Hallucinations in LLMs

1st Mariia Paik
Techical University Of Kosice

Kosice , Slovakia
mariia.paik@student.tuke.sk

2nd Heorhii Fedulov
Techical University Of Kosice)

Kosice , Slovakia
heorhii.fedulov@student.tuke.sk

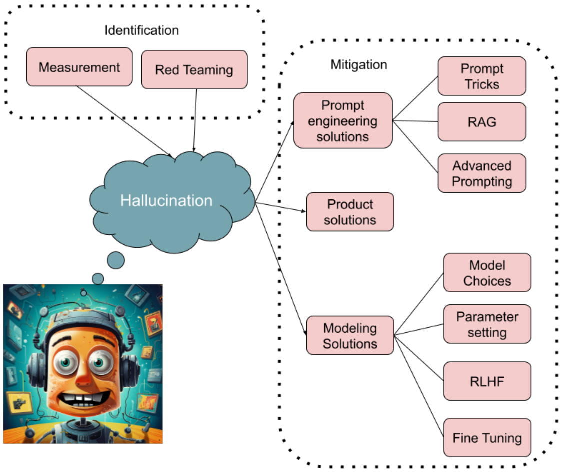

Fig. 1. A multifaceted approach to mitigating LLM hallucinations

***Abstract*—Hallucinations in large language models (LLMs) are a growing concern as these models become more widely used. This paper reviews the literature on hallucinations in LLMs, focusing on their causes, implications, and potential solutions. The paper finds that hallucinations are caused by a number of factors, including the large size and complexity of LLMs, the lack of access to external knowledge, and the pressure to generate fluent and grammatically correct text. Hallucinations can have a number of negative implications, such as the propagation of misinformation and the creation of false narratives. The paper also discusses a number of potential solutions for mitigating hallucinations, such as providing LLMs with more context and using fact-checking and bias detection techniques. The paper concludes by calling for more research on the nature and causes of hallucinations in LLMs.**

Keywords: Hallucinations , Large Language Models (LLMs) , GPT (Generative Pre-trained Transformer) , ChatGPT , Misinformation , Fact-checking , Ethical considerations Artificial intelligence , Medicine , Patient confidentiality , Autonomous systems , Text generation , Verification techniques

I. Introduction

Hallucinations in LLMs (short for Large Language Models like ChatGPT, for example) are an intriguing phenomenon that has not yet received the scientific attention it deserves. In this context, hallucination refers to the model's ability to generate information that appears to be true or based on real data but is, in fact, fiction.

In recent years, as the power and sophistication of artificial intelligence language models have grown, their use has become increasingly popular in fields ranging from content creation to decision support. However, along with the potential benefits come risks, and one such risk is the possibility of hallucinatory information from the model.

Addressing the topic of hallucinations in LLMs is especially important for several reasons:

1) Impact on society: As language models become more popular, their influence on public opinion and decision-making increases. Understanding the hallucinatory properties of LLMs can help ensure that information from such models is used responsibly and consciously.
2) Safety and reliability: Hallucinations can pose a risk in areas where the accuracy of information is critical, such as medical, legal, or financial systems.
3) Ethical considerations: When we entertain hallucinations, we may embrace incorrect or distorted perceptions of reality, which can lead to inappropriate, unethical, or even dangerous actions.
4) Education and Awareness: For end-users such as businesspeople, researchers, students, and individuals using AI in their daily lives, understanding this issue can assist in making more informed decisions when interacting with LLMs.

The aim of this paper is to review what different studies say about hallucinations in LLMs, where and how these models are used, and how such hallucinations can affect humanity. In this way, we attempt to systematize the existing knowledge on this issue, identify the main problems and challenges, and suggest possible solutions.

The paper is structured as follows: introduction, methodology, literature review, and conclusions.

II. Metodology

The following databases were used to select articles: Elsevier, Scopus, ScienceDirect, IEEE Xplore, and ACM Digital Library. The selection criteria included the keywords: 'hallucinations in LLMs' and 'GPT'. The study covers papers published between 2022 and 2023, and Windows software was used to analyze the data

III. Review

The articles were divided into thematic blocks such as: understanding and implications, hallucinations in natural language, use in jurisprudence, use in medicine, how hallucinations can be mitigated . In each block, key findings were highlighted:

A. Understanding and Implications:

The articles in this block discuss that interest in consultative agents in the fields of information retrieval, artificial intelligence and natural language processing, focus on didactic lectures supplemented by interactive exercises, provide precise definitions of hallucinations and creativity in the context of GPT models because of this there is a high need to improve the activity of consultative agents, especially after the development of ChatGPT type models. A metric based on the normalised entropy of GPT model predictions has been proposed to quantify creativity. The study shows that hallucinations can increase the creativity of the model by allowing it to consider a wider range of token sequences.Due to the sheer volume and training data, LLMs can generate extensive and deep responses even with a limited number of queries. The paper also addresses controversial issues associated with LLMs, in particular their potential to propagate human biases.

