PROTECT YOURSELF FROM AI HALLUCINATIONS: EXPLORING ORIGINS AND BEST PRACTICES

Jana Dannhoferová[1], Petr Jedlička[1]

[1]Department of Informatics, Faculty of Business and Economics, Mendel University in Brno, Zemědělská 1, 613 00 Brno, Czech Republic

ABSTRACT

Although AI-powered chat systems like ChatGPT can be trusted, we shouldn't rely on them completely. They can sometimes produce irrelevant, misleading or even false responses, known as hallucination effects. The causes can be both systemic and user related. User behavior, particularly in the area of prompt engineering, has an impact on the quality and accuracy of the result provided. Based on the literature review, we have identified the most common types of hallucination effects and provided examples in created categories. Finally, we have highlighted what users should consider when writing prompts and given recommendations for them to minimize hallucination effects in responses obtained from AI systems. Understanding how hallucinations occur can help ensure that these powerful tools are used responsibly and effectively. However, the quality of responses is always a matter of judgment, and the user's level of expertise and critical thinking is an important factor.

Keywords: artificial intelligence, AI systems, large language models, hallucination effect, text-to-text prompt engineering

JEL Code: O33, J24

1 INTRODUCTION

Artificial Intelligence (AI) is rapidly penetrating various areas of human activity. Despite the strengths of AI, generative AI models such as GPT have limitations and weaknesses. One of these is the fact that they can produce seemingly credible but incorrect, irrelevant, misleading or even false answers (Shen et al., 2023), (Xiao and Wang, 2021). This phenomenon, known as the hallucination effect, is a common problem in many large language models (LLMs) (Athaluri et al., 2023), (Bang et al., 2023), (Rohrbach et al., 2018). According to Bernard Marr (2023), hallucination in AI refers to the generation of output that may sound plausible but is either factually incorrect or unrelated to the given context. An AI hallucination is when an

https://doi.org/10.11118/978-80-7509-990-7-0049

AI makes up false information or facts that aren't based on real data or events (Keary, 2024). Hallucinations are so common that OpenAI even warns users within ChatGPT that "ChatGPT may produce inaccurate information about people, places, or facts".

1.1 Why is AI hallucination a problem?

The hallucination effect is an issue that can hinder user trust in AI systems. If the misinformation hallucinated by AI spreads like wildfire on the internet, making it appear authoritative and written by humans, it will undermine user confidence – making it difficult for users to trust information on the internet (Sushir, 2024). If AI systems produce incorrect or misleading information, users may lose trust in the technology, hindering its adoption in various sectors (Marr, 2023).

They are also ethically problematic, as inconsistencies in training data can lead to the mass dissemination of misinformation (Sushir, 2024). LLMs increase the ability to create realistic AI-generated fake content, which plays an important role in the misinformation phenomenon that our world is currently facing (Bontridder and Poullet, 2021). AI systems can expose users to legal threats if their responses are inaccurate or misleading. This is confirmed by Athaluri et al. (2023), who state that hallucinations can raise a number of ethical and legal issues and negatively affect decision-making. Poor decisions in areas such as health care can have serious consequences.

Although several studies have already described the weaknesses of AI systems, no one has yet addressed the behavior of the users of these systems, which can also affect the accuracy of the result provided. There is also a lack of categorization of types of errors, including corrective measures.

The aim of this paper is to identify the most common types of hallucination effects that can occur in responses generated by AI systems, provide examples and identify the main causes. Based on the results of the literature review, appropriate recommendations will be provided that will lead to the minimization of hallucination effects in the generated responses.

Research questions:

- What are the most common types of hallucination?
- What can users do to minimize the hallucination effect?

Categorizing the types of hallucination errors will help users of AI systems to understand what types of errors to look out for and how to modify their behavior to minimize these errors in their results.

2 LITERATURE REVIEW

2.1 LLMs

Large Language Models (LLMs) are deep learning models trained to understand and generate natural language. They use a two-stage training pipeline to learn efficiently from data. In the initial pre-training stage, LLMs use a self-supervised learning approach, which allows them to learn from large amounts of unannotated data without the need for manual annotation. In the subsequent fine-tuning phase, LLMs are trained on small, task-specific, annotated datasets to use the knowledge gained in the pre-training phase to perform specific tasks as intended by end users. As a result, LLMs achieve high accuracy on various tasks with minimal human-provided labels (Shen et al., 2023). The conversational artificial intelligence (AI) systems such as ChatGPT simulate a conversation with a human (Gupta et al., 2020). They can answer questions and provide information. The use of LLMs helps them to learn the grammar, syntax and context of different languages or subjects.

