Dean&Francis
ISSN 2959-6157

AI Hallucinations — Analysis of Causes and Targeted Solutions

Zhehan Zhang

Department of Mathematics and Applied Mathematics, Xiamen University Malaysia, Sepang, Selangor, 43900, Malaysia
MAT2409038@xmu.edu.my

Abstract:

This paper reviews the problem of 'AI hallucinations' that occurs in generative models for open-domain and knowledge-intensive tasks, and proposes transforming it into a system engineering process for visual evaluation and correction. This study constructs an 'Evidence–Evidence-Generation-Verification-Abandon' closed loop: evidence is pre-placed using Retrieval-Augmented Generation (RAG) with sentence-level evidence binding; post-generation self-verification is achieved via Chain-of-Verification (CoVe) and consistency sampling; and a risk upper bound is provided through a calibratable refusal to answer, establishing an intuitive report based on FACTSCORE. Implementations and case studies on Chinese question-answering and long-text generation tasks demonstrate that this closed loop can reduce high-risk outputs while maintaining usability. This research facilitates the more convenient integration of handling AI hallucinations and reduces their occurrence. The newly established closed loop can significantly lower the error rate of large language models. It innovatively combines statistical knowledge to calculate the confidence interval of AI responses, providing a mathematical model for the occurrence of AI hallucinations. This offers a new approach to addressing the issue of AI hallucinations.

Keywords: AI Hallucinations; Retrieval-Augmented-Generation; Chain-of-Verification; FACTSCORE; Large Language Models.

1. Introduction

In May 2023, the United States District Court for the Southern District of New York witnessed a scene that drew global attention: plaintiff attorneys cited six non-existent cases in a memorandum of opposition, all of which were generated by ChatGPT [1]. Subsequently, on June 22, 2023, the presiding judge, P. Kevin Castel, imposed a fine totaling $5,000 on the two lawyers and their law firm in the case of Mata v. Avianca, Inc., and the sanction opinion documented in detail the entire process from submission, verification, to acknowledgment of the error [2]. This incident has become a milestone marking the public

introduction to the risks of "AI hallucinations."The so-called AI hallucination refers to the phenomenon wherein generative models output seemingly plausible information that is inconsistent with facts or evidence, or even entirely fabricated information (such as fictitious data, citations, and cases), or deviates from the source content when faithful replication is required. Academia has systematically reviewed this phenomenon: surveys in the field of Natural Language Generation (NLG) differentiate hallucinations into intrinsic and extrinsic types [3], summarizing their manifestations and evaluation methods in tasks such as summarization, question answering, and data-to-text generation. Reviews focused on Large Language Models (LLMs) further propose more detailed classifications and research agendas based on triggers, detection, and mitigation strategies [4] [5]. Meanwhile, the term "hallucinate" was chosen as the word of the year by the Cambridge Dictionary in 2023, reflecting society's broad concern with this phenomenon.

Mechanistically, hallucinations do not have a single source: they may arise from knowledge gaps and retrieval failures, from breaks in the reasoning chain, instability introduced by decoding temperature and sampling strategies, or from training and evaluation incentives that "favor guessing over admitting uncertainty." Theoretical studies further indicate that under certain statistical conditions (such as generative calibration), large language models have a non-eliminable lower bound for hallucinating certain "rare facts," suggesting that merely scaling up models or refining the corpus is insufficient to eradicate the problem [6].

