A Survey of Hallucination Problems Based on Large Language Models

Xinxin Liu

School of Mathematics and Computer Science, Guangdong Ocean University, Guangdong, China

201940331004@stu.qhnu.edu.cn

Abstract. Large language models (LLM) have made significant achievements in the field of natural language processing, but the generated text often contains content that is inconsistent with the real world or user input, known as hallucinations. This article investigates the current situation of hallucinations in LLM, including the definition, types, causes, and solutions of hallucinations. Illusions are divided into different types such as factual and faithful, mainly caused by factors such as training data defects, low utilization of facts, and randomness in the decoding process. The phenomenon of hallucinations poses a threat to the reliability of LLM, especially in fields such as healthcare, finance, and law, which may lead to serious consequences. To address this issue, this article investigates methods such as managing training datasets, knowledge editing, and enhancing retrieval generation. Future research should classify and evaluate illusions more finely, explore multimodal strategies, enhance model stability, and integrate human intelligence and artificial intelligence to jointly address challenges, promoting the continuous progress of LLM.

Keywords: Large Language Models, hallucinations, Knowledge Boundaries, ChatGPT.

1. Introduction

LLM has made significant progress in the field of natural language processing with its excellent text understanding and content generation capabilities. The accumulation of large amounts of data and powerful computing power has brought great convenience to large language models such as Chat Generative Pre-trained Transformer, Bidirectional Encoder Representation from Transformers, Pathways Language Models, etc. in various fields. At the same time, the widespread application of Large language models (LLM) has exposed a fatal problem - hallucinations. The problem of hallucinations in LLM has attracted the attention of many people, that is, the answers generated by the model are incorrect or even fabricated, the logic is not rigorous or even contradictory, which seriously affects the authenticity and reliability of LLM's output answers. And in decision-making in various fields, the combination of human artificial intelligence often performs poorly because human trust in AI is inconsistent with its true abilities. Excessive trust in LLM generates erroneous answers with hallucinations, which can harm human cognitive abilities and moral concepts, ultimately leading to a crisis of trust in cutting-edge technology. Some literature suggests that the LLM hallucination problem leads to a lack of reliability in LLM, limiting its applicability. For example, in important fields such as healthcare and finance, the illusions generated by LLM may pose certain security risks and even pose significant threats to human life and property safety [1]. And it is pointed out that the current methods for assessing hallucinations require a large amount of manual labor, so effective, low-cost, and flexible strategies for alleviating hallucinations are currently the focus of scholars' research. There have been studies focusing on the detailed classification and causes of LLM hallucinations [2], and linking their causes with strategies for alleviating hallucinations, pointing out various research directions for alleviating hallucinations and existing problems, making significant contributions to the study of LLM hallucinations. Some studies have pointed out the wavering between practicality and authenticity in LLM, and proposed that LLM has deceptive behavior [3], which is often confused with hallucinations. Therefore, in future research on hallucinations, it is necessary to distinguish between honest errors and intentional deception, and divide them into different research directions. At the same time, another prominent security issue has also been pointed out, the "jailbreak phenomenon" of LLM [4], which needs to be distinguished from hallucinations and corresponding mitigation strategies need to be formulated. It can be seen that whether it is jailbreak, deception, or hallucination, they are closely related to the authenticity of LLM's output answers, and the security risks they bring are currently the focus of research. This article focuses on the issue of LLM hallucinations, investigating the classification of LLM hallucinations and the possible causes of output answer hallucinations during the data acquisition, training, and inference decoding stages of LLM. It summarizes various strategies to alleviate hallucinations and provides prospects for the future development of LLM, hoping to provide a useful reference for the development of the AI field.

2. Types of LLM hallucinations Phenomena

To enhance universality and applicability, the original LLM model classification based on the contradiction between output content and original content in specific task domains has been adjusted from internal and external hallucinations to factuality hallucination and faithfulness hallucination [2], with a wider range of applications.

