Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU33170

Research Progress on the Hallucination Problem of Large Language Models and Some Mitigation Strategies

Qifu Luo

School of International Education, Chengdu University of Technology, Chengdu, China
myyou5102@outlook.com

Abstract. Although Large Language Models (LLMs) have made significant progress in the field of natural language processing, the hallucination phenomenon has become a core issue in judging their credibility and practicality, which may cause serious consequences, especially in high-risk fields such as medicine and law. This article systematically explains the causes and current situation of LLM hallucinations and focuses on sorting out the current mainstream methods of research on alleviating hallucinations. First, the definition and classification of hallucinations are provided, and then the generation mechanism of hallucinations is explained from various aspects, such as data deviation and training goals. Then, six effective mitigation strategies on the market are introduced in detail and with emphasis. In addition, this article also introduces the main principle of evaluation and detection effectiveness using annotation to quantify hallucination tendency, and briefly describes three practical methods. Finally, he mentioned the security challenges of fabricated or misleading information generated by AI in legal, regulatory, and governance aspects, and put forward new ideas for looking at the problem of AI illusions in a positive way when looking into future research directions. Generally speaking, the problem of hallucinations in large language models consistently exists, and completely solving the hallucination problem still faces huge challenges.

Keywords: hallucination reduction, hallucination AI, large language model

1. Introduction

LaRge Language Models (LLMs) have made breakthroughs in the field of natural language processing. For example, chatGPT5.2, Grok4.0, etc., but a phenomenon occurs when it generates text. The generated text often contains content that is inconsistent with the facts or is inconsistent with logic, that is, a hallucination phenomenon. Hallucination has become an important issue in measuring the credibility and practicality of large language models. For example, when LLM is used at the core of industry applications that require high reliability, such as medical, legal, and financial, it not only affects user trust but also causes serious risks and has attracted widespread attention from academia and industry [1].

This article investigates the problem of alleviating hallucinations as its main goal. The article starts with the definition and classification of LLM hallucinations, and then explains several hallucination mechanisms and simple transitions to relief methods. Finally, this article screened out

DOI: 10.54254/2755-2721/2026.GU33170

six effective mitigation strategies. This article aims to provide more relevant information and ideas to scholars who study hallucinations.

2. Definition and classification of LMM hallucinations

2.1. Definition of hallucination

In the field of large language models, hallucination is broadly understood as generating content that is reasonable but not real and real-time. This shows that LLM can produce fabricated or false information. In short, this information is irrelevant and inconsistent with the facts. And a key difference in the medical definition of hallucinations (sensory experiences that do not correspond to external stimuli) is that in large language models, hallucinations refer to the generation of non-factual content in response to user questions, and the model often does not clarify whether its answers are true or false [2]. This feature again highlights the challenge of the lack of external validation of LLM outputs.

2.2. Classification of hallucinations

Studies have divided hallucinations into two main types: factually hallucination and faithfulness hallucination [3]. In addition, some studies have proposed a hierarchical classification method, dividing the manifestations of hallucinations into four categories: factuality errors, which generate fictitious information whose content violates objective facts. Faithfulness violations mean that the output content does not comply with user instructions. Logical inconsistencies mean that the reasoning process for generating content is inconsistent. Emergent behaviors mean that the model spontaneously generates complex behaviors that are not explicitly preset under specific conditions. In addition, its origin is traced back to three types of mechanism sources: data artifacts, which means that defects in the training data itself. Training biases mean that Preferences or flaws arise during model training. Inference failures mean problems that arise during the inference phase of the model [4].

3. The mechanism of hallucinations

Regarding the causes, research often uses a multi-level framework to discuss the causes of hallucinations, and clearly points out that analysis can be carried out at four levels: data layer, model architecture layer, training process layer, and other potential factors [5]. From the perspective of training goals and incentive mechanisms, some research clearly claims that language models are hallucinating because the training and evaluation process rewards guessing rather than acknowledging uncertainty [6]. From the perspective of internal mechanisms, some studies have also proposed the key finding that hallucinations will occur when the dominant hallucinatory associations exceed faithful associations, thus using the perspective of competitive associations to explain the generated content that deviates from fact [7]. However, in terms of multi-modal association, other studies have proposed knowledge overshadowing, describing more common knowledge that suppresses uncommon knowledge, thus leading to factual hallucinations [8]. In addition, improper application of prompt word design and thinking chain strategies, such as complex chain thinking and reasoning processes, may lead to logical breaks due to excessive cognitive load, which will enhance the tendency of hallucinations [9].

