UNIVERSITÉ Clermont Auvergne

Scientific English
Master in: Computer Science
Track: International of Computer Science
Academic year: 2023-2024

Hallucination Reduction in Generative AI Models: Approaches and Insights

Supervised by

Mrs. HENRY

Presented by

Loubna EL ATTAR & Mohammed SGHIOURI

Contents

List of Figures

List of Tables

Abstract

The global debut of ChatGPT has undeniably marked a significant milestone in the context of Generative Artificial Intelligence. This renewed interest has led to the development and deployment of cutting-edge tools like Bard, Stable Diffusion, DALL-E, Make-A-Video, Runway ML, and Jukebox. These tools show remarkable capabilities across diverse domains such as text generation, music composition, image and video creation, code generation, and scientific endeavors. However, amidst these advancements, the proliferation of Generative AI models introduces a critical challenge: the problem of hallucination. This review aims to provide a comprehensive exploration of the specific challenge of hallucinations. We will delve into different methods employed to mitigate this issue, offering insights into the current landscape and promising trajectories for the future of Generative Artificial Intelligence.

1 Introduction

In the ever-evolving landscape of Artificial Intelligence (AI), the advent of Generative AI models has sparked both excitement and apprehension. These models, exemplified by the global debut of ChatGPT, have demonstrated unprecedented capabilities across various domains, from text and music generation to image creation and code synthesis. However, with these advancements comes a formidable challenge—the issue of hallucinations, wherein these models generate content that deviates from factual reality.

To embark on a comprehensive exploration of this multifaceted landscape, this paper unfolds in two critical phases. First, we will establish the contextual background by defining Generative AI and presenting the hallucination problem. Following this, our focus shifts to an in-depth analysis of different articles and research studies dedicated to mitigating the hallucination problem in Generative AI. We will analyze various approaches, methodologies, and innovations presented in the literature, aiming to distill insights that contribute to a nuanced understanding of this challenge.

Additionally, we will engage in a comparative examination of these diverse methods, discerning commonalities, disparities, and their respective efficacies. By defining the current state-of-the-art approaches, we strive to offer a comprehensive overview of the landscape surrounding hallucination reduction in Generative AI models. Through this exploration, we aim to contribute valuable perspectives to both the scholarly and practitioner communities, fostering a deeper understanding of the challenges and possibilities that lie at the intersection of Generative AI and hallucination mitigation.

2 Contextual Background

2.1 Generative AI

Generative Artificial Intelligence (GAI) represents a significant advancement in the field of ML. It focuses on the development of algorithms and models that can autonomously create new data, such as images, audio, and text, as opposed to traditional AI, which primarily works with existing data. These advanced algorithms have demonstrated exceptional proficiency in generating content that closely resembles real-world data.

Generative Artificial Intelligence has unlocked a multitude of applications, transforming the way we interact with and generate content. From text and images to videos, 3D models, code, and even speech, the potential of Generative AI is vast.

	Model level	**System level**	**Application level**
Text Generation	X-to-text models, e.g. GPT-4 and LLaMA2	Conversational agents and search engines (ChatGPT and YouChat)	Content generation, translation and text summarization
Image/Video Generation	X-to-image models. e.g. DALL-E	Image/ Video Generation systems and bots (Midjourney)	Synthetic product and advertising visuals, educational content
Speech/Music Generation	X-to-music/speech models, e.g. MusicLM and VALL-E	Speech generation systems, e.g. ElevenLabs	AI music generation, Text-to-speech generation (news, product tutorials..)
Code Generation	X-to-code models, e.g. Codex and AlphaCode	Programming code generation systems, e.g. GitHub Copilot	Software development, code synthesis, review and documentation

Table 1: A model-based, system-based, and application-level view on GAI

However when we zero in on text, the impact is particularly profound. Text models, particularly those tailored for conversational chatbots, have led to a significant revolution in the AI landscape. The introduction of ChatGPT marked a pivotal moment in this evolution, harnessing the capabilities of Natural Language Processing and LLMs to offer an array of invaluable functions, including text summarization, writing assistance, code generation, language translation, and sentiment analysis. ChatGPT has become a central figure in Generative AI, benefiting millions of users and underscoring its profound impact.

