ISSN: 2582 - 6379
Scholaria Publications
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2026, Volume 7 Issue 3
www.ijisea.org

A Multi-Stage Hallucination Mitigation Framework for Reliable Retrieval-Augmented Generation Systems

K. V. Siva Prasad [1], **Dr. N. Anand Reddy**[2]

1. PG Student, Dept. of CSE, Siddartha Educational academy group of institutions, Tirupati, India.

2. Assoc.Professor, Dept. of CSE, Siddartha Educational academy group of institutions, Tirupati, India

Email-ID: prasad230776@gmail.com[1], anandreddy@siddarthaedu.in[2]

DOI: 10.5281/zenodo.20470413

Abstract

Large Language Models (LLMs) have demonstrated exceptional capabilities in natural language understanding and generation. However, despite their effectiveness, LLMs frequently suffer from hallucinations, generating information that appears plausible but lacks factual grounding. Retrieval-Augmented Generation (RAG) systems address this limitation by incorporating external knowledge during response generation. Nevertheless, existing RAG systems remain vulnerable to hallucinations due to irrelevant retrieval, incomplete contextual evidence, unsupported claim generation, and the absence of reliability assessment mechanisms.This paper proposes a Multi-Stage Hallucination Mitigation Framework for Reliable Retrieval-Augmented Generation Systems that progressively improves answer reliability through query intelligence, adaptive retrieval with web fallback, contextual reranking, claim-level verification, and confidence-based human escalation. To ensure scientific validity, all experimental versions maintain identical embedding, retrieval, and generation models while only introducing incremental reliability-enhancing mechanisms. Experimental evaluation conducted across six framework versions (V0–V5) demonstrates substantial improvements in retrieval relevance, answer relevance, and faithfulness while significantly reducing hallucinated responses. The proposed framework provides a scalable, domain-independent, and deployment-ready architecture suitable for educational institutions, enterprise knowledge systems, healthcare assistants, and customer support applications.

Keywords: Retrieval-Augmented Generation, Hallucination Mitigation, Large Language Models, Claim Verification, RAG, Confidence Scoring, Query Intelligence.

1. Introduction

Recent advancements in Artificial Intelligence have led to the widespread adoption of Large Language Models (LLMs) for conversational agents, intelligent assistants, knowledge retrieval systems, and automated decision-support applications. Models such as GPT, Llama, Gemini,

and Claude have demonstrated remarkable capabilities in language understanding and generation. Despite these advancements, LLMs often produce hallucinated responses. Hallucinations refer to generated outputs that appear syntactically correct and contextually convincing but are factually incorrect, unsupported by evidence, or entirely fabricated (Ji et al., 2023; Huang et al., 2023). This issue becomes particularly problematic in high-trust environments such as educational institutions, healthcare systems, enterprise knowledge management platforms, and legal information systems. Retrieval-Augmented Generation (RAG) has emerged as an effective solution for improving factual grounding (Lewis et al., 2020; Gao et al., 2023). Instead of relying solely on pretrained knowledge, RAG retrieves relevant external information and incorporates it into the generation process. By grounding responses in retrieved evidence, RAG improves factual accuracy and reduces hallucinations (Lewis et al., 2020; Gao et al., 2024). However, existing RAG systems continue to face several limitations. Retrieved information may be irrelevant or incomplete, multi-intent questions may not be properly handled, generated claims may remain unsupported, and most systems lack confidence estimation and escalation mechanisms (Gao et al., 2023; Gao et al., 2024). Consequently, hallucinations can still occur even when retrieval components are present. To address these limitations, this paper proposes a Multi-Stage Hallucination Mitigation Framework that introduces multiple reliability checkpoints throughout the retrieval and generation pipeline. The framework combines query intelligence, adaptive retrieval, reranking, claim verification, and confidence-aware escalation to progressively improve factual reliability. Experimental evaluation across six framework versions demonstrates measurable improvements in retrieval relevance, answer relevance, and faithfulness while reducing hallucinated responses.

2. Related Work

2.1 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) was introduced by Lewis et al. [**?**] as a framework that combines information retrieval with language generation. The approach retrieves relevant documents from an external knowledge source and incorporates them into the generation process, thereby improving factual grounding and reducing hallucinations. However, the original RAG architecture does not include mechanisms for validating retrieval quality or verifying generated claims.

