

Reducing Hallucination in Large Language Models using Retrieval-Augmented Generation

Mrs. P.S. Gaikwad

Dept. of AI & DS

AISSMS IOIT, PUNE

prajwal.gaikwad@aiissmioit.org

Aryan Gade

Dept. of AI & DS

AISSMS IOIT, PUNE

Aryangade1234@gmail.com

Vaishnavi Gawali

Dept. of AI & DS

AISSMS IOIT, PUNE

vaishnavirvg@gmail.com

Amit Dhuttaragi

Dept. of AI & DS

AISSMS IOIT, PUNE

amitdhuttaragi@gmail.com

Abstract — Large Language Models (LLMs) have significantly advanced natural language processing by enabling human-like text generation. However, they often produce factually incorrect yet plausible responses, known as hallucinations. This paper presents a structured approach to mitigate hallucination using Retrieval-Augmented Generation (RAG). The proposed system integrates dense retrieval, re-ranking, context augmentation, and verification mechanisms such as semantic entropy and self-consistency. Experimental insights indicate that grounding responses in external knowledge improves factual reliability. Additionally, Reinforcement Learning from Human Feedback (RLHF) and prompt engineering techniques are explored to enhance model trustworthiness.

Keywords — LLM, Hallucination, RAG, NLP, RLHF, Semantic Entropy

I. INTRODUCTION

Large Language Models (LLMs) have emerged as a powerful component of modern artificial intelligence, enabling machines to generate humanlike text and perform complex language understanding tasks. Despite their impressive capabilities, these models do not inherently verify factual correctness and instead rely on learned statistical patterns. As a result, they may produce outputs that appear convincing but are inaccurate, a phenomenon commonly referred to as hallucination.

The presence of hallucination limits the reliability of LLMs, particularly in critical domains such as healthcare, education, and finance, where accuracy

is essential. Addressing this issue has become a key research focus in recent years. One effective approach is Retrieval-Augmented Generation (RAG), which enhances model responses by incorporating relevant external knowledge during the generation process.

In this work, we propose a structured RAGbased framework combined with verification strategies to improve factual consistency and reduce hallucination in generated outputs.

II. LITERATURE REVIEW

The problem of hallucination in Large Language Models (LLMs) has been widely studied, and several techniques have been proposed to improve factual reliability.

[1] Brown et al. introduced large-scale transformer-based language models capable of fewshot learning. Their work demonstrated that increasing model size significantly improves language understanding and generation capabilities. However, the study also highlighted that such models rely on statistical patterns rather than factual verification, which can lead to incorrect but fluent outputs.

Advantages:

- Demonstrated the power of scaling in language models, establishing a foundation for modern LLMs.
- Few-shot learning reduces the dependence on large labeled datasets for downstream tasks.
- Achieved state-of-the-art performance across multiple NLP benchmarks without task-specific fine-tuning.

Limitations:

- The model has no mechanism for factual grounding, leading to hallucinated responses that are difficult to distinguish from accurate ones.

Large computational requirements make deployment resource-intensive and expensive.

- Lack of interpretability makes it challenging to diagnose or prevent specific hallucination errors.

[2] Huang et al. provided a comprehensive taxonomy of hallucination in LLMs, classifying it into intrinsic hallucination (model-generated errors) and extrinsic hallucination (errors due to missing or incorrect knowledge). The paper also discusses causes such as biased datasets, probabilistic decoding, and ambiguous prompts.

Advantages:

- Provides a structured classification of hallucination types, aiding in targeted mitigation strategies.
- Identifies root causes across data, model, and inference dimensions, enabling multilevel intervention.
- Serves as a foundational reference for future hallucination research and evaluation benchmarks.

Limitations:

- The taxonomy is descriptive in nature and does not directly provide a mitigation framework or practical solution.
- Some classification boundaries between intrinsic and extrinsic hallucinations can be ambiguous in practice.
- The study relies on a static view of hallucination and may not account for dynamic or evolving knowledge contexts.

[3] Lewis et al. proposed the Retrieval Augmented Generation (RAG) framework, which integrates a neural retriever with a generative language model. The model retrieves relevant documents and conditions generation on them, significantly improving factual accuracy in knowledge-intensive tasks.

Advantages:

- Grounds model responses in external, verifiable knowledge, directly reducing hallucination.

- Separates the retrieval component from the generative model, allowing each to be updated independently.
- Demonstrated significant improvement in open-domain question answering benchmarks over purely parametric models.

Limitations:

- The quality of generated responses is heavily dependent on the quality and coverage of the underlying knowledge base.
- Retrieval latency can slow down inference, making real-time applications challenging.
- Dense retrieval models may fail to retrieve relevant documents for ambiguous or highly specific queries.

