Progress and Challenges in Addressing Hallucination Issues in Large Language Models

Zongke Li *

Faculty of Geosciences and Engineering, Southwest Jiaotong University, Chengdu, Sichuan, China

* Corresponding Author Email: shineldrop@gmail.com

Abstract. The illusion problem of large language models refers to the phenomenon where the content generated by the model is inconsistent with the input information or objective facts, significantly limiting its reliability and safety. This paper systematically reviews the definition and classification of the illusion problem, including factual illusions and fidelity illusions, intrinsic and extrinsic illusions, as well as types of closed-domain and open-domain illusions. It also summarizes the current mainstream evaluation methods from three perspectives: data, models, and multi-task applications. Regarding mitigation strategies, the article proposes various technical paths at the data, model, and application levels, including data cleaning and augmentation, model architecture optimization, prompt engineering, and real-time retrieval-augmented generation, which effectively enhance the accuracy and consistency of the generated content. Future research should aim to establish a more refined evaluation system, promote collaborative optimization across multiple technologies, and achieve dynamic knowledge updates and lightweight deployment to strengthen the practicality and safety of large models in real-world scenarios.

Keywords: Large Language Models, Hallucination Issues, Evaluation Methods, Mitigation Strategies.

1. Introduction

Hallucination in large models refers to the phenomenon where generated text content diverges from input sources or objective real-world facts [1]. Originating in natural language processing, this concept initially described outputs unrelated to or untrue to the source content. With expanding model capabilities, its definition now encompasses factual errors and logical contradictions [2]. Hallucination issues in large language models can be systematically categorized into two types: factual hallucinations and fidelity hallucinations [3]. Based on the relationship between the large model and the input information source, hallucinations can further be classified as intrinsic hallucinations and extrinsic hallucinations [4]. Based on the model's reasoning performance in specific tasks or domains, hallucinations can be categorized as closed-domain hallucinations and open-domain hallucinations [4]. Depending on the types of contradictions that may arise during information processing, hallucinations can be classified as input-contradiction hallucinations, context-contradiction hallucinations, and factual-contradiction hallucinations [5]. With the widespread adoption of large language models, research into their hallucination issues has grown increasingly significant. The training data for large language models inevitably contains errors, biases, and outdated information. This makes models particularly prone to hallucination during text generation—confidently outputting seemingly plausible yet factually incorrect content that deviates from user intent or context [4]. Hallucinations severely undermine model reliability and safety, limiting their application in high-accuracy domains like healthcare and law, and potentially triggering ethical and legal risks.

This article systematically reviews three core aspects of the hallucination problem in large language models: its definition and classification, assessment methods, and mitigation strategies. Firstly, it clarifies the types of hallucinations, including factual/faithfulness hallucinations, internal/external hallucinations, and closed-domain/open-domain hallucinations, and points out their manifestations in different tasks. Then, it sorts out the current mainstream assessment methods from three dimensions: data text, model reasoning, and multi-task applications, and analyzes the

advantages and disadvantages of each method. Finally, it summarizes the technical paths for mitigating hallucinations at the data layer, model layer, and application layer, such as data cleaning, model structure optimization, prompt engineering, and retrieval-augmented generation. The main purpose of this article is to systematically review the current research status of hallucination problems in large models, propose a multi-level assessment and mitigation framework, and enhance the reliability and security of generated content. Its significance lies in providing a comprehensive technical review and method reference for researchers and practitioners, promoting the safer application of large language models in high-reliability requirement fields such as healthcare and law, and providing theoretical support and practical guidance for future research directions, such as dynamic knowledge update, multi-technology collaboration, and lightweight deployment.

2. Evaluation Methods for Large Model Hallucinations

2.1. Data-Text-Based Evaluation Methods

This approach quantifies the information alignment between generated text and reference text using statistical metrics (e.g., precision, recall), categorized into three types [1]: The first type uses the target text as reference, such as the PARENT metric combining precision and recall, the AUC-PR score for detecting paragraph consistency, and statistical methods based on RAG-built hallucination corpora; The second category uses only the source text as reference, such as the PARENT-T optimization variant that omits target text comparison and the Knowledge F1 metric for knowledge-based dialogue tasks; The third category employs extended text references, including methods like the chrF score based on character n-gram matching and BVSS for sentence similarity measurement, commonly used in tasks like machine translation.