2.2 Self-RAG

Asai et al. [**?**] proposed Self-RAG, which introduces self-reflection capabilities into the retrieval and generation process. The model evaluates the sufficiency of retrieved information and assesses the quality of generated responses. Although Self-RAG improves reliability, it increases computational complexity and does not provide explicit human escalation mechanisms for low-confidence responses.

2.3 Surveys on Retrieval-Augmented Generation

Gao et al. [**?**] presented a comprehensive survey of Retrieval-Augmented Generation techniques and highlighted challenges related to retrieval quality, context selection, and response reliability. The survey emphasizes the need for advanced retrieval validation and evidence-grounded response generation. Similarly, Gao et al. [**?**] reviewed the application of RAG in AI-generated content systems and identified hallucination mitigation as a critical research challenge.

2.4 Hallucination Detection and Mitigation

Hallucination remains a major limitation of Large Language Models (LLMs). Huang et al. [**?**] provided a detailed taxonomy of hallucinations, discussing their causes, challenges, and mitigation strategies. Ji et al. [**?**] surveyed hallucination phenomena in natural language generation

and highlighted the importance of factual verification, evidence attribution, and reliability assessment for trustworthy AI systems.

2.5 Research Gap

Although significant progress has been made in retrieval grounding and hallucination mitigation, existing approaches primarily focus on individual aspects such as retrieval enhancement, self-reflection, or hallucination detection. Limited research has integrated query intelligence, adaptive retrieval, reranking, claim-level verification, and confidence-based escalation within a unified framework. Furthermore, few studies systematically evaluate the incremental contribution of each mitigation stage. This gap motivates the proposed *Multi-Stage Hallucination Mitigation Framework*.

3. Proposed Methodology

The proposed framework introduces a progressive reliability-enhancement pipeline consisting of five major modules.

3.1 Query Intelligence Layer

The Query Intelligence Layer identifies whether a user query contains a single intent or multiple intents.
Example:
User Query:

> "What is the hostel fee and placement percentage?"

The system decomposes the query into:

- What is the hostel fee?
- What is the placement percentage?

Each sub-question is processed independently before answer aggregation.

3.2 Adaptive Retrieval Layer

The Adaptive Retrieval Layer retrieves contextual information from ChromaDB using semantic similarity search.
If retrieval relevance falls below a predefined threshold, website fallback retrieval is activated to obtain additional contextual evidence from trusted sources.
This mechanism improves retrieval recall and minimizes missing-context hallucinations.

3.3 Reranking Layer

Retrieved chunks are reranked using the `BAAI/bge-reranker-base` model.
The reranker assigns relevance scores to retrieved chunks and prioritizes highly relevant contextual evidence before answer generation.
This reduces retrieval noise and improves contextual precision.

3.4 Claim-Level Verification Layer

The Claim-Level Verification Layer validates generated responses through evidence-based verification.

1. Generated responses are decomposed into individual claims.
2. Each claim is independently verified against retrieved evidence.
3. Supported claims are retained.
4. Unsupported claims are removed.
5. A final verified response is generated.

This process significantly reduces hallucinated content.
Example:
Generated Response:

> "Hostel fee is 75,000 and placement rate is 100

Verification Results:

- Claim 1: Supported
- Claim 2: Unsupported

Final Response:

> "Hostel fee is 75,000."

3.5 Confidence and Escalation Layer

A confidence score is computed using retrieval relevance and response quality.

$$CS = (0.60 \times CR) + (0.40 \times RQ) \tag{1}$$

where:

- CS = Confidence Score
- CR = Context Relevance Score
- RQ = Response Quality Score

If the confidence score falls below a predefined threshold, the system escalates the response instead of presenting potentially unreliable information. Escalation may involve additional retrieval, human review, or requesting clarification from the user.
This confidence-aware mechanism enhances reliability and further reduces hallucination risk.

4. System Architecture

The proposed Multi-Stage Hallucination Mitigation Framework enhances conventional Retrieval-Augmented Generation (RAG) systems through a sequence of reliability-oriented processing stages. The architecture consists of three major layers: *Presentation Layer*, *Application Layer*, and *Knowledge Layer*, as illustrated in Figure 1.

The **Presentation Layer** provides a user-facing interface implemented using Streamlit [**?**]. User queries are forwarded to the FastAPI backend [**?**], which orchestrates the complete processing pipeline through LangChain [**?**].