[4] Ji et al. conducted a broad survey on hallucination in natural language generation, covering tasks such as summarization, dialogue generation, and machine translation. They identified faithfulness issues and proposed evaluation strategies for detecting hallucinated content.

Advantages:

- Provides a comprehensive cross-task analysis of hallucination, making findings applicable across diverse NLP applications.
- Introduces evaluation strategies that can serve as benchmarks for assessing model reliability.
- Highlights overlooked faithfulness issues in widely-used generation tasks such as summarization and translation.

Limitations:

- The scope of the survey is broad, which limits the depth of analysis for any individual task.
- Proposed evaluation strategies often require annotated data, which is costly and time-consuming to produce.
- Does not propose a unified mitigation framework applicable across all surveyed tasks.

[5] Gao et al. analysed various RAG architectures, including naive, advanced, and modular RAG. They highlighted improvements

such as query rewriting, re-ranking, and context filtering to enhance retrieval performance.

Advantages:

Provides a systematic comparison of RAG variants, enabling practitioners to select the most suitable architecture.

- Introduces modular design principles that support extensibility and customization of retrieval pipelines.
- Demonstrates that query rewriting and reranking substantially improve the relevance of retrieved content.

Limitations:

- More complex RAG architectures introduce additional computational overhead that may be prohibitive for resource-constrained environments.
- The performance gains from advanced RAG modules are often task-specific and may not generalize broadly.
- Integration complexity increases as more modules are added to the pipeline, making maintenance and debugging difficult.

[6] Farquhar et al. introduced semantic entropy as a method to detect hallucination by measuring uncertainty in generated outputs. Instead of focusing on token-level probabilities, their approach evaluates uncertainty at the semantic level.

Advantages:

- Operates without requiring external knowledge bases or labelled examples, making it highly versatile.
- Captures semantic-level uncertainty rather than surface-level token probabilities, leading to more meaningful hallucination signals.
- Can be applied post-hoc to any generative model without modifying the model architecture.

Limitations:

- Computing semantic entropy requires generating multiple responses per query, which multiplies inference costs.
- The reliability of semantic entropy as a signal is dependent on the quality of the embedding model used for comparison.

- High entropy does not always indicate hallucination; in creative or open-ended tasks, variance is expected and desirable.

[7] Shuster et al. demonstrated that retrieval-based dialogue systems significantly reduce hallucination compared to purely generative models. Their experiments showed improved factual grounding when external knowledge is incorporated. [8] Ouyang et al. introduced Reinforcement Learning from Human Feedback (RLHF), showing that fine-tuned models align better with human expectations and produce more reliable outputs.

Advantages:

- Provides empirical evidence that retrieval augmentation reduces hallucination in conversational settings.
- Dialogue-specific evaluation highlights practical deployment benefits of retrieval-based models.
- Applicable to a wide range of knowledge-intensive conversational AI applications.

Limitations:

- Performance is constrained by the scope and freshness of the knowledge base; outdated information can still lead to incorrect responses.
- Requires efficient indexing and retrieval infrastructure, which may be difficult to maintain at scale.
- Handling multi-turn context with retrieval remains a non-trivial challenge.

[8] Ouyang et al. introduced Reinforcement Learning from Human Feedback (RLHF), where models are fine-tuned using human preferences to improve alignment and response quality.

Advantages:

- Directly optimizes for human-perceived quality, making models more aligned with user expectations.
- Reduces undesirable model behaviours such as harmful outputs, verbosity, and factual inaccuracies.
- Has been demonstrated at scale in widely deployed models, confirming practical effectiveness.

Limitations:

- Requires substantial human annotation effort to create reward signals, which is expensive and time-consuming.
- Human feedback is subjective and may embed annotator biases into the fine-tuned model.

Over-optimization for human preferences can lead to reward hacking, where the model produces plausible-sounding but factually inaccurate responses.

- Requires multiple inference passes, significantly increasing computation time and cost.
- Consistently incorrect models may produce consistent hallucinations that the method cannot detect.
- Effectiveness decreases for factual topics where all sampled outputs converge on the same hallucinated answer.

[9] Min et al. proposed FActScore, a finegrained evaluation metric that breaks generated text into atomic claims and verifies each claim against a knowledge base to measure factual accuracy.

Advantages:

- Provides fine-grained hallucination evaluation at the claim level, allowing precise identification of specific errors.
- Automated verification against a knowledge base reduces reliance on human annotators.
- Can be used to compare different models or fine-tuning strategies on a standardized factual accuracy scale.

Limitations:

- Dependent on the completeness of the knowledge base; claims about rare or specialized topics may not be verifiable.
- Atomic claim decomposition may not always preserve the intended meaning of complex or nuanced statements.
- Verification can be slow for long-form generated content containing many atomic claims.

[10] Manakul et al. introduced SelfCheckGPT, a method that detects hallucination by generating multiple responses and measuring consistency across them. Inconsistent outputs indicate potential hallucination.

