Comprehensive to the Textual Hallucination in Generative AI

Yiyang Li

College of Computer Science, Sichuan University, Chengdu, Sichuan, China
2023141460145@stu.scu.edu.cn

Abstract. Generative AI has been particularly strong in many places in recent years, especially large language models that have done very well in writing articles, answering questions, and helping to learn these things. However, these models sometimes make mistakes, such as making up factual content, or giving answers that have no evidence to support or even logical confusion, which do bring a lot of trouble in actual use. This paper has carefully sorted out the current research on the illusion of generative models, and proposed a new classification method, which is based on several aspects. This paper also analyzes why these hallucinations occur, the main reason may be related to the training data, or it may be the problem of the model training, or the reasoning process is wrong, and even when people and machines interact with it. To improve this situation, there are some methods that are being tried, such as making more detailed adjustments to the model, finding ways to make the model more knowledgeable, improving the reasoning process, and combining search techniques to help generate better content.

Keywords: Textual Hallucination, Factual Consistency, Mitigation Strategies.

1 Introduction

In recent years, generative AI, commonly known as GenAI, has developed particularly rapidly, especially in technical systems based on large language models, which can be seen in many places. Whether it's writing articles, answering questions, or helping to learn, consult, or even deal with legal issues, these AIs are particularly powerful. They can write particularly natural sentences and communicate with people smoothly, which makes it easier for us to process information [1].

Some recent studies have found that hallucinations don't just occur in common tasks such as summaries, question-and-answers, and machine translation, but in different application scenarios [2]. If they are in less important situations, hallucinations may only bias the information or make users feel less useful; but in particularly critical areas, it will cause the problem to be much more serious. There are also scientific studies in which AI models sometimes invent non-existent references or give incorrect citations, which can affect the credibility of the research and may mislead other researchers into their follow-up work [3].

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_37

More critically, hallucinations from generative models often exhibit high linguistic fluency and plausible reasoning, making it difficult for ordinary users to discern false information [4]. In multi-turn dialogue scenarios, minor hallucinations from earlier exchanges can accumulate and amplify through subsequent interactions, forming chains of error [5]. This not only exacerbates the harm caused by hallucinations but also poses new challenges for human-AI collaboration.

Therefore, systematic research into the illusion of generating text is a central issue in advancing trusted and controlled AI development and a prerequisite for ensuring its secure deployment in sensitive areas such as scientific research, healthcare, and law. This article aims to organize and summarize existing research across multiple dimensions, including classification systems, causal mechanisms, assessment methods, and optimization strategies, while exploring future research directions.

2 Definition

The main discussion of the text illusions in generative models is the problem of text illusions that, in simple terms, natural language generation systems such as large language models sometimes output less reliable content. Although these are well-read and logically speaking, they may not actually match the facts, or they may not be related to the information entered. People found that the text generated by these models is doing well in terms of language expression, but often there are various problems when it comes to specific facts, such as not being accurate or inconsistent.

The first is the contradiction between the generated content and the original material. The second is the external illusion, and the second is the external illusion, although the generated content does not directly conflict with the input material, but contains some information that cannot be confirmed from the original material, which may be fictional. Some studies have mentioned this classification [6].

In simple terms, language models always predict the next word by probability when generating text. When lacking sufficient knowledge, they tend to "reasonably complete" rather than remain silent [7]. That's why hallucinations are, that's what sounds reasonable but actually inaccurate. This situation causes a lot of trouble for the practical application of AI, so people have to figure out what the illusion is before can find out how to discover it, evaluate it, and finally solve it.

3 Classification of Generative Text Hallucination Phenomena

This paper focuses on generative text illusions, which people divide into several aspects, which can help to better understand its performance and the problems that may be caused. First of all, from the content point of view, some errors appear directly in the text. For example, the illusion of fact is to make up something that does not exist at all, such as the time of a fictional character or the time of the wrong historical event. Then there is the logical illusion, which is that the reasoning is problematic, or the reasoning process may be incomplete, sometimes the conceptual confusion.

