Classification and solution of large language model hallucination

Ruoxi Zhang*

College of Electronic and Information Engineering, Tongji University, China

Abstract. Large language models (LLMs) offer significant benefits across various fields, enhancing research efficiency and innovation. However, a critical challenge is their tendency to produce "hallucinations"—content that appears fluent and reasonable but is factually inaccurate or nonsensical. This phenomenon can undermine content credibility and lead to erroneous decisions in critical areas like academic research, medical diagnosis, and news dissemination. This article provides a comprehensive overview of LLM hallucinations. It begins by defining LLMs as deep learning models trained on massive text datasets, typically using a Transformer architecture through stages like pre-training, supervised fine-tuning (SFT), and Reinforcement Learning from Human Feedback (RLHF). The paper then defines and classifies hallucinations into two main types: internal (factual) hallucinations, where the output contradicts objective facts , and external (context-inconsistent) hallucinations, where the output is logically inconsistent with the given prompt or conversational context. Furthermore, the article explores detection methods based on output content analysis, which check for features like semantic consistency and factual ambiguity. Finally, it reviews mitigation mechanisms, discussing strategies such as training on high-quality data and SFT to address factual hallucinations , and using targeted instruction fine-tuning and context-aware decoding for context-inconsistent hallucinations to improve the overall reliability of LLMs.

1 Introduction

In recent years, large language models (LLMs) represented by GPT-3, PaLM, Galactica, and LLaMA have achieved rapid development. These artificial intelligence models, based on deep learning techniques and nurtured by massive text data, are capable of understanding, generating, and analyzing human language, and have become an indispensable part of the current artificial intelligence field. Their applications are extremely broad, spanning numerous areas such as law, programming, transportation, and language, bringing significant convenience to people's daily lives and professional research.In the field of academic research, LLMs can quickly synthesize vast amounts of literature, assisting scholars in conducting literature reviews and knowledge integration, thereby identifying potential

* 2353913@tongji.edu.com

research directions and innovative points, and greatly improving research efficiency. Taking medical research as an example, by learning from a large corpus of medical literature, LLMs can help researchers quickly screen and summarize literature, providing targeted research directions and reference materials, thereby enhancing research efficiency and accuracy. In clinical diagnosis, LLMs can provide doctors with possible disease diagnosis suggestions and further examination directions based on the patient's symptom descriptions, medical history, and other information, assisting doctors in making more accurate judgments and reducing the risk of misdiagnosis and missed diagnosis. In the field of education, LLMs can provide students with personalized learning guidance, offering targeted answers and learning resources based on students' questions and learning progress, helping them better master knowledge. Furthermore, LLMs play important roles in intelligent customer service, content creation, machine translation, and other areas, bringing numerous conveniences to people's lives and work. However, while demonstrating significant advantages, LLMs also reveal a problem that cannot be ignored—the phenomenon of hallucination. When LLMs generate information, the output may appear reasonable, fluent, and logically sound in terms of grammar and semantics, yet it may contradict objective facts, or even contain fabricated information, falsified data, or inaccurate citations, seriously deviating from the truth. When answering historical questions, models might confuse the timing, figures, or details of historical events, providing incorrect answers; when dealing with scientific knowledge, they might propagate unverified viewpoints or theories, misleading users. This hallucination phenomenon not only reduces the credibility of content generated by LLMs but may also lead to serious consequences in critical fields, affecting public perception and judgment. Therefore, in-depth research into the hallucination problem of LLMs and the exploration of effective detection and mitigation methods have become urgent and important challenges in the current field of artificial intelligence. This paper focuses on the "hallucination" problem in large language models (LLMs). It begins by introducing the definition of LLMs, their training process based on the Transformer architecture, and their wide range of applications. It then clarifies the concept of "hallucination" and categorizes it into internal hallucinations and external hallucinations. Subsequently, the research expands from two aspects: detection methods and mitigation mechanisms. Finally, it concludes by summarizing the impact of the "hallucination" problem on the reliability of LLMs and the core value of related research, providing references for enhancing the applicability of LLMs in high-risk domains.

