Grounding Generative AI Models – Introduction

Holger Keibel and **Johannes Porzelt**
Karakun AG, Basel
holger.keibel@karakun.com, johannes.porzelt@karakun.com

Abstract

One fundamental shortcoming of generative AI systems such as ChatGPT is that they tend to produce hallucinatory outputs which the human user might easily take to be facts, with potentially disastrous consequences. In this paper, we briefly sketch this general problem and possible approaches to mitigating such AI hallucinations. We then summarize a conference workshop in which participants exchanged practical experience in applying such approaches.

1 Hallucinatory outputs

When Large Language Models (LLMs) are used to generate natural language answers, they tend to produce hallucinatory outputs – i.e., answers, that are incorrect or irrelevant to the question. For instance, Fraser (2024) published a hallucination generated by GPT-4 in December 2023. Asked about the "name of the first elephant to swim across the English Channel", the LLM replied that this elephant was named "Kami". Its answer went on to describe some additional information about the event.

Interestingly, as Fraser points out, while there is no such elephant and the described event did not happen, some facts mentioned in the answer are correct. This is very typical of AI hallucinations: LLMs tend to present the false information together with some correct information which makes the story more believable and therefore more dangerous.

One could argue that in this example, the question itself does not make much sense. So, the human user should not expect a meaningful answer anyway – or they should themselves infer that the answer might in part be hallucinated. Hallucinatory generative AI answers are surely more problematic and posing greater risks when they are given in response to more sensical questions.

And such hallucinatory outputs are not rare. Despite the impressive skills of LLMs, it is very likely for any user to run into hallucinations fairly quickly. You spot them easily if you ask an LLM something about your own fields of expertise – but outside these fields, you might just as easily fall for these hallucinations.

2 Dangerous hallucinations

Hallucinatory generative AI outputs do not only compromise the value produced by LLMs – even worse, they can also be dangerous if human users believe them. These dangers involve personal risks and public risks.

Personal risks can arise any time a user makes a real-life decision that is driven by some generative AI answer, without any doubt or fact-checking. Obvious examples are when the LLM gives the user bad advice on medical, financial or construction issues.

Public risks, in turn, arise when many people get caught up in such personal risks. Even more noteworthy and present is the public risk that LLMs help generate and spread misinformation, for instance, with respect to democratic elections.

In principle, it is the responsibility of the user to fact-check any generative AI output before relying on it, but this is not common practice and often impossible for users when the respective topic lies outside their field of expertise. It is, therefore, crucial to prevent hallucinations from occurring in the first place and to reduce their risks as far as possible.

3 Why hallucinations?

Why do LLMs produce hallucinatory outputs? Why do AI chatbots make things up?

Their initial pre-training involves massive data across a wide range of domains and topics – but the training tasks are language tasks by nature. The resulting models nonetheless already do capture an amazing amount of general world knowledge and specific facts – but this happens more as a side-effect of the actual learning task.

Subsequent steps such as fine-tuning and alignment (plus in part also integrating safety guardrails) make the LLM's responses more relevant and boost their correctness. But it appears that these measures can never reach far enough: No matter how much we pre-train, fine-tune and align an LLM, there is no way it will ever get to see and learn all the facts that are relevant for a given target domain and target task.

The intrinsic problem is that LLMs per se do not have access to explicit facts at run-time. They have no "awareness" of what they really do know, and they have no explicit memory of what they have seen during training and fine-tuning. They are mainly trained on processing and generating human-like text – particularly in the form of interpreting questions and generating answers. They do not necessarily *know* the answer, but they know how to formulate one, and that is what they do. Therefore, LLMs by themselves are prone to hallucinate occasionally.

LLMs are very powerful tools for understanding and generating language and in our view, they create the most value when used primarily for tasks to do just that: understand and/or generate language – but without expecting them to provide the relevant domain knowledge, too.