DOI: 10.54254/2755-2721/2026.GU33170

4. Mitigation strategies

4.1. RetrievaL-augmented Generation (RAG)

RAG combines the retrieval mechanism with the generative model to improve its performance by accessing external knowledge bases. RAG doesn't just use raw data; it also uses external databases to retrieve content and integrate it into the output, generating answers based on the retrieved documents. Research shows that the RAG enhanced model can reduce the hallucination rate from 68% to 10% [10].

4.2. Differential Penalty Decoding (DPD)

In order to solve the problem of how to track the source of hallucinations, Chen proposed a method to track the source of hallucinations through the attribution framework of internal signals, and proposed Differential Penalty Decoding (DPD) to reduce hallucinations by adjusting the posterior probability of the answer [11]. For example, ChatGPT classifies the incorrect answers generated by LLM into these categories based on the RelQA data set and obtains a new benchmark test of RelQA-Cate, under which Chen et al. studied DPD. It is not a training method, but directly imposes differential penalties on candidate answers during the generation process, so that the model is more inclined to output answers that are relatively less prone to hallucinations. The principle is to assign a penalty value to each hallucinated answer and then adjust the latter probabilities of these answers to reduce the probability of the hallucinated output being selected. This study shows that in LLaMA-7B, DPD improves by up to 7.26% on the TruthfulQA data set compared with previous SOTA strategies, and improves by 10.31% on the RelQA-Cate data set [11]. However, it is worth noting that this method relies too much on algorithms and post-processing, and the cost is high.

4.3. ScholarCopilot

ScholarCopilot is a unified framework that aims to enhance existing large-scale language models to generate professional academic articles with accurate context-sensitive citations [12]. ScholarCopilot is a RAG agent framework that determines in real time when to retrieve references by generating a retrieval template and calling the database for a query, in which the retrieved document content is introduced into the model to enhance the generation process. ScholarCopilot is based on Qwen-2.5-7B; the training content is based on 500,000 papers on arXiv. Wang's literature shows that its retrieval accuracy reaches 40.1%, higher than E5-Mistral-7B (15.0%) and BM25 (9.8%) [12]. This further reduces the occurrence of hallucinations through accurate literature searches. In a dataset of 1,000 academic writing samples, ScholarCopilot scored 16.2/25 in terms of generation quality – covering relevance, coherence, academic rigor, completeness, and novelty – significantly outperforming all existing models, including Qwen2.5-72B-Instruct (13.94). Wang et al. Ten highly knowledgeable students with experience in academic writing were invited through Human Assessment Design [12]. In terms of accuracy, interface clarity, and writing, the final average scores were 4.6/5, 4.5/5, and 4.5/5, respectively, giving a higher score than ChatGPT. However, ScholarCopilot still has limitations, because it combines the independent optimization of the retrieval model and the generation model, which will lead to misalignment of query intentions; retrieval decisions lack flexibility and context awareness: the static pipeline limits the user's control over content generation and citation needs, and currently does not involve all other disciplines except computer science. In addition, the creativity of the model needs to be continuously improved.

DOI: 10.54254/2755-2721/2026.GU33170

4.4. OMNIDPO

Compared with the visual encoder, the text encoder has significantly enhanced capabilities, which will cause text to dominate. Therefore, the model is often highly dependent on input text and ignores visual information and audio information. This is also the main cause of multimodal hallucination [13]. ODPO is a preference-alignment framework aimed at mitigating the OLLM illusion. Specifically, ODPO has two strategies. One is to build text preference samples to enhance the model's understanding of the interaction of visual and auditory information. Second, build a multi-modal preference sample enhancement model to pay attention to audio-visual information. When Chen et al. on Qwen2.5-Omni and MiniCPM-o-2.6 applied ODPO. The average performance improvement of CMM (Capability Maturity Model) benchmark is 3.48%, and the average performance improvement of AVHBench (Audio-Visual Hallucination Benchmark) is 4.23% [13]. This experiment shows that ODPO effectively improves the multi-modal hallucination problem, and ODPO also enhances the model's reasoning and question-answering capabilities in single-modal scenes. In addition, applying OMNIDPO can improve the F1 scores of models Qwen2.5Omni and MiniCPM-o-2.6, increasing by 5.82% and 2.64%, respectively [13]. ODPO is the first framework for mitigating OLLM full-modal hallucinations. Chen et al. also constructed OmniDPO-10k, a dataset for exposing and combating hallucinations [13]. However, this work only supports text, visual, and audio modalities, and has a high computational cost.

