Volume: 12 Issue: II Month of publication: February 2024

DOI: https://doi.org/10.22214/ijraset.2024.58670

PDF-Driven Q&A: A Review

Prof. Sneha R. Sontakke[1], Shreyash V. Gondane[2], Anup S. Nandedkar[3], Sahil T. Chauhan[4], Nama Z. Choudhari[5]
[1]*Assistant Professor at P R Pote College of Engineering and Management Amravati*
[2, 3, 4, 5]*Student at P R Pote College of Engineering and Management Amravati*

Abstract: Traditional methods of PDF retrieval often suffer from inefficiencies and inaccuracies due to their reliance on keyword-based search algorithms. These methods usually don't understand what words really mean and have trouble with things like different words meaning the same thing, one word meaning many things, and understanding the context. This makes the search results not very good. This project proposes a novel approach to address these shortcomings by developing a comprehensive system for efficient PDF document management and context-aware question answering. The system integrates various components, including a user-friendly interface for PDF upload, Longchain techniques for breaking down lengthy PDFs into manageable chunks, and vectorization using OpenAI's language models. These vectorized chunks are stored in the Faiss Vector Database for rapid retrieval. User queries are converted into vectors using the same language model, and a semantic search algorithm matches them against the stored PDF chunks to retrieve contextually relevant answers. By presenting these answers in a user-friendly format, the system aims to enhance accessibility and usability, enabling seamless access to pertinent information within PDF documents.

***Keywords: Chatbot, PDF retrieval Information, Longchain Technique ,Faiss Vector Database, Chunks*.**

I. INTRODUCTION

In an era marked by the exponential growth of digital information, effective management and retrieval of knowledge from vast repositories such as PDF documents pose significant challenges. To address this need, we propose a methodology for implementing a comprehensive system that enables users to upload PDF files, automatically breaks them down into manageable "chunks," and facilitates context-aware question answering. This methodology encompasses two key components: PDF Upload and Chunking, and Context-Aware Question Answering. The former focuses on providing users with a user-friendly interface for uploading PDF files and employs advanced techniques to segment lengthy documents into smaller, more digestible units. Leveraging state-of-the-art technologies such as OpenAI's language models and the Faiss Vector Database, this component ensures efficient storage and retrieval of the vectorized PDF chunks. The latter component, Context-Aware Question Answering, addresses the challenge of retrieving relevant information from the stored PDF chunks in response to user queries. By converting queries into vectors using the same model employed for PDF chunk vectorization, the system maintains consistency in representation. A semantic search algorithm is then utilized to match these vectors against the stored chunks, considering semantic similarity to deliver contextually relevant answers to the users. Through this methodology, we aim to provide a structured approach for enhancing information accessibility and user experience, empowering individuals to efficiently navigate and extract insights from large repositories of PDF documents

II. LITERATURE SURVEY

TABLE I
LITERATURE SURVEY.

Sr. No.	Author	Title	Feature	Year
1.	Sakib Shahriar And Kadhim Hayawi	Chatting With ChatGpt	ChatGPT can integrate with knowledge bases to provide relevant information based on large collections of text data, such as books, articles and web pages.	June 2023
2.	Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova	BERT: Pre-training of Deep Bidirectional Transformers	Natural Language Processing ,Deep Learning,NLU(Natural Language Understanding).	June 2019

3.	Xingxing Zhang, Furu Wei and Ming Zhou	HIBERT: Document Level Pre-training of Hierarchical Bidirectional Transformers for Document Summarization	A method to pre-train document level hierarchical bidirectional transformer encoders on unlabeled data.	July 2019

In the above given TABLE II (i.e. literature Survey) "Chatting With ChatGPT" by Sakib Shahriar and Kadhim Hayawi (2023) explores the integration of ChatGPT, a chatbot technology, with knowledge bases to provide contextually relevant information. This innovative approach leverages advanced natural language processing (NLP) techniques to enhance conversational experiences. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" by Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova introduces a groundbreaking approach in natural language processing (NLP) through deep bidirectional transformer models. This method significantly advances language understanding tasks by leveraging pre-training on large-scale corpora. "HIBERT: Document Level Pre-Training of Hierarchical Bidirectional Transformers for Document Summarization" by Xingxing Zhang, Furu Wei, and Ming Zhou presents a novel methodology for document summarization. By pre-training hierarchical bidirectional transformers on unlabeled data, it achieves effective and informative document summarization.

