JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

DocXplore: Intelligent Document Retrieval and Summarization

[1]Kinnari Vaghela, [2]Harsh Peke, [3]Dipesh Kathole, [4]Vighnesh Shevti, [5]Dr. Prashant Nitnaware

[1]Student, [2]Student, [3]Student, [4]Student, [5]Head of Department
[1]Information Technology,
[1]Pillai College of Engineering, New Panvel, India.

Abstract : *In the field of information retrieval and document processing, traditional methods for extracting and analyzing PDF content are inefficient, especially when handling multimodal data such as text, tables, and images. To overcome these limitations, Retrieval-Augmented Generation (RAG) has emerged as a promising approach, integrating retrieval-based and generative AI techniques for intelligent search and summarization. Various techniques exist, including TF-IDF, BM25, and traditional NLP-based retrieval, but they struggle with contextual understanding and multimodal data integration. Recent advancements like FAISS, ChromaDB, and CLIP embeddings enable efficient similarity-based retrieval across different data types. However, these methods alone do not generate structured, AI-powered responses. This project, titled DocXplore, will thus make the documents multimodal with the integration of embeddings, OCR (Tesseract OCR), and LLMs (Mistral-7B, Llama 3) to enhance their searchability. By combining text, table, and image processing with RAG-based retrieval, the system provides accurate and context-aware answers, making document navigation more efficient for domains like healthcare, business, legal and policy making.*

Index Terms **– RAG, LLM, Generative AI, NLP, OCR, Information Retrieval, Summarization, Vector database, Multimodal**

I. INTRODUCTION

In today's world, large amounts of data are being stored in unstructured data format pdfs can contain variable combi- nations of text, tables and images. Classic document retrieval systems use keyword-based search approaches like TF-IDF or BM25 that are context-free. The recent developments in Retrieval Augmented Generation (RAG), multimodal embeddings and the Large Language Models (LLM) present an improved alternative for extracting processing and retrieving meaningful information from complex documents

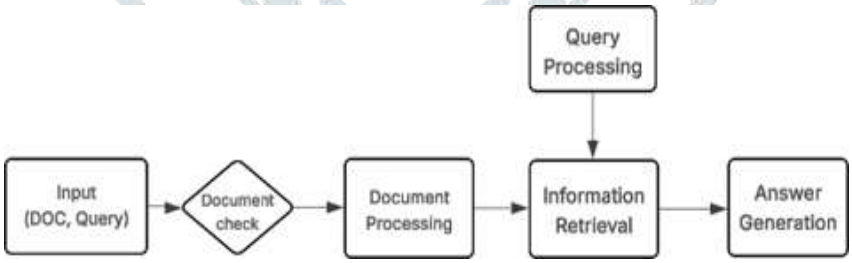

Fig. 1. Document Process Flow Diagram

A. Objectives:

The objective of this work is as follows:

To study the various information retrieval techniques and study their limitations regarding multimodal data such as text tables and images in document processing. To study retrieval-based methods like TF-IDF, BM25, FAISS and ChromaDB and their effectiveness in helping documents be searched. To study how LLM and OCR tools offer the potential for improved document understanding and retrieval information. To build a Retrieval Augmented Generation (RAG) system that would fit the multimodal embeddings and AI search techniques to be context-aware in answers. To apply performance metrics to evaluate the system and demonstrate its efficiency in domains like legal healthcare business and policymaking objectives

B. Scope

The scope of this project extends to DocXplore an RAG- based system developed to support multimodal document processing for efficient extraction retrieval and summarization of information from a pdf containing text tables and images. The system aims to increase the accessibility of documents by enabling users to ask questions and retrieve relevant information via natural language processing (NLP) and Retrieval Augmented Generation (RAG).

1) Project Goals and Deliverables:
Multimodal Data Extraction: Extracts text, tables, and images using OCR (Tesseract/EasyOCR), pdfplumber, and pdf2image.
Efficient Search and Retrieval: Uses vector embeddings (ChromaDB, FAISS, CLIP) for context-aware similarity search.
AI-Powered Answer Generation: Employs LLMs (Mistral- 7B, Llama 3) for generating human-like responses.
User-Friendly Interface: Provides a Streamlit based inter- active chatbot for ease of access.
Scalability: Designed to run locally or on cloud platforms for cost efficiency.

