READ-EASY: Automating PDF Interaction using Langchain

Amit[1,] Varsha R[2], Chetan S. Yamakanamardi[3,] Naveen M Kamagoud[4]
[1,2,3,4]*Department of Machine Learning BMS College of Engineering, Bangalore, India.*

***Abstract*—A smart, user-friendly tool called READ-EASY was created to make working with YouTube videos and PDFs easier and more engaging. READ-EASY allows you to ask questions in plain English and receive accurate responses based on a true knowledge of the content, rather than just enabling you to browse through pages or search for specific keywords. To understand spoken and written text, the platform makes use of cutting-edge technologies including Natural Language Processing (NLP), Lang Chain, and Retrieval-Augmented Generation (RAG). It organizes and extracts text from PDFs while preserving the document's original structure. It employs semantic search strategies to understand the true meaning of your searches rather than depending just on keyword matching. This makes engaging with complex content much simpler by allowing the system to offer relevant, context-aware responses. However, READ-EASY is not limited to documents. Its capacity to manage YouTube videos is an excellent capacity. As with a document, just paste a video link, and the platform will transcribe the audio, divide the content into manageable chunks, and allow you to ask questions about it. Without watching the full video, you may quickly understand the main elements of any lecture, tutorial, or interview. All of this works in the background by transforming video and PDF content into "embeddings" — numerical representations that assist the system in locating the most pertinent data using a vector database such as Pinecone. It selects the most relevant parts of your inquiry, feeds them into a large language model (LLM), and provides you with an understandable, knowledgeable answer. Everything is connected via Lang Chain, which ensures a seamless and effective procedure. To make sure it provides high-quality responses, READ-EASY goes through extensive testing using measures like accuracy and F1-score. Because it is set up on cloud platforms like AWS or Google Cloud, it can be accessed via a simple web interface at any time and from any location. READ-EASY helps you save time, acquire knowledge, and interact with content in a more intelligent and intuitive way, whether you're attempting to rapidly understand a video or digging into a complex study paper.**

***Index Terms*—Lang Chain, Vector database, Semantic Indexing, Context-Aware response, Question-Answering system.**

I. INTRODUCTION

It will be very hard to go through extensive research papers, lengthy legal contracts, technical manuals, and company reports. in this case READ-EASY will helps. user can easily interact with the PDFs and find out the suitable content they are looking for and also generate the useful responses. READ-EASY was created to work with PDFs much more quickly and easily. Rather than depending on difficult keyword searches or tedious scrolling, you may ask a question and receive exact, clear responses taken directly from the paper, exactly like you would if you were speaking to a human. Behind the scenes, READ-EASY makes use of Retrieval-Augmented Generation (RAG) NLP and Lang chain, three very advanced technologies. To put it simply, RAG takes the most pertinent information from your document, which a smart language model then uses to produce responses that are both helpful and human-like. Lang chain facilitates smooth operation by tying everything together. The key here is that READ- EASY really learns what you're asking, not just trying to find words. It detects what you mean and displays the best response based on the PDF's real content, even if your query is a little complicated.by this you can increase your work efficiency. Retrieval-Augmented Generation (RAG) and Lang chain are two very sophisticated technologies used behind the scenes in READ-EASY. Simply put, a clever language model uses the most pertinent portions of your material that RAG identifies to produce responses that are both helpful and human-like. Lang chain makes everything function properly by tying everything together. What's amazing is that READ-EASY understands what you're asking, rather than merely searching for words. Based on the PDF's real content, it determines what you mean and displays the best response, even if your inquiry is a little ambiguous or complicated redefining how we read, understand and interact with document in the digital age.