2.2 Limits of LLMs

Research has revealed significant gender and racial bias in AI systems. Some facial analysis software couldn't recognize a dark-skinned face until a person put on a white mask. When given the task of guessing the gender of a face, some systems performed significantly better on male faces than female faces (Buolamwini, 2019). Another major problem with the LLM is political bias. A team of researchers from the Technical University of Munich and the University of Hamburg provided evidence that ChatGPT has a "pro-environmental, left-libertarian orientation" (Hartmann et al., 2023). The same conclusion was reached by Fujimoto and Takemoto (2023). These results often arise from the AI model's inherent biases, lack of understanding of the real world, or limitations of the training data. In other words, the AI system 'hallucinates' information on which it has not been explicitly trained, leading to unreliable or misleading responses (Marr, 2023).

According to Athaluri et al. (2023) AI hallucination usually occurs due to adversarial examples such as varied input data that confuse the AI systems into misclassifying and misinterpreting them, resulting in inappropriate and hallucinatory output. Bang et al. (2023) they concluded that the ChatGPT suffers from hallucination problems like other AI systems and it generates more extrinsic hallucinations from its parametric memory as it does not have access to an external knowledge base. T. Sushir (2024) describes four basic types of hallucination:

- Sentence contradiction: This occurs when an LLM model produces a sentence that completely contradicts its previously asserted sentence.
- Factual contradiction: This type of hallucination occurs when the AI model presents false or fictitious information as fact.
- Prompt contradiction: This type of hallucination occurs when the output contradicts the prompt for which it generated an output.
- Random or irrelevant hallucinations: This hallucination occurs when the model produces output that is completely irrelevant to the given prompt.

Hallucinatory errors and weaknesses in AI models are usually caused by the following:

Cause	Description of cause
Misinterpretation of ambiguous input	LLMs may misinterpret ambiguous statements, leading to inaccurate or unintended responses.
Lack of contextual understanding	Chatbots may struggle to maintain the context of a conversation, resulting in responses that seem unrelated or inappropriate.
Overconfidence	An AI model that makes overly confident predictions even when faced with uncertain or ambiguous input, leading to inaccurate responses.)
Lack of common sense reasoning	AI models may lack common sense reasoning, leading to responses that seem illogical or impractical in certain situations.
Failure to recognize sarcasm or irony	AI systems may struggle to recognize sarcasm or irony in text, leading to literal interpretations and potentially incorrect responses.

Tab. 1 Causes of hallucinatory errors

Kenny Lee (2023) identified three main factors that cause LLMs to hallucinate: training data, lack of objective orientation, and inappropriately worded sentences. First, large language models have been developed through unsupervised training on large and heterogeneous datasets. These datasets come from many sources, making it difficult to ensure their impartiality and factual accuracy. The language model alone is not capable of distinguishing between truth and falsehood. Moreover, the inclusion of diverse and subjective perspectives within the training data further hampers the model's ability to discern objective truths. As a result, the model generates outputs that are likely, based on the patterns it has learned during the training process. Secondly, LLMs are susceptible to producing incorrect output when tasked with functions outside their training scope. Models such as GPT, Palm and Cohere are designed for broad natural language processing tasks. As a result, they may struggle to make accurate judgements when dealing with queries that require specialist knowledge in areas such as medicine, law and finance. Thirdly, to operate a LLM, users enter text as prompts. These prompts guide the LLM to perform certain tasks, similar to programming, but using natural language rather than programming languages. It is therefore essential that users write these prompts with the utmost precision. If the prompt is out of context, the LLM may produce an incorrect or completely unrelated response to what the user intended.

2.3 Prompt Engineering

It follows from the above that there is only one way for the user to suppress the hallucinatory effects, and that is to write the prompts appropriately. However, the quality of the prompts we give to generative AI models plays a critical role in determining the quality and relevance of the output. Clear, detailed and well-structured prompts are more likely to produce desirable results, while vague or irrelevant prompts can lead to unsatisfactory results. The methodology of designing effective requests or queries (prompts) to large language models is called prompt engineering. There are many strategies and guidelines on how to write prompts correctly when communicating with a generative AI model (Korzynski et al., 2023).

