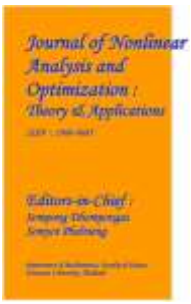
Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 15, Issue. 1 : 2024
ISSN : **1906-9685**



PDF CHAT_BOT USING GENERATIVE AI (LLMS&RAG)

P Deekshita*1,K Neeharika*1, M Sai Haritha*1, P Jagan Mohan*1,Y Sri Bindu*1
Department of AI&DS, Vignan's Institute of Information Technology, Visakhapatnam, India;
[1]deekshitaputta17@gmail.com,[1]neeharika1392@gmail.com,[1] harithamantri09@gmail.com,
pelurijaganmohan@gmail.com, [1]yerriboyenasribindu2526@gmail.com

***Abstract*—** This paper delves into the increasing craziness of planting chatbots and artificially intelligent systems (AI) in academia and disquisition. Chatbots are employed even more frequently to deliver adapted backing, attack enterprises, and develop fascinating spaces for knowledge. Chatbots empower directors to hone education while delivering actual-time support for pupils with assignments and examinations and issue explanations by outsourcing executive tasks. The creation of a PDF chatbot using Large Language Models (LLMs) in generative AI is one example of such an application. This project uses prompt-based querying, tokenization, and a vector-based component to store PDF inputs. The chatbot uses LLMs to build embedded prompts and pull text snippets from a knowledge base to respond to user inquiries. Chatbots' accessibility and flexibility enhance pupil involvement and expedite learning objectives.

Keywords— Chatbots, Artificial Intelligence (AI), Academia, Large Language models (LLMs), Generative AI, Knowledge Base**.**

I. INTRODUCTION

In the domain of education, the integration of state-of-the-art technologies such as artificial intelligence (AI) and chatbots has transformed how scholarly exchange is conducted. PDF Chatbot is one example of an innovation that facilitates Q&A dialogue based on academic papers.

This article takes a look at how a particular chatbot for PDFs is created and deployed in academic research. This, unlike other conventional bots, allows uploading multiple PDFs through which questions can be asked from within scholarly texts. Our chatbot employs sophisticated ML methods plus NLP techniques provide an exceptional interface for accessing and interacting with academic content.

Key to this PDF bot is the use of complex language models like Large Language Models (LLMs), and Recursive Aggregated Generative Models (RAG), among others. The bot applies these models when responding to very complicated queries by extracting information directly from uploaded PDFs and generating useful answers within their contexts. Additionally, our chatbot utilizes a new feature called Language chains.

Equally important, the chatbot we use incorporates the aptness of language chains, which enhances its contextual awareness and retrieval. This way, the chatbot responds with perfect cohesion and accuracy, even to ambiguous queries because it can analyze linguistic patterns and relationships within a text using language chains. LLMs make up the majority of our chatbot, which enables understanding and generating answers at an excellent scale of precision and fluency. RAG also increases the chatbot's abilities by collecting information from several documents to give contextual answers that correspond to users' questions. Furthermore, language chains are essential for enhancing context comprehension as well as information retrieval by enabling our bot to construct coherent responses based on linguistic patterns and relations within texts. These technologies are what underlay our AI-powered PDF chatbots facilitating seamless interaction with research

http://doi.org/10.36893/JNAO.2024.V15101.1727-1732

articles where useful insights can be extracted.

GENERATIVE AI

In the realm of human-machine interaction, generative AI, as seen in ChatGPT, has become a game changer. After its launch to the public, ChatGPT has received much attention and outperformed established rivals such as Google's Bard AI and Meta's Wit.ai. This research examines user expectations and competition using the analytical capability of ChatGPT on consumer queries and anticipation of future trends. We hope to give an insight into the emerging landscape of generative AI by selecting content appropriately and carrying out extensive due diligence. The use of ChatGPT by many people suggests that it is transforming various industries and societal relationships. Moreover, as different populations interact with ChatGPT, its influence on communication and dissemination of knowledge remains to be revealed. In this fast moving area, however, ChatGPT is a lodestar for innovation that rescues our relation with technology and information.

RETRIEVER-AUGMENTED GENERATION

In the world of natural language processing (NLP), Recursive Aggregated Generative Model (RAG) marks a breakthrough. So why does it matter? RAG can generate answers that have more than one point to support them, hence contextual responses from different sources that would guide in complex and informative conversations. It is an improvement on AI-powered chatbots, thus allowing them to offer more accurate and detailed explanations as they did before. By integrating with AI applications, RAG enables deep understanding and synthesis of various textual materials such that user interactions become more significant.