4.5. LVLM (Large Visual Language Model) Hallucination Revisor (LURE)

LURE (LVLM hallucination Revisor), developed by Zhou Y et al., is a lightweight and highly compatible post hoc method for correcting object illusion in LVLM [14]. It is also an object illusion corrector. Inspired by denoising autoencoders, LURE transforms latent LVLM-generated hallucinatory descriptions into accurate descriptions. The principle is to make two specific modifications to the original correct text. One is to insert additional text when describing, and the additional text may co-occur with the initial description object. Another is to encourage revisers to re-evaluate objects that replace uncertainty with placeholder labels, and ultimately use the obtained hallucination dataset to train the hallucination corrector. LURE is based on three key factors known to cause object hallucinations: co-occurrence, uncertainty, and object location, and these factors have been empirically and theoretically proven to induce hallucinations. Following this experiment, it can be seen that CHAIRI and CHAIRS were used to evaluate hallucination indicators, and it was also found that Teacher, CoT, and GPT Teacher can be significantly improved by LURE, effectively reducing the occurrence of hallucinations and hallucinations.

4.6. Counterfactual Probing

According to Feng, Counterfactual Probing is a new framework that leverages LLM's own generative capabilities to detect and mitigate hallucinations [15]. Its main principle is that in the face of many factual alternatives, a large amount of consistent knowledge should prove that the illusion is less likely to occur, and at the same time, evaluate the reliability of the model with dynamically generated counterfactual statements, that is, semantically similar but erroneous versions of the original facts. This framework is consistent with the current trend in supervised robustness testing. Counterfactual Probing provides an approach that eliminates the need to retrain external resources while providing insights that can interpret model behavior, such as explanations for why certain content is flagged as potential hallucinations, and adaptive mitigation strategies that can be

DOI: 10.54254/2755-2721/2026.GU33170

integrated with existing text generation pipelines. The experiment shows that CounterfactualProbing performs well on all indicators, especially on the TruthfulQA subset (F1: 0.816) and factual statements (F1: 0.824). The mitigation measures section gave an average 24.5% reduction in hallucination scores, and 78.5% of cases were successfully detected and intervened in [15]. In addition, this method relies too much on the quality of generating Counterfactual Probing, which may be challenging for highly specialized fields. In addition, this method focuses too much on factual hallucinations and may not be able to accurately cover other types of hallucinations.

5. Evaluate and test effectiveness

One of the core goals of the assessment and detection validity problem is to use annotations to quantify hallucination tendencies [2]. In terms of measurement and task setting, FActScore (Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation) is clearly positioned as a valid criterion for assessing the factual consistency of task outputs. Its working principle is to split the long text generated by the model into many independent "atomic facts", and then individually judge whether each atomic fact is supported by reliable knowledge sources [2].

LLM-Oasis extracts claim from Wikipedia (that is, extracts a statement based on facts) and falsifies some of the claims (modifies the facts to create content that contains illusions). Then the generated non-factual text is constructed to serve factual judgment and comparative evaluation (researchers mark one by one) [16].

In terms of unified comparison, UniFact is proposed to directly compare FV (fact verification) and HD (hallucination detection) at the instance level, thereby connecting the evaluation interfaces of "fact verification" and "hallucination detection". In the past, although the two fields had the same goals, they were incompatible because of different assessment methods. It allows LLM to answer questions in real time and record generated text and internal models for FV and HD to use for judgment. Simply put, it automatically detects factual errors in generated text [17].

6. Practical implications and future directions

In many high-risk application scenarios, hallucinations can appear in the form of seemingly believable but fabricated facts, such as legal AI tools that may generate fabricated case citations. In addition, hallucination is often described in the regulatory and governance context as AI systems generating misleading or fabricated information that deviates from the content of the source. Because this type of inaccurate output can impact critical decisions or spread misinformation, it is directly classified as a security challenge. In response to these risks, the "Practical Guidelines" strictly emphasize testing, such as explicitly "examining every citation, case, statute, rule, etc." At the same time, the guidelines also propose that risk levels should be set for AI tools and the intensity of inspection should be matched with corresponding risk use cases [18].

Facing the future, one research line regards hallucination as a structural problem related to architecture. Some people directly point out that hallucination is not an accidental defect, but a structural consequence of the transformer architecture [19]. Alternative ideas for representing space have also appeared in the same context, such as the proposal of "curved semantic space". In LLM, "meaning" is not the straight-line distance between points, but moving along a path that is constantly curved by context. Space itself can be "distorted" by the sentence being processed, leading to hallucinations. Interestingly, another route starts from positive utilization. Some studies clearly indicate that they are different from the existing reviews and re-examine the hallucination phenomenon from a positive perspective [20]. This study proposes to use the divergent and

DOI: 10.54254/2755-2721/2026.GU33170

convergent stages in cognitive science as a framework to systematically review how to transform illusions into creative resources. At the operational level, prompt engineering is also described as allowing LLM to continuously learn and adapt in context by setting specific prompts, thereby improving its ability to understand illusion generation and complex creative tasks.