2 Definitions

Before discussing the hallucination phenomenon of large language models, this article first defines and explains large language models (LLMs) to better understand their conception. It then provides a definition and explanation of the phenomenon of hallucination in large language models.

2.1 Large Language Models

Large language models (LLMs) are AI models that are based on deep learning technology and are trained on massive amounts of text data. They are capable of understanding, generating, and analyzing human language. Language models (LLMs) have become top AI systems that can process and create interrelated texts and are widely used in various tasks. Typically, LLMs use the Transformer as a language architecture[2]. Based on this architecture, they are extensively trained on a large amount of language text, allowing the model to simultaneously focus on the relationship between different words in a sentence when processing text.

2.2 Development and Training of Large Language Models

2.2.1 Pre-training

Pre-training refers to the process of initially training a model using large-scale general domain unlabeled data or weakly labeled data before optimizing model parameters on task-specific labeled data. Pre-training adopts a data-driven unsupervised paradigm and does not require manually labeled "input-output" paired data. The model constructs learning tasks independently and mines intrinsic associations and feature representations from the raw data. Pre-training includes visual language pre-training (VLP) [3], natural language processing pre-training, computer vision pre-training, speech recognition and processing pre-training, etc.

2.2.2 Supervised Fine-tuning

Supervised Fine-Tuning (SFT) [4] is a critical phase in the development of large language models, designed to adapt a pre-trained model from general language understanding to targeted task execution. This process leverages the extensive knowledge acquired during pre-training and serves as a foundational step for subsequent Reinforcement Learning with Human Feedback (RLHF) [5] by providing a human-friendly initial model. The implementation of SFT is a structured, four-step process.

It begins with the construction of a high-quality supervised dataset, which must adhere to the core principles of task alignment, annotation consistency, and diversity to ensure robust model generalization. Following dataset preparation, the model is initialized using a pre-trained LLM, and parameter tuning strategies—ranging from full parameter tuning for resource-rich scenarios to more efficient partial parameter tuning methods like LoRA—are selected based on available data and computational power.

The training process itself is an iterative cycle of input encoding, model prediction, loss calculation using a cross-entropy function, and parameter updates via backpropagation. To mitigate overfitting, regularization strategies such as data augmentation, early stopping, and dropout are employed. Finally, the fine-tuned model undergoes a multi-dimensional evaluation using both automatic metrics (e.g., BLEU, ROUGE) [6] and manual assessments [7] to verify its effectiveness and alignment with human-centric criteria. This comprehensive evaluation informs further iterative refinements until the model's performance satisfies the target task requirements.

2.2.3 Reinforcement Learning with Human Feedback

Reinforcement Learning with Human Feedback (RLHF) [5] has emerged as a pivotal technology for optimizing model behavior to better align with human expectations and values. This method integrates human feedback directly into the reinforcement learning loop, overcoming the limitations of traditional approaches that depend on predefined reward functions. Defining such functions is particularly challenging for complex, subjective tasks like text generation or dialogue system construction, which are difficult to capture with simple mathematical formulas.

RLHF addresses this by using human evaluations as the primary source of reward signals. In this framework, human annotators score or rank the model's outputs, and this feedback is used to train a reward model that predicts human preferences. The technical architecture also includes a policy model that generates actions and an optimizer, often using algorithms like Proximal Policy Optimization (PPO) [8], to update the policy model based on the reward

signals. This process drives the model's evolution toward generating outputs that meet human expectations. RLHF has demonstrated significant practical success, such as improving dialogue systems by training them to reject harmful requests and enhancing autonomous driving systems by adapting to individual user preferences for comfort and efficiency.

2.3 Hallucinations in Large Language Models

2.3.1 Definition of Hallucinations in Large Language Models

Large language models are susceptible to hallucinations. Sometimes, the text generated by the model does not conform to objective facts, lacks logical consistency, or contains non-existent information (such as fictional characters, events, data, and quotations) [9]. However, the generated text appears highly similar to the real information in terms of grammatical and semantic fluency, making it difficult for humans to quickly detect its falsity.

