© NOV 2025 | IRE Journals | Volume 9 Issue 5 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I5-1712249

Retrieval-Augmented Generation (RAG)-Based Chatbot System

HIMANK GARG[1], ANISH KUMAR[2], HIMANSHU KUMAR[3], ISHRAT ALI[4], ANUJ CHANDILA[5]
[1, 2, 3, 4, 5]*Department of Data Science (DDCS), GNIOT College, Greater Noida, India*

Abstract- ***The fast development of Large Language Models (LLMs) has highly improved abilities in natural language processing, but deploying LLMs in high-stakes contexts continues to be problematic because of inaccuracy and hallucinations. In this paper, we describe a Retrieval-Augmented Generation (RAG) based chatbot system that avoids these issues from existing models and ensures all responses are grounded and based on valid knowledge sources. The RAG system builds an architecture that scans specific documents in a domain, derives semantic embeddings that include vectors, and stores their vectors in a FAISS vector database, combining retrieval while ensuring speed and memory efficiency. When a question is passed to the RAG bot, the bot retrieves relevant context passages and contexts to condition a generative language model to yield an accurate answer that is also cited. Our implementation involves a modular pipeline and allows all knowledge to be fresh, without having to retrain the language model. We provide experimental results showing significant improvements in factual accuracy compared to baseline LLMs and improved reductions in hallucination. Conclusion: the system is constructed in a domain-free manner allowing further employment in healthcare, legal, or enterprise settings, where it is important to provide cited and verifiable information. The development of this chatbot-status represents an important milestone toward creating trustworthy Artificial Intelligence (AI) systems that exhibit generative fluency and factual reliability.***

Keywords — Retrieval-Augmented Generation, Large Language Models, Semantic Search, FAISS, Chatbot Systems, Hallucination Mitigation, Knowledge Grounding.

I. INTRODUCTION

The rise of Large Language Models (LLMs) marks a significant development in artificial intelligence, enabling exceptional possibilities for natural language understanding and generation. Models such as LLaMA,Claude etc have been shown to perform reliably across a variety of tasks including text summarization, translation, and conversation. However, LLMs show major drawbacks when used in knowledge-intensive environments where factual accuracy is a requirement.

The primary limitation is related to the static nature of LLM knowledge. They are locked as of the last model update, which means they have no means to learn or acquire updated information once deployed. Further, LLMs may create plausible but inaccurate information (hallucinations), which creates other reliability issues in high-stakes applications like healthcare, legal, financial, and enterprise. Finally, LLMs also do not allow for tracking the sources of information, which can weaken trust and adoption.

Retrieval-Augmented Generation (RAG) has shown promise as a framework to address these limitations by incorporating the advantages of two forms of neural retrieval and generative modeling. It enables real-time access to separate external knowledge bases, allowing LLMs to generate language that grounded in trusted knowledge. The RAG framework, therefore, is an improvement over traditional question-answering systems, which were only with extractive techniques or suffered from similar hallucination encounters from generative systems.

This paper introduces a comprehensive implementation of a modular RAG-based chatbot system, which uses an langchain architecture that allows it to be used in a domain specific regions. Our system overcomes significant LLM limitations with dynamic knowledge retrieval and context-aware generation. Key contributions of this work include: (1) an exploration of architecture that allows for many document types to be integrated and updated with knowledge in real-time, (2) implementation of efficient background semantic search using different embedding models and vector databases, (3) exploration of a user-friendly interface with internal citations, (4) a demonstration of iterative questioning and provision of evidence that significantly improved accuracy and reductions in hallucination across multiple domains through empirical results.

© NOV 2025 | IRE Journals | Volume 9 Issue 5 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I5-1712249

II. LITERATURE REVIEW

The integration of retrieval systems with Large Language Models marks a significant evolution in artificial intelligence, building upon decades of research in information retrieval and machine learning. This section surveys the foundational work and recent advancements that inform our approach.

A. Historical Context

Traditional question-answering systems predominantly used extractive approaches, which involved the identification and return of exact text spans in the source documents. These early systems were firmly reliable at answering factoid questions but did not have the flexibility to formulate well-structured explanations. The introduction of neural sequence-to-sequence models led to the notion of generative approaches, although these early models also struggled with factually incorrect and limited knowledge.

The notion of enhancing language models with external knowledge was first articulated as a means to combat these limitations. Initial frameworks employed additional knowledge by enriching inputs to the model with either retrieved documents or knowledge graph entities. However, these frameworks generally treated retrieval and generation as separate processes rather than a set of integrated components.

B. Practical Frameworks and Toolkits

In recent years, open-source frameworks have been created to ease implementing RAG. For example, LangChain has modular components to load documents, split text, integrate vector stores, and construct chains. LlamaIndex has built specialized tools to create and query indices over private data. These frameworks have expanded access to RAG technologies, allowing researchers and developers to quickly prototype and deploy.

C. Domain Specific Applications

RAG architectures have been deployed in multiple domains. In health care, systems have been built to curate evidence-based medical knowledge . In the legal domain, systems have been created to conduct case law research. In the enterprise space, systems have been developed for customer service automation and internal knowledge management. In all of these domains, RAG implementations have shown the benefit of integrating precise retrieval with generative flexibility.

