



IJEAST

INTERNATIONAL JOURNAL
OF ENGINEERING APPLIED SCIENCE
AND TECHNOLOGY



VOLUME : 11 ISSUE : 05 Print / Issue Publication Date: September 2026



ISSN : 2455-2143



Indexed In



WWW.IJEAST.COM

editor@ijeast.com



A HALLUCINATION-RESISTANT AGENTIC RAG SYSTEM WITH MULTI-STAGE RETRIEVAL AND RESPONSE VALIDATION

Shweta Narayan Bamne, Prof M.S. Shegokar
Department of Computer Science and Engineering
College of Engineering and Technology, Akola, India

Abstract—Large Language Models (LLMs) have shown great abilities in understanding and generating natural language. However, they can create hallucinations by providing factually incorrect or unfounded responses. To tackle this issue, this work suggests a hallucination-resistant Agentic Retrieval-Augmented Generation (RAG) system that combines multi-stage retrieval, smart re-ranking, and response validation. The system uses an agentic structure, where independent agents plan, retrieve, and refine information by utilizing external knowledge sources. A multi-stage retrieval pipeline brings together dense vector search, keyword-based retrieval, and query reformulation to enhance recall and coverage. The retrieved documents go through a re-ranking module that prioritizes contextually relevant and high confidence evidence, which reduces noise and irrelevant information. To improve reliability, we add a response validation layer that uses techniques like cross-verification, fact-checking against retrieved evidence, and confidence scoring. This ensures that the generated outputs are based on verifiable sources and lowers the risk of hallucinations. Additionally, the system includes feedback loops and iterative reasoning, allowing agents to update outputs in real-time. Experimental results show that this framework greatly boosts answer accuracy, contextual relevance, and reliability compared to traditional RAG systems. This research presents a scalable and modular setup for implementing trustworthy AI systems in important fields such as healthcare, finance, and legal applications.

Index Terms—Agentic Systems, Hallucination Mitigation, Large Language Models, Multi-Stage Retrieval, Query Reformulation, Re-Ranking, Response Validation, Retrieval-Augmented Generation (RAG), Trustworthy AI.

I. INTRODUCTION

Large Language Models (LLMs) are increasingly being used in important areas such as healthcare, finance, education, and legal support because of their ability to understand and generate natural language. Although these

models can produce highly fluent responses, they still face a serious limitation known as hallucination, where the generated output appears convincing but contains incorrect or unsupported information. In sensitive domains, even a small factual mistake can create major consequences, including wrong medical guidance, financial risk, or misleading legal interpretation. Because of this, improving the reliability of LLMs has become essential before they can be safely adopted in real-world systems where correctness is critical.

Retrieval-Augmented Generation (RAG) was originally proposed as a way to combine a pretrained language model's parametric memory with a non-parametric memory accessed through dense retrieval [11], offering a practical way to improve reliability by connecting language models with external knowledge sources. Instead of depending only on internal model memory, a RAG system first retrieves relevant documents from a knowledge base and then uses those documents while generating the final response. This approach reduces unsupported answers, but conventional RAG systems still have several weaknesses, as noted in broader surveys of the field [12]. A single retrieval step may fail to capture all relevant information when the user query is complex. Retrieved passages can also include unrelated or contradictory content, which may reduce answer quality. In addition, most traditional systems do not verify whether the final generated response is actually supported by the retrieved evidence, making the overall system less dependable for high-risk applications.