2.3.2 Classification of hallucination phenomena

Internal hallucinations. Also known as “factual hallucinations” [10], the content generated by the model directly contradicts verified objective facts or contains completely fictitious “pseudo-factual” information. When a computer has internal hallucinations, it usually fabricates non-existent entities, tampers with factual details, or makes up unfounded associations, which is the main source of LLMs risk [4]. There are also the following detailed classifications of factual hallucinations: 1) **Factual inconsistency:** The answer output by the model contradicts the known facts, but the elements involved exist in the real world. For example, when asked “Who was the first person to land on the moon?”, the model answered that it was Gagarin, not Armstrong. Gagarin is a real astronaut, but he was not the first person to land on the moon. 2) **Fact fabrication:** The model creates completely fictitious information that has no basis in the real world and cannot be verified or falsified. For example, when the model was asked to introduce the origin of unicorns, it did not point out that there is no such creature as unicorns in the world, but gave a long fabricated content about its origin. This is fabrication of facts.

External hallucinations. Also known as "context-inconsistent hallucinations"[11], this refers to the situation where the content generated by the model does not directly violate general world knowledge, but is logically inconsistent or conflicting with the current dialogue context or specific information in the input prompt. In this type of hallucination, the model fails to remember and utilize the input context, and usually suffers from problems such as forgetting context information, logical self-contradiction, and input information tampering. In a recent study, Liangke Gui et al. [12] believed that clusters of hallucinated examples belonging to the same category should be grouped together with clusters of real examples, while clusters of hallucinated examples of different categories should be separated from clusters of real examples to solve this hallucination phenomenon [5].

3 Hallucination Detection

The hallucination problem of large language models has caused problems in various fields. In practical applications, the reliability of large language models has attracted widespread attention [13]. Therefore, it is urgent to alleviate the hallucination problem of large language models and improve the authenticity of large language models. In order to achieve this effect, effective means must be taken to evaluate and detect the hallucinations of large language models. Effective detection methods [14] such as CHAIR, POPE, and MME have been

applied in various fields [6]. This paper will explore the detection methods for hallucinations from two aspects: output content analysis and hallucination mitigation mechanisms. Key related works are shown as Table 1.

Table 1: Overview of methods for alleviating hallucinations

Method	Year	Core Technology	Method limitations
Pile [10]	2021	Multi-source data construction	Analysis of Ethical Religion
SemDeDup [13]	2023	Embeddings of pre-trained models to identify semantic repetitions	Data duplication and contamination risks, and downstream task data leakage cannot be avoided
RAG [14]	2020	Introducing Neural Knowledge Retriever	Index expiration can significantly impact performance
PKG [15]	2023	Key concept identification	Dependency model output logit value

3.1 Methods based on output content analysis

Output content analysis does not rely on the internal structure of the model. It only determines whether there are "hallucination features" by analyzing the text, image, and voice features of the AI-generated content. It can be divided into text-based hallucination detection and image-based hallucination detection according to the format of the content. Text-based hallucination detection: Text-based hallucinations usually have features such as content contradictions, logical breaks, and factual ambiguity. When solving text-based hallucination detection, semantic consistency detection can be performed by calculating the semantic similarity within the text through pre-trained language models (such as BERT and RoBERTa), or factual ambiguity analysis can be performed by identifying "unsourced and unfounded" factual statements in the text. Image-based hallucination detection: Computer vision models are used to identify whether the object structure conforms to the laws of reality or detect "semantically mismatched pixel areas" in the image. It is worth mentioning that the development of large language models in recent years has gradually highlighted their potential in automatic detection [7], and researchers are using this capability to conduct comprehensive evaluation.

3.2 Hallucination Mitigation Mechanisms

At present, hallucination mitigation methods have been applied and developed. This section reviews the existing mitigation strategies and will explore the related issues of hallucination detection from the perspective of factual hallucinations and faithful hallucinations based on the classification of hallucinations proposed above.

