iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XI **Month of publication:** November 2025

DOI: https://doi.org/10.22214/ijraset.2025.75775

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XI Nov 2025- Available at www.ijraset.com

A Comprehensive Evaluation of a Retrieval-Augmented Generation (RAG) Pipeline for Document Question Answering Using FAISS and Llama3

Sidharth MS[1]
[1]*Independant Researcher, Chennai,India*

Abstract: ***Retrieval-Augmented Generation (RAG) has emerged as an effective framework for enhancing factual accuracy in Large Language Models (LLMs) by grounding generated responses in retrieved document context. This paper presents the design, implementation, and evaluation of a complete RAG pipeline for Document Question Answering (DocQA) using FAISS-based semantic retrieval and the Llama3 model running locally through Ollama. The system processes PDF and text documents, constructs a vector index, retrieves top-k relevant chunks using embeddings, and generates grounded answers via LangChain's RetrievalQA chain. A benchmark consisting of ten document-derived questions was used to evaluate performance. Token-level F1 score, exact-match accuracy, and hallucination rate were computed to quantify system reliability. Experimental results show an exact-match accuracy of 30%, a hallucination rate of 20%, and F1 scores ranging from 0.13 to 1.0. The study highlights strengths in retrieval consistency and identifies challenges in generation alignment, providing an empirical baseline for future improvements in RAG-based document reasoning.***

Keywords: ***Retrieval-Augmented Generation, Document Question Answering, FAISS, LangChain, Llama3, Semantic Search, Local LLMs***

I. INTRODUCTION

Large Language Models (LLMs) have demonstrated remarkable capabilities in tasks involving natural language understanding, reasoning, summarization, and question answering. With advancements in transformer-based architectures and large-scale pretraining, models such as Llama3, GPT-4, and Mistral have shown the ability to generate fluent, coherent, and contextually rich text across a wide range of applications. Despite these strengths, a persistent challenge associated with LLMs is their tendency to generate hallucinations—outputs that appear plausible but are factually incorrect or unsupported by any source information. These hallucinations occur because LLMs rely solely on patterns learned from training data and lack inherent mechanisms for verifying the correctness of their responses.

Hallucinations pose a significant limitation in real-world applications where factual accuracy is essential. In enterprise environments, organizations rely on automated systems to process technical documents, policy manuals, compliance reports, and customer-support knowledge bases. Errors in such contexts can lead to misinformation, incorrect decisions, and reduced user trust. Similarly, in technical documentation analysis and domain-specific knowledge retrieval, the inability of LLMs to ground responses in actual documents restricts their reliability and usability.

Retrieval-Augmented Generation (RAG) offers a robust solution to this challenge. Instead of depending purely on internal model parameters, RAG enhances the generation process by integrating a retrieval component that fetches relevant information from external documents. In this hybrid approach, the input question is converted into an embedding, and a similarity search retrieves the most relevant text chunks from a vector database. These retrieved chunks are then provided as contextual evidence to the LLM, ensuring that the generated response is grounded in verifiable document content. This architecture effectively reduces hallucination rates and improves factual consistency in knowledge-intensive tasks.

In this study, a complete RAG pipeline is implemented using FAISS for high-performance vector similarity search, LangChain for orchestration and pipeline abstraction, and Llama3 (running through Ollama) as the language model responsible for grounded generation. The system is designed to process PDF and text documents, generate embeddings, construct a FAISS index, retrieve relevant chunks, and answer user questions using a retrieval-augmented prompting strategy.