1) Conclusions::

- Hallucinations in LLMs are becoming a critical area of research because of the growing influence of counselling agents on public opinion and decision-making.
- There is a need to improve the proactive abilities of these agents so that they can actively participate in the conversation rather than simply reacting to user requests.
- Through interactive exercises, researchers and practitioners can better understand the limitations of current systems and possible ways to improve them.
- Hallucinations are an intrinsic attribute of GPT models and can occur even in well-trained models.
- It may not be possible to completely eliminate hallucinations without sacrificing other desirable characteristics such as creativity and adaptability.
- There is a trade-off between hallucinations and creativity.
- In addition to this research, an in-depth study of the vanishing gradient problem in multilayer networks is proposed to better understand how this may affect hallucinations in GPT models.
- LLM capabilities and limitations: Although LLMs like GPT-3 demonstrate an amazing ability to understand and generate relevant content, they may have inherent biases or limitations that need to be addressed.

B. Hallucinations in natural language in LLMs:

The study provides a framework for analysing and measuring the abilities of ChatGPT as a universal language model. The aim of the work was to investigate the capabilities of ChatGPT in generating effective academic text. The developed framework incorporates six principles related to artificial language processing to study the accuracy and professionalism of the text generated by the algorithm. After applying ChatGPT to generate academic texts, a critical analysis was conducted based on the proposed principles. The study revealed flaws in the performance of ChatGPT, including repetition of information, non-factual inferences, logical errors, unreliable sources, and hallucinations.

1) Conclusions::

- The framework and principles developed can be used to understand the features of academic language produced by ChatGPT.
- Despite the exceptional capabilities of ChatGPT, it has serious shortcomings in academic writing, including hallucinations and lack of pragmatic interpretation.
- ChatGPT can generate grammatically correct sentences, but is not good enough for professional academic analysis and writing.
- Using ChatGPT for academic purposes requires caution. It can serve as an aid for providing information, but researchers should not rely on the tool for writing.

C. Use in jurisprudence:

Traditionally, the automatic generation of short abstracts of judicial decisions in jurisprudence has been done using abstract summarisation techniques. However, in recent years abstract summarisation models have become popular as they are able to generate more natural and logical abstracts. Currently, there are specialised pre-trained abstract summarisation models for the legal domain. Moreover, publicly available pre-trained large language models (LLMs) such as ChatGPT are known for their ability to generate high quality text and perform text summarisation. Therefore, it is natural to ask whether these models are ready for mass use to automatically generate abstract abstracts of judgements. To investigate this question, we apply several state-of-the-art specialised abstract summarisation models and publicly available LLMs to Indian court judgements and test the quality of the generated abstracts. In addition to standard abstract quality metrics, we also check for inconsistencies and hallucinations in the abstracts. We find that abstract summarisation models typically achieve slightly higher scores than extractive models in terms of standard summarisation quality metrics such as ROUGE and BLEU. However, we often detect inconsistent or fictitious information in the generated abstract summarisations. Overall, our study shows that pre-trained abstract summarisation models and LLMs are not yet ready for fully automatic application in adjudication summarisation; a human-assisted method involving manual checking for inconsistencies is currently more appropriate.

1) Conclusions::

- Traditional extractive summarisation methods in jurisprudence are getting competition from abstract summarisation models, which are able to generate more natural and logical abstracts of judicial decisions.
- There are specialised pre-trained abstract summarisation models for the legal domain, but publicly available large

language models such as ChatGPT can also be used for text summarisation.

- The study applied various state-of-the-art summarisation models to judicial decisions in Indian courts and evaluated the quality of the generated abstracts using standard metrics.
- Abstract summarisation models in most cases performed slightly better than the extractive models in terms of standard metrics such as ROUGE and BLEU.
- However, it is important to note that inconsistencies and hallucinations were common in the generated abstract abstracts, which can pose problems in the accuracy and reliability of summarisation.
- Overall, the study indicates that pre-trained abstract summarisation models and LLMs are not yet ready for fully automatic application in adjudication summarisation, and human involvement is currently required to check for inconsistencies and hallucinations in the generated abstracts.