2.2. Model-Based Evaluation Methods

Model-based information extraction (IE) extracts knowledge (e.g., relation tuples) from generated and reference texts using end-to-end Transformers for comparison. Typical metrics include error gap rate and F1 score, suitable for detecting factual contradictions and open-domain hallucinations, though evaluation accuracy may be compromised by extraction errors [4]. Model-based inference (NLI) assesses textual entailment relationships through natural language reasoning. This encompasses sentence-level approaches (e.g., BERT/RoBERTa trained on MNLI/ANLI) and document-level methods (e.g., SummaCConv, FactCC, DocNLI), with the latter generalizable to domain-independent tasks [1]. Specific model approaches further encompass dual-model comparisons (e.g., the KoLA benchmark calculating word example probability differences via conditional/unconditional language models), direct scoring models (e.g., AlignScore, UniEval, and FEWL), and real-time detection frameworks (e.g., token-embedding-based MIND and two-stage explainable evaluation DEE).

2.3. Methods Based on Multi-Task Applications

Question-answering evaluation methods assess textual fidelity by automatically generating questions and verifying answer consistency. Examples include FEQA, which generates questions by masking summaries; QAGS, which introduces filtering mechanisms to eliminate low-quality questions; QuestEval, which combines precision and recall for comprehensive assessment; the dynamic benchmark FreshQA, supporting both strict and lenient evaluation modes; and Sac3, which enhances reliability through semantic cross-checking [4]. Classification evaluation methods rely on constructing task-specific annotated datasets (e.g., WoW for knowledge-based dialogue, Conv-FEVER for fact inconsistency detection, SUMMAC for binary classification) and employ probabilistic dynamic classification techniques (e.g., residual flow mapping) to improve accuracy in identifying hallucinations of known fact categories [4].

2.4. Chapter Summary

Hallucination evaluation methods have evolved from data statistics and model inference to multi-task applications. Core challenges include depth of semantic understanding, model generalization capability, and balancing evaluation efficiency with interpretability. Future work should integrate dynamic knowledge updates and fine-grained classification systems to further enhance reliability. As shown in Table 1, a comparison of three classification evaluation methods is presented.

Table 1. Comparison of Evaluation Methods

Classification	Advantages	Limitations
Data-Text-Based Evaluation Methods	Simple, efficient, model-agnostic, and user-friendly	Challenged in handling consistency judgments involving implicit knowledge or complex reasoning
Model-Based Evaluation Methods	Capable of capturing deep semantic relationships and overcoming missing data in text labels	High computational cost; potential misjudgments due to model hallucinations.
Multi-task application-based evaluation methods	Suitable for evaluating models with relatively homogeneous task types, achieving high accuracy on specific tasks.	Challenging to generate questions and ensure quality control, high dataset construction costs and insufficient reliability

3. Mitigation Approaches for Large Model Hallucinations

Based on the root causes of hallucinations, existing mitigation approaches can be categorized into three levels: data layer, model layer, and application layer. This paper systematically analyzes the technical principles, research progress, and practical application outcomes of each level [4].

3.1. Data-Layer Hallucination Mitigation Methods

Optimizing the data layer represents the fundamental approach to addressing hallucination issues. High-quality training data significantly reduces the probability of models learning erroneous patterns. Current data-layer methods primarily focus on two directions: data collection and data preprocessing. Regarding data collection, researchers have proposed various methods for constructing faithful datasets. Gardent et al. pioneered the extraction of data units from structured knowledge bases, leveraging template generation techniques to build a substantial faithful corpus [5, 6]. The Gabriel team took a different approach by systematically annotating erroneous summaries generated by models, establishing the first fact-consistency evaluation dataset [7].

Data preprocessing techniques primarily encompass two dimensions: data cleaning and data augmentation. Nan's team proposed an innovative cleaning method that effectively filters out unfaithful content by establishing a sentence-level matching mechanism between source documents and summaries. For data augmentation, Cao et al. developed a triplet extraction technique that automatically identifies key facts from source documents and supplements them into training data in a structured format [4].