2.1. Factuality Hallucination

Fact hallucination refers to the fabrication of facts. Inconsistent facts refer to the most common hallucination where the output answers do not match the facts of the real world. And factual fabrication refers to the fact that the answer is fabricated by the model itself, producing a wild and imaginative answer that cannot be verified based on established facts.

2.2. Faithfulness Hallucination

Faithfulness hallucination is divided into three subcategories, namely instruction inconsistency, context inconsistency, and logic inconsistency. Inconsistent instructions refer to the output answers deviating from the input instructions, specifically manifested as answers that are not relevant to the question. Inconsistent context refers to inconsistent contextual information in the output answers. Reflected in LLM ignoring user input text and giving incorrect output. Logical inconsistency refers to the occurrence of logical loopholes and contradictions in the output answers, which may manifest as contradictions between reasoning steps or between steps and results. Faithfulness hallucinations are difficult to identify because they appear reasonable but lack content support.

3. The Causes of LLM Hallucination

LLM is trained through data acquisition (pre-training, supervised fine-tuning, reinforcement learning from human feedback), and finally outputs answers through inference. The knowledge and ability of LLM come from these three steps (data, training, inference), and these three stages are also closely related to the generation of hallucination.

3.1. Data

Data is the foundation of LLM training, and the introduction of massive amounts of data enhances model performance while also bringing the risk of hallucinations.

3.1.1. Data defect. LLM relies on experience and rules rather than systematic deterministic heuristic data collection methods to collect large amounts of data, which may result in the collection of many erroneous information and biases. Hallucinations caused by erroneous information and biases can be mainly divided into three categories. Imitative lies, repetitive biases, and social biases. Imitation of lies refers to widely spread misunderstandings that LLM may consider correct and output misleading answers based on them. Repetitive bias refers to the inherent tendency of LLM to memorize training data, which is proportional to the size of the model. Words and phrases that repeatedly appear in the training dataset will be overemphasized by LLM, causing LLM to shift from generalization to memory and resulting in overly memorized sentences in the output that deviate from instructions. Social prejudice refers to the social prejudice unintentionally acquired from the Internet, which spreads to the generated content and connects two things wrongly.

In addition, there are knowledge boundary limitations, which refer to LLMs not being instructed to express their ignorance in supervised fine-tuning and pre training learning paradigms, lacking response and expression to knowledge boundaries, and often generating outdated or incorrect answers. If there is a lack of the latest facts, the internal knowledge of the trained model will never be updated. When faced with problems that require the latest knowledge, the model does not know how to explain its ignorance like a human, rather than trying to give a hallucinatory response. Therefore, it usually fabricates outdated answers based on outdated knowledge. In addition, there is a lack of professional domain knowledge. LLM, as it is trained on widely available datasets, has some limitations on obtaining data in specific domains, which often leads to the illusion of fabricating facts when faced with problems that require domain specific knowledge. This also limits the possibility of LLM being applied in professional fields such as healthcare and finance.

3.1.2. The mechanism by which LLM obtains facts is unclear, and the utilization rate of factual knowledge captured in the data is low. Excessive reliance on knowledge shortcuts refers to the tendency of LLM to use knowledge shortcuts rather than truly understanding knowledge, and to overly rely on positional proximity, co-occurrence statistics, and related document counts in pre-training data, which may lead to bias against false correlations.

Lack of knowledge recall and complex reasoning ability, LLM encounters difficulties in knowledge recall, long tail knowledge recall, and complex reasoning ability in complex scene processing, often leading to hallucinations.

3.2. Train

LLM gains knowledge and skills during the pre-training phase and further trains through supervised fine-tuning, ultimately aligning human preferences through human feedback.

3.2.1. Pre-training. During the pre-training phase, there may be hallucinations caused by architectural flaws and exposure biases [5].