II. PROPOSED WORK

The methodology for this will follow a structured approach to achieve a objective of automating the context- aware queries and information retrieval from PDF docs. The methodology has the following steps:

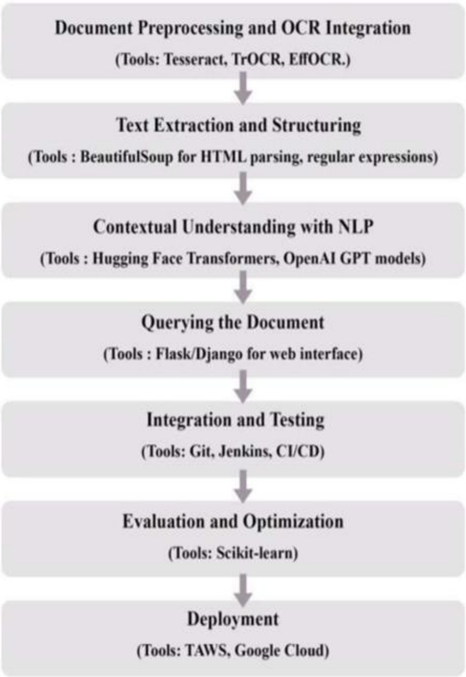

Document Preprocessing and OCR Integration
(Tools: Tesseract, TrOCR, EffOCR.)
Text Extraction and Structuring
(Tools : BeautifulSoup for HTML parsing, regular expressions)
Contextual Understanding with NLP
(Tools : Hugging Face Transformers, OpenAI GPT models)
Querying the Document
(Tools : Flask/Django for web interface)
Integration and Testing
(Tools: Git, Jenkins, CI/CD)
Evaluation and Optimization
(Tools: Scikit-learn)
Deployment
(Tools: TAWS, Google Cloud)

Fig.1 Methodology

- Document Preprocessing and OCR Integration The initial phase of a project involves processing the PDF documents, especially those with containing photos or scans, to make them a machine-readable. For this tasks, Optical Character Recognition (OCR) tools like a Tesseract and TrOCR will be a employed. These OCR systems will be converting scanned or image-based content into a text.
- Text Extraction and Structuring Following OCR, the raw text will be a extracted and structured according to a document's layout. This step will involve text the segmentation and pattern recognitions to ensure that a document's semantic relationships, such as a sections and subsections, are preserved.
- Contextual Understanding with Natural Language Processing (NLP) Once a text is structured, the next step is a apply Natural Language Processing (NLP) to understand the document contents and its contexts. Transformer-based models, such as a BERT or GPT, will be used to the extracts meaning from the text, enabling the system to a perform more than the just keyword matching.
- Querying the Document A key feature of this project is the ability to query the content of the PDF interactively. Users will input questions related to the documents, and the system will retrieve relevant answers by searching for the contextually appropriate responses. A semantic search engine, powered by NLP techniques, will rank responses based on their relevance to the query, rather than just simple keyword matching.● System Integration and Testing Once all the components OCR, NLP models, and the queries system are developed, they will be a integrated into a unified system. Integration testing will be a performed to ensure all the components work together seamlessly.
- Evaluation and Optimization After a system is integrated, it will undergo the performance evaluations using metrics like accuracy, precisions, recall, and F1- score to assess the effectiveness of a OCR and NLP components. These are evaluations will be help fine- tune the models and the documents querying mechanisms.
- Deployment Once optimized, the system will be the deployed on a cloud-based platform to ensure the scalability and accessibility. The final system will be the hosted on the platform like AWS or Google Cloud, with a user-friendly web interface where users can upload their PDFs and interacts with them in a real-time.

III. ARCHITECTURE

This diagram outlines how system processes PDFs and allows users to query them using natural language, delivering intelligent, context-aware responses with the help of the already fed data in the form of PDF or videos Here 's a step -by-step breakdown:

1. PDF Ingestion and Chunking: Input: Users upload one or more PDF files These PDFs are broken down into chunks of text (usually a few hundred words per chunk) using Lang chain's document loaders. Chunking improves semantic understanding and enables efficient search later.
2. Embedding Generation: Each text chunk is passed through an embedding model. This transforms the text into numerical vectors that capture semantic meaning.