Advantages:

- Zero-resource approach requiring no external knowledge base or additional labeled data.
- Works with black-box LLMs, making it widely applicable without model access.
- Consistency-based detection is intuitive and interpretable for downstream users.

Limitations:

III. PROPOSED SYSTEM ARCHITECTURE

The proposed system is designed to reduce hallucination in Large Language Models (LLMs) by integrating retrieval-based knowledge grounding with verification mechanisms. The architecture consists of four major components: Input Processing, Dense Retrieval, Context Augmentation, and Generation with Verification.

A. Input Processing Module

The system begins by accepting a user query Q . The query undergoes preprocessing steps including text normalization, tokenization, and removal of irrelevant symbols. To enhance retrieval performance, query expansion is performed by generating semantically similar terms.

The processed query is then transformed into a dense vector representation using a pre-trained embedding model such as Sentence-BERT:

$$q = \text{Encoder}(Q)$$

where q represents the **query embedding**.

B. Dense Retrieval Module

The retrieval module is responsible for extracting relevant information from an external knowledge base. All documents are stored as vector embeddings in a vector database such as FAISS.

Given a query embedding q , the system computes similarity scores between q and document embeddings d_i . The top- k most relevant documents are selected based on cosine similarity:

$$R = \arg \max_k \cos(q, d_i)$$

To improve precision, a cross-encoder reranking model is applied to the retrieved documents, ensuring that only the most contextually relevant information is forwarded to the next stage.

C. Context Augmentation Module

The retrieved documents are combined with the original query to construct an augmented input for the language model. A structured prompt template is used to enforce reliance on external knowledge:

Context: Retrieved Documents

Question: User Query

Instruction: Answer only based on the given context

This step ensures that the model generates responses grounded in factual data rather than relying solely on internal parametric knowledge.

Finally, the system decides whether to accept, regenerate, or flag the response based on the confidence score.

IV. APPLICATIONS

The proposed RAG-based hallucination reduction framework has broad applicability across a range of real-world domains where factual accuracy and reliability are paramount.

The following subsections outline the primary application areas.

A. Healthcare and Medical Information Systems

In the medical domain, hallucinated responses from AI systems can have life-threatening consequences. The proposed system can be

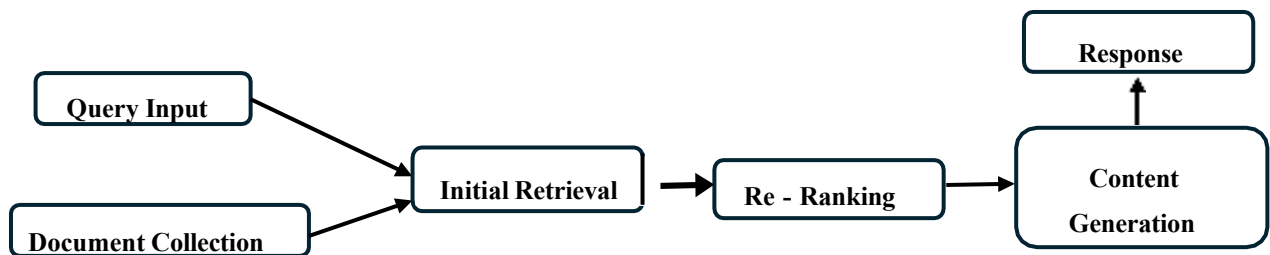


Fig. Flow Diagram

D. Generation and Verification Module

The augmented input is passed to the LLM for response generation. To ensure reliability, a multilevel verification mechanism is applied.

First, self-consistency is used, where multiple responses are generated for the same query and compared for consistency. Inconsistent outputs indicate potential hallucination.

Second, semantic entropy is computed to measure uncertainty in generated responses:

$$H = - \sum_i P(r_i) \log_{10} P(r_i)$$

where r_i represents different generated responses.

deployed in clinical decision support tools, where it retrieves information from verified medical literature, drug databases, and clinical guidelines. By grounding responses in evidence-based sources, the system can assist healthcare professionals with diagnosis support, drug interaction checks, and treatment protocol recommendations while flagging responses that exhibit high semantic entropy for further review.

B. Legal and Compliance Advisory

Legal professionals often require precise and up-to-date information regarding statutes, case law, and regulatory guidelines. A RAG-based system integrated with a legal knowledge base can provide accurate references to relevant legislation and precedents. The verification mechanisms ensure that responses are consistent with source documents, thereby reducing the risk of AI-generated legal misinformation. This is

particularly valuable in contract analysis, due diligence workflows, and compliance monitoring.