2.3.1 Multi-turn prompting

Based on the responses generated by the model. Prompts are thus structured as a series of turns or exchanges between the user and the model. This structure allows the model to consider the context of the entire conversation when generating a response, rather than just the most recent user input. This optimizes the quality and relevance of the generated responses (Bang et al., 2023). In this approach users provide input prompts and refine them over multiple turns or iterations.

2.3.2 Few-shot prompting

Few-shot prompting refers to a technique in which the user is provided with a small number of desired output examples prior to entering the task (Song et al., 2022). This approach aims to enable AI models to perform a given task with only a small number of labeled examples or prompts. In few-shot prompting, the AI model is trained or fine-tuned on a limited set of examples that demonstrate the desired task. These examples are typically provided as prompts to the model, which then learns to generalize from them to perform the task on new inputs. The focus of few-shot prompting is to adapt to a specific task or domain with minimal supervision, using the model's pre-trained knowledge to achieve task performance with few examples (Reynolds and McDonell, 2021).

2.3.3 Chain-of-Thought Prompting

Chain-of-thought (CoT) prompting focuses on facilitating multi-turn interactions or conversations between the user and the AI model, allowing for more coherent and contextually relevant responses. In chain-of-thought prompting, the prompts provided to the AI model are

structured to guide the flow of the conversation over multiple turns. The prompts are designed to build on the context established in previous turns, creating a coherent chain of thought throughout the interaction. The key feature of chain-of-thought prompting is its emphasis on maintaining context and coherence across multiple turns of the interaction, allowing for more natural and engaging conversations between the user and the AI model (Wei et al., 2022).

2.3.4 Self-Consistency

Self-consistency is an approach that simply asks a model the same prompt multiple times and takes the majority result as the final answer. It is a follow-up to CoT prompting and is more powerful when used in conjunction with it (Cheng et al., 2023), (Wang et al., 2022).

2.3.5 Temperature

The temperature parameter is a key feature in controlling the creativity and randomness of the responses generated by language models such as GPT-4. It essentially determines how conservative or adventurous the model is in its predictions. When writing prompts, users can specify the temperature parameter to adjust the level of randomness in the generated responses. The temperature value typically ranges between 0 and 1, with lower values producing more conservative and deterministic responses, and higher values producing more creative and varied outputs. Users can experiment with different temperature values to explore the creativity and variety of responses generated by the model. Lower temperatures (e.g. 0.1 or 0.2) tend to produce more predictable and coherent text that closely matches the training data. In contrast, higher temperatures (e.g. 0.7 or 1.0) introduce more randomness and variability, potentially leading to more imaginative but less reliable outputs (Mishra, 2023).

3 METHODOLOGY AND DATA

The basis of this study is a critical review of the findings that deal with the hallucination effects that can occur in the text responses provided by LLMs such as ChatGPT, Bard and Bing. By analyzing the existing research and reviewing the literature, we will attempt to map the various examples of hallucination effects and then organize them into categories. We characterize each category with a brief description and give a typical example from prompt engineering practice.

Next, we look at the causes of hallucination effects. A review of the literature shows that the causes of response errors can be both LLM and user related, particularly in the area of prompt engineering. In our study we will not look at model-side causes because we cannot control for them. However, we will focus on errors that occur on the user side. This is mainly because the user can change his or her behavior in the future, but has no way of influencing the behavior of the system itself.

The literature review also shows that there are currently a large number of different recommendations regarding the accuracy of prompts. We will try to select and provide those recommendations that will help to minimize hallucination effects. We hope that this paper will be a methodological guide for users, helping them to minimize errors in the responses generated by LLMs.

4 RESULTS

4.1 What are the most common types of hallucination?

By examining current research and reviewing the literature, we aimed to answer the research question: What are the most common types of hallucination effects? We looked for specific examples of hallucinations to identify different types and forms of hallucination effects that can manifest in responses. AI hallucinations can range from minor inconsistencies to completely false or fabricated responses, and could potentially mislead users who rely solely on the model's responses without independently verifying the facts.