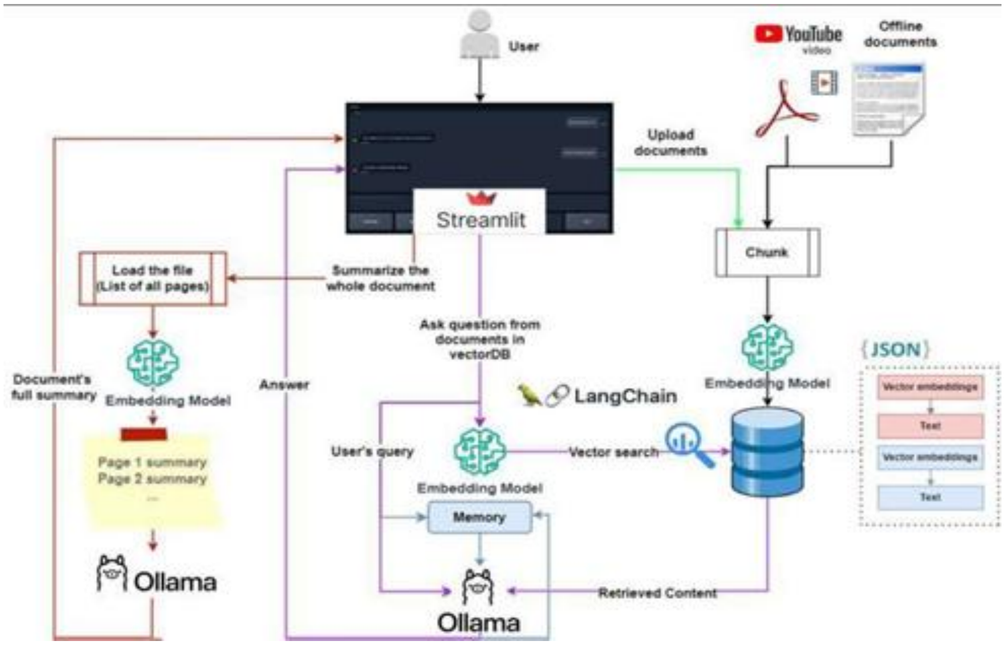


Fig.2 Architecture

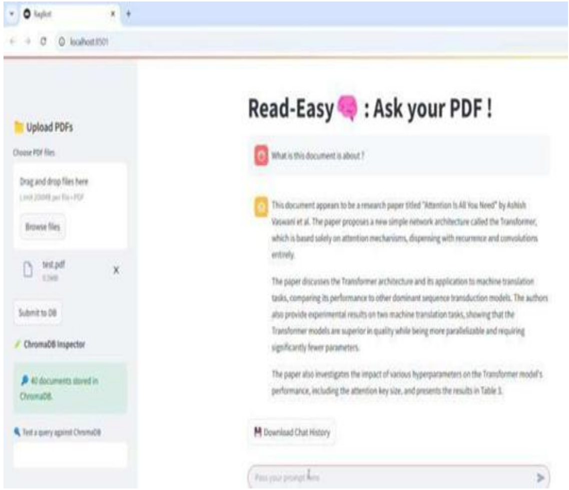
3. These vectors are stored in a vector database (in this case, Pinecone).
4. Vector Store (Knowledge Base): This is where all the embedded chunks are stored. When a user submits a query, the system doesn't search the text directly Instead, it compares vector similarity between the user's query and the stored chunks to find the most relevant information.
5. User Query and Question Embedding: The user asks a natural language question (e.g. "what is a neural network?").

The system converts the query into an embedding using the same model used for the document chunks.

6. Semantic Search: The system performs semantic similarity search between the query embedding and the stored document embeddings.
7. RAG (Retrieval-Augmented Generation): The retrieved chunks (relevant context) are passed to the LLM (Large Language Model). This forms the core of RAG, where the model uses external knowledge (retrieved chunks) to generate a grounded, accurate answer. This approach ensures the answer is both fluent and factually based on the PDF content.
8. Response Generation and Ranking: The LLM generates a response using the retrieved context. Optional: These results may be ranked or filtered based on confidence or relevance.
9. Answer Delivery: The generated answer is sent back to the user. The arrow pointing from the answer to the user indicates the final handoff of the intelligent, context-aware response

IV. PARTIAL RESULTS

The Fig. 2.1 represents implementation of the PDF interaction using Lang Chain. The system successfully enables users to upload a PDF file, extracts its text, and provides a confirmation message indicating that the extraction was successful. This ensures transparency and user confidence in the process. Once the text is extracted, users can enter a question related to the document, and the system utilizes Lang Chain and a large language model (LLM) to generate an answer based on the extracted content.



What is this document is about ?