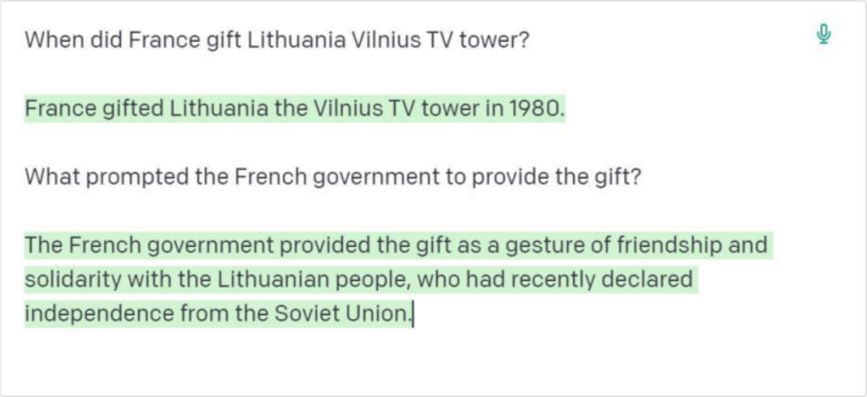
Conversational AI has garnered substantial attention in the field of artificial intelligence. These AI services, functioning as versatile chatbots, adeptly transform text prompts into corresponding text outputs. They owe their capabilities to LLMs, such as GPT-3, PaLM, Falcon, and LLaMA, characterized by their extensive training on vast text datasets. However, amidst the remarkable achievements of Generative AI in transforming the way we interact with these models, a pressing challenge emerges: hallucination.

2.2 AI Hallucination

Hallucination in the context of generative models refers to the generation of content that may not be grounded in factual reality. In the context of text generation, this can manifest as the generation of information, responses, or details that are not accurate or contextually relevant.

For instance, consider a chatbot designed for medical assistance. A hallucinating model might generate plausible-sounding medical advice or information that, in reality, is inaccurate or potentially harmful. This poses serious ethical concerns and challenges the reliability of AI-generated content.

The issue of hallucination is not confined to the medical domain alone. In various applications, from news generation to customer support, the potential consequences of generating misleading or inaccurate information are significant. Therefore, addressing the hallucination problem in Generative AI models, especially in the context of text generation, becomes imperative for ensuring the responsible and effective deployment of these technologies.



When did France gift Lithuania Vilnius TV tower?

France gifted Lithuania the Vilnius TV tower in 1980.

What prompted the French government to provide the gift?

The French government provided the gift as a gesture of friendship and solidarity with the Lithuanian people, who had recently declared independence from the Soviet Union.

Figure 1: Hallucination example [1]

3 Analysis of Articles

3.1 Factuality Enhanced Language Models for Open-Ended Text Generation [2]

This paper focuses on evaluating and enhancing the factual accuracy of large-scale pretrained language models (LMs) used for open-ended text generation. The authors introduce a new test set, FACTUALITYPROMPTS, and corresponding metrics to assess the factuality of LM-generated content. The study covers LMs with parameter sizes ranging from 126M to 530B, discovering an interesting trend where larger LMs tend to be more factual than smaller ones, contrary to previous suggestions. The research also identifies that popular sampling algorithms, such as top-p, may compromise factuality due to introduced ”uniform randomness” during text generation. To address this, the authors propose the factual-nucleus sampling algorithm, dynamically adapting randomness to enhance factuality while maintaining quality. Additionally, the paper examines inefficiencies in standard training methods related to learning correct associations between entities from factual text corpora. In response, the authors introduce a factuality-enhanced training method using TOPICPREFIX for improved awareness of facts and sentence completion as a training objective, leading to a significant reduction in factual errors.