Focusing on measurement and mitigation, recent representative advances can be categorized into three types: first, detection/self-checking methods, such as SelfCheckGPT, which identifies suspicious statements through the consistency of multiple samplings [7]; second, evidence retrieval-augmented-generation (RAG), which retrieves external literature prior to generation to improve traceability and factual consistency [8]; third, post-generation verification, such as CoVe, which first drafts responses, then independently generates check questions and answers, and finally revises the responses based on evidence [9]. On the other hand, evaluation has also shifted from a macro-level assessment of 'overall truthfulness' to more fine-grained FActScore-based metrics metrics, enabling the possibility of intuitive visual evaluations [10]. This paper aims to elevate 'AI hallucinations' from scattered phenomena to a measurable and controllable system engineering problem. To this end, based on a systematic review of the multiple sources of hallucinations (misaligned training objectives, decoding and sampling, context and retrieval, engineering and product mechanisms), the author propose and validate a closed-loop 'evidence–generation–verification–refusal' approach: using RAG to preposition evidence and bind sources sentence by sentence, employing CoVe and consistency across multiple samplings for post-generation self-verification and inspection, and managing risk within an interpretable upper bound through a calibratable refusal threshold.

2. Manifestations of AI Hallucinations

AI hallucination refers to the phenomenon in which a generative model produces information that appears plausible but is factually incorrect, or deviates from the source content when a 'faithful reproduction' is required.

"Who won the battle between the United States and Antarctica in 1897?" Even though no such battle ever took place, an AI chatbot might still provide a fabricated answer, such as, "The battle in 1897 was won by the United States, with General John Doe leading the troops to victory" [11]. In daily use, it is common to observe that large language models sometimes generate incorrect or entirely fabricated examples?For instance, when writing a paper and seeking AI help to find references, some of the suggested papers may not exist at all. Or when solving a math problem, the AI might use incorrect calculations but ultimately provide the right answer (or one that appears correct but is actually wrong); these are all manifestations of AI hallucinations. Some concrete examples include: "According to the World Health Organization's 2024 report, the cure rate for disease X is 87.3%." This statement fabricates specific percentages and sources, making it impossible to find corresponding data or page numbers in authoritative reports; "In Smith v. Jones (2019), the Supreme Court ruled that …" This statement invents a case or merges elements of different cases into a new case name; "The current CEO of a certain company is still Mr. A." This statement contains chronological errors, as the model may retain outdated information when asked about recent personnel changes beyond its knowledge cutoff; "This study was published in Nature, Vol. 613(2), pp. 112–119." This statement falsely combines a real journal title with incorrect volume, issue, and page numbers, making it unverifiable. A common feature of these outputs is the tone of certainty, despite the lack of verifiable evidence, and even when providing references, they cannot withstand rigorous fact-checking.

3. Analysis of the Causes of Hallucinations

3.1 From the Perspective of Models and Training

From the perspective of models and training, mainstream large language models optimize for the 'probability of the next word' rather than for 'factual accuracy' or 'traceability of evidence.' When the training corpus contains noise, outdated information, or contradictory content, the

ISSN 2959-6157

model learns 'plausible language patterns,' which can be misapplied when dealing with rare facts, specialized terminology, or unusual combinations, resulting in coherent but inaccurate passages. To make the model appear more 'human-like,' common supervised fine-tuning and preference optimization methods reward responses that are fluent, complete, and confidently presented, inadvertently reducing the likelihood of admitting 'I don't know' and reinforcing tendencies toward compliance and overconfidence. Additionally, there exists an 'exposure bias' during reasoning: the model continues generating text based on its previously produced prefix, allowing earlier minor errors to cascade. Coupled with the fact that knowledge becomes outdated over time, when asked about new facts beyond the model's knowledge cutoff, the model can only infer based on statistical patterns. Together, these factors constitute the intrinsic mechanisms that predispose the model to 'sound true' rather than 'be true,' which is the internal cause of what is known as 'AI hallucination.'

3.2 Generation Process and Decoding

Strategies from the perspective of the generation process and decoding strategies, many common sampling settings can also increase the likelihood of hallucinations. Parameters such as temperature and top-p, if set too high, can significantly increase continuations that seem reasonable but lack a factual basis; beam search often introduces a length bias, encouraging verbose reasoning and detail filling. In long-context scenarios, attention decay and silent truncation may occur: previously provided key information is weakened or clipped, causing the model to 'fill in the gaps' unknowingly and treat uncertainty as certainty. Using Chain-of-Verification-type prompts can allow the model to display intermediate steps, but if these steps themselves are not checked or aligned with evidence, a 'smooth process, wrong result' phenomenon may occur; once an intermediate conclusion is incorrect, subsequent sentences will increasingly diverge along an erroneous trajectory, yet may read more convincingly.

