluable in various fields, from academic research to creative writing. Despite these models' abilities, they don't understand the content they produce and sometimes create confusing or unsuitable outputs. So, Human supervision is essential to guide these models and ensure their results align with the intended purpose (Najjar, 2023).
Large Language Models (LLMs) are the driving force behind generative artificial intelligence. These models consist of highly complex layers of neural networks called Transformers, which can handle various tasks in natural language processing, such as text generation, summarization, translation, answering questions, and text classification. These models learn from vast amounts of textual data and use complex algorithms to identify patterns and relationships between words and concepts.
One of the most known examples of these models is GPT-3, which is the core of the ChatGPT model developed by OpenAI. GPT-3 can generate text that resembles human writing. The term "large" refers to the huge size of the model's training data set and hyperparameters, which are sometimes measured in petabytes. Hyperparameters are the memory and knowledge that the model has learned during training. They determine the model's ability to solve tasks, such as predicting the next piece of text. (Al-Khalifa, 2023).

Despite the advancements in deep learning and natural language processing (NLP) "A field of artificial intelligence and linguistics that studies the problems inherent in the processing and manipulation of natural language, with an aim to increase the ability of computers to understand human languages." (IBM Corporation.2024). A big issue with the AI-Generated output is the phenomenon of "**Hallucinations".** AI responses are still prone to hallucinate unintended, irrelevant, or incorrect text (Ji et al., 2022). Hallucinations which stand for "A response from a foundation model that includes off-topic, repetitive, incorrect, or fabricated content" (IBM Corporation.2024).

The **Causes** of these hallucinations are rooted in the way AI models are trained. When prompted to generate text, a model relies on its training data to produce a response. However, if the model does not have enough relevant information to draw from, it may resort to fabricating details, leading to outputs that are inaccurate or misleading. These hallucinations not only degrade the performance of AI systems but also undermine the trust users place in these technologies, especially in critical real-world scenarios (Ji et al., 2022).

"**Prompt Engineering"** is a field focused on developing and crafting commands directed at generative artificial intelligence. This process involves methods for

DOI: 10.70000/cj.2024.73.622

effectively and systematically communicating with language models, such as ChatGPT, to achieve desired results. Designing prompts requires understanding various factors, including the language models used, the context, the purpose of the prompt, and how the AI interprets the given command.
As Saudi Data and Artificial Intelligence Authority,2023 mentioned This simple approach can help when prompting to gain the best results:

- **Clarify the Context of the Request:** Provide a clear background and purpose for the request to set the stage for what you want to achieve.
  Researcher's Comment: Providing context helps to layout the request, ensuring that the response is relevant and aligns with your goals.

- **Define the Model's Personality:** Specify the type of persona or style you want the model to adopt in its responses. For example, should the tone be formal, informal, technical, or educational?
  Researcher's Comment: Defining the persona helps tailor the response to your needs, making it more applicable to your specific situation.

- **Use Specific Symbols:** Using symbols or markers to highlight key points in the input that need focus. These can be bullet points, numbers, or special characters.
  Researcher's Comment: Symbols help organize information clearly, making it easier to identify and address important aspects of the request.

- **Request Structured Outputs:** Asking for a structured format output such as lists, tables, or detailed reports, to ensure clarity and ease of access.
  Researcher's Comment: Structured outputs help in analyzing and interpreting the information effectively, which provides a clear and organized response.

- **Verify Input Accuracy:** Check the accuracy of the information provided as an input. Ensuring that data and details are correct helps avoid errors or misunderstandings in the response.
  Researcher's Comment: Accurate inputs are crucial for generating reliable outputs. Verification prevents potential issues and enhances the quality of the response.

- **Provide Successful Examples:** Offer clear examples of similar requests handled successfully to guide the model on how to approach the task.

DOI: 10.70000/cj.2024.73.622

Researcher's Comment: Examples serve as practical guides, showing how to progress similar requests and setting expectations for the desired outcome.

- **Outline Required Steps:** Define steps one by one to complete the request. Provides a clear plan of action for each stage of the process.
  Researcher's Comment: A well-defined process helps push the workflow and ensures that all necessary steps are followed, leading to more effective results.

- **Check the Outputs:** After receiving the response, make sure that it meets the requirements. Ensure that the results align with your expectations.
  Researcher's Comment: Reviewing the outputs is crucial as it ensures they are accurate and meet the required standards, confirming that the request has been properly addressed.

- **Use Specific References:** Provide specific references or sources to use for additional support or information related to the request.
  Researcher's Comment: References add credibility and provide a base for the exploration, helping to substantiate the information and findings.

- **Apply Iterative Methods:** Use iterative approaches to review and revise information. Reassess and adjust the input as necessary to improve accuracy and effectiveness.
  Researcher's Comment: Iterative methods are essential for developing the prompting technique and refine responses and enhancing accuracy.

Overall, prompt engineering is a developing skill, acts as a bridge between humans and artificial intelligence, helps presenting commands or questions in a way that ensures the AI produces the desired outcomes.

DOI: 10.70000/cj.2024.73.622