International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

# Lightweight Retrieval-Augmented Generation System for CPU-Only Document Question Answering

**Pratik Halnor, Om Kale, Vishal Gore, Abhishek Kahar, Devyani Jadhav**
Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning)
Sanjivani University,Kopargaon, India.

**Abstract- Retrieval-Augmented Generation (RAG) improves the factual accuracy of Large Language Models by grounding responses in external documents. However, most existing systems rely on dense em-beddings, vector databases, and GPU-based computation, making them unsuitable for low-resource environments. This paper presents a lightweight RAG system designed specifically for CPU-only environments. The system integrates PDF text extraction and Optical Character Recognition (OCR) using PyMuPDF and Tesseract, followed by a keyword-based retrieval mechanism. The retrieved context is then passed to a language model API for response generation. Experimental evaluation demonstrates that the system achieves an accuracy of 83.3% with an average response time of approximately 2.2 seconds. The results highlight that efficient document intelligence systems can be developed without heavy computational requirements.**

**Keywords- Lightweight RAG, Document Question Answering, OCR, CPU-Only Systems, Information Retrieval.**

## I. INTRODUCTION

Document-based question answering has become increasingly important in academic and real-world applications. Large Language Models (LLMs) have improved natural language under-standing significantly, but they often generate incorrect or outdated information when used alone [1].

Retrieval-Augmented Generation (RAG) helps overcome this issue by combining document retrieval with response generation. Many modern RAG systems use embeddings and vector databases such as FAISS [2], [3]. While these methods improve semantic understanding, they require high computational resources.

With the growing need for efficient systems, especially in edge and low-resource environments, lightweight approaches are becoming more relevant [4]. This work presents a CPU-efficient RAG system that avoids heavy dependencies while maintaining practical accuracy.

This work explicitly avoids the use of vector databases and embedding models, focusing instead on a lightweight retrieval mechanism suitable for CPU-only systems.

## II. LITERATURE REVIEW

Recent studies show that RAG significantly improves document understanding and factual accu-racy in question answering systems [5]. Hybrid retrieval methods that combine sparse and dense techniques have also been explored to improve performance [1].

Some works integrate knowledge graphs into RAG pipelines to enhance domain-specific reasoning [6]. Others focus on handling structured data such as tables and images in technical documents [7].

OCR plays a key role in extracting information from scanned documents. Modern approaches combine OCR with AI models to improve text recognition [8], [9]. However, OCR errors can negatively affect downstream tasks such as question answering [10].

Lightweight retrieval approaches are also gaining attention as alternatives to computationally expensive neural methods [11], [12].

## III. RESEARCH GAP

- Most RAG systems rely on vector databases and embeddings [2]
- High computational cost limits deployment in CPU-only systems
- Limited work focuses on lightweight retrieval-based RAG pipelines
- Few systems combine OCR and efficient retrieval in a unified approach

## IV. METHODOLOGY

### A. System Overview

The proposed system consists of document processing, retrieval, and response generation.

### B. Flowchart of Proposed System

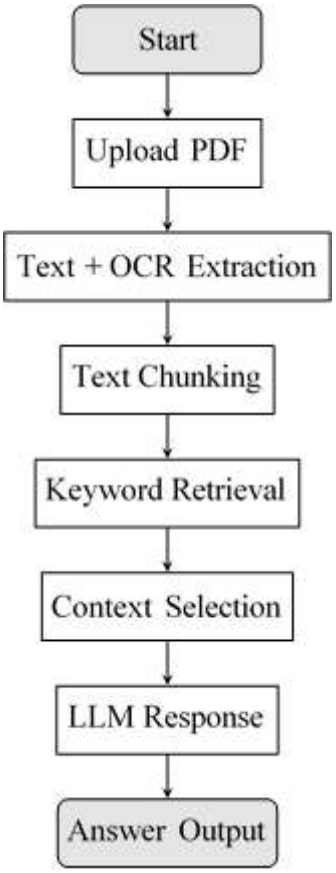


Fig. 1: Flowchart of the proposed lightweight RAG system.

### C. User Interface

The system provides an interactive web-based interface that allows users to upload PDF doc-uments and perform query-based interactions. The interface includes a sidebar for document management and a chat-based interaction panel for question answering.


Fig. 2: User Interface of the Lightweight RAG System showing document upload and query-response interaction

### D. Document Processing

PDF documents are processed using PyMuPDF, and OCR is applied using Tesseract for scanned content.

