Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ30007

Hallucination Detection in Large Language Models by Classifying Information Sources

Yifei Zhang

School of Computer Science and Technology, Harbin Institute of Technology, Weihai, China
2023212108@stu.hit.edu.cn

Abstract. In recent years, large-scale language models have made significant progress in natural language processing, providing excellent performance for applications such as text generation and information retrieval. However, at the same time, the problem of "hallucinations" has become increasingly prominent. Hallucinations primarily refer to inconsistencies between the content provided by the model and the facts or context, which can seriously undermine the model's reliability. Hallucination output from a model can lead to disastrous consequences. Consequently, a large amount of research has recently emerged on hallucination detection, encompassing both interdisciplinary research and diverse approaches to constructing hallucination detection models. This paper systematically reviews the latest work in hallucination detection. This paper firstly reviews the definition and classification of hallucinations and use this to evaluate the capabilities and characteristics of hallucination detection models. Next, this paper categorizes existing hallucination detection models based on the information source of the hallucinations, and further analyzes the detection mechanisms, scope of application, and accuracy of each hallucination detection model. Compared with existing reviews, this paper clarifies the categories of existing research and provides a promising direction for the development of comprehensive and improved models in the future.

Keywords: Large Language Models, Hallucination Detection, Information Source Classification, Factual Consistency, Trustworthy AI

1. Introduction

In recent years, the rapid development of large language models in the field of natural language processing has profoundly impacted people's daily lives. However, their generated results often exhibit "hallucinations," which appear plausible but contradict facts or logic. Without verification of authenticity, these hallucinations can easily be mistaken for correctness. This not only makes it difficult for ordinary users to distinguish authenticity and distorts online information, but can also lead to reduced efficiency and market volatility in high-risk scenarios such as industrial production and financial markets, jeopardizing automation processes and safety. Therefore, identifying and mitigating hallucinations has become a key research topic in natural language processing.

In the past three years, hallucination detection technology has continued to develop and can be divided into three main categories. First, the external knowledge-based approach uses an existing

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.TJ30007

database to check the generated content against it to identify any discrepancies with the data. The FactAlign model uses knowledge graph alignment technology to effectively identify hallucinations in fields such as medicine [1]. Second, methods based on internal signals perform real-time monitoring by analyzing the hidden states or gradients during model generation, and can efficiently detect hallucinations without external knowledge. Representative examples include the MIND framework, which uses internal states to achieve detection under unsupervised conditions [2]. Third, self-consistency methods identify potential hallucinations by comparing output consistency through multiple samplings. A typical example is SAC^3, which introduces semantic-aware consistency testing and significantly improves detection reliability in black-box scenarios [3]. These methods have shown important application value in medical, financial, and other high-risk decision-making scenarios.

Based on the above analysis, this paper proposes an information source classification method to systematically explore and compare existing hallucination detection models. Specifically, this paper first sorts out the concept, classification, and generation mechanism of hallucination to help understand its essence; then, it classifies detection models according to information source, summarizes their principles, datasets, evaluation criteria, and applicable scenarios, and reveals the advantages and limitations of different methods, especially the challenges in cross-domain applications. Unlike existing reviews, this paper emphasizes the guiding value of information source classification in future model integration and optimization, and proposes related research challenges and development directions, in order to provide a reference for the reliable implementation of large language models.

2. Research methods

2.1. Classification of hallucinations

In the process of generating large language models, hallucination phenomena are generally divided into two categories, namely internal hallucinations and external hallucinations.

Internal hallucinations typically arise when models generate outputs without the support of external knowledge bases. In these cases, the model relies on its internal language model to generate content. While the generated text is grammatically and contextually sound, it may contain errors in logic, semantics, or reasoning. This type of hallucination is characterized by local contextual coherence coupled with global semantic deviation.

External hallucinations primarily occur when a model relies on external knowledge or information sources for its generation. In this case, while the model accesses knowledge bases or documents, the generated content is inconsistent with real-world facts, resulting in erroneous or inaccurate output. This type of hallucination is characterized by generated content that appears authoritative or reasonable when citing external information, but in reality, conflicts with real data or facts, easily misleading users' understanding of external knowledge.