7. Conclusion

The LLM illusion problem is a multi-dimensional and multi-level complex challenge, and its solution requires collaborative efforts from multiple dimensions, such as data, models, reasoning, knowledge, and applications. At present, a relatively complete cognitive framework has been formed, and the probability of AI hallucinations is being reduced step by step. However, there are huge challenges in completely solving the hallucination problem. Because AI is currently the most powerful assistant for humans, only by completely reducing the occurrence of hallucinations can humans more fully trust and rely on AI. In order to achieve a safer and more reliable artificial intelligence system, the LLM hallucination problem still has a long way to go. Overall, this article is produced according to the logical sequence structure of studying hallucinations. It starts with the definition of hallucinations, explains hallucinations, and compares medical hallucination diseases, then provides two classification methods of hallucinations, and then mentions various hallucination production mechanisms to help readers understand and lead to the focus of this article and methods to solve hallucination problems. Finally, the impact of textual illusion explains why this problem needs to be solved if people are to fully trust AI. In addition, the article ends with a new way of looking at hallucinations in a positive light. The hallucination problem has exposed the shortcomings of today's large-scale language models in terms of reliability. Its complete solution still requires continuous efforts from multiple dimensions, such as data, models, and applications.

References

[1] Alansari, A., & Luqman, H. (2025). Large language models hallucination: A comprehensive survey (arXiv: 2510.06265). arXiv.

[2] Cossio, M. (2025). A comprehensive taxonomy of hallucinations in large language models (arXiv: 2508.01781). arXiv.

[3] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., & Peng, W. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems. Advance online publication.

[4] Zhang, W., Zhang, C., Gu, C., Li, J., & Wang, X. (2024, October). Hallucination in large language models: From mechanistic understanding to novel control frameworks [Paper presentation]. 2024 7th International Conference on Universal Village (UV), IEEE.

[5] Lu, M. (2025). Exploring the causes of hallucinatory phenomena and coping strategies in the big language model. ITM Web of Conferences, 78, 04018.

[6] Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, Y. (2025). Why language models hallucinate (arXiv: 2509.04664). arXiv.

[7] Sun, Y., Gai, Y., Chen, L., & Zhang, M. (2025). Why and how LLMs hallucinate: Connecting the dots with subsequence associations (arXiv: 2504.12691). arXiv.

[8] Zhang, Y., Li, S., Qian, C., & He, B. (2025). The law of knowledge overshadowing: Towards understanding, predicting, and preventing LLM hallucination (arXiv: 2502.16143). arXiv.

[9] Liu. (2025). Prompt engineering and chain-of-thought strategies in LLM reasoning (inappropriate application may increase hallucination).

[10] Zhang, Y. (2025). A retrieval-augmented generation framework with retriever and generator modules for enhancing factual consistency. Applied and Computational Engineering.

[11] Chen, Y., Li, Z., You, S., & Wang, D. (2025). Attributive reasoning for hallucination diagnosis of large language models. Proceedings of the AAAI Conference on Artificial Intelligence, 39(22), 23660–23668.

DOI: 10.54254/2755-2721/2026.GU33170

[12] Wang, Y. B., Ma, X. G., Nie, P., & [Others]. (2025). ScholarCopilot: Training large language models for academic writing with accurate citations (arXiv: 2504.00824v2). arXiv.
[13] Chen, J. Z., Sun, C., Zhang, T. S., & [Others]. (2025). OMNIDPO: A preference optimization framework to mitigate omni-modal hallucination (arXiv: 2509.00723v1). arXiv.
[14] Zhou, Y. Y., Cui, C. H., Yoon, J. H. (2024). Analyzing and mitigating object hallucination in large vision-language models. 12th International Conference on Learning Representations (ICLR), Vienna, Austria.
[15] Feng, Y. J. (2025). Counterfactual probing for hallucination detection and mitigation in large language models. 42nd International Conference on Machine Learning (ICML), Honolulu, HI, United States.
[16] Sciré, A., Bejgu, A. S., Tedeschi, S., & Navigli, R. (2024). Truth or mirage? Towards end-to-end factuality evaluation with LLM-Oasis (arXiv: 2411.19655). arXiv.
[17] Liu, [Initial]. (2025). UniFact: Unified factuality evaluation for hallucination detection and fact verification (instance-level comparison framework).
[18] National Center for State Courts. (2026). A legal practitioner's guide to AI & hallucinations. https: //www.ncsc.org/resources-courts/legal-practitioners-guide-ai-hallucinations
[19] Ackermann, R., & Emanuilov, S. (2025). How large language models are designed to hallucinate (arXiv: 2509.16297). arXiv.
[20] Jiang, X., Tian, Y., Hua, F. (2024). A survey on large language model hallucination via a creativity perspective (arXiv: 2402.06647). arXiv.