Several recent studies have attempted to improve this limitation. Agent-based RAG methods have shown that iterative reasoning can reduce hallucinated content [1]. Multi agent retrieval systems improve factual grounding by adjusting search behavior dynamically [2]. Some researchers have demonstrated that explicit response verification can significantly lower unsupported generation [3]. Transparent validation methods also help increase trust in AI-generated content [4]. Feedback-driven pipelines have been proposed to continuously refine responses [5]. Cross-encoder models have improved ranking quality by identifying stronger supporting documents [6]. Recent reviews further confirm that validation is one of the most



important components in hallucination prevention [7]. Hybrid retrieval methods that combine sparse and dense search have improved document recall in knowledge intensive tasks [8]. Natural Language Inference has also been successfully used for checking factual consistency [9], while confidence-based scoring methods help detect uncertain answers before they reach the user [10]. Broader taxonomies of hallucination in natural language generation further motivate the need for grounding and verification mechanisms of this kind [13], and earlier work has shown that retrieval augmentation alone can meaningfully reduce hallucination in conversational settings [14].

Even with these improvements, some important problems remain unsolved. Many existing systems still depend on limited retrieval pipelines that may overlook useful evidence. In some approaches, re-ranking is weak or separated from the overall workflow. Validation often happens only once and does not support correction when the response is inaccurate. As a result, the integration of multi-stage retrieval, intelligent ranking, and continuous validation inside a single agentic framework has not been fully developed. Challenges such as computational cost, excessive refusal of correct answers, and adaptation across different domains also continue to affect system performance.

To address these limitations, this work proposes a hallucination-resistant Agentic RAG framework that combines multi-stage retrieval, intelligent re-ranking, and response validation within a unified architecture. The system is organized into three specialized agents. The retrieval agent improves search coverage by expanding the query and combining BM25 for lexical matching, Dense Passage Retriever for semantic similarity, and Reciprocal Rank Fusion to merge results from both methods. The re-ranking agent then applies a cross encoder model to identify the most reliable evidence and remove weak documents. After response generation, the validation agent breaks the answer into smaller factual claims and checks each claim against the retrieved evidence using Natural Language Inference with entailment, contradiction, and neutral classification. If unsupported content is detected, the system either regenerates the answer or safely declines to respond. Through this iterative design, the proposed framework aims to deliver more dependable AI-generated responses for practical use in healthcare, finance, and legal environments.

II. METHODOLOGY

This section explains how our Hallucination-Resistant Agentic RAG system is designed and how each part of the system works together. It covers the overall architecture, the role of each agent, the retrieval and ranking process, validation strategy, and the feedback mechanism used to improve response quality.

A. System Architecture

The system follows a modular and agent-based design so that each step can be controlled and verified. It consists of multiple components that operate in sequence but can also interact iteratively when needed.

- **Query Processing Agent:** This module takes the user query and creates multiple variations using prompting techniques. These variations help in improving search coverage and also assist in understanding the intent and context of the query.
- **Multi-Stage Retrieval Module:** Retrieval is performed in three steps. First, BM25 is used to match keywords. Second, Dense Passage Retrieval (DPR) captures semantic meaning. Finally, Reciprocal Rank Fusion (RRF) merges both results to select documents that perform well across both approaches.
- **Re-Ranking Agent:** Retrieved documents are further filtered using a cross-encoder model such as MiniLM or DeBERTa. This model evaluates each query-document pair more precisely, ranks them, and keeps only the top relevant documents while removing weaker ones.
- **Generator Agent:** This component generates the final answer using the selected documents as context. It follows strict instructions to rely only on the provided evidence and avoids answering when sufficient information is not available.
- **Response Validation Agent:** This is the key component for reducing hallucination. It breaks the generated answer into smaller factual claims and checks each claim against the retrieved documents using a Natural Language Inference (NLI) model.
- **Feedback and Iteration Module:** If the validation step detects errors, the system sends feedback to the generator to revise the answer. This process can repeat a limited number of times before returning a safe fallback response.
- **Confidence Scoring Layer:** A final confidence score is calculated using validation results, retrieval quality, and model output signals. Responses with low confidence are rejected.

B. Agentic Workflow

The workflow starts with query expansion, followed by document retrieval and re-ranking. The generator then produces an answer based only on the selected documents. After that, the validation agent checks each claim. If all checks pass, the answer is returned along with a confidence score. If not, the system attempts to regenerate the response using feedback. After a fixed number of attempts, it returns a safe refusal. This process ensures that every response is supported by evidence.