Zhang et al. proposed knowledge graph fusion techniques, leveraging structured knowledge bases to provide verified entity-relationship information. Concurrently, they employed credibility assessment frameworks (e.g., CAG) to hierarchically filter retrieval results based on data source relevance, timeliness, and authority. For query optimization, researchers developed query expansion techniques (Query2Doc) and decontextualization processing. The former clarifies query intent by generating pseudo-relevant documents, while the latter resolves anaphoric ambiguity in dialogues through rewriting. Furthermore, self-generated context methods (e.g., GENREAD) utilize LLMs to directly generate relevant documents, circumventing noise introduced by traditional retrieval. As shown in Figure 1, the data layer specifically focuses on multi-source data fusion and validation, employing Retrieval-Augmented Generation (RAG) to ensure information comprehensiveness and

timeliness. These techniques reduce noise and conflicts at the information input source, providing cleaner, more reliable data foundations for subsequent processing [3, 8].

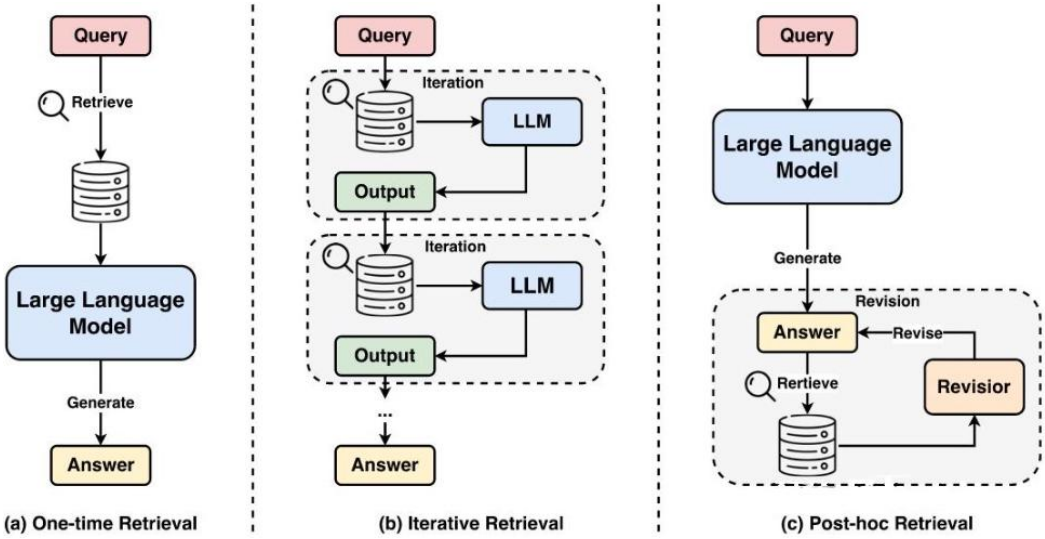

Figure 1. Three RAG Retrieval Approaches [3]

3.2. Methods Model-Level Hallucination Mitigation Methods

Model-level optimization represents the core technical approach to addressing hallucination issues, primarily encompassing three aspects: model architecture refinement, training strategy improvements, and post-processing techniques.

Regarding model architecture, researchers have innovated across both encoder and decoder dimensions. Xiao's team introduced a confidence penalty mechanism, significantly reducing the generation of low-confidence content. Within a multi-task learning framework, Nan's team enhanced generated accuracy through knowledge sharing mechanisms. In reinforcement learning, Cao et al. designed entity-aware reward functions to provide finer-grained supervision signals for generation quality (as shown in Figure 2. Zhang et al. enhanced long-text information extraction efficiency by constructing multi-level document representations through hierarchical retrieval systems (e.g., RAPTOR) and recursive clustering. The position-agnostic reasoning method (PAM-QA) specifically addresses the "middle curse" problem by enabling models to equally consider information across all contextual positions through reinforcement training [8].