2) Users of the Product:
Healthcare & Researchers - Retrieval of critical patient data and medical reports.
Business Analysts - Analyzing financial reports and market research.
Government & Policy Makers - Searching regulatory policies and public documents.
Journalists - Extracting key details from large reports.
Legal Professionals - Quick access to case laws, contracts,and compliance documents.
By integrating cutting-edge AI and retrieval techniques, DocXplore enhances the efficiency and accuracy of document search, making it valuable across multiple industries.

II. Literature Survey

This chapter reviews applicable methods in multimodal document retrieval and processing. Present strategies are surveyed, their virtues and vices indicated, and innovative developments influencing RAG-based multimodal systems discussed. Evaluation of techniques regarding text, table, and picture-based document retrieval is also presented. Retrieval- based approaches tend to advance document search relevance and accuracy by enriching information extraction, embedded storage, and vector-based matching similarities.

A. Literature review

1) Conversational Text Extraction with Large Language Models Using Retrieval-: Augmented Systems **Year:** 2024
Authors: S. Roy, M. Goswami, N. Nargund, S. Mohanty and P. K. Pattnaik

Large Language Model (LLM) usage in conversational text extraction has exhibited encouraging advancements in docu- ment retrieval and information processing. The method utilizes sentence embeddings and LLM-based retrieval to improve the accuracy and contextual understanding of extracted text. It is especially beneficial in academic, legal, and corporate documents, where AI can be used for question answering, summarization, and knowledge extraction. Nonetheless, the research is mostly concerned with text-based document pro- cessing, and thus its applicability in situations involving multimodal retrieval is limited. The system does not handle images, tables, or other non-textual features, and this limits its capability to offer a full document analysis. This weakness renders it less efficient in industries where documents include a combination of text and visuals, including medical reports, technical manuals, and industrial datasets.

2) Artificial Intelligence Text Processing Based on Retrieval-Augmented Generation. **Year:** 2024 **Authors:** Bogdan-S. Posedaru, Florin-V. Pantelimon, Mihai-N. Dulgheru, Tiberiu- M. Georgescu

Artificial Intelligence (AI) text processing based on Retrieval-Augmented Generation (RAG) has been extensively investigated to improve accuracy and relevance for business and educational use cases. This method relies on embeddings, ChatGPT, LangChain, and vector databases to fetch and generate semantically relevant text-based responses. By combining retrieval capabilities with generative AI, the system is able to provide more contextually relevant responses, enhancing efficiency in document analysis, automated summarization, and knowledge retrieval. However, a significant limitation of this approach is its inability to process multimodal data. The system does not establish relationships between text and images, making it ineffective for applications requiring image- text alignment and document analysis that involves visual elements.

3) Beyond Text: Optimizing RAG with Multimodal Inputs for Industrial Applications **Year:** 2024 **Authors:** Monica Riedler, Stefan Langer

To extend the capabilities of RAG beyond text, researchers have optimized it with multimodal inputs for industrial appli- cations. This approach enhances retrieval accuracy by incorporating GPT-4 Vision, LLaVA, CLIP, OpenAI embeddings, LangChain, and ChromaDB, enabling the system to process both text and images. The integration of multimodal AI significantly improves automated visual inspections, industrial automation, and quality control, where images play a critical role. However, while this system excels in image retrieval, it faces challenges in extracting meaningful relationships between images and textual descriptions. The lack of deep semantic alignment between different data formats limits its effectiveness in cases where text-image interactions are crucial for accurate analysis and decision-making.

4) Robust Multimodal RAG Pipeline for Documents Containing Text, Tables, and Images **Year:** 2024 **Authors:** Pankaj Joshi, Aditya Gupta, Pankaj Kumar, Manas Sisodia

To address the challenges of multimodal retrieval, a Robust Multi-Model RAG Pipeline has been proposed for processing documents containing text, tables, and images. This system leverages OpenAI models, Gemini, and Quadrant DB to en- hance document chunk retrieval and facilitate better knowl- edge extraction from complex data formats. By integrating multiple AI models, the system improves text comprehension, structured data processing, and multimodal content retrieval, making it highly useful in domains like finance, legal research, and technical documentation. However, despite its advance- ments, the system faces difficulties in semantic retrieval, struggling to establish meaningful connections between text, tables, and images. This limitation hinders its ability to fully contextualize multimodal content, highlighting the need for further improvements.

