Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ29621

Detection and Mitigation of Factual Hallucinations in Large Language Models: A Comparative Review of the Timing and Effectiveness of External Retrieval, Post-hoc Verification, and Evaluation Methods

Bingchen Zhou

Institute of School of Artificial Intelligence and Advanced Computing, Xi'an Jiaotong-Liverpool University, Suzhou, China
Bingchen.Zhou22@student.xjtlu.edu.cn

Abstract. Large language models (LLMs) excel at reasoning and generation but remain susceptible to factual hallucinations. This survey categorizes recent progress around the timing of intervention: First, through retrieval augmentation and explicit reasoning during generation; second, through post-hoc verification and self-correction after generation. Regarding the former, this survey reviews IRCoT, SELF-RAG, ReAct, and Atlas, highlighting how combining retrieval with thought chaining can reduce multi-layer errors and improve answer grounding in knowledge-intensive question answering. Regarding the latter, this study examines RARR, Chain-of-Verification (CoVe), CRITIC, and Reflexion, which explore evidence, generate verification questions, or revise outputs through repeated criticism while minimizing stylistic bias. This work also summarizes detection and evaluation practices (e.g., SelfCheckGPT, FACTSCORE) that quantify attribution, support, and completeness. Across various approaches, this study analyzes effectiveness, computational cost, latency, and robustness to noisy retrieval, identifying failure modes such as retrieval dependency, ungrounded verification, and domain-shift gaps. This survey provides practical guidelines: for constrained question answering and multi-step reasoning, a combined retrieval-and-reasoning approach is most effective; whereas post-hoc validation is more suitable for long-form generation and reporting, and for citing external sources. Finally, this work proposes open problems: a unified cost-aware controller that can adjust intervention frequency; stronger evidence attribution; richer benchmarks beyond open-domain question answering; and a clearer trade-off between efficiency and realism for more reliable and scalable hallucination suppression.

Keywords: Large language models, Factual hallucinations, Retrieval enhancement, Verification and self-correction

DOI: 10.54254/2755-2721/2026.TJ29621

1. Introduction

In recent years, large language models such as Gemini and ChatGPT have achieved human-level performance in tasks such as question answering, dialogue, and programming. However, this is accompanied by the problem of “factual hallucination”: the model sometimes confidently outputs fictitious, outdated, or unverified facts. This hallucination not only undermines user trust, but also hinders the application of the model in high-risk fields such as medicine and law. The causes of hallucination include incomplete or biased pre-training corpus, probability diffusion in the decoding process, and lack of immediate access to external knowledge [1]. Comparison of retrieval enhancement (intervention during generation) and self-correction (intervention after generation) methods. Retrieval enhancement and chain reasoning methods (intervention during generation): This type of method embeds external retrieval or tool calls in the process of LLM generating answers to provide real-time factual evidence. For example, a retrieval-enhanced model can query the knowledge base or the Internet multiple times during generation to obtain evidence related to the current reasoning step. The advantage of this method is that it can significantly reduce the model's tendency to make up for random facts and improve the factual accuracy and freshness of the answers [2]. Especially in knowledge-intensive tasks, the introduction of retrieval can make up for the non-existent or outdated knowledge in the model parameters. Applicable scenarios include open-domain question answering, multi-hop reasoning, and other tasks that require the latest real-world information [3]. However, its limitation lies in its high dependence on the retrieval module and the quality of the retrieval content: if the retrieval results are irrelevant or even incorrect, the model may still produce hallucinations or be misled. In addition, frequent retrieval increases the inference overhead and may reduce the generation efficiency. How to balance the retrieval frequency and generation speed becomes a challenge. Therefore, when time or computing resources are limited, or the model has mastered sufficient domain knowledge, excessive retrieval may be counterproductive. Post-verification and self-correction methods (post-generation intervention): This type of solution introduces additional verification and correction steps after the model produces a preliminary answer. For example, after the model generates an answer, it can extract the verifiable assertions in it, retrieve evidence through a search engine, and let the model correct the original answer accordingly. The advantage of this method is that it is less invasive to the original model and can be used as a post-processing module to improve the credibility of the answer. It allows the model to self-reflect or use tools to discover and correct errors, thereby strengthening factual consistency without interfering with the initial creation process. Especially in scenarios that require high accuracy (such as legal, medical Q&A, or long summaries), post-verification can help filter and correct hallucinations. Its limitation is that it increases the reasoning delay and complexity: it requires additional query and model inference steps, and its real-time performance is poor. If there are too many errors in the initial answer, the post-verification may be unable to recover, or excessive modifications may cause the content to deviate from the original meaning. In addition, such methods sometimes rely on the model's own critical ability, and small models may find it difficult to reliably identify and correct their own errors [1]. Therefore, this method is suitable for scenarios with extremely high accuracy requirements and tolerance for a certain amount of delay, but is not suitable for strong real-time tasks such as instant conversations. This review focuses on the hallucination reduction technologies that have emerged in the past three years. The purpose is to sort out their core ideas, experimental results and limitations, compare the advantages and disadvantages of various methods, and provide a systematic reference for subsequent research.