4.1.1 Unrealistic Cause and Effect

However, AI systems can produce results that express unrealistic cause-and-effect relationships if the training data on which they have been trained contains biases or inaccuracies. For example, if an AI system trained on flawed data incorrectly concludes that "eating ice cream prevents sunburn", this would be a hypothetical example of expressing an unrealistic cause--and-effect relationship. Srinivasan and Chander (2021) give an example of unrealistic cause and effect when "a child wearing sunglasses is labeled as a failure, a loser, a nonstarter, an unsuccessful person".

4.1.2 Historical Revisions

Bernard Marr (2023) gives examples of LLMs inadvertently revising historical events. To the question, "When did Leonardo da Vinci paint the Mona Lisa?" he received the answer: "Leonardo da Vinci painted the Mona Lisa in 1815." This was incorrect because the Mona Lisa was painted between 1503 and 1506, or perhaps as late as 1517. When asked, "Tell me a fact about George Washington," he received the answer: "George Washington is known for inventing the cotton gin". These claims are unrelated, as Eli Whitney, not George Washington, invented the cotton gin.

4.1.3 Unsupported Claims

LLMs could make unsubstantiated claims without providing evidence. In a promotional video released by Google in February 2023, its AI chatbot Bard made a false claim. It incorrectly stated that "the James Webb Space Telescope has captured the first image of a planet outside our solar system" (Sushir, 2024), (Keary, 2024). However, this claim was inaccurate. LLMs can produce information that doesn't match the temporal sequences. For example, in the current conflict between Israel and Gaza, both the Bard and Bing systems incorrectly claimed that a ceasefire had been declared, probably based on news from May 2023. Bard then backtracked and said: "No, I am not sure that is correct. I apologize for my earlier response," but also made up casualty figures for two days into the future (Gillham, 2023).

4.1.4 Racial and Gender Bias

AI systems have been found to perpetuate racial bias in predicting recidivism rates. One notable example is the COMPAS system, which predicts that black defendants are at higher risk of reoffending than they actually are, while the opposite is true for white defendants (Cossins, 2018).

4.1.5 Misleading Information

Users may also receive information that is misleading or leads to incorrect conclusions. AI-generated writing was suspected when the Microsoft Start travel pages published a guide to places to visit in the Canadian capital, Ottawa. While there were errors in the details of some locations, most of the comments about the article were about how it included the Ottawa Food Bank as a tourist hotspot, encouraging readers to visit on an empty stomach (Gillham, 2023).

Amazon's Kindle Direct Publishing sold what appeared to be AI-written guides to foraging for edible mushrooms. One e-book encouraged the collection and consumption of legally protected species. Another mushroom guide included instructions that contradicted accepted best practices for identifying mushrooms that are safe to eat (Gillham, 2023).

4.1.6 Geographical Errors

AI-generated text can sometimes contain geographical errors, highlighting the importance of critically evaluating and fact-checking information generated by such models, especially when it comes to issues of geographical accuracy. Answers may include instances where places are incorrectly associated with different regions or countries, incorrect geographical boundaries or relationships between places, fictitious or non-existent places, or confusion between similar names. An example is the prompt "Name three cities in the United States" and the response "New York, Los Angeles, Toronto" (Lutkevich, 2023).

4.1.7 Random Output

Sometimes the output contradicts the prompt for which it generates output. LLMs can produce output that is completely irrelevant to the prompt given. For example, if the prompt is "Write an invitation to my friends for my birthday party". The model might generate output such as "Happy anniversary, Mum and Dad". (Sushir, 2024). A. Riaz (2023) confirms that AI systems can generate stories or narratives based on given prompts or data. However, due to limitations in understanding context or logical coherence, the stories generated may have nonsensical or illogical plots, resembling hallucinatory narratives.

4.1.8 Reasoning Errors

This type of error occurs when an AI system fails to apply correct logical reasoning or common sense to a problem. Reasoning errors are a significant challenge in AI, particularly for LLMs, which often struggle with tasks that require an understanding of the world that humans take for granted (Richardson and Heck., 2023).

We have divided the types of AI hallucinations into several categories, as shown in the Table 2.

It is important to recognize that all these examples illustrate the inherent limitations and potential risks associated with LLMs, particularly in terms of their ability to generate accurate and contextually appropriate information. AI hallucinations are a significant barrier to the reliability and accuracy of AI-generated content. Users are advised to approach the output of such models with caution and critical evaluation. Mitigating these challenges requires a comprehensive strategy that includes improved context awareness and user education.