C. Multi-Stage Retrieval Mechanism

The retrieval process is designed to improve both coverage and accuracy. In the first stage, BM25 is used for keyword based matching, which works well for direct queries. In the second stage, Dense Passage Retrieval maps queries and documents into vector space, helping capture meaning and relationships that keyword methods may miss. In the final stage, Reciprocal Rank Fusion combines both outputs. The scoring formula is:

$$RRF_{score} = \sum_{i=1}^n \frac{1}{k + rank_i}$$

where $k = 60$, and $rank_i$ is the rank of a document in each retrieval list. This method gives higher importance to documents that appear consistently across methods. If needed, the system can also reformulate queries to improve retrieval results.

D. Re-Ranking Strategy

The re-ranking step uses a cross-encoder model that processes the query and document together. This allows the model to understand detailed relationships between words and context. Each document receives a relevance score, and only the top-ranked ones are kept. This step reduces noise and ensures that the generator works with reliable information.

E. Response Validation Framework

Validation is performed in three steps. First, the generated answer is split into smaller claims. Second, each claim is compared with retrieved documents using an NLI model that classifies them as supported, contradicted, or neutral. Third, all results are combined to decide whether the answer is valid. If any claim is unsupported or contradicted, the response is marked as unreliable.

F. Natural Language Inference Integration

The NLI model is used to verify factual correctness. It has been trained on large datasets and can identify whether a claim is supported by evidence. To improve performance, the system uses batch processing and caching. Threshold values are chosen carefully to balance between accepting correct answers and rejecting incorrect ones.

G. Feedback Loops and Iterative Refinement

When validation fails, the system provides detailed feedback to the generator, including which parts of the answer were incorrect. The generator then tries to fix the answer using only the given documents. This process is repeated up to two times. If the system still cannot produce

a verified answer, it returns a safe response indicating uncertainty. This approach improves accuracy while maintaining user trust.

H. Implementation Considerations

To make the system practical, several optimizations are applied. Retrieval steps are parallelized to reduce delay. Frequently used results are cached. The NLI model runs on GPU for faster validation. The system supports multiple language models and is packaged using Docker for easy deployment. Parameters such as thresholds can be adjusted depending on the application domain.

III. IMPLEMENTATION

This section describes how the system is built in practice, including the tools, configurations, and deployment setup.

A. Technology Stack

The system is developed using Python 3.10. Different LLM options are supported, including LLaMA, GPT-based models, and Mistral. Retrieval is handled using frameworks like LangChain and LlamaIndex. BM25 is used for keyword search, while dense retrieval is supported through embedding models and FAISS. Re-ranking is done using cross-encoder models, and validation uses an NLI model. The system is exposed through a FastAPI interface. Docker is used for deployment, and Redis helps in caching to improve performance.

B. Component Configuration

Each component is tuned for balanced performance. Query expansion generates multiple variations. Retrieval fetches documents from both sparse and dense methods and merges them. Re-ranking filters the top results. The generator uses controlled parameters to produce concise answers. The validation module checks multiple claims efficiently using batch processing.

C. Integration Approach

A central controller manages the entire pipeline, from query processing to final output. Each module is implemented separately, allowing flexibility and easy updates. Parallel processing is used where possible to improve speed. The system also supports different data formats and maintains state during feedback iterations.

D. Knowledge Base Preparation

Documents are divided into smaller chunks with slight overlap to maintain context. These chunks are indexed for both keyword and semantic search. Metadata is stored alongside each document to support filtering. The system also supports updates without needing full reprocessing.



E. Implementation Challenges and Solutions

During development, several challenges were identified. Re-ranking initially caused delays, which were reduced using smaller models and batching. Validation also added latency, which was improved through caching and parallel execution. False refusals were reduced by adjusting thresholds. Claim extraction quality was improved by using better models. Infinite loops during regeneration were avoided by limiting retries.