III. METHODOLOGY

A. Methodology for Implementing PDF Upload and Chunking

1) *User Interface for PDF Upload:* Develop a user-friendly interface that allows users to easily upload PDF files. This interface should be intuitive and accessible, enabling users to select and upload their desired PDF documents effortlessly.
2) *Longchain Techniques for Chunking:* Implement Longchain techniques, such as those provided by libraries like PyPDF2 or PDFMiner, to partition lengthy PDF documents into smaller, manageable "chunks." These techniques identify logical divisions within the PDF, such as paragraphs or sections, to create more digestible segments for processing.
3) *Vectorization with OpenAI:* Utilize OpenAI's language models, such as GPT-3, to convert each chunk of PDF content into high-dimensional vectors. This involves transforming the textual information within the chunks into numerical representations that capture semantic meaning and context, enabling advanced analysis and processing.
4) *Faiss Vector Database Integration:* Integrate the Faiss Vector Database to efficiently store the vectorized PDF chunks. Faiss is a powerful library known for its scalability and fast retrieval capabilities, making it well-suited for managing large volumes of high-dimensional vectors. By leveraging Faiss, the system can store and retrieve vectorized PDF chunks effectively, enabling efficient processing and analysis.

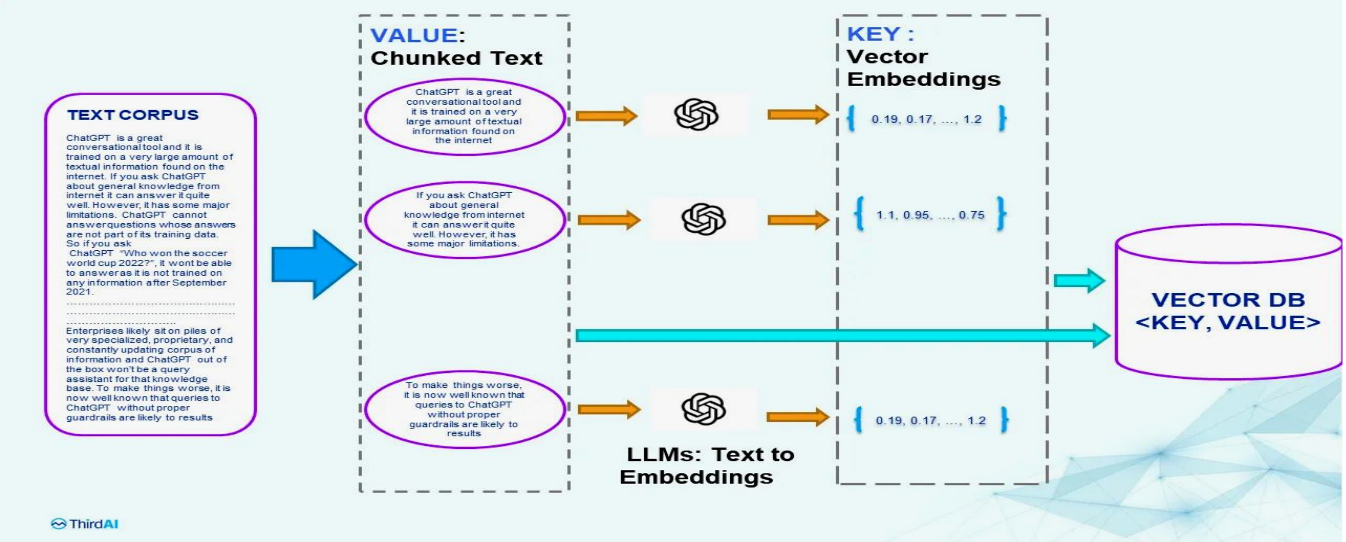
TEXT CORPUS
ChatGPT is a great conversational tool and it is trained on a very large amount of textual information found on the internet. If you ask ChatGPT about general knowledge from internet it can answer it quite well. However, it has some major limitations. ChatGPT cannot answer questions whose answers are not part of its training data. So if you ask ChatGPT "Who won the soccer world cup 2022?", it wont be able to answer as it is not trained on any information after September 2021.
Enterprises likely sit on piles of very specialized, proprietary, and constantly updating corpus of information and ChatGPT out of the box won't be a query assistant for that knowledge base. To make things worse, it is now well known that queries to ChatGPT without proper guardrails are likely to results
VALUE: Chunked Text
ChatGPT is a great conversational tool and it is trained on a very large amount of textual information found on the internet
If you ask ChatGPT about general knowledge from internet it can answer it quite well. However, it has some major limitations.
To make things worse, it is now well known that queries to ChatGPT without proper guardrails are likely to results
LLMs: Text to Embeddings
KEY : Vector Embeddings
{ 0.19, 0.17, ..., 1.2 }
{ 1.1, 0.95, ..., 0.75 }
{ 0.19, 0.17, ..., 1.2 }
VECTOR DB <KEY, VALUE>
ThirdAI

Figure 1.1: You need to store both text and vector embedding in the database with vectors being the KEY. The process requires an LLM to convert text chunk to vectors. The LLM should be the same for querying

The above figure (i.e Figure 1.1 and Figure 1.2) is taken from Anshumali Shrivastava,(June 2023). Understanding the Fundamental Limitations of Vector-Based Retrieval for Building LLM-Powered Chatbots—(Part 1/3).ThirdAI Blog. Retrieved from [Medium].