Firstly, there are architectural flaws, as LLM modeling typically adopts a converter-based architecture that follows the paradigm established by GPT. This architecture has certain flaws, manifested in the inadequacy of one-way representation, which only utilizes context from a single direction and predicts the generation of subsequent tokens in order from left to right based on the currently processed token. This hinders its ability to capture complex contextual dependencies, In addition, in terms of attention, as the sequence length increases, the distribution of attention weights at various positions tends to be dispersed. The soft attention mechanism has limitations, which affect the model's ability to accurately capture key information and increase the probability of hallucinations. The difference between training and inference leads to exposure bias [5]. By using the chain rule to decompose the language model into a linear chain form, and then training the model parameters on the training corpus by maximizing the log likelihood, the teacher enforces a maximum likelihood estimation training strategy. This strategy provides the real value token as input, but relies on the conditional reflection context and the token generated by oneself for prediction during the generation process, without using the real context. There is an error between the context in the generation stage and the real context, and any error can cause a chain reaction, causing the generated content to deviate and increasing the possibility of hallucinations.

3.2.2. Alignment. Supervised fine-tuning utilizes high-quality instructions and responses to endow the LLM with the ability to follow user instructions, and then aligns human preferences through reinforcement learning from human feedback, enabling the LLM to generate high-quality and harmless answers. However, there is also a risk of hallucinating during this stage.

The ability and description are inconsistent, and the inherent functionality of LLM may not be consistent with its described functionality. When the need to align data exceeds these predefined ability boundaries, LLM will be trained to generate content beyond its own knowledge boundaries, amplifying the risk of hallucinations.

Internal beliefs and outputs are inconsistent, and the internal beliefs of models trained through human feedback reinforcement learning are not always perfectly aligned with the final generated output, which may result in a response that caters to human flattery. This tendency may be driven by both humans and preference models, generating false answers. In paper [3], it is mentioned that authenticity and social norms may be an implicit expectation of the input, but LLM may not include prior conditions, which may lead to erroneous outputs that contradict authenticity. Considering the different interactive environments of LLM, the situation is more complex, which is also seen as an implicit form of deception of LLM.

3.3. Inference

The shortcomings of inference inference inference strategies also have an impact on the occurrence of hallucinations. Likelihood traps caused by random sampling and limitations of top-level representation.

The likelihood trap caused by random sampling, which requires LLM to be creative and maintain diversity in generated content, is currently the mainstream decoding strategy used by LLM. However, incorporating randomness into the decoding strategy may lead to unexpected low-quality text output from high likelihood sequences, a phenomenon known as the likelihood trap [2]. But the pursuit of diversity increases the parameters for controlling the randomness of the model output, which will make the distribution of tokens more uniform and increase the likelihood of some uncommon tokens at the tail being sampled. This tendency to sample low-frequency tokens due to increased sampling temperature will exacerbate the risk of hallucinations. The limitation of the top-level representation is that LLM uses it to predict the next token during the decoding process. This limitation is manifested in insufficient contextual attention, as the generation model using the encoder decoder architecture is overconfident [2] and overly focused on locally generated content, often sacrificing loyalty to the input contextual content and achieving fluency. Additionally, the lack of attention mechanisms in the unidirectional architecture exacerbates LLM's output, often leading to faithfulness hallucinations. In addition, most language models use the Softmax layer to operate on the representation of the last layer of the model, but the non-linear function softmax reduces hardware utilization and is affected by hardware unfriendly non-linear operations [6], resulting in low computational parallelism. Transformer based LLMs often use Softmax as a key component of the self attention mechanism, so the softmax bottleneck increases the risk of hallucinations during the LLM generation stage.

4. Strategies to alleviate hallucinations

The most fundamental and effective strategy is to suppress hallucinations based on their categories and causes. This article investigates strategies and directions for reducing hallucinations related to data, training, and inference stages.