2.2. Mainstream methods, datasets, and evaluation criteria

Currently, mainstream large language model hallucination models can be divided into three categories based on the source of information: external knowledge-based, internal signal-based, and self-checking-based. Information is used to build models through knowledge bases, model signals, and multiple consistency.

DOI: 10.54254/2755-2721/2026.TJ30007

2.2.1. External knowledge-based methods

Hallucination detection models based on external knowledge bases aim to identify situations where the content generated by large language models is inconsistent with external facts, thereby reducing the occurrence of external hallucinations. Such models typically combine information retrieval and generation mechanisms, enhancing generation capabilities by introducing a knowledge base. The core principle is that the model first retrieves document fragments related to the input query from the knowledge base, then feeds these fragments into the generation model along with the query to make the answer more accurate. Simultaneously, the model performs consistency verification, comparing the generated content with the retrieved information to identify any discrepancies with the facts.

The most typical model among external knowledge methods is RAG. Because external knowledge bases can prevent model overfitting, RAG models are often used in open datasets such as RAGTruth [4]. This dataset contains a large number of question-answer pairs, mainly in scenarios that require retrieval and generation. Each question is accompanied by multiple relevant literature fragments to provide sufficient information to help the model generate accurate answers.

For hallucination detection models based on external knowledge bases, the following evaluation indices are usually adopted, such as accuracy rate, that is, to determine the degree of consistency between the model's output and the facts. Generation consistency refers to the consistency between the output of the detection model and the retrieved information.

2.2.2. Internal signal-based methods

Hallucination detection models based on internal signals are primarily used to determine whether large models are generating hallucinations of "internal" knowledge and reasoning. These models typically use the large model's own representations, hidden layers, and attention information to determine whether there are any contradictions with reality. The core idea of this detection method is to track the various feedback the model receives when making predictions for each token, such as the model's internal uncertainty confidence, inconsistencies, ambiguity, or logical clarity. This allows the model to detect where the erroneous information originates, thereby controlling the generation of hallucinations in the large model.

RSV is an example of a model based on internal signals. During this process, after proposing various candidate outputs for the answer, the model will use an internal reward mechanism to strengthen the model's answer. As mentioned above, if the results given by the model are very uncertain or the model parameters do not meet the requirements, the output will be modified. The RSV model introduces a reward mechanism and uses the idea of self-verification to control hallucinations. It is often used in the MedQA dataset for medical question answering and the SWAG dataset, which contains a large amount of common sense reasoning. This type of problem that requires difficult reasoning to complete often uses the RSV model [5].

In this type of model, internal signals include model parameters such as attention weights and neural network activations to represent uncertainty. Attention weights can generate confidence information, verifying whether the model is focusing on content irrelevant to the problem; while hidden activations describe the number of active neurons at different layers in the model and their activity levels.

DOI: 10.54254/2755-2721/2026.TJ30007

2.2.3. Self-consistency-based methods

Unlike models based on external knowledge bases that require the introduction of a database, and unlike models based on internal signals that require consideration of the changes in parameters at each layer during model processing, models based on self-consistency generate answers to the same question multiple times and compare the consistency between these answers to determine the possibility of hallucinations. The core principle is that if the model maintains a high degree of consistency in its multiple outputs for the same question, the output can be considered to be highly credible. On the contrary, if there are significant differences in the output, it is considered that the output has hallucination problems [6].

This idea was first proposed by Wang et al. to improve the reliability of model reasoning through multiple sampling comparisons [6]. This method was subsequently introduced to the hallucination detection task and developed into models such as the SAC³ model. The SAC³ model effectively reduces the hallucination problem caused by the inaccuracy of single generation by combining self-consistency and contrastive learning, thereby improving reliability in scenarios such as long text generation and multi-step reasoning tasks [3].