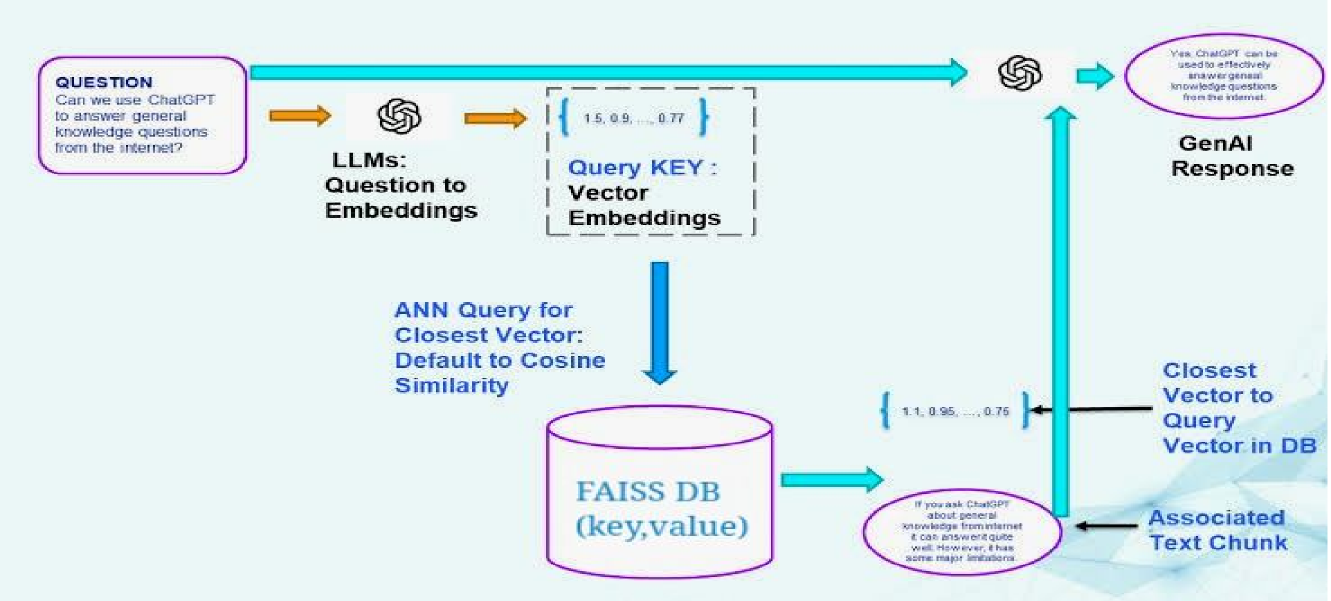
QUESTION
Can we use ChatGPT to answer general knowledge questions from the internet?
LLMs: Question to Embeddings
1.5, 0.9, ..., 0.77
Query KEY : Vector Embeddings
ANN Query for Closest Vector: Default to Cosine Similarity
FAISS DB (key,value)
1.1, 0.95, ..., 0.75
Closest Vector to Query Vector in DB
If you ask ChatGPT about general knowledge from internet it can answer quite well. However, it has some major limitations.
Associated Text Chunk
Yes, ChatGPT can be used to effectively answer general knowledge questions from the internet.
GenAI Response

Figure 1.2: The Q &A Phase

B. *Methodology for Context-Aware Question Answering*

1) *User Query Interface*: Develop an intuitive user interface that enables users to input questions or queries into the system. This interface should provide a seamless and user-friendly experience, allowing users to easily interact with the system and submit their queries.
2) *Conversion of Queries into Vectors*: Utilize the same OpenAI model employed for PDF chunk vectorization to convert user queries into vectors. By using consistent vector representations across queries and PDF chunks, the system ensures that queries are appropriately matched with relevant document segments during the search process.
3) *Semantic Search Algorithm:* Implement a semantic search algorithm capable of performing vector matching against the Faiss-stored PDF chunks. This algorithm should consider the semantic similarity between the query vector and the vectorized PDF chunks, enabling the system to retrieve contextually relevant answers that accurately address the user's query.
4) *Answer Retrieval and Presentation:* Retrieve the contextually relevant answers from the Faiss Vector Database and present them to the user in a clear and user-friendly format. This presentation format could include ranked lists of answers or highlighted excerpts from the original PDF documents, ensuring that users can easily review and comprehend the information provided by the system.