C. Educational Platforms and E-Learning

AI tutoring systems powered by LLMs are increasingly used for personalized learning. Hallucinated explanations, however, can reinforce misconceptions in students. The proposed framework can be integrated into e-learning platforms to ensure that factual explanations, solved examples, and concept summaries are grounded in verified educational content. The context augmentation module can be configured to retrieve domain-specific textbooks or curated course materials, improving the pedagogical quality of AI-generated responses.

D. Customer Support and Enterprise Knowledge Management

Large organizations maintain extensive internal documentation, including product manuals, FAQs, policy documents, and technical guides. Deploying the proposed system in enterprise chatbots ensures that customer-facing and employee-facing AI assistants retrieve accurate information from internal knowledge bases rather than hallucinating product specifications or policy details. The reranking module further ensures that the most relevant internal documents are prioritized for each query.

E. News Summarization and Fact-Checking

The media industry faces significant challenges from AI-generated misinformation. The proposed system can be applied to news summarization pipelines to ensure that generated summaries are grounded in source articles. Additionally, by integrating fact-checking databases and verified news archives into the knowledge base, the system can flag responses that contradict established facts. This application is particularly relevant for combating the spread of disinformation in automated content generation workflows.

F. Financial Analysis and Investment Advisory

Financial professionals require accurate, real-time information about market conditions, regulatory changes, and economic indicators. A RAG-based system integrated with financial data sources, SEC filings, and research reports can assist analysts in generating grounded investment summaries and risk assessments. The self-consistency verification mechanism helps identify conflicting signals in financial data, alerting analysts to areas requiring deeper manual review.

G. Scientific Research Assistance

Researchers working on literature reviews, hypothesis generation, and experimental design can

benefit significantly from an AI assistant that retrieves and synthesizes information from verified scientific publications. The dense retrieval module can be connected to scientific databases such as PubMed, arXiv, or IEEE Xplore to ensure that generated summaries and recommendations are grounded in peer-reviewed research rather than model-generated approximations.

CONCLUSION

Large Language Models (LLMs) have demonstrated remarkable capabilities in natural language understanding and generation; however, the issue of hallucination remains a significant limitation affecting their reliability and real-world applicability. This paper presented a structured approach to mitigate hallucination by integrating RetrievalAugmented Generation (RAG) with additional verification mechanisms.

The proposed system enhances traditional RAG architecture by incorporating dense retrieval, crossencoder re-ranking, context augmentation, and verification techniques such as self-consistency and semantic entropy. By grounding the generated responses in externally retrieved knowledge, the system reduces reliance on parametric memory, thereby improving factual accuracy. Furthermore, the inclusion of verification mechanisms ensures that uncertain or inconsistent outputs are identified and handled appropriately.

The study also analysed existing literature, including approaches such as RLHF, FActScore, and SelfCheckGPT, and demonstrated how combining retrieval and verification leads to a more robust and trustworthy system. Compared to conventional generation-based models, the proposed framework provides improved consistency, reduced hallucination, and better alignment with factual information.

Despite these improvements, challenges such as retrieval latency, dependency on knowledge base quality, and handling ambiguous queries still persist. Future work can focus on integrating realtime data retrieval, multi-hop reasoning, adaptive confidence scoring, and hybrid knowledge systems to further enhance performance.

In conclusion, the proposed hybrid approach of combining retrieval and verification represents a promising direction toward building reliable and trustworthy AI systems for critical applications.

REFERENCES

- [1] T. Brown *et al.*, “Language Models are FewShot Learners,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 1877–1901, 2020.
- [2] L. Huang *et al.*, “A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions,” *arXiv preprint arXiv:2311.05232*, 2023.
- [3] P. Lewis *et al.*, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” *NeurIPS*, pp. 9459–9474, 2020.
- [4] Z. Ji *et al.*, “Survey of Hallucination in Natural Language Generation,” *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.
- [5] Y. Gao *et al.*, “Retrieval-Augmented Generation for Large Language Models: A Survey,” *arXiv preprint arXiv:2312.10997*, 2024.
- [6] S. Farquhar *et al.*, “Detecting Hallucinations in Large Language Models Using Semantic Entropy,” *Nature*, vol. 630, pp. 625–630, 2024.
- [7] K. Shuster *et al.*, “Retrieval Augmentation Reduces Hallucination in Dialogue Systems,” *Findings of EMNLP*, pp. 3784–3803, 2021.
- [8] L. Ouyang *et al.*, “Training Language Models to Follow Instructions with Human Feedback,” *NeurIPS*, vol. 35, pp. 27730–27744, 2022.
- [9] S. Min *et al.*, “FActScore: Fine-Grained Atomic Evaluation of Factual Precision in Long Form Text Generation,” *EMNLP*, pp. 12076–12100, 2023.
- [10] P. Manakul *et al.*, “SelfCheckGPT: ZeroResource Black-Box Hallucination Detection for Generative LLMs,” *EMNLP*, pp. 9004–9017, 2023.