II. LITERATURE

Our work builds upon the bedrock principles of chatbot technology, exploiting Using NLP and ML approaches to increase learner engagement as well as task efficiency because these are connected with a changing context of AI-driven conversation systems [1]. The main focus of the article by Mansurova, Nugumanova, and Makhambetov was on the creation of a chatbot for blockchain that can respond to inquiries. The paper focuses on techniques and approaches used towards designing a good chatbot that deals with questions within the complex domain of blockchain technology. In this study, advanced natural language processing methods have been utilized to close the gap between users and blockchain knowledge, thereby facilitating seamless communication and knowledge sharing [2]. The paper written by Gyorgy Molnar and Zoltan Szuts is concerned with how chatbots are widely used in modern online interaction, especially in education. Chatbots, as educational aids, combine interactive technology with artificial intelligence; they do not only enhance but also simplify learning processes. In this paper, through a brief examination of the theoretical and historical background, it explicates how chatbots can help simplify various tasks and facilitate flow of information across the educational sector [3]. Chatbots are now complex computer programs that require experts in different technical fields. The technical brief describes the software engineering issues involved in creating advanced chatbots. Participants can design their own bots using Xatkit, an open-source chatbot development platform. [4].

In this article, we present an alternative way based on Retrieval Augmented Generation (RAG) of developing chatbots which makes use of Frequently Asked Questions (FAQs). It proves that home grown retrieval embedding models trained with the infoNCE loss outperforms general public-purpose embeddings. This essay is also the first to develop a cost minimization and efficiency maximization Reinforcement Learning (RL)-oriented optimization approach in the RAG pipeline thereby proving its applicability outside FAQ bots[5].

This paper investigates the vital role of Large Language Models (LLMs) in promoting AI development, particularly in designing chatbots for specific industries like therapy. In addition, the study leverages LLMs to examine novel approaches such as Reinforcement Learning from Human Feedback that can be used to improve the performance of chatbot and also points out that LLMs have a major influence on future conversational AI[6]. Improving domain adaption of search augmentation generation (RAG) models to answer open domain questions by Shaman and others [7]. A complete approach to building extraction-based chatbots using neural networks and solved many problems while implementing a

chatbot for real-world response retrieval, beginning with an unprocessed chatlog sample [8]. In this survey, the paper reviews chatbot technology in recent time, which is adding AI (Artificial Intelligence) and NLP (Natural Language Processing), on a broad basis. The trial uncovers how such companies are using chatbots for virtual customer support instead of focusing on main limitations and issues as far as their implementation is concerned. The findings from the trial serve to point out future areas of research that should be undertaken with an emphasis on new ways of improving chatbot efficiency in order to accomplish this[9]. Chatbot Growth prospects via Computational Intelligence. In this paper authors clearly explained how chatbots influence business aspects[10]. Moral implications applying chatbots and computation intelligence to research and education addresses challenges and proposes solutions Through qualitative research, it provides insight into the transformative potential of in this technology and ethical considerations. The study highlights the need for a proactive adaptation and ethical framework to leverage the benefits of AI to mitigate potential risks[11].

III. METHODOLOGY

There has been a tech revolution in academia. Information availability and individualization were restricted in the past since lecturers lectured and pupils depended on tangible materials. With the help of technology, students may now collaborate online, access digital content, and even use AI tutors (chatbots). Teachers use technology to create intriguing courses of study, use AI to make themselves more effective, and customize learning experiences. Technology has an enormous influence on how we teach and learn. Education is changing as a result of the advent of AI chatbots. By responding to inquiries and generating engaging educational content, these chatbots provide students with individualized assistance. They also automate administrative work for instructors and offer real-time help with homework, tests, and topics. This enables instructors to concentrate on instructing.

This is an example of how a PDF chatbot that uses large language models (LLMs) may function:

A learner poses this query: The learner asks a query about the content of a PDF while interacting via the chatbot and uploading it first.

Tokenization and knowledge base: The chatbot retrieves pertinent data from the PDF and stores it in a knowledge base. Tokens are the smallest fragments (chunks) into which this data is divided.

Prompt creation: Based on the student's inquiries and the acknowledged tokens (chunks), the chatbot uses LLMs to generate specific prompts (a query or command).

Answer generation: Using the knowledge base as an asset, the LLM then renders most of its credentials to deliver text snippets as an answer.

Enhanced engagement: By rendering concerns clearer for learners to grasp, this adaptable, appealing approach optimizes outcomes for learning.