5) Building Customized Chatbots for Document Summarization and Question Answering using Large Language Models using a Framework with OpenAI, Lang chain, and Streamlit **Year:** 2024 **Authors:** Sangita Pokhrel, Swathi Ganesan, Tasnim Akther, Lakmali Karunarathne

This research proposes a comprehensive framework to build custom chatbots driven by large language models (LLMs) for document summarization and answering user queries. The framework, based on technologies such as OpenAI, LangChain, and Streamlit, arms users to tackle information overload by quickly summarizing their insights from such lengthy documents. In this research, the architecture, implementation, and practical applications of the framework were discussed, emphasizing enhancing productivity and information retrieval. Using a step-by-step approach, this study has shown developers how they can use the framework to create end-to-end document summarization and question-answering applications while the proposed framework has shown great promise, many limitations are still present. First, the dependence on OpenAI API creates a dependency on third-party services that may introduce cost, data privacy, and accessibility concerns. Second, the quality and formatting of input documents could affect model performance and ultimately impact the accuracy of summaries and answers. Third, the current system might not effectively deal with domain-specific terminology or multilinguality unless it undergoes some fine-tuning or additional training. Finally, real-time performance may be hindered based on document size and system resources, thus causing a poor user experience in applications with large scales or time-sensitive conditions.

B. Summary of Literature Survey

This survey identifies gaps in existing retrieval and AI- based document processing methods. While multimodal RAG techniques improve search accuracy, they require large-scale training and high processing power. To address these issues, this project DocXplore integrates text, table, and image retrieval with LLM-powered query synthesis, optimizing docu- ment interaction and search efficiency.

III. System Architecture

A. Existing System Architecture

Traditionally, document search systems have relied on keyword-based techniques like TF-IDF and BM25. While these are useful for basic text searches, they fall short when it comes to handling content that includes images, tables, or other non-textual elements. Although modern vector search tools like FAISS and ChromaDB offer improved relevance through embedding-based search, they don't produce structured, AI-generated responses, nor do they fully integrate multimodal data.

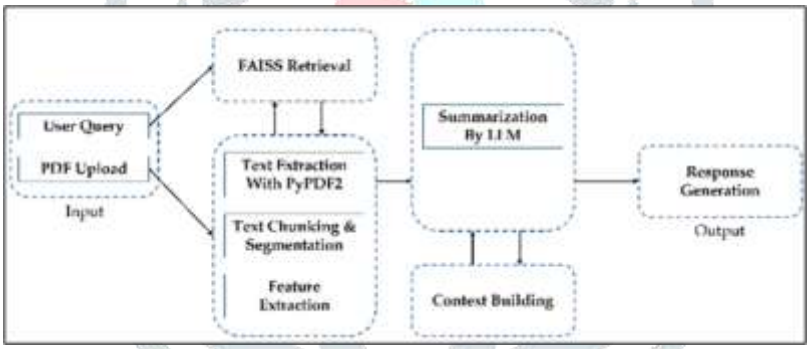

Fig. 2. Existing System Architecture

In this existing system, the user will first upload a pdf. This pdf will be processed by various techniques such as text extraction, chunking, segmentation and feature extraction. Meanwhile, the user will ask a query to the system. Then the relevant information will be extracted from the pdf using FAISS retrieval. Then through context building and Summarization with the help of LLM models response will be generated.

Limitations

• Lack of multimodal processing: The system primarily focuses on text-based retrieval and does not efficiently process images or tables.
• Limited contextual understanding: While FAISS retrieval helps in searching documents, it lacks deeper contextual understanding required for intelligent responses.
• Basic feature extraction: The current approach does not leverage advanced embeddings (such as CLIP or OpenAI embeddings) for improving retrieval accuracy.

B. Proposed System Architecture

The proposed system is an improvement upon the traditional Retrieval-Augmented Generation (RAG) concept, supporting retrieval of PDF and CSV document-based information through text, tables, and images. The limitations of the previous system are addressed by the integration of multimodal embeddings and hybrid retrieval.

1) Workflow

The workflow consists of the following components:

1. Document Upload & Preprocessing
• The system enables users to upload PDFs or CSV files.
• A document checker identifies whether the uploaded file is a PDF or CSV and routes it accordingly.

2.Multimodal Data Extraction
For PDFs:
• Image Retrieval: Extracts images from PDFs, generates image embeddings, and stores them in ChromaDB.
• Text Retrieval: Extracts raw text, converts it into text embeddings, and stores it in ChromaDB.
• Table Retrieval: Identifies tabular data, extracts structured content, converts it into table embedding and stores it in ChromaDB.
For CSVs:
• Only table retrieval is performed, with table embeddings stored in ChromaDB.