4.1. Reduce hallucinations related to data

The most effective way to alleviate the illusion related to data is to manually manage the training dataset, collect high-quality and effective data, and clean up low-quality and erroneous data. However, the large size of the dataset makes data filtering more difficult, so an efficient and low-cost strategy has become the direction of future research. In addition, research has found that sampling factual data during the pre training phase can effectively reduce the likelihood of hallucinations occurring. To reduce bias, the repetitive bias caused by the model's excessive memory tendency can be removed. The method of directly identifying the same string is affected by the complexity of data and can be queried in a short period of time through the construction of suffix arrays. However, approximate repeated syntactic overlapping items can be identified using hash based techniques [2].

The social bias that the model unintentionally obtains from the Internet can also be weakened by adjusting the data set.

The two mainstream methods for alleviating knowledge boundaries are knowledge editing [2], which directly modifies the parameters of the model to narrow the gap between the model and the required knowledge, and directly updates and strengthens the knowledge implementation.

Another strategy relies on retrieval enhanced generation [7], which processes non parametric external knowledge sources and introduces relevant information into the generation process through retrieval techniques. This method does not directly modify the internal parameters of the model, but instead propagates the retrieved documents to the input text to generate the desired output for the user, indirectly enhancing the model's knowledge capability.

In addition, research has proposed a confidence based knowledge boundary expression called CoKE, which teaches LLMs to use their internal signals to express knowledge boundaries and reject unanswerable questions [8]. This training setting helps the model better learn to express knowledge boundaries semantically, thereby enhancing its generalization ability.

4.2. Relieve training related hallucinations

One way to alleviate the shortcomings caused by unidirectional architecture is to use the bidirectional autoregressive method BATGPT [2]. Being able to predict the next token based on all previously observed tokens, while taking into account both past and future contextual information. By comprehensively capturing the dependency relationships between texts in two directions, LLM generates more accurate and rich texts.

In addition, a study has proposed a method called Faithful Finetuning [9], which emphasizes the fidelity of the response by carefully designing an explicit loss function during the fine-tuning process, and has a certain effect on improving model fidelity.

Besides, there is a method called skepticism modeling in research that effectively enhances the model's ability to estimate its uncertainty by combining token and logits information for self estimation, due to curiosity about whether human self doubt emotions can spread to LLM. Construct data on the perception of skeptical emotions, conduct continuous pre training, and improve the self estimation ability of LLM. Its generalization ability to other tasks has also been validated through out of domain experiments [10].

4.3. Reduce hallucinations related to inference

Reducing hallucinations related to reasoning is mainly divided into fact enhancement and independent decoding. Fact enhancement emphasizes the accuracy of facts to ensure the accuracy of LLM output, while independent decoding uses fact kernel sampling algorithm [2]. Dynamically adjusting the kernel probability and resetting it at the beginning of each new sentence balances the accuracy and diversity of generated content.

In addition, there is a strategy of using inference time intervention [2] to guide the true response of LLMs, because the activation space of LLMs has a clear structure with accurate facts. This method identifies the direction in the activation space related to accurate facts and adjusts the activation along the truth related direction during the inference process.

There is also a strategy called DoLa [2] for dynamically selecting and comparing different levels of logic, which delves into how to improve the authenticity of LLM decoding process from the perspective of factual knowledge storage, in order to enhance the authenticity of decoding.

And post editing decoding uses the self correction capability of LLM to improve the original generated content.

The problem caused by the bottleneck of softmax can be replaced by the hardware friendly software hardware collaborative design algorithm ConSmax [6] with learnable parameters, which helps to avoid data synchronization and improve the computational parallelism in softmax.

5. Challenges and prospects

By selecting strategies based on the causes, the problem of LLM hallucinations has been alleviated, but there are still many issues that require continuous research and investigation in this field. Multimodality is an important direction for the development of LLM, but the fusion of other modalities will inevitably encounter many obstacles. This article investigates the illusion mitigation strategies of large visual language models in the multimodal direction.