In terms of task performance, the hallucination problem is that the answer given by the model when doing a specific task is not quite right. Sometimes the model does not follow the user's requirements, and the generated content has nothing to do with the task. This is the command deviation illusion, the specific manifestation is to deliberately avoid the task requirements, say the problem, or to pull too far.

The problem of hallucinations in everyday life is more common, mainly contrary to common sense as everyone knows or to get the basic facts wrong. In some professional fields, the problem of hallucinations is much more serious, especially in medical, legal, and financial industries. The hallucinations in these professional fields will violate industry regulations, may also use professional terms, and more frighteningly, some wrong guidance, such as doctors giving wrong diagnoses or lawyers providing unreliable legal opinions, which can cause great trouble [8].

There are some deficiencies in the training data itself that lead to hallucinations, such as the lack of some knowledge in the data, the lack of good data quality, or the lack of timely information updates. The structure and reasoning methods of the other model also bring hallucination problems, such as the fact that attention mechanisms can sometimes go wrong, and the decoding strategy is flawed, such as greedy decoding or sampling methods.

This paper mainly studies from several different aspects, including the characteristics of the content itself, the performance of actual use, the specific application and the generative principle. Through such a multi-faceted analysis, people can establish a relatively complete classification system. This classification framework can not only help us better understand and identify the illusion problems generated by AI, but also can be used to evaluate and solve these problems, which is also a good foundation and analytical tool for future research.

4 Mechanistic Analysis of Generative Text Hallucination Phenomena

There are many reasons for the hallucinatory problem of generative text, which can look at in four ways. First, the impact of data, then the problems that may occur during the training of the model, and some of the characteristics of the reasoning mechanism itself, and finally the impact of human and machine interaction.

When an AI encounters a knowledge blind zone, it does not choose to be silent, but rather invents seemingly reasonable content, which is called an illusion. Another problem is that the training data itself may not be clean enough, there are all kinds of noise and deviations, and the AI will amplify these problems after learning this data. Additionally, as model knowledge is frozen at the training point in time, temporal lag effects lead to the generation of outdated or inaccurate information [9].

The paper goes on to discuss the problem of bias in the training process, which is also a major cause of AI hallucinations. When training models, people usually use the most likely method to estimate this method, and its main role is to make the generated text look more natural, but this may ignore the accuracy of the facts, which is contradictory. In the reinforcement learning stage, if the human feedback data used for

training is itself problematic, such as containing subjective views or wrong information, and even some feedback may prefer those that are written in detail but are not very accurate. In areas that require particularly rigor, such as health care or law, models can easily go wrong in these places, producing hallucinatory content that does not correspond to reality [10].

The third problem is that there are problems with the reasoning process that also hallucinates AI. The decoding methods used today have some drawbacks, such as greedy decoding, although simple and direct, but easy to get stuck in the same place, and the generated content becomes repetitive and monotonous. Although bundled search can be more efficient, sometimes leaks those low-probability but may sometimes be the correct answers.You may miss some of the less common but correct answers. Another problem is that the current model basically has no way to check in real time when it comes to generating content, and there is a lack of automatic error correction [11].

Blurred, unclear, or misleading prompts can easily trigger an unexpected or fictional content of the model, making the quality of the cub design critical to reducing hallucinations. Overestimating the model's ability also prompts it to produce "confident nonsense" beyond its knowledge boundaries. In multi-round interactions, minor hallucinations or understanding deviations from earlier rounds accumulate as contextual input for future generations, amplifying errors [12, 13].

From these four aspects of analysis, people can see that the illusion of generative text is not a separate problem, but is caused by the combination of factors such as data, training process, reasoning method and interaction. These factors interact to form a complex system. This discovery is helpful for us to study how to detect and optimize text generation later, and it provides some theoretical basis.

5 Comprehensive Evaluation Framework for Generative Text Hallucinations

This paper focuses on how to better evaluate the hallucinatory problems that may occur in generative text, which makes great sense for ensuring the reliability and security of the model output. This article first describes the accuracy of checking the accuracy by splitting the text into the smallest de facto unit, by breaking the generated content into small facts that can be verified separately, and then comparing them to the knowledge graph or the authoritative information in professional databases. The measurement standard mainly depends on how the accuracy of the facts is, whether the fictitious characters or events can be found, and whether the timeline is consistent.These situations are more evident in particularly fast areas of expertise.