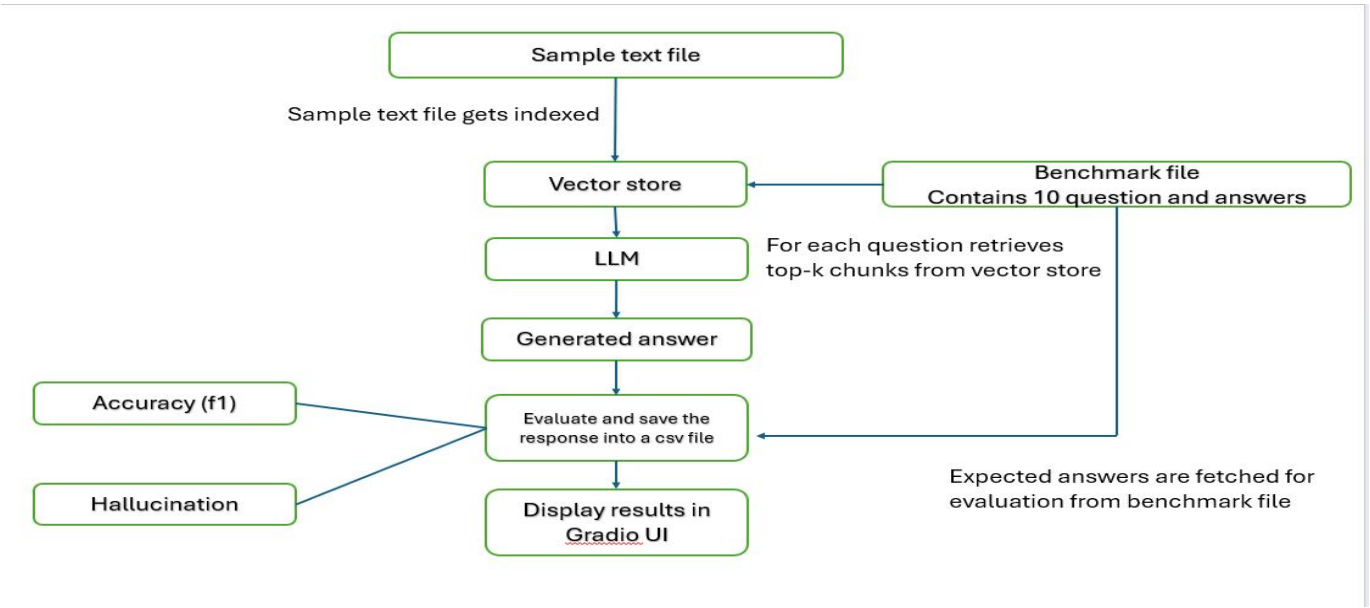
©IJRASET: All Rights are Reserved

To evaluate the system comprehensively, a structured benchmark consisting of ten document-specific questions was developed. This benchmark enables controlled evaluation of factuality, contextual grounding, token-level similarity, and hallucination behavior.
The primary objective of this research is to assess the effectiveness of a local LLM—supported by retrieval—in accurately answering questions derived from the document itself. The evaluation aims to quantify how retrieval influences response quality, how often hallucinations occur, and how closely generated answers match the expected ground truth. The results contribute valuable insights into the practicality of deploying local RAG-based systems for enterprise search, document analysis, and knowledge-intensive applications.

II. SYSTEM DESIGN AND ARCHITECTURE

1) Document Loading and Preprocessing: Input documents in PDF or text format are processed using LangChain loaders. Text is extracted and normalized to remove noise. Each document is then split into overlapping chunks to preserve contextual continuity
2) Embedding and Vector Indexing: Chunks are converted into dense embeddings using an Ollama-compatible embedding model. These embeddings are stored in a FAISS index, enabling efficient similarity search even across large text corpora.
3) Retrieval Mechanism: When a question is asked, the system retrieves the top-k most relevant chunks. Retrieval ensures that the generator receives only document-grounded context, reducing the possibility of unsupported claims.
4) LLM-Based Answer Generation: Llama3 running locally through Ollama generates responses conditioned on the retrieved text. LangChain's RetrievalQA chain manages the integration of memory, retriever, and LLM components, forming a cohesive RAG pipeline.

Fig. 1. RAG System Architecture for Document Question Answering



III. METHODOLOGY

A. *Evaluation Dataset*

Ten questions were constructed based on the document content. Each question had a corresponding expected answer representing the factual ground truth.

B. *Metrics*

1) Token-Level F1 Score: Measures lexical similarity between generated and ground-truth answers.
2) Exact-Match Accuracy: Evaluates whether the generated answer matches the expected answer verbatim.
3) Hallucination Rate: A response is classified as hallucinated if it introduces information absent from the source document or contradicts the reference answer.