F. Deployment Configuration

The system is deployed on a cloud machine with sufficient CPU, RAM, and GPU resources. Different modules are allocated specific resources to ensure smooth operation. The API includes rate limiting and logging for monitoring. Health checks are used to maintain system stability.

IV. EVALUATION AND RESULTS

This section presents how the system was tested and how it performs compared to other approaches.

A. Experimental Setup

Experiments were conducted on a cloud environment with GPU support. Each test was repeated multiple times, and average results were reported to ensure consistency.

B. Datasets

Three datasets were used for evaluation: a general-domain dataset, a domain-specific biomedical dataset, and a dataset designed to trigger hallucinations. This helped evaluate performance across different scenarios.

C. Baseline Systems

The system was compared with multiple baseline methods, including basic RAG and other advanced reasoning-based approaches. All systems used the same model and data to ensure fair comparison.

D. Evaluation Metrics

Performance was measured using hallucination rate, factual accuracy, answer correctness, and response time, following evaluation practices established for automated RAG assessment [15]. These metrics provide a complete view of system reliability.

E. Quantitative Results

The proposed system showed clear improvement over baseline methods. It significantly reduced hallucinations, improved factual correctness, and achieved higher answer accuracy. Although the response time increased slightly due to additional validation steps, it remained within acceptable limits.

F. Ablation Study

Each component was tested independently to measure its contribution. Removing any major component led to worse performance, confirming that all parts of the system are important.

G. Qualitative Analysis

The system performed well in identifying incorrect or unsupported information and either corrected or rejected such responses. Some limitations were observed in handling complex reasoning and ambiguous queries.

H. Discussion

Overall, the system achieves a strong balance between accuracy and reliability. While it introduces additional computational cost, the improvement in trustworthiness makes it suitable for high-risk applications. Future improvements can focus on reducing latency and improving validation accuracy further.

V. CONCLUSION

In this work, we presented a hallucination-resistant agentic Retrieval-Augmented Generation (RAG) system that combines multi-stage retrieval, improved re-ranking, and response validation within a unified framework. The system integrates BM25-based keyword retrieval, semantic retrieval methods, and Reciprocal Rank Fusion to ensure better coverage of relevant information, followed by a cross-encoder re-ranking step to select high-quality evidence. To enhance reliability, a validation mechanism based on natural language inference is used, where generated answers are broken into smaller claims and verified against retrieved documents to detect unsupported or incorrect information. The agent-based design further enables iterative refinement through feedback loops, allowing the system to revise responses when validation fails or safely refuse when sufficient evidence is not available. Experimental results show that the proposed approach significantly reduces hallucination rates compared to standard RAG systems while maintaining acceptable response time. However, the system introduces additional computational cost and still faces challenges in handling complex reasoning and improving claim-level validation accuracy. Future work will focus on optimizing efficiency, enhancing claim extraction, and extending validation capabilities to support more complex reasoning tasks, ultimately contributing to the development of more reliable and trustworthy AI systems for real-world applications.

VI. REFERENCES

- [1]. Kumar A., Finley S., and Thompson R. (2026). Agentic Retrieval Augmented Generation for Hallucination Mitigation in Large Language Models, arXiv preprint, DOI: arXiv:2603.07379.