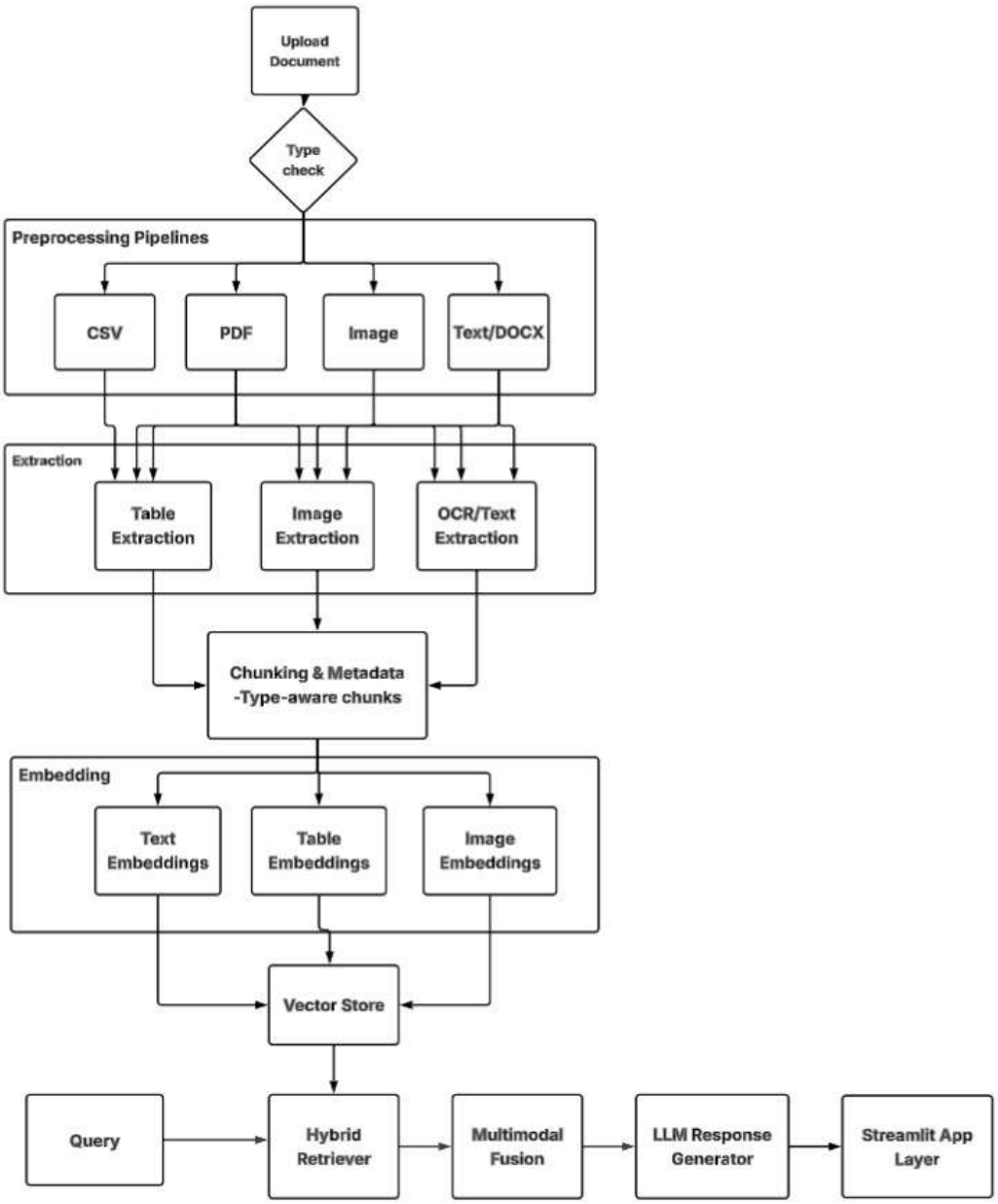

Fig. 3. Proposed System Architecture

3. Query Processing & Hybrid Retrieval
• The system processes the user's query and retrieves relevant text, images, and tables from ChromaDB.
• Hybrid retrieval combines multiple modalities (text, tables, images) to provide contextually rich search results.
4. Multimodal Fusion & Response Generation
• The retrieved data undergoes multimodal fusion, ensuring seamless integration of text, tables, and images.
• A multimodal LLM processes the fused data to generate structured, informative, and context-aware responses.
• The final response is generated based on relevant extracted information.