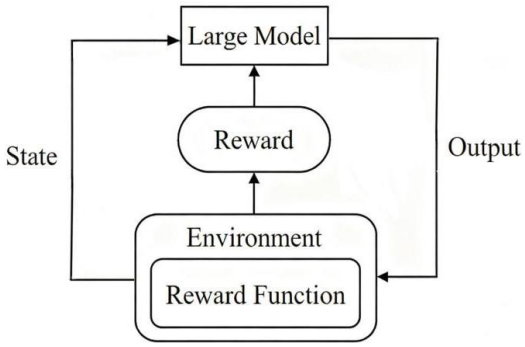

Figure 2. Schematic of reinforcement learning approaches [4]

Regarding training strategies, researchers developed diverse prompt engineering techniques, including Chain of Thought (CoT) for stepwise reasoning guidance and Programmatic Thinking (PoT) for invoking external interpreters to perform precise computations. Fine-tuning approaches like the ALCE framework train models to generate cited answers, forcing explicit links between outputs and retrieved content. For knowledge conflict resolution, the COMBO framework effectively balances internal knowledge and external evidence by generating parameter-based and retrieval-based answers in parallel and synthesizing them. These innovations not only enhance model fidelity to context but

also strengthen complex reasoning and fact-checking capabilities, yielding more reliable outputs [9]. He et al. successfully reduced hallucination rates through a swarm intelligence-based approach [10].

Post-processing techniques garner significant attention for their flexibility. As shown in Figure 3, Chen's team developed an iterative correction algorithm that automatically detects and replaces erroneous entities in generated text, achieving an 8% accuracy improvement in news summarization tasks. Zhou et al. utilized external expert models to filter hallucinated content [4].

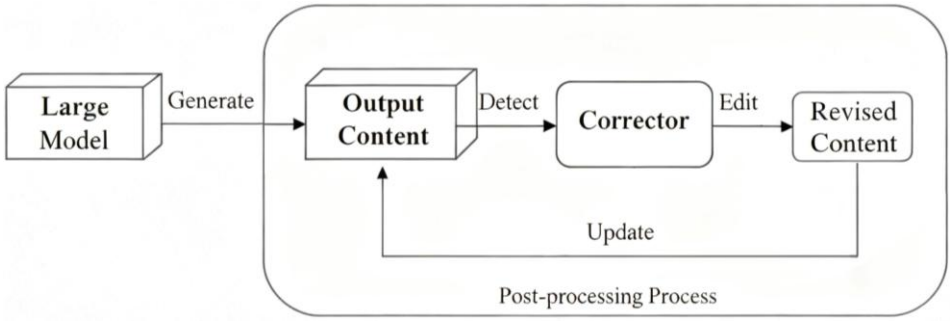

Figure 3. Schematic of post-processing methods [4].

The link strategy of the model-layer can be seen in Figure 4:

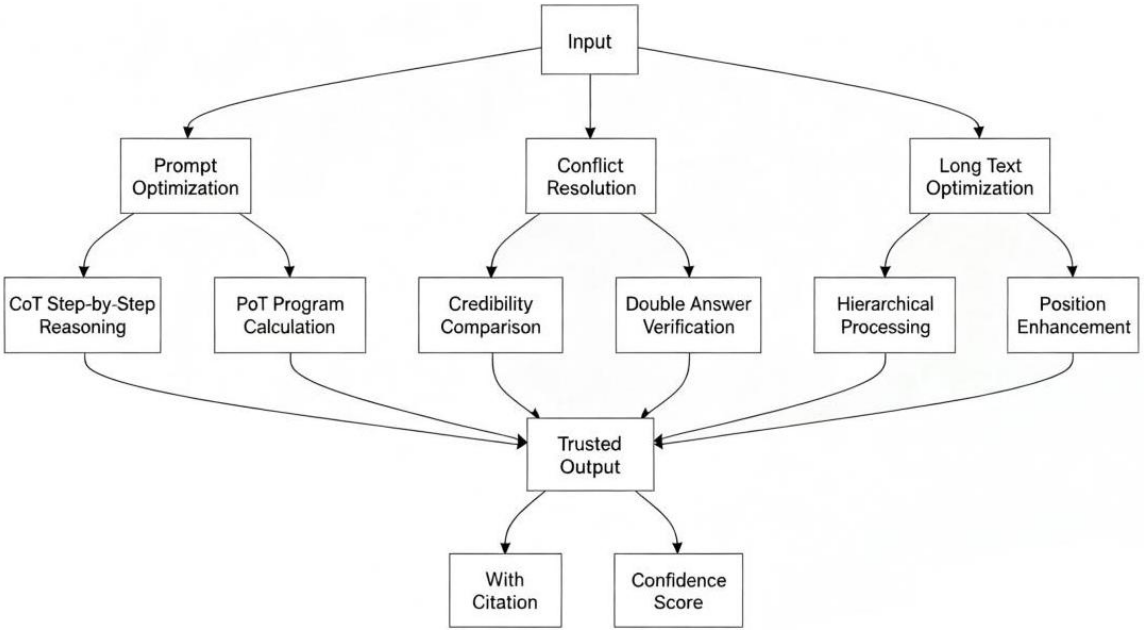

Figure 4. Model-layer mitigation methods [4].

3.3. Application-Layer Hallucination Mitigation Methods

Application-level optimizations provide practical solutions for deploying large models, primarily encompassing prompt engineering and fact-based guidance.

Prompt engineering has evolved rapidly from early simple instructions to current multi-stage interactive prompts. Chain-of-Thought (CoT) techniques show particularly notable progress, with the latest self-validating CoT automatically detecting logical contradictions during reasoning. Role-playing prompts are also becoming increasingly refined, controlling hallucination rates below 5% in specialized domains through precise identity setting and scenario constraints. While these techniques are user-friendly, they demand high professional expertise from prompt designers [9].

Fact-guidance techniques are advancing toward real-time and dynamic capabilities. The keyword attention mechanism developed by Li's team enables real-time monitoring of factual deviations during generation. For knowledge augmentation, Shuster et al.'s dynamic retrieval-generation framework achieves efficient integration of external knowledge [4]. In knowledge question-answering systems, Zhang et al. proposed a framework for user-knowledge base interaction, employing language models

to achieve automatic question-knowledge alignment, thereby associating stored information with user queries. They combined active detection, PAL (Prompt-Activated Language), and RARR (Refined Answer Retrieval and Refinement) to dynamically correct generated content [8]. The link strategy of the application layer can be seen in Figure 5.