Because self-consistent models don't rely on domain-specific knowledge, they can be applied to unknown domains with zero samples, addressing a major pain point of models based on external knowledge bases. Furthermore, by generating multiple candidate answers, self-consistent models address the drawback of internal signal-based models, which rely on a single deterministic reasoning path, thereby increasing output diversity.

3. Experimental results

In recent years, hallucination detection using large language models has made significant progress. In the medical field, the RAG model has been widely used and has demonstrated high practical value. The MedRag model proposed by Xiong et al. significantly outperforms traditional generative models in processing complex medical knowledge by introducing an external knowledge base and a generative enhancement mechanism [7]. Clinical data show that the GPT-4+RAG model, combined with international medical guidelines, has an accuracy rate of 96.4% in preoperative medical assessment, significantly higher than the 86.6% of traditional experts, demonstrating the practical value of models based on external knowledge bases [8].

In addition, hallucination detection methods based on model internal signals have also achieved results. The white-box method EGH proposed by Hu et al. combines embedding and gradient information to surpass SelfCheckGPT on multiple datasets [9]. Zhang et al.'s ICR Probe uses hidden state dynamics to further improve detection accuracy, making it particularly suitable for multi-task scenarios [10].

Finally, self-consistency-based detection methods such as SelfCheckGPT proposed by Manakul et al. performed well in sentence-level hallucination detection on the WikiBio dataset, with an AUC-PR significantly higher than multiple baseline methods, verifying the effectiveness of generating and comparing consistency through multiple sampling to identify potential hallucinations [11].

In summary, weak generalization and poor detection of complex hallucinations are the two major bottlenecks of current LLM hallucination detection technology. Future research can further improve the robustness and applicability of the detector through strategies such as cross-domain knowledge integration, multi-model result weighting, text segmentation, and information aggregation, laying the foundation for the reliable deployment of large language models in practical applications.

DOI: 10.54254/2755-2721/2026.TJ30007

4. Challenges and prospects

Although large language model hallucination detection technology has made some progress in recent years, significant challenges still exist. First, limited generalization capacity remains a central challenge. Many existing detection models are mainly trained for specific domains or datasets, and their performance is significantly reduced in cross-domain or cross-task scenarios. For example, the hallucination detection method based on model internal signals proposed by Preiß et al. performs well on a single benchmark, but its generalization ability across different datasets and models is still limited [12]. To solve this problem, the model needs to introduce a cross-domain knowledge base or fuse different categories of models together and use different weights to solve problems in different fields.

Furthermore, the current model's detection effect on complex hallucinations still makes it difficult to apply them in practice, especially when dealing with scenarios involving long texts. The current model has difficulty extracting effective information from long texts, which in turn leads to a decrease in the accuracy of the model. However, there are still some models worth learning from at present. For instance, the model proposed by Liu et al. effectively identifies inconsistent information between paragraphs by introducing multiple rounds of internal reasoning, thereby improving the accuracy of long text illusion detection [13]. This method indicates that the model can segment long texts and then aggregate the information of the segmented texts. In this way, the problem of long text detection can be solved to a certain extent.

Overall, weak generalization ability and poor performance in long text detection are the main difficulties of current hallucination detection models. To apply large language models in high-risk fields such as industrial production, these problems must be solved. Future models can introduce multiple models to detect specific domains to enhance generalization ability. At the same time, a cutting mechanism can be introduced to split long texts into multiple short texts to improve accuracy in long text scenarios.

5. Conclusion

This paper comprehensively studies hallucinations generated by large-scale language models and their detection methods. It proposes a classification method for information sources in the hallucination problem of large-scale language models and categorizes existing hallucination detection models into three categories: exogenous, endogenous, and self-attainment. The paper comprehensively discusses the principles, datasets, evaluation metrics, and application scenarios of each method. It finds that exogenous methods can effectively eliminate the frequency of external hallucinations in industries such as healthcare and finance, ensuring the consistency of content with facts. Endogenous methods, based on the model's internal credibility information and hidden states, can effectively and explainably detect internal hallucinations at certain levels. Self-attainment-based methods, by comparing multiple different generation and reasoning paths, can generate more generative content and higher-confidence content. This paper reveals the respective strengths of each method, helping users to better select the detection method appropriate for their specific situation.