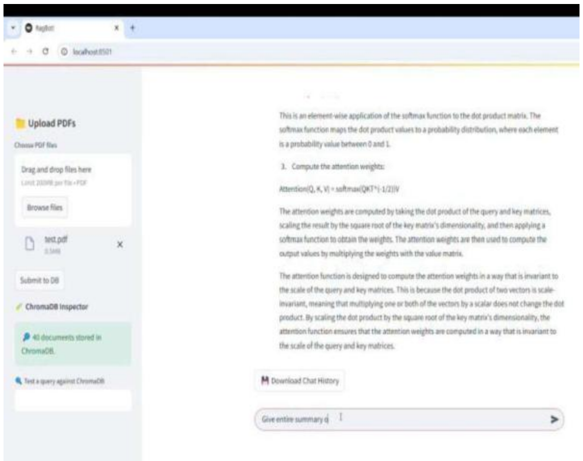
This document appears to be a research paper titled "Attention Is All You Need" by Ashish Vaswani et al. The paper proposes a new simple network architecture called the Transformer, which is based solely on attention mechanisms, dispensing with recurrence and convolutions entirely.

The paper discusses the Transformer architecture and its application to machine translation tasks, comparing its performance to other dominant sequence transduction models. The authors also provide experimental results on two machine translation tasks, showing that the Transformer models are superior in quality while being more parallelizable and requiring significantly fewer parameters.

The paper also investigates the impact of various hyperparameters on the Transformer model's performance, including the attention key size, and presents the results in Table 3.

Fig 2.1 Question-Answering System

In the given example, we uploaded a PDF titled Increasing_Global_Warming.pdf and asked What is global warming? The system responded with a definition based on the document's content, discussing the rise in global temperatures,CO_2 emissions and other tributing factors. This showcases the project's ability to retrieve and summarize information accurately.



This is an element-wise application of the softmax function to the dot product matrix. The softmax function maps the dot product values to a probability distribution, where each element is a probability value between 0 and 1.

3. Compute the attention weights:

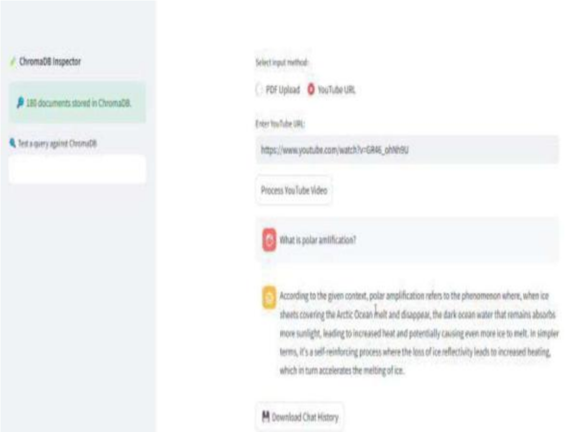
Attention(Q, K, V) = softmax(QKT^(-1/2))V

The attention weights are computed by taking the dot product of the query and key matrices, scaling the result by the square root of the key matrix's dimensionality, and then applying a softmax function to obtain the weights. The attention weights are then used to compute the output values by multiplying the weights with the value matrix.

The attention function is designed to compute the attention weights in a way that is invariant to the scale of the query and key matrices. This is because the dot product of two vectors is scale-invariant, meaning that multiplying one or both of the vectors by a scalar does not change the dot product. By scaling the dot product by the square root of the key matrix's dimensionality, the attention function ensures that the attention weights are computed in a way that is invariant to the scale of the query and key matrices.

Give entire summary o

Fig 2.2 Summarization of PDF


ChromaDB Inspector
180 documents stored in ChromaDB.
Test a query against ChromaDB
Process YouTube Video
Download Chat History

Select input method:

○ PDF Upload ● YouTube URL

Enter YouTube URL:

https://www.youtube.com/watch?v=GR46_ohNh9U

What is polar amlification?

According to the given context, polar amplification refers to the phenomenon where, when ice sheets covering the Arctic Ocean melt and disappear, the dark ocean water that remains absorbs more sunlight, leading to increased heat and potentially causing even more ice to melt. In simpler terms, it's a self-reinforcing process where the loss of ice reflectivity leads to increased heating, which in turn accelerates the melting of ice.