II. METHODOLOGY

A. Document Pre- Processing

The system is able to accept documents of the following formats: PDF, DOCX, and TXT. The parsing of document content will use the libraries PyPDF2 and python-docx to extract the raw text of the document.

The chunking strategy of the text uses a sliding window approach, with a 10-20% overlap between chunks, which gives a balance of context and retrieval accuracy. The chunk size can also be modified depending on multiple factors, including the characteristics and length of the document, and the suggested requirements.

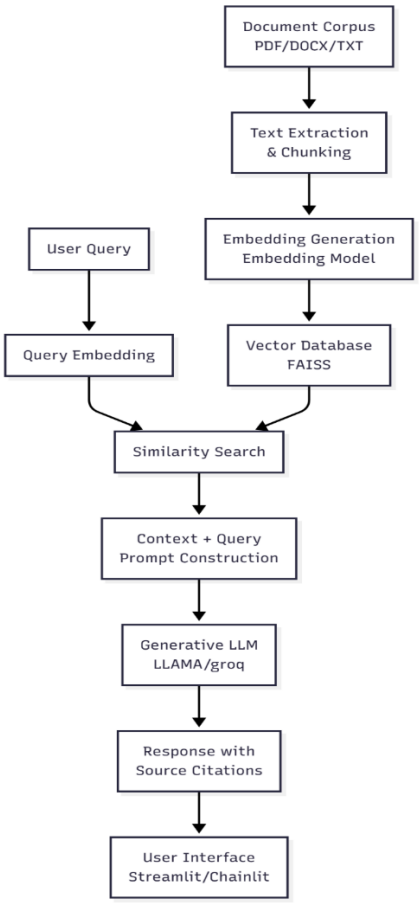
B. Embedding Chunks

Each of the chunks of text is represented as fixed-dimensional vectors using embedding models. The type of embeddings models that were enumerated is shown and described in another section of the paper.

C. Vector Databases

Using FAISS (Facebook AI Similarity Search) index used for efficient nearest-neighbor fast search in high dimensional space [16], the embeddings are stored in the index.

© NOV 2025 | IRE Journals | Volume 9 Issue 5 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I5-1712249



III. RESULTS

A. Performance Analysis

The results of the experiment highlight shows consistent improvement of the performance of the proposed RAG-based system in comparison to the baselines. The RAG-based system outperformed baseline models across all the evaluated domains in terms of factual grounding and overall response quality. Responses produced by the RAG system could be verified directly against the knowledge base..

B. Effects

The most significant benefit of the RAG architecture was its ability to remove model hallucinations effectively. While there was no generation of facts, the retrieval mechanism was the source of grounding constraints that limited the responses based on context retrieved. This resulted in significantly fewer fabricated or unsupported claims in response options across every evaluation domain.

C. Limitations

While effective, a number of limitations were observed. First, the performance of the system relies significantly on the quality and quantity of knowledge base. If there are gaps in the sources, or the responder does not find enough information, then responses will be inherently incomplete. However, the explicit citation mechanism provides transparency as users can view citations and know that there may be limitations to the response. Second, the chunking strategy creates a tension between the need for context persistence and retrieval precision, which required careful tuning to maintain some fidelity and performance.

IV. CONCLUSION

This research demonstrates that the RAG-based architecture enhances factual reliability in conversational AI. Future work maycinclude improved chunking techniques, hybrid retrieval strategies, and model fine-tuning for better alignment with retrieved context.

REFERENCES

[1] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998-6008.

[2] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171-4186.

[3] J. Maynez et al., "On faithfulness and factuality in abstractive summarization," in *Proc. ACL*, 2020, pp. 1906-1919.

[4] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. NeurIPS*, 2020, pp. 9459-9474.

[5] R. McDonald et al., "WikiReading: A novel large-scale language understanding task over Wikipedia," in *Proc. ACL*, 2018.

[6] M. Lewis et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. ACL*, 2020, pp. 7871-7880.

[7] P. Rajpurkar et al., "SQuAD: 100,000+ questions for machine comprehension of text," in *Proc. EMNLP*, 2016, pp. 2383-2392.

[8] V. Karpukhin et al., "Dense passage retrieval for open-domain question answering," in *Proc. EMNLP*, 2020, pp. 6769-6781.

[9] M. Lewis et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. ACL*, 2020, pp. 7871-7880.

[10] K. Guu et al., "REALM: Retrieval-augmented language model pre-training," in *Proc. ICML*, 2020, pp. 345-356.

[11] "LangChain Documentation," [Online]. Available: https://docs.langchain.com/

[12] A. Johnson et al., "Medical question answering with retrieval-augmented generation," in *Journal of Medical Systems*, 2023.

[13] S. Chen et al., "Legal document analysis using hybrid retrieval-generation models," in *Proc. ICAIL*, 2023.

[14] M. Thompson et al., "Enterprise knowledge management with RAG systems," in *Proc. KDD*, 2023.

[15] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. EMNLP-IJCNLP*, 2019, pp. 3982-3992.

[16] M. Douze et al., "Faiss: A library for efficient similarity search," *Journal of Machine Learning Research*, vol. 24, no. 1, pp. 1-6, 2023.