Mechanisms for mitigating factual hallucinations: For factual hallucinations, high-quality data training methods can be used to train with higher-quality, cleaner, and factually accurate datasets. Errors, contradictions, and inaccurate information in web crawled data can be cleaned up. Supervised fine-tuning can also be used to teach the model how to follow instructions and generate accurate responses through manual annotation. Furthermore, the self-correction ability of large language models can be used to achieve the purpose of hallucination mitigation through self-learning of language models [8].

Mechanisms for mitigating faithful hallucinations: The mitigation of faithful hallucinations requires attention to the internal consistency of the generated content, while also paying attention to consistency with the provided context. To address the fidelity illusion, targeted instruction fine-tuning is used, and a large amount of "instruction-context-faithful answer" data is used to fine-tune the model; reinforcement learning and reward models are used; CAD and other methods are used to perceive context decoding. At each step of the

model generating text [9], the algorithm is used to restrict it to only select words that are highly relevant to the source content, or to force the generated text to contain key phrases extracted from the source content; modal alignment is performed through contrastive learning [10].

4 Conclusion

This paper has provided a comprehensive examination of the hallucination phenomenon in large language models, a critical challenge that tempers their otherwise transformative potential. We have defined and classified hallucinations into factual and context-inconsistent categories, identifying them as a significant impediment to the reliability of LLMs in high-stakes domains. Through an analysis of model training paradigms, including Supervised Fine-Tuning (SFT) and Reinforcement Learning with Human Feedback (RLHF), we have contextualized hallucinations within the broader effort to align model outputs with factual accuracy and human values.

References

1. W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen, “A survey of large language models,” arXiv preprint arXiv:2311.03099, 2023.
2. A. Chernyavskiy, D. Ilvovsky, P. Nakov, Transformers: “the end of history” for natural language processing?, in: Machine Learning andKnowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13 – 17, 2021,Proceedings, Part III 21, Springer, 2021, pp. 677 – 693. 1
3. Z. Gan, L. Li, C. Li, L. Wang, Z. Liu, and J. Gao, "Vision-Language Pre-Training: Basics, Recent Advances, and Future Trends," Foundations and Trends® in Computer Graphics and Vision, vol. 14, no. 3 – 4, pp. 163 – 352, 2022, doi: 10.1561/0600000105.
4. J. Li, J. Chen, R. Ren, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen, "The Dawn After the Dark: An Empirical Study on Factuality Hallucination in Large Language Models," arXiv preprint arXiv:2401.03205, Jan. 2024.
5. L. Gui, A. Bardes, R. Salakhutdinov, A. Hauptmann, M. Hebert, and Y.-X. Wang, "Learning to Hallucinate Examples from Extrinsic and Intrinsic Supervision," *arXiv preprint arXiv:*, 2024.
6. LiangmingPan,MichaelSaxon,WendaXu,DeepakNathani,XinyiWang,andWilliamYang Wang.2023. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. ArXiv preprint abs/2308.03188 (2023). https://arxiv.org/abs/2308.03188
7. Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2023. TrustingYour Evidence: Hallucinate Less with Context-aware Decoding. ArXiv preprint abs/2305.14739 (2023).https://arxiv.org/abs/2305.14739
8. Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, andShikun Zhang. 2023. Hallucination Augmented Contrastive Learning for Multimodal Large Language Model. arXivpreprint arXiv:2312.06968 (2023)

9. T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., "Language models are few-shot learners," 2020.
10. OpenAI. (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.
11. Chowdhery, A., Narang, S., Devlin, J., et al. (2023). PaLM: Scaling Language Modeling with Pathways. Journal of Machine Learning Research, 24, 240:1–240:113.
12. Touvron, H., Martin, L., Stone, K., et al. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. GenAI, Meta.
13. Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al. (2021). The pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027.
14. Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). Retrieval Augmented Language Model Pre-Training. In Proceedings of the 37th International Conference on Machine Learning (ICML 2020) (Vol. 119, pp. 3929–3938). PMLR.
15. Varshney, N., Yao, W., Zhang, H., Chen, J., & Yu, D. (2023). A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation. arXiv preprint arXiv:2307.03987.