DOI: 10.54254/2755-2721/2026.TJ29621

2. Concepts and data sets

Large language models sometimes confidently output fabricated, outdated, or unverified facts. Factual hallucinations refer to generated content that doesn't correspond to objective facts. These include fabrication hallucinations, where the model creates nonexistent people, dates, or events, such as misquoting years or fabricating awards; detail distortion hallucinations, where the model exaggerates or distorts real facts, such as changing a person's birthplace from London to New York; and overgeneralization or overconfidence, where the model asserts facts without sufficient supporting evidence, such as misinterpreting unknown information as either false or true.

Common datasets include open-domain, knowledge-intensive question-answering datasets such as WikiBio, HotpotQA, and 2WikiMultihopQA, which are used to evaluate the factuality of chain-based reasoning models. TruthfulQA and FACTOR are used to assess the truthfulness of short answers or long paragraphs and contain numerous "trap questions" that are prone to error. Tasks such as Wikidata lists and MultiSpanQA are used to evaluate the precision of list-based questions. When constructing these datasets, they are usually labeled with three categories: “true”, “false” or “unknown”. However, some evaluations only count “true/false” and ignore “unknown”, which may underestimate the “unverifiable/overgeneralized” type of hallucinations [1].

3. Retrieval enhancement and chain reasoning

3.1. IRCoT

Interleaving Retrieval with Chain-of-Thought (IRCoT): IRCoT alternates the retrieval process with chain-of-thought steps. The model retrieves relevant information after each intermediate reasoning step, thereby gradually narrowing the knowledge gap [2]. This alternating strategy provides a basis for each step of the model's reasoning, greatly improving the accuracy of multi-hop question answering and significantly reducing the phenomenon of hallucinations, making each conclusion in the reasoning chain more verifiable [2]. Studies have shown that compared with one-time retrieval, IRCoT reduces factual errors by about 40–50% in complex QA tasks, demonstrating the effectiveness of step-by-step retrieval in reducing knowledge hallucinations.

3.2. SELF RAG

Self-Reflective Retrieval-Augmented Generation (Self-RAG): Self-RAG is a self-reflective retrieval-augmented generation framework. It trains the model to adaptively decide when retrieval is needed and inserts special “reflection” markers during the generation process to self-evaluate [3]. Specifically, the model can retrieve multiple times as needed (or skip), and after obtaining a document, it first determines whether the material is relevant, and then continues to generate based on the retrieval results, while critically reflecting on the generated content [3]. Experiments have shown that Self-RAG significantly improves the authenticity and citation accuracy of the model output, and outperforms strong baseline models such as ChatGPT in tasks such as open-domain question answering, reasoning, and fact verification [3]. This shows that by allowing the model to learn when to access external knowledge and how to review its own answers, the proportion of hallucinations in long-form generation can be effectively reduced.