2) Key Advantages of the Proposed System
• Multimodal Support: Enables retrieval and processing of text, tables, and images, unlike the existing system, which is primarily text-based.
• Efficient Retrieval: Uses ChromaDB for similarity based retrieval, improving speed and accuracy.
• Hybrid Retrieval Mechanism: Ensure better contextual understanding by combining multiple modalities.
• Enhanced User Experience: Generate more comprehensive, accurate, and structured responses.

3) Sample Dataset Used
To support the development of a document summarization and question-answering system, a diverse and representative dataset was curated from various domains. The documents used include:

• Healthcare PDFs: Clinical guidelines, patient reports, medical journals, and pharmaceutical documents.
• Legal Papers: Case laws, legal judgments, policy briefs, and regulatory documents.
• Educational Paper: Lecture note, textbooks, curriculum material
• Research Articles: Peer-review scientific publications publications and research summaries from multiple disciplines.

C. Hardware and Software Specifications

The experiment setup is carried out on a computer system which has the different hardware and software specifications as given respectively

Processor	Intel Core i7 (2.8 GHz)
Storage(SSD)	512 GB
RAM	16 GB
GPU(optional)	NVIDIA RTX 3090

Table 1. Hardware Specification

Operating System	Windows / Linux
Programming Language	Python 3.10
Libraries	pdfplumber, PyMuPDF, pdf2image pandas, TesseractOCR, hugging face, chromaDB, CLIP, Streamlit
Database	ChromaDB (for storing embeddings)

Table 2. Software Specifications

IV. Applications

DocXplore can be helpful in many real-life situations where people need to work with a lot of documents. Here's how it can be used:
• Business and Corporate Use: It can quickly go through long business reports, financial documents, and market research, and give you the important points - saving time and effort for business analysts and managers.
• Students and Researchers: If you're doing a project or thesis, DocXplore can help by summarizing big research papers, organizing useful information, and making it easier to understand complex topics.
• Education: Teachers and students can use it to pull out key points from textbooks or lecture materials. It makes studying and preparing for lessons much faster.
• Human Resources (HR): HR teams often deal with many resumes and reports. DocXplore helps by picking out important details from these documents so they can make better hiring or management decisions.
• Legal Work: Lawyers and legal teams can use it to quickly find specific laws, contract details, or case studies from big piles of legal documents. It saves them hours of searching.
• Healthcare: Doctors and healthcare professionals can use it to understand patient reports, medical research, and clinical data more easily, helping them make better decisions.
In conclusion, DocXplore helps people find what they're looking for in long, complicated documents - whether it's text, tables, or images - quickly and smartly.

V. Summary

This report documents the extensive investigation conducted into the diverse document retrieval and processing methodologies being implemented in RAG-based multimodal systems. Several methods were discussed and compared, namely content-based retrieval, embedding-based search, and hybrid recommendation methods. The limitations of classical methods, such as TF-IDF and BM25, were addressed, along with how vector embeddings, multimodal processing, and AI-assisted retrieval systems allow for gains in accuracy and efficiency.
DocXplore is a system that integrates innovations in text, table, and image extraction, semantic search via FAISS/ChromaDB, and AI answers via LLM inference (Mistral-7B, Llama 3). The hybrid retrieval enhances itself by means of multimodal embedding, structured document processing, and contextual summarization. The comparative study of various retrieval methods shows that RAG-based retrieval is much superior to classical methods.
Overall, this report offers insight into state-of-the-art document retrieval techniques, hybrid AI systems, and their real-world applications, and how DocXplore improves search efficiency, document accessibility, and AI-based query answers

VI. References

1. S. Roy, M. Goswami, N. Nargund, S. Mohanty and P. K. Pattnaik, "Conversational Text Extraction with Large Language Models Using Retrieval-Augmented Systems," 2024 6th International Conference on Computational Intelligence and Networks (CINE), Bhubaneswar, India, 2024, pp. 1-6, doi: 10.1109/CINE63708.2024.10881808.
2. TY - JOUR AU - Posedaru, Bogdan-Stefan AU - Pantelimon, Florin AU - Dulgheru, Mihai-Nicolae AU - Georgescu, Tiberiu-Marian PY - 2024/07/03 SP - 209 EP - 222 T1 - Artificial Intelligence Text Processing Using Retrieval-Augmented Generation: Applications in Business and Education Fields VL - 18 DOI - 10.2478/picbe-2024-0018 JO - Proceedings of the International Conference on Business Excellence ER -