Fig 2.3 Youtube video Summarization

Apart from answering user queries, it also has the capability to summarize entire PDFs as shown in the above fig 2.2, This feature is particularly useful for users who need a quick overview of a lengthy document without going through it entirely. The system processes the extracted text and generates a concise summary highlighting the key points, making it easier for users to grasp the main ideas. This functionality is beneficial for researchers, students, and professionals who frequently deal with large volumes of textual data and need to extract relevant insights efficiently. Future improvements could include customizable summarization levels and multi-document summarization.

However, as a partial implementation, some areas need refinement. For example, the response appears slightly cut off at the end, indicating that the answer generation process may need better text handling. Accuracy may also be increased by including a more structured summarizing technique, managing multi-part queries, and increasing awareness of context. Stability and reliability in practical applications will be ensured by additional testing with different document types.

V. CONCLUSION

READ-EASY was created to make working with PDF documents less complicated. Instead of searching through countless pages or trying to discover the right keywords, READ-EASY allows you to simply ask what you're looking for in simple, common language, and it provides you with exact, clear responses directly from the document. With the use of advanced technologies like Retrieval-Augmented Generation (RAG), Lang chain, and intelligent Natural Language Processing (NLP), it can comprehend the questions you ask rather than only looking for words. READ-EASY lets you get to the point more quickly, whether you're reading technical manuals, academic papers, or difficult reports. How is it even better? Because it fits easily with text-based PDFs, it is incredibly helpful for researchers, educators, students, and anyone else who works with a lot of digital content. Its modular, adaptable architecture allows it to simply expand with you and adjust to your changing demands. In summary, READ-EASY is more than just a PDF reader; it's like having a smart helper that can find, shrink, and clarify what you need at the right time. Because of its emphasis on precision and respect to the original text, READ-EASY is especial suitable for high-stakes work where correct is crucial, such as legal research or academic studies.

In short, READ-EASY is not a tool for viewing documents it's a assistant that helps you interact with them more efficiently. By reading and understanding PDFs, it saves time, boosts productivity, and supports fast, more informed decisions. It is a clear example of how AI can make every day work easy and more intuitive. As it continues to grow, features like voice control, searching across multiple documents, and multilingual support could make READ-EASY even more helpful and user-friendly, turning it into a go-to solution for smarter document handling.

VI. FUTURE ENHANCEMENTS

Even though READ-EASY already offers an efficient and effective method of interacting with PDF documents there are many other features and functions that can be added to the current model which will boost its efficiency and models overall performance.

- Multi-Document Querying In our day-to-day life the information about one domain is stored in multiple PDFs. Our current model works on only one PDF. By adding multi -document querying to the model, we can use multiple PDFs to generate the required response.
- Voice-Based Interaction Adding voice-based interaction will increase the capabilities of the model. User can use the model in a handsfree way and it will increase the accessibility for the user.
- Multilingual Support: We can train the model on different languages, so that it can understand multiple languages and provide response in the required language. It will also help in the language translation.
- Domain-Specific Optimization The model can be trained on a specific domain dataset. Which will help the model to boost its performance, because of the lower size of the dataset.
- User Feedback Loop & Learning: By adding the user feedback function to the system, the efficiency and accuracy can be increased. It will be helpful to generate good responses.
- Data Visualization and Analytics: Adding features like data visualization will help the end user to understand the pattern and trend present in the data in an easy way.
- Collaboration Features: READ-EASY will be helpful for the people working in a group.
- Enhanced Security and Privacy Controls: User authentication can be added to protect the confidential data.

REFERENCES

[1] Zhang, X., Zhang, H., & Liu, L. (2023). A comprehensive survey of document-level relation extraction (2016-2023). arXiv preprint.

[2] Cheng, Y., & Jiang, X. (2021). TrOCR: Transformer-based optical character recognition with pre-trained models. arXiv preprint.

[3] Hu, W., et al. (2022). EffOCR: Efficient and modular OCR framework for real-world scenarios. International Journal of Computer Vision.

[4] Lee, S., & Lee, M. (2021). Multi-task learning for document- based question answering. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing

[5] Zhang, X., et al. (2023). Towards scalable and explainable AI for PDF processing. Elsevier AI Journal.

[6] Hugging Face. (n.d.). Transformers library.

[7] Tesseract OCR. (n.d.). Tesseract OCR GitHub repository