### E. Retrieval Module

Instead of embeddings, the system uses a keyword-based retrieval approach [11].

$$Score(d_i) = \sum_{w \in Q} freq(w, d_i)$$

### F. Design Choice: Eliminating Vector Databases

Unlike conventional RAG systems that rely on dense embeddings and vector databases such as FAISS, the proposed system adopts a lightweight keyword-based retrieval strategy. This design decision is motivated by the need to reduce computational overhead and enable deployment in CPU-only environments.

Vector databases require embedding generation, indexing, and similarity search operations, which increase memory usage and processing time. In contrast, the proposed system performs direct keyword matching over text chunks, eliminating the need for embedding computation and vector indexing.

This approach significantly reduces system complexity while maintaining acceptable retrieval performance for document-

based question answering tasks. Although semantic retrieval methods provide deeper contextual matching, the results demonstrate that keyword-based retrieval can still achieve effective performance in low-resource scenarios.

### G. Response Generation

The selected content is passed to a language model API to generate answers [1].

## V. RESULTS AND DISCUSSION

### A. Experimental Setup

The system was evaluated on a CPU-only system with no GPU support using 9 document-based queries.

### B. Results

The system achieved an overall accuracy of 83.3%.

TABLE I: System Performance

| Metric | Value |
|---|---|
| Accuracy | 83.3% |
| Average Response Time | 2.2 sec |
| Total Queries | 9 |

### C. Discussion

The results demonstrate that lightweight retrieval methods can effectively support document question answering. The system balances efficiency and performance, making it suitable for deployment in low-resource environments [5].

## VI. CONCLUSION

This paper presents a lightweight RAG system designed for CPU-only environments. By elimi-nating embeddings and vector databases, the system reduces computational requirements while maintaining acceptable accuracy. The integration of OCR improves document accessibility.

Future work will focus on improving retrieval ranking and expanding evaluation datasets.

## REFERENCES

1. L. Lin and X. Zhu, “Adaptive retrieval enhancement for open-domain question answering,” in International Conference on Knowledge Science, Engineering and Management. Springer, 2025, pp. 118–133.
2. M. Douze et al., “The faiss library,” IEEE Transactions on Big Data, 2025.
3. Herlawati et al., “Comparative study of chatbot architectures using llms and rag,” in 2025 International Conference on Informatics and Computing. IEEE, 2025, pp. 1–6.
4. C. O¨ zkan and S. S¸ ahin, “Ai applications in real-time edge processing: Leveraging artificial intelligence for enhanced efficiency,” 2025.
5. W. Ke, Y. Zheng, Y. Li, H. Xu, D. Nie, P. Wang, and Y. He, “Large language models in document intelligence: A comprehensive survey,” ACM Transactions on Information Systems, vol. 44, no. 1, pp. 1–64, 2025.
6. E. Burgin, S. Dutta, and M. Wang, “Quark: Llm-based domain-specific question answering using retrieval augmented generation and knowledge graphs,” in Proceedings of the International Conference on Recent Advances in Natural Language Processing, 2025, pp. 210–217.
7. S. Sobhan and M. A. Haque, “Llm-assisted question-answering on technical documents using structured data-aware retrieval augmented generation,” arXiv preprint arXiv:2506.23136, 2025.
8. K. Yesugade, A. K. Yesugade, S. Kadam, A. Bombe, E. Ghanwat, and S. Gogawale, “Ai-based text extraction from images and pdf documents,” in 2025 IEEE Pune Section International Conference (PuneCon). IEEE, 2025, pp. 1–6.
9. R. Murugan, P. Deivendran, D. S. Kumari, B. Lokesh, P. Nirmal, and S. C. Keerthy, “Ai-powered ocr for handwritten documents with low quality and degradation,” MSRDG International Journal of Computer Scientific Technology & Electronics Engineering, vol. 1, pp. 1–10, 2025.
10. B. Piryani, J. Mozafari, A. Abdallah, A. Doucet, and A. Jatowt, “Evaluating robustness of llms in question answering on multilingual noisy ocr data,” in Proceedings of the 34th ACM International Conference on Information and Knowledge Management, 2025, pp. 2366–2376.
11. K. A. Mekonnen, Y. Tang, and M. de Rijke, “Lightweight and direct document relevance optimization for generative information retrieval,” in Proceedings of the 48th International ACM SIGIR Conference, 2025, pp. 1327–1338.
12. C. M. Nguyen, H. A. P. Tran, P. T. Thai, H. D. Nguyen, and V. T. Nguyen, “Lightweight semantic search for low-

resource languages," in International Conference on Similarity Search and Applications. Springer, 2025, pp. 119–131.