3. Riedler M, Langer S. Beyond Text: Optimizing RAG with Multimodal Inputs for Industrial Applications. arXiv preprint arXiv:2410.21943. 2024 Oct 29. DOI - https://arxiv.org/abs/2410.21943
4. P. Joshi, A. Gupta, P. Kumar and M. Sisodia, "Robust Multi Model RAG Pipeline For Documents Containing Text, Table & Images," 2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2024, pp. 993-999, doi: 10.1109/ICAAIC60222.2024.10574972.
5. TY - JOUR AU - Pokhrel, Sangita AU - Ganesan, Swathi AU - Akther, Tasnim AU - Mapa Senavige, Lakmali Shashika Karunarathne PY - 2024/04/08 SP - 70 EP - 86 T1 - Building Customized Chatbots for Document Summarization and Question Answering using Large Language Models using a Framework with OpenAI, Lang chain, and Streamlit VL - 6 DOI - 10.36548/jitdw.2024.1.006 JO - Journal of Information Technology and Digital World ER -
6. A. Oussaleh Taoufik and A. Azmani, "AI-Enhanced Techniques for Extracting Structured Data from Unstructured Public Procurement Documents," 2024 8th International Symposium on Innovative Approaches in Smart Technologies (ISAS), İstanbul, Turkiye, 2024, pp. 1-8, doi: 10.1109/ISAS64331.2024.10845583.
7. H. Yang, M. Zhang and D. Wei, "IRAG: Iterative Retrieval Augmented Generation for SLU," 2024 20th IEEE International Colloquium on Signal Processing & Its Applications (CSPA), Langkawi, Malaysia, 2024, pp. 30-34, doi: 10.1109/CSPA60979.2024.10525270. keywords: {Training;Predictive models;Decoding;Iterative methods;Iterative decoding;Automatic speech recognition;Spoken Language Understanding;Retrieval Augmented Generation;Large Language Models}
8. Haohao Luo, Ying Shen, and Yang Deng. 2023. Unifying Text, Tables, and Images for Multimodal Question Answering. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 9355–9367, Singapore. Association for Computational Linguistics.
9. Y - BOOK AU - Barochiya, Madhuri AU - Makhijani, Pratishtha AU - Patel, Hetul AU - Goel, Parth AU - Patel, Bankim PY - 2024/12/04 SP - 69 EP - 75 T1 - Evaluating RAG Pipeline in Multimodal LLM-based Question Answering Systems VL - DOI - 10.1109/ICACRS62842.2024.10841620 ER -
10. U. Vallabhaneni, Y. Wutla, T. Dichpally, V. R. Reddy Ch, M. R. Gone and P. L. Kumari, "Mining Mate: A Chat Bot for Navigating Mining Regulations Using LLM Models," 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2024, pp. 888-892, doi: 10.1109/ICACCS60874.2024.10716988.
11. H. S, A. Paramarthalingam, S. Sundaramurthy and S. Cirillo, "A Study on Word Embeddings in Local LLM-based Chatbot Applications," 2024 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), Sakhir, Bahrain, 2024, pp. 232-237, doi: 10.1109/3ict64318.2024.10824498.
12. TY - JOUR AU - M.Sapkal, Kunal AU - R.Gharge, Vinayak AU - J.Kadam, Sanika AU - N.Mulik, Rutuja AU - Bhosale, Prof PY - 2024/09/01 SP - 129 EP - 134 T1 - VEDA-VISION GPT-AN AI-POWERED MULTILINGUAL DOCUMENT PROCESSING AND INTERACTION PLATFORM VL - 09 DOI - 10.33564/IJEAST.2024.v09i05.015 JO - International Journal of Engineering Applied Sciences and Technology ER -
13. M. Barochiya, P. Makhijani, H. N. Patel, P. Goel and B. Patel, "Evaluating RAG Pipeline in Multimodal LLM-based Question Answering Systems," 2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS), Pudukkottai, India, 2024, pp. 69-75, doi: 10.1109/ICACRS62842.2024.10841620.
14. T. B.M, R. H.V, R. S, R. Salam and M. Pai, "TextVerse: A Streamlit Web Application for Advanced Analysis of PDF and Image Files with and without Language Models," 2024 Asia Pacific Conference on Innovation in Technology (APCIT), MYSORE, India, 2024, pp. 1-6, doi: 10.1109/APCIT62007.2024.10673559.
15. Y. Kong et al., "Document Embeddings Enhance Biomedical Retrieval-Augmented Generation," 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Lisbon, Portugal, 2024, pp. 962-967, doi: 10.1109/BIBM62325.2024.10822781.