4.2 What can users do to minimize the hallucination effect?

While the AI systems themselves cannot be influenced by users, the input they provide as prompts can. It's therefore crucial to provide clear and specific prompts, while unclear, inaccurate, inconsistent or contradictory prompts should be avoided (Sushir, 2024), (Lutkevich, 2023).

4.2.1 How to prevent AI hallucinations generally

While artificial intelligence has notable strengths, generative AI models such as ChatGPT have limitations and vulnerabilities. Hallucinations are considered an inherent part of LLMs. However, there are ways to reduce hallucinations. First, company owners must ensure that the AI model's training datasets are regularly updated and expanded to account for and keep up with cultural, political, and other evolving events (Sushir, 2024). In addition, AI hallucination can certainly be minimized by improving training inputs through the inclusion of diverse, accurate, and contextually relevant datasets, as well as frequent user feedback and the involvement of human reviewers to evaluate the outputs generated by an AI system Athaluri et al.

Name of category	Short description	Example of response
Unrealistic Cause and Effect	LLMs might suggest unrealistic cause-and-effect relationships.	"A child wearing sunglasses is labeled as a failure, loser, nonstarter, unsuccessful person." (Srinivasan and Chander, 2021)
Historical Revisions	LLMs may inadvertently revise historical events.	"Leonardo da Vinci painted the Mona Lisa in 1815." (Marr, 2023)
Unsupported Claims	LLMs could make unsupported assertions without providing evidence.	"The James Webb Space Telescope had taken the first image of a planet outside our solar system." (Sushir, 2024)
Racial and Gender Bias	LLMs perpetuate racial biases in predicting recidivism rates.	"Black defendants pose a higher risk of recidivism." (Cossins, 2018)
Misleading Information	LLMs might provide information that is misleading or leads to incorrect conclusions.	"The Ottawa Food Bank is a tourist hotspot" (Gillham, 2023)
Geographical Errors	LLMs may provide inaccurate information about locations.	"New York, Los Angeles and Toronto are three cities in the United States." (Lutkevich, 2023)
Random Output	LLMs generate completely irrelevant output to the given prompt.	"Happy anniversary, Mom and Dad." (Sushir, 2024)
Reasoning Errors	LLMs produce outputs that defy common sense.	Prompt: "When I was 6, my sister was half my age. Now I'm 70. How old is my sister?" Output: "35" (Richardson and Heck, 2023)

Tab. 2 Types of errors in the LLM responses

(2023). Providing users with information about how the AI model works and its limitations can help them understand when to trust the system and when to seek additional verification (Marr, 2023). However, apart from solutions to this problem on the part of AI systems, there are several ways in which users can avoid or minimize the occurrence of hallucinations when communicating with LLMs.

4.2.2 Best practices for prompt writing to minimize hallucinations

Providing clear and specific prompts, along with the relevant context, is essential when interacting with an AI system. This clarity helps to guide the system towards the intended output, thereby increasing the accuracy and relevance of the response. Including detailed context in prompts allows the AI to better understand the nuances of the request, minimizing the likelihood of generating irrelevant or incorrect information. Therefore, users should carefully consider the information they provide to ensure it is comprehensive and relevant. By doing so, they can significantly improve the efficiency and effectiveness of the AI's performance. Korzynski et al. (2023) state that effective prompting may include:

- Context – includes information about the role the model is to play in the task, or any necessary information about the situation that may justify it. Example: "You are a human resources manager in a trading company."
- Instruction – the task to be performed. Example: "Write an email to the customer offering new products".
- Input data – data and facts that the model should use to complete the task. Example: keywords to include in the response.

- Expected output format – information about the format and type of output in which the answer is to be provided. Examples: “Generate a CSV file. Generate Python code”.

The use of idiomatic expressions, colloquial slang, or overly technical terminology can also obscure contextual understanding when interacting with AI systems. Users can adjust the temperature parameter, which controls the degree of randomness in the output. A higher temperature setting increases the variety and creativity of the text, allowing outputs that may exceed the expectations set by the input. However, this increase in randomness also increases the likelihood of producing responses that do not strictly follow the input patterns, potentially resulting in outputs that could be perceived as incorrect or misleading in certain situations. Conversely, a lower temperature setting will produce more consistent and predictable outputs that closely match the input patterns. This reduces the likelihood of producing outputs that are incorrect or misleading.