Parameter adaptation plays a pivotal role in fine-tuning large language models (LLMs) to align their generated content with user intent while mitigating biases acquired during pre-training. This process involves various innovative strategies such as contrastive learning optimization where the generation probability of negative samples is reduced at the span level. By introducing contextual knowledge background that contradicts the model's intrinsic biases, the influence of prior knowledge is effectively diminished. Parameter adaptation also explores flexible sampling to align with user requirements, as observed by Lee et al. (2022) [2], who introduce the factual-nucleus sampling algorithm to balance faithfulness, quality, and diversity in generation.

3.2 SELF-REFINE: Iterative Refinement with Self-Feedback [3]

In an effort to enhance language models for hallucination mitigation, a promising avenue involves incorporating relevant information from large textual databases. For instance, one approach divides the input sequence into chunks and retrieves analogous documents to address hallucination concerns. Recognizing the limitations of scaling in improving memory for factual knowledge in the long tail, it selectively retrieves non-parametric memories as needed for enhanced performance. This method leverages integrated external knowledge and automated feedback to refine the truthfulness score of generated answers and it iteratively analyzes future content to retrieve pertinent documents for sentence re-generation, particularly when tokens exhibit low confidence.

3.3 Learning to summarize from human feedback [4]

As language models advance in sophistication, leveraging evaluation feedback becomes a pivotal strategy to enhance the quality of generated text and minimize hallucinations. To operationalize this concept, the solution consists on predicting human-preferred summarizations and employs them as rewards for fine-tuning summarization strategies using reinforcement learning. On the other hand, to enhance system reliability, we can strategically select questions to refuse to answer, to improve system reliability through reinforcement learning from human preferences. Beyond direct learning from feedback, incorporating self-evaluation functions becomes crucial in filtering candidate generated texts.

3.4 Comparative Analysis

In comparing the three papers, several key themes emerge in their approaches to improve language model performance. The common thread across all three is the recognition of the challenges related to factuality, hallucination mitigation, and the incorporation of user and evaluation feedback. However, they diverge in their specific methodologies and emphasis on different aspects of language model improvement.

Parameter adaptation is a central theme in the first 2 papers. Lee et al. (2022)introduce innovative strategies such as contrastive learning optimization and flexible sampling to align language models with user requirements and improve factuality [2]. On the other hand, the second paper explores parameter adaptation for large language models, employing methods like contextual knowledge background introduction and factual-nucleus sampling algorithm to balance faithfulness, quality, and diversity in generation. [3] Leveraging External Knowledge is a crucial aspect addressed in the second and third papers. The second paper emphasizes incorporating relevant information from large textual databases to enhance language models for hallucination mitigation. It employs strategies like selective retrieval of non-parametric memories and integrating external knowledge with automated feedback to refine truthfulness scores [3]. Similarly, the third paper focuses on leveraging external knowledge and assessment feedback for enhancing language model performance. It explores strategies such as predicting human-preferred summarizations and strategically selecting questions to improve system reliability through reinforcement learning from human preferences [4].

4 Discussion

In the analysis of the papers—*Factuality Enhanced Language Models for Open-Ended Text Generation*, *Parameter Adaptation for Large Language Models*, and *SELF-REFINE: Iterative Refinement with Self-Feedback*—we observe that the methods proposed are not entirely orthogonal but could complement each other based on the requirements of specific tasks in practical applications.

Parameter adaptation strategies, as discussed in the first two papers, aim to fine-tune language models to align with user intent and mitigate biases acquired during pre-training. These techniques focus on optimizing the generation process and enhancing the factuality of the generated content. On the other hand, leveraging external knowledge, as emphasized in the second and third papers, brings in additional context and information from large textual databases to address hallucination concerns and improve the truthfulness of generated answers.