4 Ways to ground generative AI outputs

To prevent hallucinations from happening this often and from posing such great risks, generative AI outputs should be grounded in relevant facts. At the very least, any generative AI output should be presented with a value roughly quantifying the confidence that this output is correct.

Existing approaches range from integrating domain-specific knowledge directly into the models (typically by means of *retrieval-augmented generation, RAG*) to applying post-generation filtering techniques (*automated fact-checking*) to making the generation of the AI output transparent to the user (*explainable AI*), e.g., by providing reasons or sources.

Especially for RAG, a lot of progress has been made over the past 12 months and it is getting very popular – to the extent that it looks like becoming a standard approach in the field.

5 Workshop

The best way to learn about these approaches to grounding generative AI systems is by practical examples. Therefore, we organized a workshop at SwissText 2024 that aimed at bringing together professionals from both academia and industry that could share their experience in applying such approaches in real-life projects.

In this workshop, three speakers presented papers applying LLMs generatively in a range of domains and use cases. Hallucinations pose substantial challenges in all three cases, and the papers describe different approaches to mitigating them.

In the first paper, Gishamer and Arwadi (2024) apply LLMs in the context of ticket routing in customer support, and they mitigate hallucinations by means of RAG against a fixed set of outputs, in conjunction with supervised learning approaches.

Secondly, Zhang (2024) uses LLMs for building teaching assistants. Here, hallucinations are minimized also via RAG, but in this case using multimodal knowledge graphs (KGs) as source data. These KGs in turn were in part created with the help of LLMs which proposed additions to the KGs which were then manually validated.

Finally, Schneider and Spitale (2024) apply LLMs within the ethically sensitive domain of euthanasia decisions. They address hallucinations with an *explainable AI* approach, by querying the LLM itself in a series of yes/no questions.

For further details, we encourage you to read the three papers.

In a closing panel discussion, various questions were debated. For example: In a RAG approach, how can further domain knowledge be integrated into the system and how can certain types of errors be prevented? The answers revolved around prompt engineering, fine-tuning and syntactic guardrails.

Another question concerned the possible danger of circular information flow in a setup where LLMs use KGs which in turn were informed by LLMs (see second paper, Zhang 2024). Zhang confirmed that this is a true challenge which he aims to control by applying the LLM at different levels of information in both steps.

For RAG in general, the practical experience of participants was that much of the quality hinges on

the retrieval step. Therefore, the discussion revolved around ways to filter out irrelevant retrieval results so they would not be sent to the LLM.

Participants also had different opinions on the question of how well the term "hallucinations" is chosen with respect to generative AI outputs.

6 Conclusion

The three workshop papers illustrated for academic and industry projects how LLMs can be used for generative tasks while minimizing or uncovering the hallucinations they might make.

Two papers demonstrated that *RAG* combines the strengths of retrieval and generation and that it has already started to create significant value in real-life applications. Nevertheless, many practical challenges remain and probably have to be solved for each application independently.

The *explainable AI* approach described in the second paper uses the LLM not only for generating answers but also for evaluating itself, by means of yes/no questions. This is a promising path of research where the greatest challenge seems to be the volatile nature of LLM answers.

References

Colin Fraser. 2024. Hallucinations, errors, and dreams. Available online at: https://medium.com/@colin.fraser/hallucinations-errors-and-dreams-c281a66f3c35 (accessed 30 May 2024), Medium.

Flurin Gishamer and Alexander Khalil Arwadi. 2024. Practical Strategies for Enhancing Reliability of GenAI Systems in Customer Operations: An Overview. In *Proceedings of the 9th SwissText Conference*. https://www.swisstext.org/

Gerold Schneider and Giovanni Spitale. 2024. Evaluating Transformers on the Ethical Question of Euthanasia. In *Proceedings of the 9th SwissText Conference*. https://www.swisstext.org/

Xiaokun Zhang. 2024. Knowledge Graphs Enhanced Retrieval-Augmented Generation for Eliciting Higher-Order Thinking. Paper presented at *the 9th SwissText Conference*. https://www.swisstext.org/