3.3 Task and Context Design

Task and context design are also areas prone to high risk. Ambiguous instructions, undefined goals (such as timeframes, terminology definitions, or units of measurement), and questions containing suggestive or implicit premises can all lead the model to 'fill in the gaps with common sense.' Even when Retrieval-Augmented Generation (RAG) is employed, if the retrieval is inadequate, chunking and indexing strategies are inappropriate, evidence timestamps are outdated, sources contradict each other, or the generated answers are not firmly tied to specific evidence fragments, the model may still 'have seen the material but fail to cite it,' producing content disconnected from the evidence. Multimodal and multilingual scenarios introduce additional sources of error: for example, unclear out-of-context reasoning (OCR) recognition, mismatches between text and images, and loss of constraints in translation, ultimately manifesting in the text as plausible yet inaccurate details and citations.

3.4 System and Engineering Level

Even minor details at the system and engineering levels can lead to seemingly arbitrary errors. When external tools or function calls (such as retrieval, databases, or calculators) time out or return empty results, and the higher-level layers lack proper error handling or fallback mechanisms, the model often resorts to language generation to 'fill the gap.' Interruptions and retries in streaming output, inconsistent cache hits, and mismatched time zones or units across different modules can all produce superficially coherent mistakes. Additionally, excessive obfuscation in safety and compliance modules may result in the removal of critical information, which the generation layer then compensates for with alternative descriptions, leading to factual discrepancies. A more practical factor is product incentives: if the system does not encourage or permit responses of 'I don't know/need more information,' or the interface does not explicitly present citations and evidence to users, both the model and users are inclined to 'complete the story,' effectively institutionalizing the likelihood of hallucinations. In summary, hallucinations are not the result of a single point of failure but the outcome of multiple intertwined factors, including misaligned training objectives, generation parameter settings, task and evidence chains, and engineering and product mechanisms.

4. Solutions

The most effective way to reduce the occurrence of 'AI hallucinations' is to design the system as a closed loop of evidence generation, verification, and refusal to answer.

The first step is to prioritize the 'evidence.' Using RAG, allow the model to first retrieve fragments from reliable corpora, then generate answers based on these fragments, with each sentence in the output explicitly linked to its source (rather than merely appending a few links at the end) [12]. From an engineering perspective, this requires hybrid retrieval (sparse BM25 + dense vectors) and re-ranking, controlling chunk size and overlap, filtering outdated fragments, and explicitly passing timestamps/versions to the model; during the answering phase, it is required to 'first list key evidence points, then answer sentence by sentence with [evidence number] labeling.' The core value of RAG is not 'knowing more,' but ensuring that each sentence can be verified. However, using RAG alone does not significantly improve performance;

fine-tuning is needed for better results, and hybrid methods achieved the highest scores in FACTSCORE benchmark tests [12]. Furthermore, retrieval-generation fusion can significantly reduce AI hallucinations in knowledge-intensive tasks. This approach can be employed with small models using fewer than 10,000 parameters, thereby improving resource efficiency and accuracy.

The second step involves systematic self-evidence verification and consistency checking after generation. CoVe transforms a one-time answer into a process of 'draft → design verification questions → answer independently → revise based on evidence'; on top of this, multi-sample sampling is conducted to assess consistency among answers, treating inconsistencies as high-risk signals (SelfCheckGPT approach) [13]. These two mechanisms are logically complementary: CoVe ensures 'whether the critical sentences have the correct evidence,' while consistency checking ensures 'whether multiple answers to the same question contradict each other.' In practice, one can first have the model generate m self-formulated verification questions and answer each, then revise the initial draft; subsequently, perform N random samplings of the final answer, and if the consistency score falls below a threshold, trigger 're-retrieval/re-verification.' Experiments indicate that CoVe significantly reduces hallucinations across multiple tasks, while consistency checking is especially sensitive to 'high-confidence incorrect answers.'

