© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1722609

Self-Reflective Large Language Models for Reducing AI Hallucinations: A Novel Framework for Reliable Generative AI

PRITI SHARMA[1], SACHIN SHARMA[2]
[1, 2] *Department of Engineering and Technology, Jagannath University, Jaipur, India*

Abstract- ***Large Language Models (LLMs) have become a central technology for generative artificial intelligence, but their tendency to produce fluent yet factually unsupported information remains a major barrier to dependable deployment. This paper proposes a self-reflective framework in which an LLM generates an answer, identifies claims that may be uncertain, performs an internal verification stage, and revises the response before delivery. Unlike a single-pass generation process, the proposed approach separates generation, claim inspection, evidence-oriented verification, and response refinement. The framework is designed to reduce unsupported claims while preserving useful information and acceptable response latency. The paper presents a research-oriented evaluation methodology using factuality, unsupported-claim rate, answer completeness, calibration, and computational overhead as evaluation dimensions. The proposed framework can be integrated with retrieval-augmented generation, external knowledge sources, or domain-specific validation modules. The study argues that self-reflection should be treated not merely as prompt engineering but as a structured reliability layer for generative AI systems.***

Keywords - large language models, generative ai, ai hallucination, self-reflection, factuality, retrieval-augmented generation, llm reliability, artificial intelligence

I. INTRODUCTION

Generative artificial intelligence has rapidly changed the way people interact with information systems. Large Language Models can summarize documents, answer questions, generate software code, assist learning, and support professional workflows. Their usefulness comes from their ability to predict and organize language at scale. However, linguistic fluency does not guarantee factual correctness. An LLM may generate an answer that appears confident while containing unsupported, outdated, or fabricated information.

This phenomenon is commonly described as hallucination. It becomes particularly problematic in education, research, software engineering, legal information, business intelligence, and other settings where users expect dependable answers. Conventional improvements such as larger models, additional training data, and retrieval mechanisms can reduce some errors, but they do not eliminate the problem. A generated response may still contain claims that are weakly supported or inconsistent with available evidence.

This paper proposes a self-reflective reliability framework. The central idea is to introduce an explicit verification stage between initial generation and final response. The model is asked to inspect its own answer, identify potentially unreliable claims, verify them against available evidence when possible, and revise the final response. The approach aims to convert a one-pass generation process into a controlled generation-and-verification pipeline.

II. PROBLEM STATEMENT

The principal research problem is the generation of factually unsupported content by LLMs despite high linguistic quality. Existing systems often optimize for helpfulness and coherence, whereas factual reliability requires an additional mechanism for uncertainty detection and verification. The research question addressed by this work is: How can a structured self-reflection layer reduce unsupported factual claims in LLM-generated responses without causing excessive loss of useful information or computational cost?

III. RESEARCH OBJECTIVES

The objectives of this research are to: (1) develop a structured self-reflection framework for LLM response generation; (2) identify potentially unsupported claims before final response delivery; (3) incorporate evidence-oriented verification into the response pipeline; (4) define measurable criteria for evaluating hallucination reduction; and (5) establish a practical framework that can be combined with retrieval-augmented generation and domain-specific knowledge sources.

IV. LITERATURE REVIEW

Research on hallucination has shown that LLMs can generate plausible content that is not adequately grounded in source information. Retrieval-Augmented Generation (RAG) addresses part of this challenge by supplying external information to the model during generation. However, retrieval alone does not guarantee that every generated claim is supported by retrieved evidence. The model can misunderstand evidence, combine unrelated passages, or introduce additional unsupported statements.

Self-reflection and iterative reasoning methods provide another direction. In these approaches, a model evaluates or critiques an intermediate response before producing a final answer. Related work on self-correction suggests that language models can sometimes identify and repair their own errors when provided with suitable feedback or verification procedures. However, self-reflection is not universally reliable; a model can also repeat an incorrect belief during self-evaluation.

Therefore, a robust architecture should not depend exclusively on internal reflection. It should combine self-critique with evidence checking where external evidence is available. This paper positions self-reflection as a reliability layer rather than a complete replacement for retrieval, human review, or specialized verification systems.

V. RESEARCH GAP

Three gaps are identified. First, many hallucination-reduction approaches focus primarily on retrieval or prompt design rather than treating generation and verification as separate stages. Second, factuality is often evaluated using a single metric, which can hide trade-offs between correctness, completeness, and response cost. Third, internal self-reflection can become circular when the same model generates and verifies an incorrect claim without independent evidence. These gaps motivate a hybrid framework in which self-reflection is combined with claim-level verification and evidence grounding.

VI. PROPOSED SELF-REFLECTIVE FRAMEWORK

The proposed framework consists of five stages.

Stage 1 — Initial Generation: The LLM receives a user query and produces a preliminary response.

Stage 2 — Claim Extraction: The preliminary response is divided into factual claims. Claims containing dates, numerical values, named entities, causal statements, or externally verifiable assertions receive higher verification priority.

Stage 3 — Self-Reflection: The model reviews each claim and assigns a qualitative reliability status such as supported, uncertain, or potentially unsupported. The model is instructed to identify ambiguity and avoid assuming that fluent wording implies correctness.

Stage 4 — Evidence Verification: When an evidence source is available, uncertain claims are compared against retrieved passages, structured databases, or domain-specific knowledge. Claims without adequate support are revised, qualified, or removed.

Stage 5 — Final Response Refinement: The verified claims are assembled into the final response. Where evidence remains insufficient, the system explicitly communicates uncertainty rather than inventing details.