IV. LIMITATION

1) *Technical Limitations*: AI-powered chatbots may encounter technical limitations such as processing speed, memory constraints, or compatibility issues with certain document formats or languages.
2) *Handling of Non-Textual Content:* PDFs can include non-textual content such as images, graphs, and charts, which may contain valuable information relevant to the user's query. However, the chatbot may lack the capability to analyze and interpret such content, limiting its ability to provide comprehensive answers based solely on textual information extracted from the PDF
3) *Integration Complexity:* Integrating AI document readers into existing workflows and systems can be complex and time-consuming. Compatibility issues, data synchronization, and customization requirements may arise during the integration process, requiring specialized technical expertise and resources

V. FUTURE SCOPE

The future scope of the chatbot could be to enhance its ability to extract and understand information from PDF documents provided by users. This could involve implementing advanced natural language processing techniques to accurately interpret the content, allowing the chatbot to provide more insightful and tailored responses based on the information within the PDFs. Additionally, incorporating machine learning algorithms could help the chatbot to learn and improve its ability to extract suitable information over time, making it even more effective at assisting users with their queries.

VI. CONCLUSION

PDF-driven question-and-answer (Q&A) retrieval systems play a pivotal role in modern information management by efficiently extracting and categorizing data from documents. Leveraging advanced algorithms, these systems automate the processing of PDF files, identifying pertinent information and structuring it for easy access and analysis. However, challenges such as initial investment requirements, data dependency, and technical limitations pose obstacles to their implementation. Despite these challenges, PDF-driven Q&A retrieval systems offer invaluable solutions, streamlining information retrieval processes and enhancing efficiency within organizations. By automating tasks, reducing manual intervention, and facilitating rapid access to relevant data, these systems contribute to heightened productivity, informed decision-making, and optimized resource utilization. Continuous advancements in artificial intelligence and natural language processing further augment their capabilities, reinforcing their status as indispensable tools for modern enterprises.

REFERENCES

[1] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). DOI: 10.18653/v1/N19-1423

[2] Shahriar, S., & Hayawi, K. (2023). Let's Have a Chat! A Conversation with ChatGPT: Technology, Applications, and Limitations. Artificial Intelligence and Applications. DOI: 10.47852/bonviewAIA3202939.

[3] Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. HIBERT: Document Level Pre-training of Hierarchical Bidirectional Transformers for Document Summarization. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5059–5069, Florence, Italy. Association for Computational Linguistics

[4] Yadav, D., Desai, J., & Yadav, A. K. (2022). Automatic Text Summarization Methods: A Comprehensive Review. arXiv. Retrieved from https://doi.org/10.48550/arXiv.2204.0184

[5] Lin, J., Pradeep, R., Teofili, T., & Xian, J. (2023, August 29). Vector Search with OpenAI Embeddings: Lucene Is All You Need. arXiv:2308.14963v1 [cs.IR].

[6] Pandiarajan, S., Yazhmozhi, V. M., & Praveen Kumar, P. (2015). Semantic Search Engine Using Natural Language Processing. In Proceedings of the International Conference on Advanced Computing and Communication Systems (pp. 641-649). DOI: 10.1007/978-3-319-07674-4_53.

[7] Shaikh, A., More, D., Puttoo, R., Shrivastav, S., & Shinde, S. (2019). A Survey Paper on Chatbots. International Research Journal of Engineering and Technology (IRJET), 06(04), Volume: 06 Issue: 04.

[8] Shrivastava, A. (June 2023). Understanding the Fundamental Limitations of Vector-Based Retrieval for Building LLM-Powered Chatbots (Part 1/3). ThirdAI Blog. Retrieved from [Medium].

[9] D. Hingu, D. Shah, and S. S. Udmale, "Automatic Text Summarization of Wikipedia Articles," 2015 International Conference on Communication, Information & Computing Technology (ICCICT), Mumbai, India, Jan. 16-17.

[10] X. Yang, K. Yang, T. Cui, M. Chen, and L. He, "A Study of Text Vectorization Method Combining Topic Model and Transfer Learning," ISJ Theoretical & Applied ScienceE, vol. 1, no. 2, pp. XX-XX, Year.

[11] K. Singh and M. Shashi, "Vectorization of Text Documents for Identifying Unifiable News Articles," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 10, no. 7, 2019.