In practical scenarios, a hybrid approach that combines both parameter adaptation and external knowledge incorporation may offer a more robust solution. For instance, a language model could benefit from fine-tuning its parameters to align with user preferences while simultaneously leveraging external knowledge to ensure a broader understanding of the context.

While the papers provide valuable insights into specific aspects of language model enhancement, it is essential to consider the potential synergies between these approaches. The challenge lies in finding a balance that optimally utilizes both strategies to achieve improved language model performance across various tasks. Ultimately, the integration of parameter adaptation and external knowledge incorporation could contribute to more versatile and reliable language models in real-world applications.

5 Future Directions

The exploration of factuality enhancement, parameter adaptation, and leveraging external knowledge in language models provides a foundation for future research directions. As we delve into the potential avenues for improvement, several ideas emerge:

- **Hybrid Approaches:** Investigate the development of hybrid approaches that seamlessly integrate factuality enhancement strategies with parameter adaptation and external knowledge incorporation. Combining these methods could lead to more robust language models capable of addressing diverse challenges.

- **Task-Specific Adaptation:** Explore adaptation strategies tailored to specific domains or tasks, such as medical diagnosis or legal document summarization. Dynamic adjustment of model parameters based on task characteristics could optimize performance for specialized applications.

- **Explainability and Transparency:** Integrate explainability and transparency in language models. Develop models that not only generate high-quality content but also provide insights into their decision-making processes through interpretation and visualization techniques.

- **Continual Learning:** Explore the concept of continual learning, where language models adapt and learn continuously over time. Investigate models that accumulate knowledge from new data streams, adapt to evolving language nuances, and dynamically refine their parameters.
- **Ethical Considerations:** Address ethical considerations surrounding language models. Explore methods to mitigate biases, ensure fairness in model outputs, and incorporate ethical guidelines into the training process. Align language models with ethical and societal values.

These proposed directions represent only a fraction of the potential avenues for advancing language models. As the field continues to evolve, researchers have the opportunity to contribute to the development of more sophisticated, ethical, and versatile language models that can positively impact various domains.

6 Conclusion

In this review, we explored the landscape of language models with a focus on factuality enhancement, parameter adaptation, and leveraging external knowledge. The analyzed papers—*Factuality Enhanced Language Models for Open-Ended Text Generation*, *Parameter Adaptation for Large Language Models*, and *SELF-REFINE: Iterative Refinement with Self-Feedback*—provided valuable insights into the challenges and potential solutions in advancing language models.

The investigation into factuality enhancement strategies highlighted the importance of addressing factual accuracy in open-ended text generation. The introduction of novel test sets and metrics, as seen in the first paper, demonstrated a dedicated effort to measure and improve factuality, with implications for a wide range of applications.

Parameter adaptation emerged as a pivotal aspect in fine-tuning large language models, aligning them with user intent, and mitigating biases. The contrastive learning optimization and flexible sampling strategies discussed in the second paper showcased innovative approaches to enhance the generation process.

The third paper delved into the incorporation of external knowledge, emphasizing the relevance of integrating information from large textual databases to address hallucination concerns. The iterative refinement process with self-feedback offered a promising avenue for refining the truthfulness of generated answers.

As we look to the future, potential research directions include hybrid approaches that combine these methods seamlessly, task-specific adaptation for specialized domains, and the integration of explainability and transparency in language models. Continued exploration of continual learning and ethical considerations will also shape the evolution of language models.

In conclusion, the field of language models is dynamic and evolving, with ongoing efforts to improve their capabilities and address challenges. The reviewed papers contribute valuable perspectives and methodologies, paving the way for future advancements in language model research. As researchers explore the proposed future directions, we anticipate a more nuanced, adaptable, and ethically grounded generation of language models that can cater to diverse applications and societal needs.

References

[1] Chatgpt's bard ai answers: Hallucination, 2023. Accessed: December 8, 2023.

[2] Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. Factuality enhanced language models for open-ended text generation, 2023.

[3] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023.

[4] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, 2022.