This PDF chatbot has used large language models (LLMs) to bridge the gap between student queries and the information provided in the PDF. The LLM does more than just search for appropriate phrases when a learner poses a question. Instead, it acts as a very significant practitioner. It deciphers the query posed by the student, breaks down the PDF material into commemorative pieces to understand the surroundings, and also creates a unique recommendation for itself. With the help of this inquiry, the LLM may extract the most important information from the knowledge base and condense it into a succinct answer for the student. Essentially, the LLM functions as the chatbot's brain, providing accurate responses and encouraging the learner to go more into peripherals.

The mathematical equation used for the evaluation

Cosine similarity is used to look at how identical the replies are to one another. It consists of **an** n-dimensional **cosine** angle **generated** by two vectors. Vectors are compared using a cosine distance measure. Finding out if two vectors point in the same general path is how it is computed.

In the analysis of texts, it is also used to gauge document similarity. Millions of attributes may be found in a document, and each one of them can be utilized to find out how often a word occurs in the text. Assume that A and B are the two vectors that need to be compared. We have created a similarity

function using the cosine measure.

Although it is not stated clearly, the chatbot could get facts via cosine similarity. The way the cosine similarity functions in the background is as follows: the PDF contents as well as the learner's question are transformed into vectors, which are numerical representations. The connections and implications between words are recorded by these vectors. Next, the cosine similarity—numerical value ranging from 0 to 1—is computed to ascertain the degree of resemblance between the question's vector and the vectors of other PDF parts (each represented by a separate vector). At last, the chatbot extracts the section of the PDF that has the greatest similarity score, presuming that this portion contains the most pertinent response for the learner.

$$similarity(A,B) = cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sum_{i=1}^{n} B_i^2}.$$

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \|\vec{b}\|}$$

$$\|\vec{a}\| = \sqrt{a_1^2 + a_2^2 + a_3^2 + \cdots + a_n^2}$$

$$\|\vec{b}\| = \sqrt{b_1^2 + b_2^2 + b_3^2 + \cdots + b_n^2}$$

$$\vec{a} \cdot \vec{b} = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$

where the Euclidean norm of vector Q=(q1,q2,...,qn) is indicated by ||Q||, and the Euclidean norm of vector R is indicated by ||R||.

One computes the angle of the cosine produced by vectors Q and R. Both vectors are identical of one another and there is no match if the cosine value is 0. The cosine value is 1 when the two vectors match.

STEPS IN PDF CHATBOT USING GENERATIVE AI

Step 1:Loading the Document:

Uploading the input pdf files to the model from an external source.

Step 2: Text Pre-processing:

All the text analysing,semantic and syntatic analysis,text preprocessing are all done here.

Step 3: PDF Loaders:

Load the data to the model for text preprocessing using PyPDF loaders.

Step 4:Tokenization:

The splitted data is processed to separate it into smaller components known as chunks. And that chunk_size=10000,chunk_overlap=1000.

Step 5:Embedding:

Each block is then transformed into a vector embedding, which is a high-dimensional representation of the document info.

Step 6:Vector Store:

These embeddings are saved in a vector store called Faiss in the form of matrices.

Step 7:User Prompt:

The prompt has been provided by the user through streamlit interface Embedding

Step 8:Similarity:

Cosine similarity is performed between two vectors to check the similarity,if the two vectors are same then the cosine value is 1 otherwise, if no match, the cosine value is 0.

Step 9:Praquet:

It converts the vectors into packets and transport those data packets according to the request.

fig. Architecture view of pdf chatbot using generative AI

Algorithm ----------

Step 1:

Fleeting crucial libraries

result in st as streamlit

from PyPDF2 minute PdfReader

Step 2:

Characterize work to extricate substance from PDFs

Step 3:
Characterize tokenization work

Step 4:
Characterize vector capacity work

Step 5:
Characterize client input work

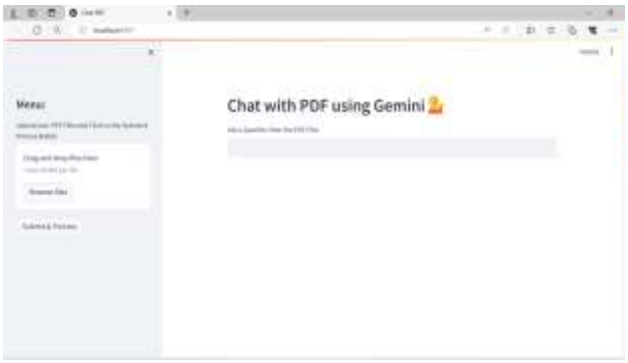
Step 6:
Characterize Streamlit UI work

Step 7:
Handle PDF exchange and planning

Additional taking care of steps (tokenization, vector capacity) can be included here
Sidebar menu for uploading PDF records

Step 8:
Essential work
within the occasion that title == "essential":
crucial()

Results This section presents the outcome of the proposed system.