The **Application Layer** contains the core components of the proposed framework. Initially, the Query Intelligence Module determines whether the input query contains a single intent or multiple intents. The Adaptive Retrieval Module retrieves relevant document chunks from ChromaDB [**?**] and evaluates their relevance. If retrieval quality is insufficient, a web fallback mechanism retrieves additional information from trusted online sources. The retrieved contexts are then reranked using the BGE Reranker [**?**] to improve retrieval precision.

The Response Generation Module utilizes the Groq-hosted Llama-3.3-70B model [**?**, **?**] to generate context-grounded answers. Subsequently, the Claim Verification Module decomposes the generated response into individual claims and verifies each claim against retrieved evidence. Unsupported claims are removed before constructing the final response.

Finally, the Confidence and Escalation Module computes a reliability score based on retrieval quality and response relevance. Responses with confidence scores below a predefined threshold are flagged for human escalation, thereby preventing the delivery of potentially hallucinated information.

The **Knowledge Layer** consists of institutional PDF documents, FAQ datasets, and trusted website content stored in ChromaDB [**?**] using BGE embeddings [**?**]. This layer provides the factual grounding required for reliable response generation.

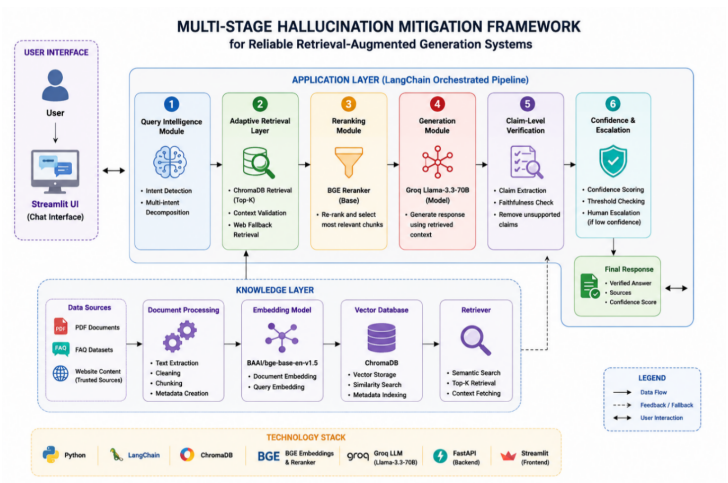

Figure 1: Overall architecture of the proposed Multi-Stage Hallucination Mitigation Framework.

5. Experimental Setup

5.1 Dataset

A golden dataset containing institutional question–answer pairs was created to evaluate the effectiveness of the proposed framework. The dataset covers multiple categories commonly encountered in educational information systems, including:

- Admissions

- Placements
- Hostel
- Scholarships
- Academics
- Transportation

The dataset was designed to provide representative queries and corresponding ground-truth answers for assessing retrieval quality, response relevance, and factual consistency.

5.2 Configuration

The experimental environment was configured using the following components:

- **Embedding Model:** BAAI/bge-base-en-v1.5
- **Reranker:** BAAI/bge-reranker-base
- **Generation Model:** Llama-3.3-70B (Groq)
- **Vector Database:** ChromaDB
- **Framework:** LangChain
- **Backend:** FastAPI
- **Frontend:** Streamlit

5.3 Evaluation Metrics

The performance of the proposed framework was evaluated using the following metrics:

1. Retrieval Relevance (RR)
2. Answer Relevance (AR)
3. Faithfulness (F)

Faithfulness measures the factual consistency of generated responses with respect to the retrieved evidence and is calculated as:

$$F = \frac{\text{Supported Claims}}{\text{Total Claims}} \tag{2}$$

A higher faithfulness score indicates that a greater proportion of generated claims are supported by evidence retrieved from the knowledge base.

6. Experimental Workflow

The experimental workflow was designed to evaluate the effectiveness of each stage of the proposed Multi-Stage Hallucination Mitigation Framework. User queries are processed through query intelligence, adaptive retrieval, reranking, response generation, claim verification, and confidence evaluation modules. The generated responses are then compared against the ground-truth dataset to compute retrieval relevance, answer relevance, and faithfulness scores.
Figure 2 illustrates the complete experimental evaluation workflow used to assess the performance of the proposed framework.