The architecture therefore follows the sequence: Query → Generate → Extract Claims → Reflect → Verify → Revise → Final Response. The framework can operate with or without external retrieval, although evidence-grounded verification is expected to provide stronger reliability.

VII. PROPOSED ALGORITHM

Input: User query Q; language model M; optional evidence source E.

1. Generate preliminary response R using M(Q).
2. Extract factual claims C = {c1, c2, …, cn} from R.
3. For each claim ci, estimate an uncertainty score ui.
4. Select claims with ui above a verification threshold.
5. Retrieve or access relevant evidence for selected claims when E is available.
6. Compare each selected claim with available evidence.
7. Mark claims as supported, contradicted, or unresolved.
8. Rewrite, qualify, or remove contradicted and unresolved claims according to predefined rules.
9. Generate final response R*.
10. Perform a final consistency check between R* and the verified claim set.

Output: Reliability-enhanced response R*.

The important design principle is that uncertainty should trigger additional verification rather than automatic acceptance. The system should prefer an explicitly qualified statement over a confident but unsupported statement.

VIII. EVALUATION METHODOLOGY

A controlled evaluation can compare a conventional single-pass LLM with the proposed self-reflective pipeline. A test set should contain factual question-answering tasks from multiple domains, including general knowledge, science, technology, and time-sensitive information. The same queries should be processed using both approaches.

The primary evaluation metrics should include factual accuracy, unsupported-claim rate, evidence alignment, answer completeness, uncertainty calibration, and latency. Factual accuracy measures whether claims are correct. Unsupported-claim rate measures the proportion of factual claims lacking adequate evidence. Evidence alignment measures whether claims are consistent with the retrieved or authoritative information. Completeness measures whether the system retains relevant information after verification. Latency and computational cost quantify the practical overhead of reflection and verification.

Human evaluation can be conducted by independent reviewers using a predefined rubric. Automated metrics should be used as supporting measurements rather than as the sole basis for determining factual reliability.

IX. EXPECTED RESULTS AND DISCUSSION

The proposed framework is expected to reduce unsupported factual claims compared with single-pass generation, particularly for questions where evidence can be retrieved. The largest benefit is expected from claim-level verification because the system does not need to regenerate the entire response blindly. Instead, it concentrates verification effort on claims that are uncertain or externally verifiable.

A trade-off is expected between reliability and computational cost. Additional reflection and verification stages require more model calls, retrieval operations, or processing time. The framework may also reduce answer completeness if its filtering rules are excessively conservative. Consequently, threshold selection should balance hallucination reduction with usefulness.

Another important limitation is correlated error. If the same model is responsible for generation and reflection, it may fail to recognize an incorrect assumption. Independent evidence sources, specialized verification models, deterministic tools,

or human review can reduce this risk. Thus, self-reflection should be considered one component of broader trustworthy-AI architecture.

X. APPLICATIONS

The framework can be applied to academic research assistants, educational chatbots, enterprise knowledge systems, software-development assistants, technical support systems, and document question-answering platforms. In academic environments, the system could flag unsupported claims before producing a literature summary. In software engineering, verification could be applied to API names, configuration values, and code explanations. In enterprise settings, retrieved organizational documents could serve as evidence for claim validation.

XI. LIMITATIONS

The proposed framework is conceptual and methodological rather than a report of a completed benchmark experiment. Its effectiveness depends on the quality of the underlying language model, evidence sources, claim extraction process, and verification rules. Self-reflection may introduce additional errors and computational overhead. External sources can also be incomplete, contradictory, or outdated. Therefore, future empirical studies should test the framework across models, domains, languages, and evidence conditions.

XII. FUTURE SCOPE

Future research can extend the framework through independent verifier models, multimodal evidence verification, domain-specific fact-checking agents, confidence calibration, and adaptive verification thresholds. Another promising direction is multi-agent verification, where separate agents perform claim extraction, evidence retrieval, contradiction analysis, and final auditing. Research can also investigate lightweight small language models for the verification stage to reduce cost and latency.

XIII. CONCLUSION

This paper presented a self-reflective framework for reducing hallucinations in Large Language Models. The framework introduces claim-level inspection, uncertainty identification, evidence verification, and final response refinement into the generation process. Its main contribution is the separation of fluent generation from reliability checking. Rather than assuming that a model's first answer is sufficiently reliable, the proposed architecture creates an explicit opportunity to identify and correct unsupported claims. The framework is suitable for integration with RAG, domain knowledge bases, and independent verification systems. Future empirical evaluation is required to quantify its performance and determine the optimal balance between factuality, completeness, and computational cost.

REFERENCES

[1] Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems.

[2] Ji, Z., et al. (2023). Survey of Hallucination in Natural Language Generation. ACM Computing Surveys.

[3] Madaan, A., et al. (2023). Self-Refine: Iterative Refinement with Self-Feedback. Advances in Neural Information Processing Systems.

[4] Shinn, N., Cassano, F., Labash, B., Gopinath, A., & Yao, S. (2023). Reflexion: Language Agents with Verbal Reinforcement Learning. Advances in Neural Information Processing Systems.

[5] Asai, A., et al. (2024). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. International Conference on Learning Representations.

[6] Manakul, P., Liusie, A., & Gales, M. (2023). SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. Proceedings of EMNLP.

[7] OpenAI. (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.

[8] Touvron, H., et al. (2023). LLaMA: Open and Efficient Foundation Language Models. arXiv preprint arXiv:2302.13971.