STM JOURNALS

Journal of
Computer Technology & Applications

ISSN: 2229-6964 (Online)
ISSN: 2347-7229 (Print)
Volume 16, Issue 1, 2025
January–April
DOI (Journal): 10.37591/JoCTA

https://journals.stmjournals.com/jocta

Research JoCTA

Developing a RAG-PDF Reader Using Instructor XL and Falcon 7B

Daisy Vyas, Prachi Sharma, Saroj Prajapat*, Saumya Mittal

Abstract

This study outlines the development of a Retrieval-Augmented Generation (RAG) application, designed to efficiently extract, retrieve, and synthesize insightful responses from complex PDF documents. Leveraging advanced models like Instructor XL for generating high-quality semantic embeddings and Falcon 7B for sophisticated language generation, this system provides a robust solution for document comprehension in academic, research, and professional environments. By implementing efficient PDF text processing, embedding storage with FAISS for rapid similarity-based retrieval, and real-time response generation, the application transforms unstructured data into accessible, contextually accurate information. A user-friendly interface, built with Streamlit, enables seamless document interaction, allowing users to upload PDFs, submit queries, and receive accurate, coherent responses in real-time. This study also discusses potential enhancements for scalability across multiple domains, including multilingual support, making the system a versatile tool for research, education, and industry applications. Key features include enhanced document interaction, efficient information retrieval and generation, providing significant value in data-rich environments where precise information access is critical.

Keywords: Application, development, *RAG,* NLP, LLMs, Instructor XL, Falcon 7B

INTRODUCTION

In recent years, advances in Natural Language Processing (NLP) have driven transformative changes across a wide range of applications, from conversational agents and chatbots to complex text summarization and automated content generation. At the forefront of these advancements are Large Language Models (LLMs), powerful architectures designed to generate human-like responses, analyse language, and perform nuanced reasoning tasks. LLMs like OpenAI's GPT-3, Google's BERT, and Meta's Falcon 7B have redefined our understanding of what machines can achieve in language understanding and generation [1]. Trained on extensive datasets, these models exhibit an impressive ability to handle various language-related tasks, positioning them as essential tools for industries and researchers alike [2].

However, despite the significant capabilities of LLMs, they face inherent limitations. One of the primary challenges lies in the models' reliance on their training data [3]. LLMs are generally trained on static datasets, which means they lack real-time knowledge updates and access to constantly evolving information [4]. As a result, they are unable to respond effectively to queries that require up-to-date or highly specific information beyond their training set. For instance, querying a traditional LLM about the latest developments in a specialized field, such as legal or medical documentation, can yield limited or outdated information [5]. Furthermore, LLMs struggle with the complexity of handling vast, heterogeneous

***Author for Correspondence**
Saroj Prajapat
E-mail: sarojprajapat21.set@modyuniversity.ac.in

Student, Department of Computer Science and Engineering, Mody University of Science & Technology, Laxmangarh, Rajasthan, India

Received Date: November 13, 2024
Accepted Date: December 10, 2024
Published Date: December 31, 2024

Citation: Daisy Vyas, Prachi Sharma, Saroj Prajapat, Saumya Mittal. Developing a RAG-PDF Reader Using Instructor XL and Falcon 7B. Journal of Computer Technology & Applications. 2025; 16(1): 45–49p.

databases or document repositories, where critical information may be embedded within long, unstructured text [6].

To overcome these limitations, researchers have developed Retrieval-Augmented Generation (RAG) frameworks, which enhance LLMs by integrating a retrieval mechanism [7]. The RAG architecture addresses the static knowledge limitation by adding a retrieval component that can search large databases or text corpora in real time, extracting relevant information to answer a user's query more accurately. This architecture represents a hybrid approach: while retrieval enables the model to locate and filter through relevant information, the generation component, powered by LLMs, crafts responses that are not only factually accurate but also contextually coherent and human-like [8]. RAG, therefore, allows for a dynamic interaction with both static and constantly updated datasets, offering a robust solution for complex, information-dense environments [9].