1: User Interface

In the above screen, the Learner can upload a document or a file. Here learners have to upload files in PDF format.

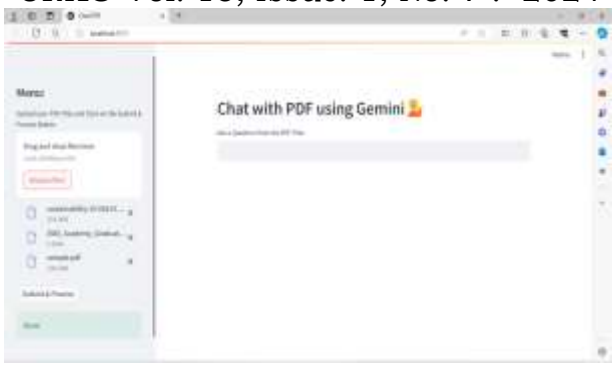


Fig. 2: Multi_pdf upload

The learner uploads the desired documents by clicking the explore file button first. After that, they click the submit and process buttons, which cause the procedure to run automatically in the background and produce a done message.

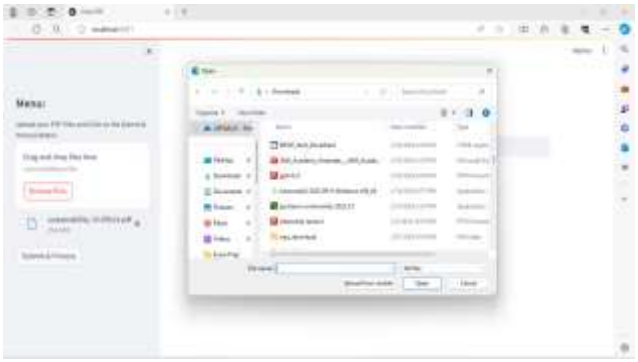


Fig. 3: Browsing PDFs from system

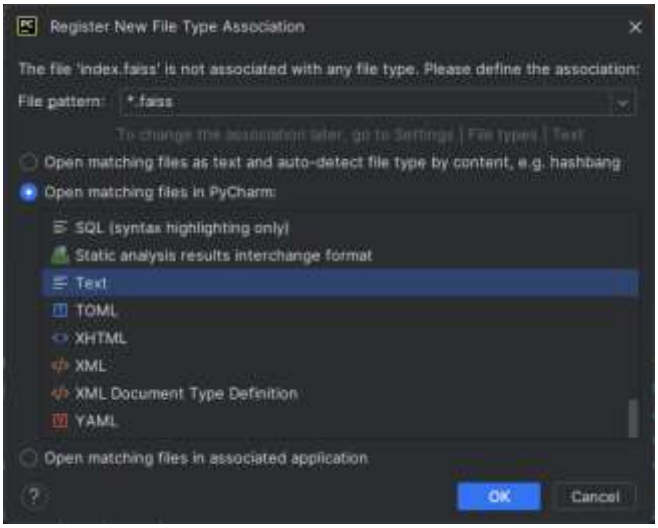


Fig. 4: Vector Storage (FIASS)

http://doi.org/10.36893/JNAO.2024.V15101.1727-1732

The above page shows that the uploaded document is tokenized into smaller chunks and stored as vectors in the FIASS vector storage.

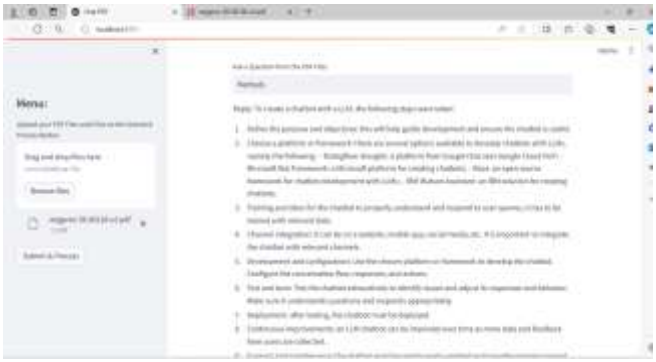


Fig. 5: Output generation page

The above-mentioned page serves as an example of how the model provides users with an intuitive, interactive interface to assist them in resolving any issues they may be having with the uploaded PDF file.