D. Use in medicine:

This review highlights the optimism surrounding the use of LLMs, such as ChatGPT, in medicine. However, there are concerns about reliability. Research has shown that ChatGPT-3.5 makes significant mistakes in bibliographic research. This includes "hallucinations" of text, where the model produces information that does not match the original data. Another study explored the ability of LLMs (Bing Chat, ChatGPT 3.5 and 4.0) to answer ophthalmology-style exam questions. The results showed that ChatGPT-4.0 and Bing Chat perform comparably to human respondents but are also prone to hallucinations and illogical conclusions. The potential of LLMs in medical tasks is emphasized, but there are also discussions on issues related to associations and bias in training data. It is recommended that models be trained on specialized medical data and adjusted to ensure safety and accuracy. The application of LLMs in urology is considered, highlighting their potential for differential diagnosis, surgeon and patient training. However, authors also point to the issue of "hallucinations" in model responses and the need to ensure patient confidentiality. Lastly, the use of chatbots based on large language models (LLMs) in medicine is explored. The authors evaluated the ability of chatbots, such as ChatGPT, to answer administrative, medical, and complex medical questions. They also discussed the potential application of these technologies for managing incoming messages in electronic health records (EHR).

1) Conclusions::

- ChatGPT 4.0 demonstrated better performance than ChatGPT 3.5 for administrative and uncomplicated medical questions as well as complex medical questions.
- Potential applications of these chatbots include managing incoming messages in the electronic health record (EHR), which can reduce the workload of medical staff.
- Obtaining regulatory approval for the use of such technology in medicine requires extensive supervisory training by experts in the relevant field, establishing measures to prevent "hallucinations" and ensure privacy, and demonstrating that these chatbots can compete with (or even outperform) human experts.
- Integrating LLMs-based chatbots with electronic medical records can help with patient communication and patient education.
- Continued refinement and discussion of its ethical and legal aspects are necessary for the successful use of this technology in medicine.

E. How to improve text generation and why it's important:

The study discusses the possibility of solving the hallucination problem in LLM using formal methods. Formal methods have previously been used in cyber-physical and autonomous systems to provide strong guarantees about the behaviour of the system and to verify that it conforms to expected specifications. However, the limited scalability of these methods has limited their application. The study proposes that LLM can be used as an effective but under-reliable search tool in autonomous tasks and formal methods can be used to detect inconsistencies and eliminate hallucinations.

1) Conclusions::

- Large language models such as ChatGPT and GPT-4 have great potential for natural language text generation, but are prone to hallucinations, making their responses factually incorrect.
- Formal methods can be used to identify and eliminate hallucinations in LLM responses. This helps to ensure the correctness and reliability of the answers, especially in mission critical applications.
- The proposed methodology involves an iterative process where LLMs refine their responses using formal methods until correct responses are achieved.
- Experiments conducted on example planning tasks have shown successful reduction of hallucinations and obtaining correct solutions using the proposed methodology.
- Further research in this area includes automatically extracting specifications from query text and using more advanced formal verification techniques to improve the robustness of LLM in mission-critical applications.
- ChatGPT can be a useful tool in scientific writing if the authors have conducted an extensive literature review and provided brief notes for each reference. The model can assemble logically related text from these notes, which can speed up the writing process.
- The use of large language models in scientific writing raises debates in terms of ethics and acceptability, and risks creating false experts in the medical field. Therefore, it is suggested that policies and practices for the evaluation of scientific manuscripts should be modified to maintain high standards of scientific integrity and that

artificial intelligence detectors should be introduced into the peer review process.