Based on existing research, the paper outlines the main future research directions for hallucination detection models: Based on existing hallucination models, since all existing models are difficult to adapt to cross-domain or complex scenarios, simply weighting or unweighting each different detection model may achieve very good results. By separating model information sources by category, this paper provides a prototype for an ideal multi-model fusion approach for researchers working in this area. This allows for the organic integration of the strengths of various existing

DOI: 10.54254/2755-2721/2026.TJ30007

methods, enhancing the ability to distinguish between internal and external hallucinations. In summary, this paper enriches theoretical understanding of LLM hallucination presentation and lays the foundation for a robust and reliable hallucination detection framework, possessing significant practical application value and potential for future development.

References

[1] Mohamed, R., Zahran, A., Amin, A., Abdelaal, A., and Altantawy, M. (2024) FactAlign: Fact-Level Hallucination Detection and Classification Through Knowledge Graph Alignment. In Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024), pages 79–84, Mexico City, Mexico. Association for Computational Linguistics.

[2] Su, W. H., Wang, C. Y., Ai, Q. Y., Hu, Y. R., Wu, Z. J., Zhou, Y. J. and Liu, Y. Q. (2024) Unsupervised Real-Time Hallucination Detection based on the Internal States of Large Language Models. In Findings of the Association for Computational Linguistics: ACL 2024, pages 14379–14391, Bangkok, Thailand. Association for Computational Linguistics.

[3] Zhang, J., Li, Z., Das, K., Malin, B., and Kumar, S. (2023) SAC3: Reliable Hallucination Detection in Black-Box Language Models via Semantic-aware Cross-check Consistency. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 15445–15458, Singapore. Association for Computational Linguistics.

[4] Niu, C., Wu, Y., Zhu, J., Xu, S., Shum, K., Zhong, R., Song, J., and Zhang, T. (2024) RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.

[5] Liu, Z., Zhang, T., Chen, Z., Huang, T., & Wang, H. (2023, August). Learning to Verify: Reinforced Self-Verification for Text Generation. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), pages 1905–1919.

[6] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., and Zhou, D. (2022) Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv: 2203.11171.

[7] Xiong, G., Jin, Q., Lu, Z., and Zhang, A. (2024) Benchmarking Retrieval-Augmented Generation for Medicine. In Findings of the Association for Computational Linguistics: ACL 2024, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.

[8] Ke, Y. H., Jin, L., Elangovan, K., et al. (2025) Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. npj Digital Medicine, 8, 187. https: //doi.org/10.1038/s41746-025-01519-z

[9] Hu, X., Zhang, Y., Peng, R., Zhang, H., Wu, C., Chen, G., and Zhao, J. (2024) Embedding and Gradient Say Wrong: A White-Box Method for Hallucination Detection. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 1950–1959, Miami, Florida, USA. Association for Computational Linguistics.

[10] Zhang, Z., Hu, X., Zhang, H., Zhang, J., and Wan, X. (2025) ICR Probe: Tracking Hidden State Dynamics for Reliable Hallucination Detection in LLMs. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 17986–18002, Vienna, Austria. Association for Computational Linguistics.

[11] Manakul, T., Zhang, Y., et al. (2023, June). SelfCheckGPT: Improving Hallucination Detection with Self-consistency on WikiBio. In Proceedings of the 41st International Conference on Machine Learning (ICML 2023), pages 2840–2852.

[12] Preiß, M. (2025) Hallucination Detection with the Internal Layers of LLMs. arXiv preprint arXiv: 2509.14254.

[13] Liu, Z., Guo, J., Zhang, H., Chen, H., Bu, J., and Wang, H. (2025, July) Long-form Hallucination Detection with Self-elicitation. In Findings of the Association for Computational Linguistics: ACL 2025, pages 4082–4100.