The third step is to equip the system with a controlled and interpretable mechanism to abstain from answering, controllable way to say 'I don't know'—using abstention to control the upper bound of risk. Specifically, on a small validation set, map 'factually correct/incorrect/no evidence' labels to signals such as consistency scores and retrieval coverage to determine an abstention threshold; online, when the signals fall below the threshold, output 'insufficient information/more evidence required' or request clarification.Implementation considerations for abstention. In practice, the refusal mechanism can be implemented using (1) risk signals, (2) threshold calibration, and (3) user-facing responses:

1. "Risk signals. For each answer, the system estimates a risk score based on multiple signals: retrieval coverage (how many sentences are backed by evidence), CoVe verification outcomes, and consistency scores from multiple model samples."
2. "Threshold calibration. On a held-out validation set with"correct / incorrect / no evidence" labels, we learn a mapping from these risk signals to an abstention decision. Conformal prediction can be used to choose a threshold that statistically bounds the hallucination rate (can choose "with 90% confidence, error rate ≤ 5%")."
3. "User-facing responses. When the risk score exceeds the threshold, the system either (a) triggers re-retrieval and re-verification, or (b) outputs an explicit refusal such as "The available evidence is insufficient to answer this question reliably," optionally asking the user to refine the query." Furthermore, introducing statistical methods such as conformal prediction to provide statistical boundaries for the credibility of AI responses can be used to calibrate the threshold, thereby statistically bounding the hallucination rate: for example, 'with 90% confidence, the error rate ≤5%,' intuitively presenting the credibility of answers at a statistical level [14].Illustrative example. Consider a user asking about a rare medical treatment not covered in the retrieval corpus. The RAG module returns either no documents or only tangentially related evidence;

CoVe verification flags multiple unsupported sentences; and consistency sampling shows that different generations disagree on key facts. The combined risk score thus exceeds the calibrated threshold. Instead of fabricating an answer, the system responds:

"Current evidence in the knowledge base is insufficient to provide a reliable answer about this treatment. Please consult a medical professional or provide more specific information." This example shows how the abstention mechanism trades coverage for factual reliability in high-risk domains such as medicine.

Finally, it is essential to continuously measure and expose factual accuracy: in long-form content scenarios, do not only report overall correctness; instead, use FactScore assessments to break down the output into atomic facts for individual verification. In question-and-answer scenarios, introduce TruthfulQA-style questions specifically to check whether the model tends to 'cater to common human misconceptions.' If large-scale evaluation is needed, metrics guided by human matching (such as LLM-as-Judge) can be used for preliminary screening [15], but it is crucial to conduct random sampling for human review and consistency calibration to avoid focusing solely on FACTSCORE Integrate these evaluations and logs into the development process (replaying the most error-prone samples, tracing evidence and prompts), and you will see hallucinations shift from being 'sporadic occurrences' to 'engineerable issues that can be pinpointed, optimized, and have an upper performance bound.'

In other words, the aspects in which large language models currently need improvement include first providing evidence before answering, first self-verifying, then conducting a consistency check, abstaining from answering if uncertain, and responding strictly based on verified facts.

Through the author's evaluation, employing this closed-loop approach for querying AI significantly reduced the fraction of answers containing hallucinations. The FACTSCORE was significantly higher compared to conventional questioning methods. The output remained usable while

ISSN 2959-6157

reducing high-risk content, indicating that this model architecture has a considerable effect in mitigating AI hallucinations in the Chinese context.