The next logical consistency assessment is mainly to find out the reasoning problems that may exist in the article. People use the language model that has been trained, plus some rule templates and algorithms that specialize in check for logical contradictions, so that can find problems such as causal inversion, inconversion, or classification errors. In the evaluation, people mainly look at several aspects, such as how many points the logical consistency is scored, whether the reasoning process is effective, and

whether the whole argument is complete enough, which can help determine whether the article is logical and structurally reasonable [14].

Third, the instruction compliance assessment focuses on the alignment of the generated content with user input or task requirements. Common methods include semantic similarity calculations and key element extraction to quantify whether the output satisfies the core task, exhibiting semantic bias or containing irrelevant content. Key metrics include task completion rate, semantic bias, and irrelevant content ratio, which is especially important in multiple rounds of interaction scenarios.

In professional areas and in some high-risk situations, it is particularly important to do a good job of security assessment. People can establish a risk content classification system, which can identify content that may be harmful and divide them into different categories, such as some content may not be in line with the facts, some may violate regulations, and some may involve ethical issues. By detecting these indicators, these indicators can be used to provide security for medical, legal and financial industries to use these models.

In general, this evaluation framework is mainly considered from four aspects, namely, whether the content conforms to the facts, whether the reasoning process is inconsistent, whether the content is generated according to the requirements, and whether the results are safe. Through these four dimensions, people can more systematically check the possible illusions in the model output, and also find some theoretical basis for improving the reliability and security of the model output [15].

6 Research on Optimization Methods for Generative Text Hallucination Issues

The study of optimizing the generation of text illusions is designed to improve the factuality and reliability of the model output. This can be approached by three types of methods: knowledge enhancement for fine-tuning orientation, optimization of inference processes, and enhanced generation of retrieval.

The main research in this paper is to improve the ability of the model itself from the most basic place, so that the output of the model itself is more accurate and reliable. Specifically, the researchers will use some specially filtered human preference data, through which the model will learn to understand human feedback, so that the model can produce more logical and factual answers. In addition, the method of confrontation training is deliberately to give the model some examples of errors, and help the model to learn to find and correct its own mistakes.It consumes a lot of computer resources, as mentioned in the paper [16, 17].

This paper focuses on the real-time control problem of inference optimization in the generation process, mainly to prevent particularly serious error output. Specifically, people can optimize the generation effect by adjusting the decoding strategy, such as lowering the temperature parameter, so that the randomness of the output will become smaller. Probability filtering can also be used to filter out the less reasonable candidate words. For those models that are not so sure, the "may be inaccurate" can be added to

let the user know that this information may be risky.At points, their role is limited, after all, the ability of the model is determined by what it learns.

The final point of this paper is the Retrieval-Augmented Generation (RAG) technology, which is the ability to use an external authoritative knowledge base to help generate content, so that what is generated is more credible and can be found. Specifically, when the user questions a question, the system will automatically search for a variety of structured and unstructured knowledge bases, find the most useful pieces of information, and then feed the information to the model, so that the model can generate an answer based on this evidence.The quality is good. Another problem is that sometimes the retrieved content and the generated content may not work well together, and even contradictions may occur, as is mentioned in the literature [18].

In general, the three methods of fine-tuning-oriented knowledge enhancement, inference process optimization, and retrieval-enhancing generation have their own characteristics. The first method is mainly to find ways to improve the ability of the model itself and solve the problem from the root cause, while the latter two methods pay more attention to how to improve the reliability of the output by adjusting or introducing external knowledge in real time, which is equivalent to dealing with surface phenomena.

7 Conclusion

Generative text illusions are one of the major challenges facing contemporary generative AI systems, especially in large language models. The popularity and diversity of hallucinatory phenomena has attracted great attention from academia and industry. This paper provides a comprehensive review of generative text illusions across multiple dimensions, including definitions, classifications, causal mechanisms, assessment methods, and optimization strategies.