The hallucinations of large-scale visual language models combine visual perception and language processing capabilities, breaking through the limitations of traditional pre trained multimodal models in combining vision and language. However, it also responds to inconsistent image content, including misidentification of objects, attribute distortion, and semantic relationship confusion. However, LVLM is prone to misguidance, excessive reliance on language priors, and insufficient defense against malicious user input, which greatly increases the risk of significant performance degradation and hallucinations. And in current research, most discussions focus on illusions at the object level. It should be noted that even if visual elements are correctly recognized, the logical reasoning process of LVLM may still have defects, and this aspect has not been fully explored. In traditional LLM, effective methods for suppressing hallucinations, such as RAG not mechanically transferring to LVLM, may lead to more severe hallucinations. Therefore, by studying the introduction of active retrieval to enhance large visual language models to alleviate hallucinations, LVLM can be enhanced with external knowledge [7]. It can be seen that eliminating the hallucinations phenomenon in LVLM requires continuous research and efforts to ensure that these models can unleash their potential in more practical application scenarios.

6. Conclusion

LLM has made contributions to the development of AI, but the problem of hallucinations is becoming increasingly serious. The phenomenon of hallucinations not only manifests as incorrect understanding and expression of facts, but also has the potential to mislead users and trigger a crisis of trust. Therefore, the essence and causes of hallucinations, as well as how to effectively reduce hallucinations in LLM, have become the focus of research. This article delves into the categories of LLM hallucinations and analyzes the underlying causes, exploring diverse hallucination relief strategies tailored to different stages.

The limitations of LLM need to be taken seriously. While enjoying the convenience brought by AI, potential risks and challenges should not be ignored. AI technology needs to be regulated and viewed rationally by the public. The paper hopes that various disciplines can work together to promote the healthy and sustainable development of LLM.

References

[1] Zhao, Y., et al. (2024). Security Status and Challenges of Large Language Models. Computer Science, 51 (01): 68-71.

[2] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ArXiv, abs/2311.05232.

[3] Su, Z., Zhou, X., Rangreji, S., Kabra, A., Mendelsohn, J., Brahman, F., & Sap, M. (2024). AI-LieDar: Examine the Trade-off Between Utility and Truthfulness in LLM Agents.

[4] Mei, L., Liu, S., Wang, Y., Bi, B., Mao, J., & Cheng, X. (2024). "Not Aligned" is Not "Malicious": Being Careful about Hallucinations of Large Language Models' Jailbreak. *ArXiv, abs/2406.11668*.

[5] Arora, K., Asri, L.E., Bahuleyan, H., & Cheung, J.C. (2022). Why Exposure Bias Matters: An Imitation Learning Perspective of Error Accumulation in Language Generation. *Findings*.

[6] Liu, S., Tao, G., Zou, Y., Chow, D., Fan, Z., Lei, K., Pan, B., Sylvester, D., Kielian, G., & Saligane, M. (2024). ConSmax: Hardware-Friendly Alternative Softmax with Learnable Parameters. *ArXiv, abs/2402.10930*.

[7] Qu, X., Chen, Q., Wei, W., Sun, J., & Dong, J. (2024). Alleviating Hallucination in Large Vision-Language Models with Active Retrieval Augmentation. *ArXiv, abs/2408.00555*.

[8] Chen, L., Liang, Z., Wang, X., Liang, J., Xiao, Y., Wei, F., Chen, J., Hao, Z., Han, B., & Wang, W. (2024). Teaching Large Language Models to Express Knowledge Boundary from Their Own Signals. *ArXiv, abs/2406.10881*.

[9] Hu, M., He, B., Wang, Y., Li, L., Ma, C., & King, I. (2024). Mitigating Large Language Model Hallucination with Faithful Finetuning. *ArXiv, abs/2406.11267*.

[10] Wu, Y., Wang, Y., Chen, T., Liu, C., Xi, N., Gu, Q., Lei, H., Jiang, Z., Chen, Y., & Ji, L. (2024). Alleviating Hallucinations in Large Language Models with Scepticism Modeling.