DOI: 10.54254/2755-2721/2026.TJ29621

3.3. ReAct

In the paradigm of "intervention during generation", ReAct alternates chained thinking with external action/retrieval, so that each step of reasoning is constrained and corrected by new evidence, thus having a stronger inhibitory effect on "knowledge gap type" and "process" illusions. Compared with pure CoT or action alone, ReAct demonstrates higher authenticity and robustness in knowledge-intensive and interactive tasks, and complements self-consistent CoT (CoT-SC): when pure reasoning is inconsistent or evidence is insufficient, the retrieval-observation loop reduces the propagation of assumptions, and when retrieval fails or the environmental noise is high, falling back to CoT-SC can maintain output quality. The cost is the introduction of additional calls and tokens step by step, and the overhead increases approximately linearly with the number of reasoning steps. Therefore, the "authenticity gain/increased overhead" depends on whether the task is multi-hop, whether the available knowledge is externalized, and the latency budget [4].

3.4. Atlas

Atlas is a retrieval-enhanced language model proposed by Meta. When generating answers, it retrieves relevant fragments from a large external text corpus and feeds them to the model along with prompts. Research has shown that retrieval-enhanced architectures like Atlas can reduce the tendency to fabricate false information in conversations and question-answering [5]. By outsourcing knowledge updates to retrieval indexes, Atlas demonstrates strong real-time learning capabilities, achieving excellent results in few-shot learning while significantly reducing hallucinations in conversational agents [5]. This confirms that the introduction of a retrieval module can improve the factual consistency of LLM outputs.

4. Post-verification and self-correction

4.1. RARR

RARR is a "research and revise" framework that aims to enhance the verifiability of answers after the fact [6]. Specifically, for the key information in the model's initial answer, RARR first automatically generates queries to search for relevant evidence (the research phase), and then uses the retrieved information to conduct targeted editing of the original answer (the revision phase) to introduce reliable sources and delete false content [6]. This method does not require retraining the original model, but instead polishes the answer through post-output editing so that each sentence has a source to support it. Experiments show that compared with answers without post-processing, the answers revised by RARR significantly improve the supportability of the content while maintaining the original meaning, and effectively reduce the proportion of unfounded statements [6]. This shows that the process of using search engines to verify and having LLMs revise their own answers can reduce hallucinations without sacrificing fluency.

4.2. Reflexion

Reflexion belongs to the verification/self-correction paradigm of "post-generation intervention": after a failed or uncertain answer, the model converts the environment or evaluation signal into an actionable natural language reflection and writes it into "episodic memory" to guide the planning and rewriting of subsequent rounds; its framework is composed of the executor (Actor) - evaluator (Evaluator) - self-reflection model (Self-Reflection), which can continuously reduce the process and

DOI: 10.54254/2755-2721/2026.TJ29621

knowledge illusions in multiple rounds of trial and error without fine-tuning the weights, and achieve significant improvements in tasks such as decision-making, reasoning and code generation. Compared with “in-generation” retrieval + chain reasoning (such as IRCoT/SELF-RAG) that uses “step-by-step citation” to pre-constrain, Reflexion uses one or a few a posteriori reflections and rewriting to achieve low-coupling correction, which is suitable for scenarios where evidence is difficult to obtain online or where delay/budget is more sensitive; the two complement each other in terms of effectiveness timing and authenticity-cost, and can be combined and selected according to task multi-hop, retrievability and delay budget [7].

4.3. CoVe

Chain of Verification (CoVe) proposes a four-step process of "generation, questioning, independent answering, and revision": the model first generates a draft, then plans verification questions, answers these questions independently to avoid mutual influence, and finally generates verified output based on the answers. On the Wikidata and Wiki Category list tasks, CoVe improved the precision of Llama 65B from 0.17 to 0.36 and significantly reduced hallucination answers (negative samples were reduced from 2.95 to 0.68) [8]; in the MultiSpanQA closed-book question answering, F1 increased from 0.39 to 0.48; in the long text generation task, CoVe's FACTSCORE increased from 55.9 to 71.4, an increase of 28%. At the same time, factoring the verification step (factored CoVe) and adding a "revision step" further improved performance.