A notable limitation of Large Language Models (LLMs) is their tendency to produce inaccurate results in tasks requiring multi-step reasoning, such as arithmetic or logic problems. However, the accuracy of these models improves significantly when they are given multiple examples (few-shot learning), instructed to break down tasks into sequential steps (chain of thought) and to aggregate their results. This can significantly reduce the hallucinatory effects. Another effective strategy for improving accuracy is the self-consistency method. This technique is based on the premise that complex reasoning problems can often be approached in different ways but still lead to the same correct answer. Introduced by Wang et al. (2022), this method has been shown to significantly improve performance on arithmetic and common sense reasoning tasks across several large language models of different scales. In addition, the self-evaluation technique allows users to distinguish between correct and incorrect answers. By asking a model to generate responses along with probabilities of their correctness, users can use these well-calibrated probabilities to filter out likely incorrect responses. This technique assumes that the model is generally aware of its own knowledge limitations, thus allowing for more reliable filtering of the generated output (Kadavath et al., 2022).

5 CONCLUSION

AI systems have made significant progress, but they are not error-free. It is crucial for users to recognise these limitations in order to accurately assess AI-generated responses. Ultimately, the evaluation of such results must rely on human judgment. The user’s expertise and robust critical thinking skills are particularly important. Our research aimed to identify the most common risks associated with the use of AI systems. We identified the most common types of hallucinatory effects in text responses and provided guidelines for users to reduce errors in the responses they receive from these systems. An open question is whether it is possible to completely eliminate or correct hallucinations in AI.

Acknowledgements
This paper was supported by the project CZ.02.1.01/0.0/0.0/16_017/0002334 Research Infrastructure for Young Scientists, which is co-founded by the Operational Programme Research, Development and Education.

REFERENCES

ATHALURI, S., MANTHENA, S., KESAPRAGADA, V., YARLAGADDA, V., DAVE, T. and DUDDUMPUDI, R. T. S. 2023. Exploring the boundaries of reality: Investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus*, 15(4), e37432. DOI: 10.7759/cureus.37432

BANG, Y., CAHYAWIJAYA, S., LEE, N., DAI, W., SU, D., WILIE, B., LOVENIA, H., JI, Z., YU, T., CHUNG, W., DO, Q. V., XU, Y. and FUNG, P. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. *arXiv preprint arXiv*: 2302.04023. DOI: 10.48550/arXiv.2302.04023

BONTRIDDER, N. and POULLET, Y. 2021. The role of artificial intelligence in disinformation. *Data & Policy*, 3, e32. DOI: 10.1017/dap.2021.20.

BUOLAMWINI, J. 2019. Artificial intelligence has a problem with gender and racial bias: Here's how to solve it. *Time* [Online]. Available at: https://time.com/5520558/artificial-intelligence-racial-gender-bias/ [Accessed 2024, February 19].

CHENG, F., ZOUHAR, V., ARORA, S., SACHAN, M., STROBELT, H. and EL-ASSADY, M. 2023. RELIC: Investigating large language model responses using self-consistency. *arXiv preprint arXiv*: 2311.16842. DOI: 10.48550/arXiv.2311.16842

COSSINS, D. 2018. Discriminating algorithms: 5 times AI showed prejudice. *Newscientist* [Online]. Available at: https://www.newscientist.com/article/2166207-discriminating-algorithms-5-times-ai-showed-prejudice/ [Accessed 2024, February 19].

FUJIMOTO, S. and TAKEMOTO, K. 2023. Revisiting the political biases of ChatGPT. *Frontiers in Artificial Intelligence*, 6, 2023. DOI: 10.3389/frai.2023.1232003

GILLHAM, J. 2023. *AI hallucination factual error problems* [Online]. Available at: https://originality.ai/blog/ai-hallucination-factual-error-problems [Accessed 2024, February 19].

GUPTA, A., HATHWAR, D. and VIJAYAKUMAR, A. 2020. Introduction to AI chatbots. *International Journal of Engineering Research and Technology,* 9, 255–258. DOI: 10.17577/IJERTV9IS070143

HARTMANN, J., SCHWENYOW, J. and WITTE, M. 2023. The political ideology of conversational AI: Converging evidence on ChaGPT's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv*: 2301.01768. DOI: 10.48550/arXiv.2301.01768

KADAVATH, S., CONERLY, T., ASKELL, A., HENIGHAN, T., DRAIN, D. et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:* 2207.05221. DOI: 10.48550/arXiv.2207.05221.