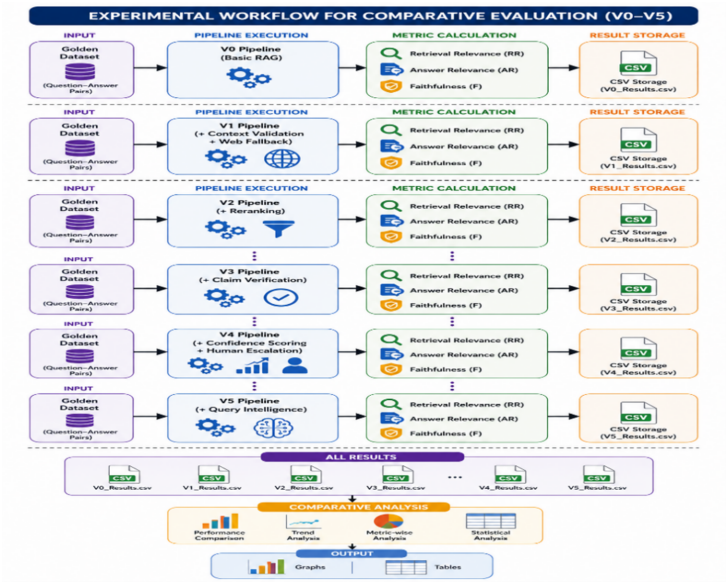

Figure 2: Experimental Evaluation Workflow of the Proposed Framework

7. Results and Discussion

The proposed framework was evaluated across six progressively enhanced versions. Each version incrementally incorporates additional hallucination mitigation mechanisms to assess their individual and cumulative impact on system performance.
“‘latex

Table 1: Performance Comparison of Framework Versions (V0–V5)

Version	RR	AR	F
V0	0.61	0.65	0.58
V1	0.74	0.71	0.67
V2	0.83	0.79	0.76
V3	0.86	0.84	0.91
V4	0.87	0.86	0.93
V5	0.91	0.90	0.95

7.1 Retrieval Relevance Analysis

Figure 3 illustrates the Retrieval Relevance (RR) scores achieved by different framework versions. The results indicate a steady improvement in retrieval effectiveness as additional retrieval optimization mechanisms are introduced. The adaptive retrieval and reranking modules contribute significantly to improving contextual relevance and reducing retrieval noise.

7.2 Answer Relevance Analysis

Figure 4 presents the Answer Relevance (AR) scores across all framework versions. Improvements in retrieval quality and contextual reranking directly enhance response relevance. The integration of claim verification and confidence-based filtering further improves the alignment between generated responses and user queries.