This study presents the development of a RAG-based PDF reader that is specifically designed to simplify and enhance document analysis [10]. Our application combines the power of two advanced models: Instructor XL and Falcon 7B. Instructor XL is a state-of-the-art embedding model known for its ability to capture rich semantic meanings in text, making it ideal for generating dense vector representations of PDF content. By creating these vectors, Instructor XL allows for effective information retrieval within large and complex document datasets. Falcon 7B, on the other hand, serves as the language generation model. With its impressive generative capabilities, Falcon 7B synthesizes responses that are clear, contextually relevant, and engaging for the user. This combination allows our system to provide users with precise, articulate answers to queries about PDF content, making it particularly valuable for professionals in fields like research, law, and education, where large volumes of unstructured data are prevalent.

Our project aims to streamline the interaction between users and PDF documents by enabling a RAG-powered platform where users can upload complex documents, pose queries, and receive insightful responses within seconds. By leveraging Instructor XL's embedding capabilities and Falcon 7B's generative prowess, this PDF reader not only retrieves relevant information but also constructs coherent and user-friendly responses. Furthermore, we have developed an intuitive user interface using Streamlit, allowing users to upload PDF documents, type in their questions, and receive immediate answers. This interface bridges the gap between the RAG model's technical depth and the end user's need for simplicity and accessibility, ensuring that both technical and non-technical users can benefit from the system's advanced functionalities.

What is RAG?

RAG blends information retrieval and language generation by first identifying relevant content within a large corpus, then feeding it to a language model for a synthesized response. Traditional retrieval techniques like TF-IDF, while effective, often lack context-sensitive interpretation. RAG's LLM integration allows it to generate contextually accurate answers, proving particularly valuable in complex information retrieval scenarios such as legal, medical, or technical document analysis.

Explanation of LLMs

LLMs are sophisticated architectures trained on diverse datasets, capable of performing complex NLP tasks. However, they are limited to their training data, which can become outdated. Through RAG, the LLM retrieves relevant, up-to-date data, leading to more precise and context-aware outputs. This integration addresses limitations in LLMs, making them suitable for real-time, information-dense environments.

Models Used

Instructor XL

Instructor XL generates high-dimensional vector embeddings of textual data, capturing semantic relationships crucial for effective retrieval. For this project, it translates PDF text chunks into searchable vector representations, enhancing search precision.

Falcon 7B

Falcon 7B is an open-source LLM with 7 billion parameters, designed for nuanced response generation. Its architecture is optimized to deliver high-quality, contextually rich responses, which makes it ideal for this application's requirements.

METHODOLOGY

Our approach focuses on integrating PDF processing, embedding generation, retrieval, and response generation through the following steps:

- PDF Text Extraction and Preprocessing Using libraries like PyPDF2 and pdfplumber, the text is segmented for embedding. These segments are vectorized via Instructor XL, making them searchable by relevance.
- *Embedding model:* Instructor XL: Instructor XL generates vector representations that capture the meaning of each text chunk, crucial for the next steps of indexing and searching.
- *Language model:* Falcon 7B: Once relevant content is identified, Falcon 7B takes over, synthesizing responses that are coherent, accurate, and human-like.
- *Embedding storage and retrieval using FAISS:* FAISS is employed to manage and quickly search embeddings, facilitating real-time query responses.
- *User-friendly interface:* We developed a user interface with Streamlit that allows PDF uploads, query input, and response display, making the system accessible and user-friendly.

IMPLEMENTATION STEPS

Data Preparation and Embedding

To enable accurate and efficient retrieval, the PDF content is first pre-processed and divided into smaller, manageable chunks. These text chunks are then passed through Instructor XL, which generates dense semantic embeddings that capture the contextual meaning of the content. These embeddings, serving as compact representations of the text, are subsequently stored in FAISS, a vector database optimized for similarity-based searches. By storing the embeddings in this format, the system ensures that relevant information can be quickly retrieved based on user queries, even within large documents.