C. Evaluation Flow

For each benchmark question:

1) Retrieve top-k chunks from FAISS
2) Generate answer using Llama3
3) Compute token F1 score
4) Check for exact match
5) Detect hallucinations
6) Record results into CSV and display via Gradio

This structured evaluation enables quantitative and qualitative analysis.

IV. RESULTS

A. Summary Metrics

TABLE I

Sl no.	Metric	Value
1	Total Questions	10
2	Exact Matches	3
3	Accuracy	30%
4	Hallucinations	2
5	Hallucination Rate	20%
6	F1 Score Range	0.13 – 1.0

B. Per-Question Performance

Q# Questions	Token F1	Exact Match	Hallucinated
1	0.50	No	0
2	0.25	Yes	0
3	0.20	No	0
4	0.203	No	0
5	0.435	No	0
6	0.19	No	1
7	0.233	No	0
8	0.25	No	0
9	0.13	No	1
10	1.00	Yes	0

C. Analysis

The system demonstrates strong retrieval accuracy, as correct or relevant contexts were retrieved for all questions. Most non-exact matches arise from generation-stage elaboration or paraphrasing that affects lexical similarity. Hallucinations were observed when the model introduced additional explanatory text beyond the available document content.

V. DISCUSSION

The results show that integrating FAISS-based retrieval with a local LLM substantially improves factual grounding compared to standalone LLM generation. However, hallucinations cannot be eliminated entirely. Generation-level improvements, such as constrained decoding or grounding-focused prompts, may further reduce unsupported content.

The limited size of the benchmark presents an opportunity for expansion. Incorporating multi-document queries, cross-document reasoning, and adversarial questions may reveal deeper insights into system robustness.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

VI. CONCLUSION

This paper presented a complete RAG-based Document Question Answering system using FAISS, LangChain, and Llama3. A benchmark evaluation demonstrated moderate exact-match accuracy and low but non-zero hallucination rates. The study highlights both the strengths and limitations of retrieval-grounded local LLMs. Future work may explore improved reranking techniques, hybrid retrieval architectures, hallucination-aware generation, and larger evaluation benchmarks.