F. Training States of Large Language Models:

The attributes and behaviors of LLMs are deeply intertwined with their training processes. LLMs undergo three primary training stages: pre-training, supervised fine-tuning (SFT), and reinforcement learning from human feedback (RLHF). Analyzing these stages provides insight into hallucination origins in LLMs, as each stage equips the model with specific capabilities. Pre-training. Pre-training is generally considered a crucial stage for LLM to acquire knowledge and skills (Zhou et al., 2023a). Language models, during pre-training, aim to predict the next token in a sequence autoregressively. Through selfsupervised training on extensive textual corpora, the model acquires knowledge of language syntax, world knowledge, and reasoning abilities, providing a robust foundation for subsequent fine-tuning tasks. Besides, recent research (Sutskever, 2023; Delétang et al., 2023) suggests that predicting subsequent words is akin to losslessly compressing significant information. The essence of language models lies in predicting the probability distribution for upcoming words. Accurate predictions indicate a profound grasp of knowledge, translating to a nuanced understanding of the world. Supervised Fine-Tuning. While LLMs acquire substantial knowledge and capabilities during the pre-training stage, it's crucial to recognize that pretraining primarily optimizes for completion. Consequently, pre-trained LLMs fundamentally served as completion machines, which can lead to a misalignment between the next-word prediction objective of LLMs and the user's objective of obtaining desired responses. To bridge this gap, SFT (Zhang et al., 2023d) has been introduced, which involves further training LLMs using a meticulously annotated set of (instruction, response) pairs, resulting in enhanced capabilities and improved controllability of LLMs. Furthermore, recent studies (Chung et al., 2022; Iyer et al., 2022) have confirmed the effectiveness of supervised fine-tuning to achieve exceptional performance on unseen tasks, showcasing their remarkable generalization abilities.

IV. Experiments

To thoroughly assess the impact of the tagged context prompts on question answering using generative language models, we conducted various experiments involving multiple permutations of context, question, and engine. This comprehensive approach allowed us to investigate the influence of inserting tags into source contexts on the generation of hallucinated information and the LLMs' ability to validate responses using random "known good" tags that could not be synthesized by the models otherwise. Each experiment iteration included the following components:

- No-context Question: We used one of the pre-generated questions which did not include any supporting context. To encourage the model to provide citations, we appended the instruction "Provide details and include sources in the answer." to the question.
- Tagged-context Question: We incorporated the pre-generated context tagged as described earlier. To ensure a comprehensive evaluation, each context was applied across all questions, making it relevant for one group of questions and not relevant for the others.
- Generative Language Model Selection: We tested the prompts on a variety of recent generative language models, including GPT-4, GPT-4-0314, GPT-3.5-turbo, GPT-3.5-turbo-301, text-003davinci, davinci-instruct-beta, and curie-instruct-beta. The results of each experiment were stored for further analysis

For this study, we generated a total of 3,430 prompt-response pairs throughout the experiments. These responses were divided into different categories based on the context provided to the generative language models. The breakdown of these pairs is as follows:

- No-context Questions (1,715): These prompt-response pairs centered on questions that did not include any context and relied entirely on the generative model's inherent knowledge to produce appropriate answers.
- Relevant-context Questions (245): This category comprised question prompts that were accompanied by context directly applicable to the inquiries. Consequently, these contexts aimed to guide or enhance the generative model's response capabilities.
- Mismatched-context Questions (1,470): In this set of prompt-response pairs, the questions were paired with contexts that did not align with their subject matter. This enabled us to study how mismatched information could impact LLMs' performance and if they would still adhere correctly to their training without generating hallucinations

V. Conclusion

In conclusion, hallucinations in large language models (LLMs) are a complex and multifaceted issue with far-reaching implications for society. As LLMs become more widely used, their ability to generate realistic but factually incorrect information poses a significant risk to public discourse and decision-making. The studies reviewed in this paper have identified a number of factors that contribute to hallucinations in LLMs, including the large size and complexity of these models, the lack of access to external knowledge, and the pressure to generate fluent and grammatically correct text. The implications of hallucinations in LLMs are manifold. These models can be used to propagate misinformation, create false narratives, and undermine trust in institutions. In addition, hallucinations can have a negative impact on the development of critical thinking skills and the ability to distinguish between fact and fiction. A number of potential solutions for mitigating hallucinations in LLMs have been proposed. These include providing LLMs with more context, using fact-checking and bias detection techniques, and developing methods for evaluating the trustworthiness of LLM-generated information. However, it is important to note that there is no easy solution to the problem of hallucinations in LLMs. As these models become more powerful, it is likely that they will continue to generate realistic but factually incorrect information. Therefore, it is essential to develop a critical understanding of the limitations of LLMs and to be aware of the potential risks associated with their use. In addition to the research directions outlined in the reviewed studies, further research is needed to:

- Develop a deeper understanding of the causes of hallucinations in LLMs
    - Currently, there is limited understanding of what leads to hallucinations in LLMs. Researchers have identified a number of potential factors, including:
        - The large size and complexity of LLMs. LLMs are trained on massive datasets of text and code, which may contain errors and biases. This can lead to LLMs generating information that is not consistent with reality.
        - he lack of access to external knowledge. LLMs are typically trained on internal data, which may be incomplete or outdated. This can lead to LLMs generating information that is not current or accurate.
        - The pressure to generate fluent and grammatically correct text. LLMs are often evaluated on their ability to generate text that is grammatically correct and coherent. This can lead to LLMs generating text that is persuasive but is not factually accurate.
- Develop more effective methods for detecting and mitigating hallucinations
    - Currently, there are a number of methods for detecting and mitigating hallucinations in LLMs. These methods include:
        - Fact-checking. Fact-checking systems can be used to verify the accuracy of information generated by LLMs.
        - Reliability assessment. Reliability assessment systems can be used to assess the level of trust in information generated by LLMs.
- Develop methods for educating users about the limitations of LLMs
    - User education. Curriculums and materials can be used to educate users about how to use LLMs safely and responsibly.
    - Transparency. LLM developers should be transparent about the limitations of their models.

References

[1] Leippold, Markus. ”Thus spoke GPT-3: Interviewing a large-language model on climate finance.” *Finance Research Letters* 53 (2023): 103617.

[2] J. Lee, Minhyeok. ”A mathematical investigation of hallucination and creativity in gpt models.” *Mathematics* 11.10 (2023): 2320.

[3] Liao, Lizi; Yang, Grace Hui; Shah, Chirag. ”Proactive Conversational Agents in the Post-ChatGPT World” (2023)

[4] Deroy, Aniket; Ghosh, Kripabandhu; Ghosh, Saptarshi ”How Ready are Pre-trained Abstractive Models and LLMs for Legal Case Judgement Summarization?” (2023)

[5] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, Pascale Fung. ”Survey of Hallucination in Natural Language Generation” (2023)

[6] Mahyoob, Mohammad (57222084210); Algaraady, Jeehaan (57223980669); Alblwi, Abdulaziz ”A Proposed Framework for Human-like Language Processing of ChatGPT in Academic Writing” (2023)

[7] Au Yeung, Joshua; Kraljevic, Zeljko; Luintel, Akish; Balston, Alfred; Idowu, Esther Dobson, Richard J., Teo, James T. (”AI chatbots not yet ready for clinical use” (2023)

[8] McGowan, Alessia (58088693300); Gui, Yunlai (58504026200); Dobbs, Matthew; Shuster, Sophia; Cotter, Matthew; Selloni, Alexandria; Goodman, Marianne; Srivastava, Agrima; Cecchi, Guillermo A. ; Corcoran, Cheryl M. ”ChatGPT and Bard exhibit spontaneous citation fabrication during psychiatry literature search” (2023)

[9] Javid, Mohamed; Reddiboina, Madhu; Bhandari, Mahendra. ”Emergence of artiflcial generative intelligence and its potential impact on urology” (2023)

[10] Jha, Susmit; Jha, Sumit Kumar; Lincoln, Patrick; Bastian, Nathaniel D.; Velasquez, Alvaro; Neema, Sandeep. ”Dehallucinating Large Language Models Using Formal Methods Guided Iterative Prompting” (2023)

[11] Metze, KonradinFlorindo, João B. et al. ”Bibliographic Research with ChatGPT may be Misleading: The Problem of Hallucination” (2023)

[12] Louis Z. Cai, Abdulla Shaheen, Andrew Jin, Riya Fukui, Jonathan S. Yi, Nicolas Yannuzzi, Chrisfouad Alabiad. ”Performance of Generative Large Language Models on Ophthalmology Board–Style Questions” (2023)

[13] Anand Athavale MD, Jonathan Baier BS, Elsie Ross MD, MSc, Eri Fukaya MD, PhD. ”The potential of chatbots in chronic venous disease patient management” (2023)

[14] Philip Feldman, James R. Foulds, Shimei Pan. ”Trapping LLM Hallucinations Using Tagged Context Prompts ” (2023)

[15] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, Ting Liu. ”A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions”

[16] Florian Leiser, Sven Eckhardt, Merlin Knaeble, Alexander Maedche, Gerhard Schwabe, Ali Sunyaev. ”From ChatGPT to FactGPT: A Participatory Design Study to Mitigate the Effects of Large Language Model Hallucinations on Users” (2023