Efficient Retrieval

FAISS (Facebook AI Similarity Search) plays a critical role in handling large-scale embedding data, allowing the system to perform rapid and accurate similarity-based searches. When a user submits a query, FAISS enables quick retrieval of the most relevant text embeddings, connecting user questions to the most pertinent sections of the document. This process significantly enhances response time and ensures that even complex or nuanced queries are matched with appropriate document content, making the retrieval process both effective and efficient.

Response Generation

Once the most relevant text chunks are retrieved, they are processed by Falcon 7B, a large language model capable of producing coherent and contextually relevant responses. Falcon 7B synthesizes the retrieved information into a single, articulate answer, ensuring the response is not only accurate but also easy for the user to understand. The model's advanced generative capabilities allow it to handle complex questions and provide insightful, human-like responses, making the interaction with the system more intuitive and productive.

User Interaction Interface

To facilitate easy access and interaction, we developed a user-friendly interface using Streamlit, an open-source app framework for machine learning and data science applications. This web-based interface enables users to upload PDF documents, submit their queries, and receive responses in real time. The intuitive design and smooth functionality of the interface ensure an accessible and seamless experience, making it suitable for users across various industries and technical backgrounds. By integrating document upload, query input, and answer retrieval into a single platform, the interface

enhances the overall usability of the system, providing a streamlined and interactive environment for document analysis.

RESULT AND EVALUATION

Our system has demonstrated strong performance in effectively answering complex document queries, with Instructor XL significantly enhancing retrieval accuracy and Falcon 7B generating coherent, contextually relevant responses. Key metrics, including accuracy, relevance, and response time, were carefully evaluated, and the results indicated consistently high levels of precision and coherence in the answers provided. The model's ability to retrieve relevant information quickly and generate well-structured responses highlights its effectiveness in diverse document types, ranging from academic papers and technical documentation to legal texts and research reports. This adaptability underscores the system's broad applicability and its substantial potential as a valuable tool for handling complex, unstructured content across various fields.

DISCUSSION

Our application highlights the significant value that Retrieval-Augmented Generation (RAG) systems bring to environments where nuanced and thorough document analysis is essential. By integrating Instructor XL and Falcon 7B with FAISS, our RAG-based PDF reader achieves rapid, contextually accurate retrieval and generation, setting a new standard for interaction with complex, unstructured text. Instructor XL's semantic embedding capabilities allow the system to parse and organize document content into a structured, searchable format, while Falcon 7B's language generation enables the creation of clear, relevant, and human-like responses. This combination allows users to access critical information quickly and reliably, transforming how they engage with intricate text sources.

The application is especially valuable in fields where precise information extraction is crucial, such as law, medicine, research, and academia. In legal research, for example, where understanding specific clauses and precedents is vital, the PDF reader can extract highly relevant sections from lengthy legal documents, providing professionals with the insights they need to make informed decisions. In the medical field, the system could assist healthcare professionals by retrieving relevant findings, case studies, or treatment protocols from medical research documents, thereby supporting evidence-based practice and improving patient outcomes.

Moreover, our approach represents a scalable and adaptable solution for any domain that depends on timely and accurate information retrieval. The use of FAISS ensures that our system can handle large volumes of data efficiently, making it suitable for real-time applications in industries where speed and reliability are paramount. FAISS's indexing capabilities enable rapid similarity search, allowing the system to match user queries with the most relevant document sections in seconds. This efficiency is particularly advantageous for users who require immediate responses to complex queries, such as researchers working under tight deadlines or corporate teams preparing reports.

Future improvements to the system could involve supporting additional document types beyond PDFs, such as Word documents, PowerPoint presentations, and HTML-based web content, broadening its usability across different formats. Expanding the system to accommodate multiple languages would also enhance accessibility, enabling users across various linguistic backgrounds to benefit from the platform's capabilities. Additionally, optimizing the user interface for industry-specific use cases could further tailor the experience, with customizations for fields like legal research, financial reporting, or academic analysis. This would allow the application to provide a more targeted and intuitive user experience, catering to the unique needs of professionals in different domains.