4.4. CRITIC

Self-Correcting with Tool Interactive Critiquing (CRITIC) allows LLM to use external tools such as search engines and code interpreters to check and modify outputs like humans. The process includes generating an initial answer, calling external tools for fact checking or code running, forming critical feedback, and then generating a corrected answer based on the feedback [9]. The authors tested ChatGPT, text davinci 003, and LLaMA 2 series models in three types of tasks: open question answering, mathematical program generation, and toxic content reduction. The results show that CRITIC improved F1 by 7.7 points in question answering tasks, an absolute improvement of 7.0% in mathematical reasoning tasks, and a reduction of toxic content probability by 79.2% [9]. Relying on model self-correction (without access to tools) has limited effect and may even worsen. This method emphasizes the importance of external feedback for model self-improvement.

5. Hallucination detection and evaluation metrics

5.1. SelfCheckGPT

SelfCheckGPT is a zero-resource, black-box detection method: it randomly samples the output of LLM and compares the consistency of multiple generated results. If a sentence is inconsistent between different samples, it is judged as hallucination. The method proposes five consistency metrics: BERTScore, automatic question answering, n-gram language model, first-order natural language inference (NLI), and self-assessment using LLM. Experiments show that SelfCheckGPT achieves an AUC PR of 83.21 for detecting non-factual sentences on the WikiBio dataset, which is significantly higher than the random baseline of 27.04; in the more difficult "non-factual" setting, the AUC PR is 38.89. Compared with the gray-box method that relies solely on the probability of the model itself, SelfCheckGPT has higher indicators and does not require access to probability

DOI: 10.54254/2755-2721/2026.TJ29621

distributions or external databases. This method is applicable to black-box LLM and can be extended to various generation tasks [10].

5.2. FACTSCORE

FACTSCORE uses "atomic facts" as units to calculate the proportion of facts supported by a specified knowledge source, replacing coarse-grained binary classification evaluations and used to granularly characterize the accuracy of long-form text generation.

InstructGPT/ChatGPT/PerplexityAI scored only 42.5%/58.3%/71.5% in human evaluations of biographical data, revealing that significant factual errors still exist even with retrieval. The paper also proposed an automatic estimator of "retrieval + strong model", with an error of less than 2% for human evaluations, which can be used to compare the authenticity improvement and cost of "retrieval during generation + chained reasoning" and "verification/self-correction after generation" on a large scale [11].

6. Challenges and prospects

6.1. Limitations

6.1.1. Limitations of retrieval enhancement and chain reasoning methods

Such methods are highly dependent on the quality of retrieval: irrelevant or low-quality evidence can mislead the model into generating incorrect content. The model may not strictly follow the retrieved evidence and may ignore or even contradict it. Errors in the reasoning chain may also propagate and lead to incorrect final answers. In addition, although the interweaving of multiple rounds of retrieval and reasoning reduces hallucinations and improves the reliability of answers, it also increases computational overhead and response delays, and the evaluation and interpretation of output credibility remain difficult.

6.1.2. Limitations of post-verification and self-correction methods

In the absence of external feedback, model self-verification is often ineffective [12]; the introduction of tools such as search can provide factual evidence [9], but when the evidence is unreliable, it can introduce misleading information. Such methods can often correct obvious factual errors, but it is difficult to simultaneously modify the part of the reasoning chain that relies on the error, which can easily lead to contradictory answers. At the same time, increasing the verification iteration step means higher computational cost and response delay. There is a lack of unified standards in terms of evaluation and interpretation, and while reducing hallucinations, the task accuracy may be slightly reduced.

6.2. Outlook

6.2.1. Multitasking and versatility

Currently, many anti-hallucination methods have only been validated on a few tasks (such as knowledge question answering and summarization). Some require fine-tuning for specific domains, and their cross-task generalizability is unclear. Future research is needed to investigate the transferability of these techniques across different tasks and domains, as well as how to design more

DOI: 10.54254/2755-2721/2026.TJ29621

general anti-hallucination strategies to ensure that models maintain high confidence in multi-task scenarios.

6.2.2. Insufficient evaluation benchmarks and metrics

A unified benchmark for hallucinations remains lacking. While benchmarks such as TruthfulQA and HaluEval exist for evaluating the authenticity of model generation, their coverage of domains and tasks remains limited, and the definition of "hallucination" is subjective. Future work should develop more comprehensive evaluation frameworks and datasets, encompassing diverse scenarios such as open question answering, dialogue, and multi-turn reasoning. This should also establish fine-grained hallucination classification and quantitative metrics to objectively compare the effectiveness of different methods.