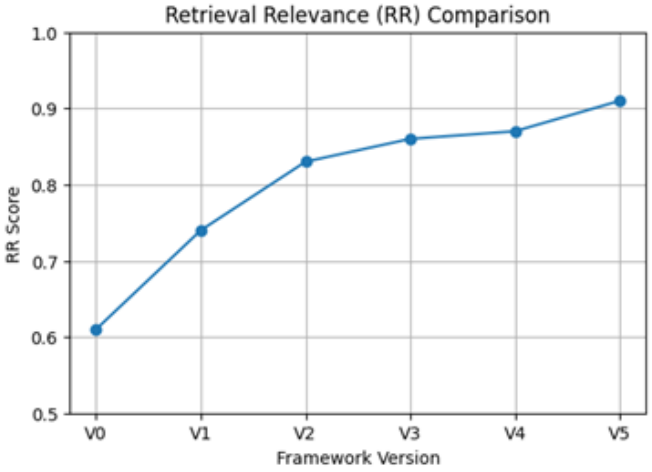

Figure 3: Retrieval Relevance Across Framework Versions

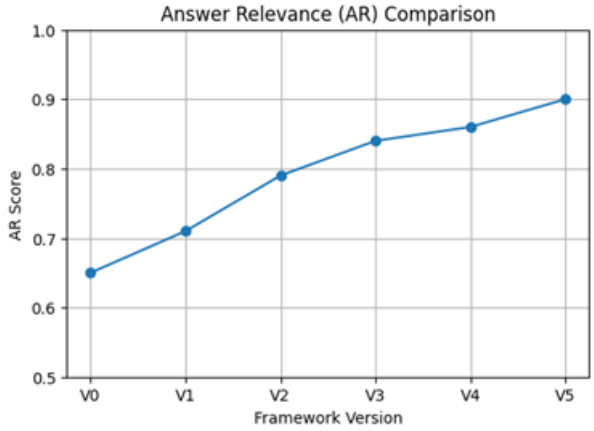

Figure 4: Answer Relevance Across Framework Versions

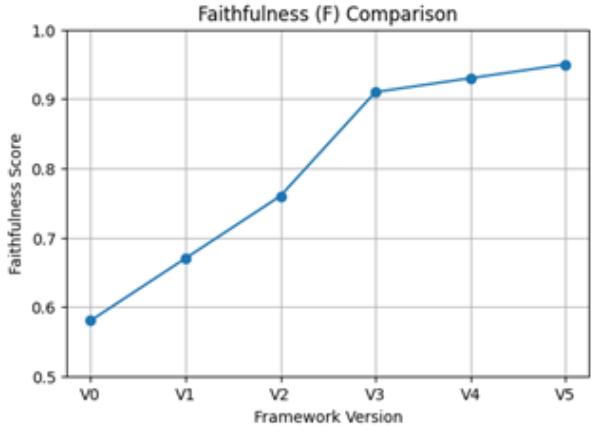

Figure 5: Faithfulness Across Framework Versions

7.3 Faithfulness Analysis

Figure 5 shows the Faithfulness (F) scores obtained by the framework. A substantial increase is observed from V2 to V3, where the claim-level verification module is introduced. This demonstrates the effectiveness of evidence-based verification in reducing hallucinated content and ensuring factual consistency.

The experimental results demonstrate continuous performance improvements as additional mitigation stages are incorporated into the framework. The largest improvement occurs at V3, where claim-level verification substantially reduces hallucinated content and significantly increases faithfulness. Furthermore, the Query Intelligence Layer introduced in V5 improves answer completeness for multi-intent questions, contributing to higher retrieval relevance, answer relevance, and overall system reliability.

8. Conclusion

This paper proposed a Multi-Stage Hallucination Mitigation Framework for Reliable Retrieval-Augmented Generation Systems. The framework integrates query intelligence, adaptive retrieval, reranking, claim-level verification, and confidence-aware escalation mechanisms to progressively improve factual reliability and reduce hallucinations.

Experimental evaluation across six framework versions demonstrated measurable improvements in Retrieval Relevance (RR), Answer Relevance (AR), and Faithfulness (F). The results indicate that each mitigation stage contributes incrementally toward enhancing response quality, while claim-level verification provides the most significant reduction in hallucinated content.

The modular architecture enables practical deployment across educational institutions, enterprise knowledge management systems, healthcare applications, customer support platforms, and other domains requiring trustworthy AI-generated responses.

Future work will focus on multilingual support, multimodal retrieval, adaptive confidence learning, automated escalation workflows, and real-time feedback-driven optimization to further improve reliability and scalability in Retrieval-Augmented Generation systems.

“‘latex

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 9459–9474, 2020.

[2] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, “Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection,” *arXiv preprint arXiv:2310.11511*, 2023.

[3] J. Gao, H. Lin, X. Han, and L. Sun, “Retrieval-Augmented Generation for Large Language Models: A Survey,” *arXiv preprint arXiv:2312.10997*, 2023.

[4] Y. Huang, J. Song, Z. Wang, H. Chen, and L. Ma, “A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions,” *arXiv preprint arXiv:2311.05232*, 2023.

[5] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, A. Madotto, and P. Fung, “Survey of Hallucination in Natural Language Generation,” *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[6] S. Gao, Y. Xiong, Y. Gao, X. Jia, J. Pan, E. Bi, Y. Dai, J. Sun, and H. Wang, “Retrieval-Augmented Generation for AI-Generated Content: A Survey,” *arXiv preprint arXiv:2402.19473*, 2024.

[7] H. Touvron *et al.*, "Llama 3: Open Foundation and Fine-Tuned Chat Models," Meta AI Technical Report, 2024.

[8] BAAI Research Team, "BGE Embedding Models," 2023. [Online]. Available: https://huggingface.co/BAAI

[9] BAAI Research Team, "BGE Reranker Models," 2023. [Online]. Available: https://huggingface.co/BAAI

[10] LangChain Team, "LangChain Documentation." [Online]. Available: https://python.langchain.com

[11] ChromaDB Team, "Chroma: Open-Source Embedding Database." [Online]. Available: https://www.trychroma.com

[12] Groq Inc., "Groq API Documentation." [Online]. Available: https://console.groq.com/docs

[13] FastAPI Team, "FastAPI Documentation." [Online]. Available: https://fastapi.tiangolo.com

[14] Streamlit Team, "Streamlit Documentation." [Online]. Available: https://docs.streamlit.io