In summary, our application demonstrates the transformative potential of RAG-based systems for document analysis, providing a streamlined and intelligent solution for users who need fast, accurate, and contextually aware responses. By continuing to expand its functionality and refine its capabilities,

this system has the potential to become an invaluable tool across various industries, ultimately bridging the gap between raw data and meaningful insights.

CONCLUSION

This study introduces a cutting-edge Retrieval-Augmented Generation (RAG)-based PDF reader that combines the advanced capabilities of Instructor XL and Falcon 7B to streamline document interaction. By integrating powerful semantic embedding and language generation technologies, our application addresses the challenges associated with analysing complex, unstructured documents, providing users with an accessible, real-time tool for extracting and understanding critical information. Instructor XL's robust embedding capabilities enable the system to process and index vast amounts of text, allowing for precise retrieval of relevant information, while Falcon 7B's generative strengths facilitate the creation of coherent, contextually accurate responses. Together, these models lay the foundation for an efficient and intuitive approach to document comprehension, with broad implications for industries that rely heavily on detailed document analysis, such as research, education, law, and healthcare.

Our work highlights the potential of RAG-based systems to transform document-based workflows, empowering users to interact with complex information sources in a more meaningful way. This application demonstrates significant progress toward making high-quality NLP tools accessible to users with varying levels of technical expertise. Additionally, the system's user-friendly interface built with Streamlit enhances usability, making it easy for users to navigate large PDF files, pose questions, and retrieve insights in real time.

Future efforts on this project will concentrate on several crucial aspects to improve its flexibility and influence. First, refining the model's performance through additional fine-tuning and optimization will improve response accuracy and processing speed. Another priority is to enhance language support, ensuring that users from diverse linguistic backgrounds can effectively utilize the system. Lastly, enhancing scalability to handle larger datasets and more specialized domains will open up new possibilities for application across industries. By advancing in these areas, this RAG-based PDF reader has the potential to become an indispensable tool for efficient, comprehensive document analysis in both professional and academic contexts.

REFERENCES

1. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, *et al*. Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv Neural Inf Process Syst. 2020; 33: 9459–74.
2. Izacard G, Grave E. Leveraging passage retrieval with generative models for open domain question answering. arXiv Preprint arXiv:2007.01282. 2020 Jul 2.
3. Almazrouei E, Alobeidli H, Alshamsi A, Cappelli A, Cojocaru R, Debbah M, et al. The Falcon series of open language models. arXiv Preprint arXiv:2311.16867. 2023 Nov 28.
4. Chen M, Tworek J, Jun H, Yuan Q, Pinto HP, Kaplan J, et al. Evaluating large language models trained on code. arXiv Preprint arXiv:2107.03374. 2021 Jul 7.
5. Guu K, Lee K, Tung Z, Pasupat P, Chang M. Retrieval augmented language model pre-training. In: Proceedings of the International Conference on Machine Learning; PMLR. 2020 Nov 21; 3929–38.
6. Karpukhin V, Oğuz B, Min S, Lewis P, Wu L, Edunov S, *et al*. Dense passage retrieval for open-domain question answering. arXiv Preprint arXiv:2004.04906. 2020 Apr 10.
7. Reimers N. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. arXiv Preprint arXiv:1908.10084. 2019.
8. Li Y, Wu J, Luo X. BERT-CNN based evidence retrieval and aggregation for Chinese legal multi-choice question answering. Neural Comput Appl. 2024 Apr; 36(11): 5909–25.
9. Liu Y, Lu W, Cheng S, Shi D, Wang S, Cheng Z, *et al*. Pre-trained language model for web-scale retrieval in Baidu search. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021 Aug 14; 3365–75.
10. Xiao S, Liu Z, Han W, Zhang J, Shao Y, Lian D, *et al*. Progressively optimized bi-granular document representation for scalable embedding-based retrieval. In: Proceedings of the ACM Web Conference. 2022 Apr 25; 286–96.