KEARY, T. 2024. *AI Hallutinations* [Online]. Available at: https://www.techopedia.com/definition/ai-hallucination [Accessed 2024, February 19].

KORZYNSKI, P., MAZUREK, G., KRZYPKOWSKA, P. and KURASINSKI, A. 2023. Artificial intelligence prompt engineering as a new digital competence: Analysis of generative AI technologies such as ChatGPT. *Entrepreneurial Business and Economics Review,* 11(3), 25–37. DOI: 10.15678/EBER.2023.110302

LEE, K. 2023. *Understanding LLM Hallucinations and how to mitigate them* [Online]. Available at: https://kili-technology.com/large-language-models-llms/understanding-llm-hallucinations-and-how-to-mitigate-them [Accessed 2024, February 19].

LUTKEVICH, B. 2023. *AI hallucination* [Online]. Available at: https://www.techtarget.com/whatis/definition/AI-hallucination. [Accessed 2024, February 19].

MARR, B. 2023. *ChatGPT: What Are Hallucinations And Why Are They A Problem For AI Systems* [Online]. Available at: https://bernardmarr.com/chatgpt-what-are-hallucinations-and-why-are-they-a-problem-for-ai-systems/ [Accessed 2024, February 19].

MISHRA, A. N. 2023. *Hallucination in Large Language Models* [Online]. Available at: https://medium.com/@asheshnathmishra/hallucination-in-large-language-models-2023-f7b4e77855ae [Accessed 2024, February 19].

REYNOLDS, L. and MCDONELL, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. Paper presented at: *the Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. DOI: arXiv:2102.07350

RIAZ A. 2023. *29 Mind-Blowing Examples of AI Hallucinations* [Online]. Available at: https://vividexamples.com/examples-of-ai-hallucinations/ [Accessed 2024, February 19].

RICHARDSON, C. and HECK, L. 2023. Commonsense Reasoning for Conversational AI: A Survey of the State of the Art. arXiv preprint arXiv: 2302.07926. DOI: 10.48550/arXiv.2302.07926

ROHRBACH, A, HENDRICKS, L. A., BURNS, K., DARRELL, T. and SAENKO, K. 2018. Object Hallucination in Image Captioning. *arXiv preprint arXiv*: 1809.02156. DOI: 10.48550/arXiv.1809.02156

SHEN, Y., HEACOCK, L., ELIAS, J., HENTEL, K. D., REIG, B., SHIH, G. and MOY, L. 2023. ChatGPT and Other Large Language Models Are Double-edged Swords. *Radiology*, 307(2). DOI: 10.1148/radiol.230163

SONG, W., LIYAN, T., AKASH, M., JUSTIN, F. R., GEORGE, S., YING, D. and YIFAN, P. 2022. Trustworthy assertion classification through prompting. *Journal of Biomedical Informatics*, 132, 104139. ISSN 1532-0464. DOI: 10.1016/j.jbi.2022.104139

SRINIVASAN, R. and CHANDER, A. 2021. Biases in AI Systems. *Communications of the ACM*, 64(8), 44–49. DOI: 10.1145/3464903

SUSHIR T. 2024. What is AI Hallucination, and Can it Be Fixed? *Geekflare* [Online]. Available at: https://geekflare.com/ai-hallucination/ [Accessed 2024, February 19].

WANG, X., WEI, J., SCHUURMANS, D., LE, Q., CHI, E., NARANG, S., CHOWDHERY, A. and ZHOU, D. 2022. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv preprint arXiv*: 2203.11171. DOI: 10.48550/arXiv.2203.11171

WEI, J., WANG, X., SCHUURMANS, D., BOSMA, M., ICHTER, B., XIA, F., CHI, E. H., LE, Q. V. and ZHOU, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint arXiv*: 2201.11903v6. DOI: 10.48550/arXiv.2201.11903

XIAO, Y. and WANG, W. Y. 2021. On Hallucination and Predictive Uncertainty in Conditional Language Generation. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*. Main Volume, 2734–2744. Association for Computational Linguistics.

Contact information

Jana Dannhoferová: e-mail: jana.dannhoferova@mendelu.cz
Petr Jedlička: e-mail: petr.jedlicka@mendelu.cz