INTELLIGENT DOCUMENT QUESTION ANSWERING SYSTEM USING LARGE LANGUAGE MODELS AND RETRIEVAL-AUGMENTED GENERATION

Abhinanth H.B, Ashva A, B. Tharunesh, P. Sudha

Dhanalakshmi Srinivasan University, Tamil Nadu, India

Abstract

The rapid growth of unstructured digital documents has created challenges in extracting relevant and accurate information efficiently. Traditional keyword-based search systems fail to understand semantic context and often return irrelevant results. Large Language Models (LLMs) have demonstrated strong capabilities in natural language understanding and generation; however, they may produce hallucinated or unverified responses when operating independently. This paper proposes an Intelligent Document Question Answering System that integrates LLMs with Retrieval-Augmented Generation (RAG) to provide accurate, context-aware, and reliable answers grounded in source documents. The system processes documents by generating embeddings, storing them in a vector database, retrieving relevant content through similarity search, and generating responses using a language model. Experimental evaluation shows improved answer relevance, reduced hallucination, and enhanced retrieval accuracy compared to conventional search systems. The proposed system can be applied in enterprise knowledge management, legal analysis, healthcare documentation, and academic research.

Index Terms—LLM, RAG, NLP, Vector Database, Document QA

I. Introduction

The exponential growth of digital content has led to massive collections of unstructured documents such as PDFs, reports, research papers, manuals, and contracts. Extracting specific information from these documents manually is time-consuming and inefficient. Traditional search systems rely on keyword matching, which lacks semantic understanding and contextual reasoning. Recent advancements in Natural Language Processing have introduced Large Language Models capable of generating human-like responses. However, standalone language models are prone to hallucination and lack access to domain-specific or proprietary knowledge bases. To address these limitations, Retrieval-Augmented Generation integrates external document retrieval mechanisms with language models. This paper presents the design and implementation of an Intelligent Document Question Answering System using LLM and RAG. The system retrieves relevant document content using vector similarity search and generates accurate answers grounded in the retrieved context.

II. Related Work

Early document retrieval systems relied on TF-IDF and BM25 ranking mechanisms. While effective for keyword matching, they failed to capture semantic meaning. Neural embedding models later introduced contextual similarity using dense vector representations. Large Language Models such as GPT and LLAMA significantly improved natural language generation. However, these models operate on fixed training data and cannot dynamically access updated documents. Retrieval-Augmented Generation bridges this gap by

combining vector-based retrieval with generative models. Recent studies demonstrate that RAG-based systems improve factual accuracy and reduce hallucination compared to standalone LLMs. Our proposed system builds upon these advancements by integrating document preprocessing, embedding generation, vector storage, and contextual answer generation.

III. Problem Formulation

Let the document collection be represented as D = {d1, d2, ..., dn}. Each document is divided into chunks ci. The objective is to retrieve relevant chunks and generate an accurate answer.

Embedding Representation: E(q), E(ci) ∈ ℝ^k

Cosine Similarity Equation: Sim(q, ci) = (E(q) · E(ci)) / (||E(q)|| ||E(ci)||)

Answer Generation: a = LLM(q, R)

IV. Dataset Description

The dataset comprises 120 documents in PDF and TXT formats. These documents are segmented into approximately 26,400 chunks, each with 500–1000 tokens and 150-token overlap. The embedding dimensions used are 768 or 1024.

Parameter	Value
Documents	120
Formats	PDF, TXT
Chunks	26,400
Embedding Dimension	768 / 1024

V. System Architecture

The proposed system consists of two primary workflows: Offline Document Processing and Online Query Processing. Offline processing includes document collection, text extraction, cleaning, chunking, embedding generation, and storage in a vector database. Online processing involves user query embedding, similarity-based retrieval, context injection into the LLM, and grounded answer generation.

VI. Methodology

The methodology follows a structured pipeline: document collection → preprocessing → chunking → embedding → storage → query processing → retrieval → answer generation. Pre-trained embedding models are used to convert text chunks into dense vector representations. These embeddings are stored in a vector database such as FAISS or ChromaDB. Similarity search techniques like cosine similarity are applied for retrieval. The retrieved document chunks are passed to the LLM as contextual input, which generates responses strictly based on the retrieved context.

VII. Implementation Details

The system is implemented using Python. Key tools and libraries include LangChain for orchestration, FAISS or ChromaDB for vector storage, OpenAI GPT or LLAMA for language modeling, PyPDF2 for document parsing, and Streamlit for the user interface. The backend handles embedding generation and retrieval, while the frontend allows users to upload documents and ask questions interactively.

VIII. Results and Discussion

The system was evaluated using academic PDFs and technical manuals. Evaluation metrics included answer relevance, retrieval accuracy, response time, and hallucination reduction. The RAG-based system demonstrated higher contextual relevance, reduced incorrect answer generation, faster retrieval using vector similarity search, and improved user satisfaction compared to traditional keyword-based systems.

IX. Applications

1. Enterprise Knowledge Management

2. Legal Document Analysis

3. Healthcare Documentation

4. Academic Research Assistance

5. Customer Support Automation

X. Advantages of the Proposed System

• Context-aware responses

• Reduced hallucination

• Scalable architecture

• Efficient document retrieval

• Supports natural language queries

XI. Limitations and Future Work

Despite its advantages, the system has certain limitations. Performance depends on embedding quality, and LLM inference is computationally expensive. The system currently has limited multilingual support. Future work includes multilingual document processing, real-time document updates, hybrid retrieval models, and fine-tuned domain-specific LLMs.

XII. Conclusion

This paper presented an Intelligent Document Question Answering System using Large Language Models and Retrieval-Augmented Generation. By combining semantic retrieval with generative modeling, the system provides accurate and context-aware answers grounded in source documents. Experimental evaluation demonstrates improved reliability and efficiency compared to traditional search methods. The proposed solution offers practical applications across multiple domains and represents a significant advancement in document intelligence systems.

References

[1] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," 2020.

[2] T. Brown et al., "Language Models are Few-Shot Learners," 2020.

[3] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers," 2018.

[4] H. Touvron et al., "LLaMA: Open and Efficient Foundation Language Models," 2023.

[5] OpenAI, "GPT-4 Technical Report," 2023.

[6] IJACSA, "International Journal of Advanced Computer Science and Applications," Vol. 15, No. 3, 2024.