6.2.3. Unclear trade-off between efficiency and performance

Many anti-hallucination methods increase accuracy at the expense of increasing inference steps or model complexity, and the efficiency-performance trade-off between the two is still unclear. For example, while retrieval enhancement and multiple rounds of self-correction improve answer reliability, they incur additional computational and time overhead. Whether this is worthwhile in actual systems needs to be weighed against application requirements. Future research should focus more on optimizing the efficiency of methods, such as intelligently determining when to invoke retrieval/verification to achieve maximum hallucination mitigation at minimal cost, and exploring the quantitative relationship between model size, computing power budget, and hallucination rate.

7. Conclusion

To address these issues and challenges, this review provides a new, structured perspective. Based on a comprehensive literature review, this paper constructs a comparative framework for the first time, focusing on intervention timing (during and after generation) and efficiency impact, systematically analyzing the performance of different anti-hallucination methods in a horizontal and vertical manner. This approach, based on intervention phases, helps readers clearly understand when to employ which strategies and the trade-offs between response speed and accuracy. Through this structured comparison, this paper reveals differences in the effectiveness and costs of existing methods, providing insights for unified evaluation. Furthermore, this review highlights many aspects not systematically addressed in previous work (such as multi-task generalization and evaluation criteria), and proposes future research directions, providing valuable references for subsequent scholars. In summary, this review provides a comprehensive overview of the development context and key insights in the field of large language model anti-hallucination, establishing an analytical framework that focuses on the trade-off between effectiveness timing and efficiency. This innovative perspective will help promote more efficient and reliable research on LLM hallucination mitigation.

References

[1] Li, J., Chen, J., Ren, R., Cheng, X., Zhao, W. X., Nie, J.-Y., & Wen, J.-R. (2024). The Dawn After the Dark: An Empirical Study on Factuality Hallucination in Large Language Models. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Bangkok, Thailand, 10879–10899.

[2] Trivedi, H., Balasubramanian, N., Khot, T., & Sabharwal, A. (2023). Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), Toronto, Canada, 10014–10037.

DOI: 10.54254/2755-2721/2026.TJ29621

[3] Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). SELF-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. arXiv: 2310.11511.
[4] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. R., & Cao, Y. (2023). ReAct: Synergizing Reasoning and Acting in Language Models. International Conference on Learning Representations (ICLR), Kigali, Rwanda.
[5] Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Dwivedi-Yu, J., Joulin, A., Riedel, S., & Grave, E. (2023). Atlas: Few-shot Learning with Retrieval-Augmented Language Models. Journal of Machine Learning Research, 24(251), 1–43.
[6] Gao, L., Dai, Z., Pasupat, P., Chen, A., Chaganty, A. T., Fan, Y., Zhao, V. Y., Lao, N., Lee, H., Juan, D.-C., & Guu, K. (2023). RARR: Researching and Revising What Language Models Say, Using Language Models. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), Toronto, Canada, 16477–16508.
[7] Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language Agents with Verbal Reinforcement Learning. Advances in Neural Information Processing Systems 36 (NeurIPS), New Orleans, LA, USA.
[8] Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., & Weston, J. (2024). Chain-of-Verification Reduces Hallucination in Large Language Models. Findings of the Association for Computational Linguistics (ACL), Bangkok, Thailand, 3563–3578.
[9] Gou, Z., Shao, Z., Gong, Y., Shen, Y., Yang, Y., Duan, N., & Chen, W. (2024). CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing. International Conference on Learning Representations (ICLR).
[10] Manakul, P., Liusie, A., & Gales, M. J. F. (2023). SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Singapore, 9004–9017.
[11] Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-T., Koh, P. W., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023). FActScore: Fine-Grained Atomic Evaluation of Factual Precision in Long-Form Text Generation. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Singapore, 12076–12100.
[12] Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., & Zhou, D. (2024). Large Language Models Cannot Self-Correct Reasoning Yet. International Conference on Learning Representations (ICLR), Vienna, Austria.