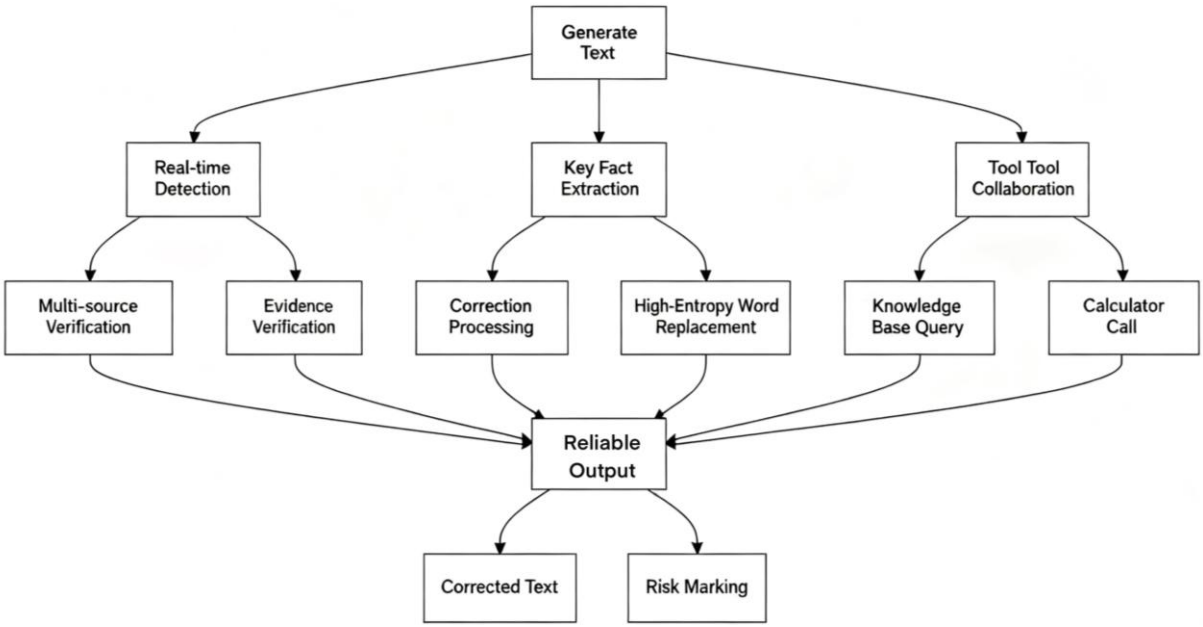

Figure 5. Application-Layer Mitigation Methods [8].

3.4. Chapter Summary

As shown in Table 2, various methods for alleviating hallucinations each have their own advantages: the data layer enhances data quality from the source, but has high construction costs; the model layer directly optimizes model performance but needs to consider a balance with computational efficiency; the application layer is flexible and easy to use, but depends on external interventions. Through multi-layer collaborative optimization, the reliability and practicality of large language models will be further improved.

Table 2. Comparison of Hallucination Mitigation Methods

Classification	Advantages	Limitations
Data Layer Mitigation Approach	Building faithful datasets to ensure data quality	High construction costs, lack of generalizability
Model-Layer Mitigation Approach	Flexible handling of complex scenarios, convenient iterative optimization	Challenges in ensuring timeliness and balancing generation performance with hallucination levels
Application-layer mitigation methods	High efficiency, easy to use	Relies on external knowledge, lacks precise alignment with downstream tasks

4. Conclusion

Future research on large model hallucinations will focus on balancing creativity and authenticity, developing controllable generation techniques for flexible regulation. Evaluation systems require greater refinement, achieving precise localization and automated verification through dependency analysis and knowledge graphs. Data and models must be co-optimized, adopting few-shot learning and causal reasoning modules to enhance efficiency and interpretability, while establishing continuous learning mechanisms to update knowledge. At the application level, lightweight fine-tuning platforms and visual interaction systems should be built to support domain-specific model development. Concurrently, industry standards and governance frameworks must be established to unleash innovation potential while ensuring safety. These directions will propel large models from

passively mitigating hallucinations to actively harnessing them, transforming them into intelligent partners that blend reliability and creativity.

References

[1] Ji ZW, et al. Survey of hallucination in natural language generation. ACM Computing Surveys, 2023, 55 (12): 248.

[2] Zhang S, Pan LM, Zhao JZ, Wang WY. The knowledge alignment problem: bridging human and external knowledge for large language models. arXiv: 2305.13669, 2024.

[3] Huang L, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 2025, 43 (2).

[4] Liu ZY, Wang PJ, Song XB, Zhang X, Jiang BB. A survey on hallucination in large language models. Journal of Software, 2025, 36 (3): 1152 – 1185.

[5] Zhang Y, Li YF, Cui LY, Cai D, Liu LM, Fu TC, Huang XT, Zhao EB, Zhang Y, Chen YL, Wang LY, Luu AT, Bi W, Shi F, Shi SM. Siren's song in the AI ocean: a survey on hallucination in large language models. arXiv: 2309.01219, 2023.

[6] Gardent C, Shimorina A, Narayan S, Perez-Beltrachini L. Creating training corpora for NLG micro-planning. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Vancouver: Association for Computational Linguistics, 2017: 179 – 188.

[7] Gabriel S, Celikyilmaz A, Jha R, Choi Y, Gao JF. GO FIGURE: a meta evaluation of factuality in summarization. In: Proceedings of the Association for Computational Linguistics: ACL - IJCNLP 2021. Association for Computational Linguistics, 2021: 478 – 487.

[8] Zhang W, Zhang J. Hallucination mitigation for retrieval-augmented large language models: a review. Mathematics, 2025, 13 (3): 856.

[9] Kumar K. Geotechnical Parrot Tales (GPT): harnessing large language models in geotechnical engineering. arXiv: 2304.02138, 2023.

[10] He J, Shen Y, Xie RF. Recognition and optimization of hallucination phenomena in large language models. Computer Applications, 2025, 45 (3): 709 – 714.