- [2]. Nguyen T., Chen M., and Zhang L. (2025). Multi-Agent Frameworks for Factual Accuracy in LLM-Based Question Answering, arXiv preprint, DOI: arXiv:2501.09136.
- [3]. Pisano S., Patterson J., and Williams K. (2025). Hallucination Detection and Correction in Retrieval-Augmented Language Models, arXiv preprint, DOI: arXiv:2510.25518.
- [4]. Le H., Truong T., and Nguyen D. (2025). Explainable AI in Agentic Retrieval-Augmented Generation: A Comprehensive Framework, arXiv preprint, DOI: arXiv:2506.21931.
- [5]. Sarpatwar K., Mehta R., and Sharma A. (2024). Self-Correcting RAG Pipelines with Iterative Feedback Loops, Hugging Face Papers, DOI: arXiv:2404.08189.
- [6]. Arya V., Dave P., and Marathe M. (2024). Cross-Encoder Re-Ranking for Improved Document Selection in RAG Systems, arXiv preprint, DOI: arXiv:2402.10612.
- [7]. Dhamange N., Somy S., and Deshmukh R. (2025). Towards Explainable AI in Agentic Retrieval-Augmented Generation: A Systematic Review, ResearchGate, Publication No. 397518736.
- [8]. Johnson M., Lee P., and Kumar S. (2025). Multi-Stage Retrieval Architectures for Knowledge-Intensive NLP Tasks, in IEEE Transactions on Knowledge and Data Engineering, vol. 36, no. 8, (pp.4120-4135).
- [9]. Patel R., Singh A., and Gupta N. (2024). Natural Language Inference for Fact Verification in High-Stakes RAG Applications, in IEEE Access, vol. 12, (pp.107692-107708).
- [10]. Williams T., Brown S., and Davis L. (2025). Confidence Scoring and Uncertainty Estimation for Hallucination Detection in LLM-Generated Content, in IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 11, (pp.111830-111845).
- [11]. Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Kuttler H., Lewis M., Yih W., Rocktaschel T., Riedel S., and Kiela D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, in Advances in Neural Information Processing Systems 33 (NeurIPS 2020).
- [12]. Gao Y., Xiong Y., Gao X., Jia K., Pan J., Bi Y., Dai Y., Sun J., Wang M., and Wang H. (2024). Retrieval-Augmented Generation for Large Language Models: A Survey, arXiv preprint, DOI: arXiv:2312.10997.
- [13]. Ji Z., Lee N., Frieske R., Yu T., Su D., Xu Y., Ishii E., Bang Y., Madotto A., and Fung P. (2023). Survey of Hallucination in Natural Language Generation, in ACM Computing Surveys.
- [14]. Shuster K., Poff S., Chen M., Kiela D., and Weston J. (2021). Retrieval Augmentation Reduces Hallucination in Conversation, in Findings of the Association for Computational Linguistics: EMNLP 2021.
- [15]. Es S., James J., Espinosa-Anke L., and Schockaert S. (2023). RA GAS: Automated Evaluation of Retrieval Augmented Generation, arXiv preprint, DOI: arXiv:2309.15217.

IJEAST

INTERNATIONAL JOURNAL OF ENGINEERING APPLIED SCIENCE AND TECHNOLOGY



AUTHORS

 Shweta Narayan Bamne

 Prof M.S. Shegokar



ABOUT IJEAST

International journal of Engineering Applied Science and Technology (IJEAST) is a peer-reviewed, open access journal that publishes high - quality research paper in the field of Engineering, Applied Science and Technology.

IJEAST aims to provide a platform for researchers, academicians, and professionals to share their innovative ideas, research findings, and practical experiences with the global scientific community.

JOURNAL METRICS



H-INDEX

33



i10-INDEX

154



IJEAST
CITATIONS

9,647

(IJEAST CITATIONS)



PEER-REVIEWED
& OPEN ACCESS

PUBLISHED
MONTHLY



For more information, visit our website

www.ijeast.com



PEER REVIEWED

All submission are rigorously peer reviewed to ensure quality.



AUTHORS ARE OUR PRIORITY

We value our authors and recognize their contribution to the advancement of research and knowledge.



OPEN ACCESS

Free and unrestricted access to research for all.



GLOBAL REACH

Connecting researchers and professionals worldwide.



TIMELY PUBLICATION

We ensure a swift and efficient publication process.



editor@ijeast.com



www.ijeast.com



India



2455-2143