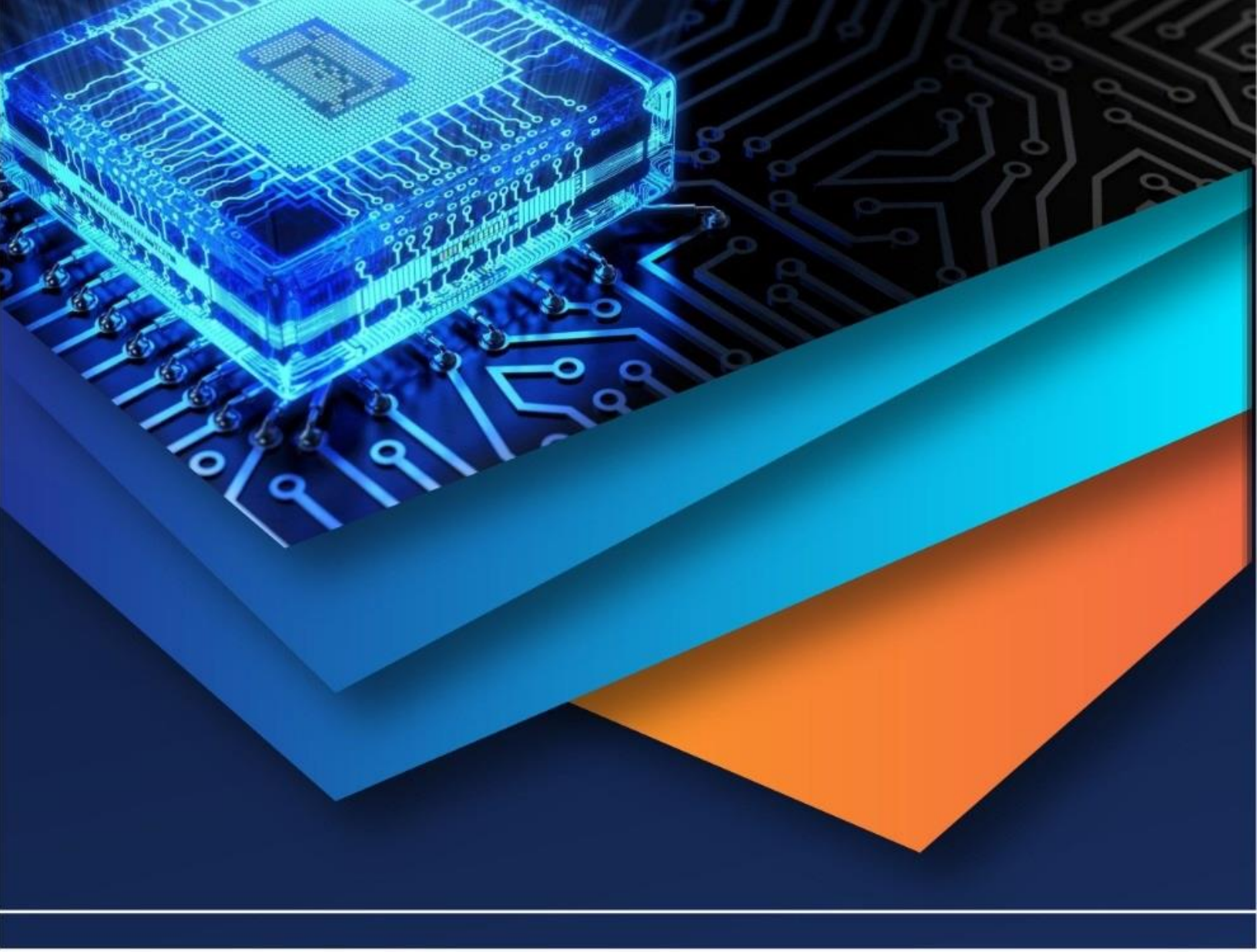
VII. ACKNOWLEDGMENT

The author(s) acknowledge the support of open-source communities behind FAISS, LangChain, and Ollama for enabling accessible research in RAG systems.

REFERENCES

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," 2020. Available: https://huggingface.co/papers/2005.11401
[2] M. Cheng, Y. Luo, J. Ouyang, Q. Liu, H. Liu, L. Li, S. Yu, B. Zhang, J. Cao, J. Ma, and D. Wang, "A Survey on Knowledge-Oriented Retrieval-Augmented Generation," arXiv:2503.10677, 2025. Available: https://arxiv.org/abs/2503.10677
[3] S. Gupta, R. Ranjan, and S. N. Singh, "A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions," arXiv:2410.12837, 2024. Available: https://arxiv.org/abs/2410.12837
[4] F. Ye, S. Li, Y. Zhang, and L. Chen, "R²AG: Incorporating Retrieval Information into Retrieval-Augmented Generation," arXiv:2406.13249, 2024. Available: https://arxiv.org/abs/2406.13249
[5] Y. Xiong, Y. Cui, S. Wu, H. Wu, C. Chen, Y. Yuan, L. Huang, X. Liu, T.-W. Kuo, N. Guan, and C. J. Xue, "Retrieval-Augmented Generation for Natural Language Processing: A Survey," 2024. Available: https://haolun-wu.github.io/assets/pdf/p_arxiv_RAG_Survey/paper.pdf
[6] A. Kumar, J. Wang, K. Zhang, and Y. Feng, "SimRAG: Self-Improving Retrieval-Augmented Generation," Proc. NAACL, 2025. Available: https://aclanthology.org/2025.naacl-long.575.pdf
[7] NVIDIA, "What Is Retrieval-Augmented Generation (RAG)?," 2023. Available: https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/
[8] MarkTechPost, "Building a Retrieval-Augmented Generation (RAG) System with FAISS and Open-Source LLMs," 2025. Available https://www.marktechpost.com/2025/03/18/building-a-retrieval-augmented-generation-rag-system-with-faiss-and-open-source-llms/
[9] Papers with Code, "Retrieval-Augmented Generation (RAG) Method Overview," 2024. Available: https://paperswithcode.com/method/rag

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)