The main discussion in this paper is the problem of generative text illusions, in short, that the content generated by AI sometimes deviates from the facts, or lacks sufficient evidence to support it, and even logically wrongs. People can look at the problem from several different perspectives, such as the characteristics of the content itself, the performance of the AI to complete the task, see their application in different scenarios, and see how they are generated.

In fact, there are many reasons for hallucinations, such as problems with the data itself, deviations in the model training, imperfect reasoning, and the impact of human and machine interactions. These generative models sometimes output less real content without careful inspection, especially when they encounter knowledge blind spots, or training goals and actual requirements, or when the decoding strategy is not good enough.

In general, the problem of generative text illusion is really important in AI systems, and we can't ignore it. If can study this problem in theory and improve the technology at the same time, it's possible to solve this problem. In this way, generative AI can be used by more people and safely applied to more demanding complex areas.

References

1. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Fung, P.: Survey of Hallucination in Natural Language Generation. ACM Comput. Surv. 55(12), 1–38 (2023)
2. Bang, Y., Cahyawijaya, S., Lee, N., Ji, Z., Su, D., Fung, P.: A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. arXiv preprint arXiv:2309.05922 (2023)
3. Walters, W.H.: Fabrication and Errors in the Bibliographic Citations Produced by ChatGPT-3.5 and ChatGPT-4. Sci. Rep. 13, 13399 (2023)
4. Maynez, J., Narayan, S., Bohnet, B., McDonald, R.: On Faithfulness and Factuality in Abstractive Summarization. In: Proc. 58th Annual Meeting of the Association for Computational Linguistics, pp. 1906–1919 (2020)
5. Rathkopf, C.: Hallucination, Reliability, and the Role of Generative AI in Science. Philos. Sci. 91(5), 789–812 (2024)
6. Thomson, C., Reiter, E.: A Gold Standard Methodology for Evaluating Accuracy in Data-to-Text Systems. In: Proc. 13th International Conference on Natural Language Generation, pp. 158–168 (2020)
7. Goyal, T., Durrett, G.: Annotating and Modeling Fine-Grained Factuality Errors in Summarization. In: Proc. 2021 Conference of the North American Chapter of the Association for Computational Linguistics, pp. 1449–1462 (2021)
8. Kryściński, W., McCann, B., Xiong, C., Socher, R.: Evaluating the Factual Consistency of Abstractive Text Summarization. In: Proc. EMNLP 2019, pp. 933–939 (2019)
9. Zhang, T., Roller, S., Goyal, N., Artetxe, M., Bach, S.H., et al.: Instruction Tuning with GPT-4: Analyzing Data Quality and Model Generalization. arXiv preprint arXiv:2305.01045 (2023)
10. Bai, Y., Ziegler, D., et al.: Constitutional AI: Aligning Language Models with Human Intentions. arXiv preprint arXiv:2212.08073 (2022)
11. Holtzman, A., Buys, J., Du, L., Forbes, M., Choi, Y.: The Curious Case of Neural Text Degeneration. In: International Conference on Learning Representations (ICLR) (2020)
12. Reynolds, L., McDonell, K.: Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. arXiv preprint arXiv:2102.07350 (2021)
13. Lin, C., Hilton, J., Evans, O.: TruthfulQA: Measuring How Models Mimic Human Falsehoods. In: Proc. ACL 2022, pp. 3210–3252 (2022)
14. Pagnoni, A., Balachandran, V., Tsvetkov, Y.: Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics. Trans. Assoc. Comput. Linguist. 9, 807–824 (2021)
15. Ouyang, L., Wu, J., Jiang, X., Almeida, D., et al.: Training Language Models to Follow Instructions with Human Feedback. Adv. Neural Inf. Process. Syst. 35, 27730–27744 (2022)
16. Gao, L., Biderman, S., Black, S., et al.: The Pile: An 800GB Dataset of Diverse Text for Language Modeling. arXiv preprint arXiv:2101.00027 (2021)
17. Lewis, P., Oguz, B., Rinott, R., Riedel, S., et al.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Adv. Neural Inf. Process. Syst. 33, 9459–9474 (2020)
18. Thorne, J., Vlachos, A.: Automated Fact-Checking: Task Formulations, Methods and Future Directions. In: Proc. NAACL-HLT 2018, pp. 1–15 (2018)