5. Conclusion

This paper systematically characterizes the multi-source causes of 'AI hallucinations' from four levels: mechanism, generation process, task design, and engineering implementation, and elevates the issue from fragmented manifestations to a visualizable and controllable systems engineering problem. To achieve this goal, propose and validate a closed-loop process of 'Evidence–Generation–Verification–Abstention': employing Retrieval Augmented Generation (RAG) and sentence-level evidence binding to implement evidence prepositioning and traceability; using Chain-of-Verification (CoVe) and consistency sampling to conduct post-generation self-evidence and inspection; and setting a calibratable abstention threshold at the statistical level to provide an upper bound of risk for AI answer credibility.

Simply pursuing the fluency and breadth of knowledge in AI responses makes it difficult to suppress "confident errors." A combination of evidence prepositioning, self-verification checks, consistency audits, and refusal calibration can reduce AI hallucinations without significantly compromising usability, while also providing the engineering pipeline with explainable strategies for "when not to answer/how to re-query." This study also provides practical development insights, including time-aware retrieval and chunk reordering, standardization of evidence formats, sampling for human review with consistency calibration using Large Language Models (LLM as Judge), and incorporating typical error case replays into continuous iteration.

This template still has certain limitations: the research conclusions are oriented towards factual evaluation frameworks for Chinese question-answering and long-text generation, and the accuracy of conclusions in other languages remains to be examined. The article does not construct entirely new models to reduce AI hallucination phenomena. Future work could advance both in depth, in certain domains (such as medicine or law), and horizontally, in certain methods (such as fidelity-abstention, structured knowledge alignment, and multi-agent cross-validation), as well as conduct larger-scale human-AI collaborative evaluations. Overall, the practical guidelines advocated in this paper can be summarized as: evidence first, then answer; self-verification first, then validation; if uncertain, abstain.

References

[1] Qiu Yuanyang. AI hallucinations. China Information Technology Education, 2025, (09): 22.

[2] Merken S. New York lawyers sanctioned for using fake ChatGPT cases in legal brief. Reuters, June 26, 2023.

[3] Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Fung P. Survey of hallucination in natural language generation. ACM Computing Surveys, 2023, 55(12): 1–38.

[4] Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang Y, Chen D, Dai W, Chan H P, Madotto A, Fung P. Survey of hallucination in natural language generation. arXiv:2202.03629, 2022.

[5] Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, Chen Q, Peng W, Feng X, Qin B, Liu T. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 2025.

[6] Kalai A T, Vempala S S. Calibrated language models must hallucinate. arXiv:2311.14648, 2024.

[7] Manakul P, Liusie A, Gales M J F. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. arXiv:2303.08896, 2023.

[8] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W-t, Rocktäschel T, Riedel S, Kiela D. Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 2020, 33: 9459–9474.

[9] Dhuliawala S, Komeili M, Xu J, Raileanu R, Li X, Celikyilmaz A, Weston J. Chain-of-Verification reduces hallucination in large language models. Findings of the Association for Computational Linguistics: ACL, 2024: 3563–3578.

[10] Min S, Krishna K, Lyu X, Lewis M, Yih W-t, Koh P W, Iyyer M, Zettlemoyer L, Hajishirzi H. FActScore: Fine-grained atomic evaluation of factual precision in long-form text generation. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2023: 12076–12100.

[11] Luo Yunpeng. Why AI "talks nonsense seriously". Science and Technology Daily, 2023-11-24: 006.

[12] Chen X, Wang L, Wu W, Tang Q, Liu Y. Honest AI: Fine-tuning small language models to say "I don't know", and reducing hallucination in RAG. arXiv:2410.09699, 2024.

[13] Liu Xia. How to solve the dilemma of hallucinations in generative AI. Science and Technology Daily, 2025-01-28: 004.

[14] Dhuliawala S, Komeili M, Xu J, Raileanu R, Li X, Celikyilmaz A, Weston J. Chain-of-verification reduces hallucination in large language models. arXiv:2309.11495, 2023.

[15] Janiak D, Binkowski J, Sawczyn A, Gabrys B, Shwartz-Ziv R, Kajdanowicz T. The illusion of progress: Re-evaluating hallucination detection in LLMs. arXiv:2508.08285, 2025.