The below pages show the evaluation done manually for the vectors stored in the vector storage

Consider following vectors

$$a:[1,1,0]$$

$$b:[1,0,1]$$

$$\text{Norm of vector } a \text{ is} = \|\vec{a}\| = \sqrt{1^2+1^2+0^2} = \sqrt{2}$$

$$\text{Norm of vector } b \text{ is} = \|\vec{b}\| = \sqrt{1^2+2^2+1^2} = \sqrt{2}$$

$$\vec{a}\cdot\vec{b} = \sum_{i=1}^{n} a_i b_i = a_1b_1 + a_2b_2 + a_3b_3 = 1\times1+1\times0+0\times1 = 1+0+0 = 1$$

$$\cos\theta = \frac{1}{\sqrt{2}\times\sqrt{2}} = \frac{1}{2} = 0.5$$

Fig. 6 a): Manual Evaluation of vectors

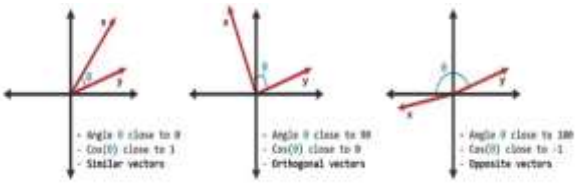


Fig. 6 b): Cosine angle variations

In the above Fig. 6, shows the manual evaluation of the vectors stored and the cosine angle variations.

Note: A number between -1 and 1, which denotes complete dissimilarity, 0 implies no similarity, and 1 shows perfect similarity, is obtained from the graphs mentioned above.

IV. CONCLUSIONS

Conventional PDFs might be difficult to learn from and stagnant. To solve this, this project builds a PDF chatbot that is driven by large language models (LLMs). Students can immediately ask questions by uploading PDF. To comprehend the student's question and the PDF's content, the chatbot makes use of LLMs. After that, it analyzes the PDF content and customizes the answer to fit the particular query. With this individualized approach, students receive immediate support inside the PDF's context, leading to better comprehension. The chatbot also frees up instructors' time for more participatory teaching methods by automating duties. Considerably, this chatbot driven by LLM has the ability to completely change how students read and absorb material from PDFs.

Furthermore, in terms of future work, the system may be developed to the advanced summarization model or can develop the model by integrating with learning management systems to work directly with learning online courses also or can also develop the model as a multilingual support model that may help learners from any regional language.

REFERENCES

[1] Harsha Pariyani, Anshika Sinha, Preeti Bhat, Roshni Rote, and Asst. Prof. N. A. Mulla. n.d. “LITERATURE SURVEY OF VARIOUS CHATBOTS.” Accessed Mar 25, 2024. https://www.researchgate.net/publication/362517384_LITERATURE_SURVEY_OF_VARIOUS_CHATBOTS

[2] Mansurova, Nugumanova, and Makhambetova, Z. 2023. “DEVELOPMENT OF A QUESTION ANSWERING CHATBOT FOR BLOCKCHAIN DOMAIN.” *Scientific Journal of Astana IT University* 15 (Sep): 27–40.

[3] G. Molnár and Z. Szüts, "The Role of Chatbots in Formal Education," 2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY), Subotica, Serbia, 2018, pp. 000197-000202, doi:

10.1109/SISY.2018.8524609. keywords: {Education;Natural language processing;Task analysis;Facebook;Machine learning;History},
[4] G. Daniel and J. Cabot, "The Software Challenges of Building Smart Chatbots," 2021 IEEE/ACM 43rd International Conference on Software Engineering: Companion Proceedings (ICSE-Companion), Madrid, ES, 2021, pp. 324-325, doi: 10.1109/ICSE-Companion52605.2021.00138. keywords: {Buildings;Chatbot;Software;Software engineering;chatbot;voicebot;bot}
[5] Mandar Kulkarni, et al. "Reinforcement Learning for Optimizing RAG for Domain Chatbots." 2024, https://doi.org/10.48550/arXiv.2401.06800
[6] D. Bill and T. Eriksson, 'Fine-tuning a LLM using Reinforcement Learning from Human Feedback for a Therapy Chatbot Application', Dissertation, 2023.

[7] Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, Suranga Nanayakkara; Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering. Transactions of the Association for Computational Linguistics 2023; 11 1–17. doi: https://doi.org/10.1162/tacl_a_00530
[8] Kristen Moore, Shenjun Zhong, Zhen He, Torsten Rudolf, Nils Fisher, Brandon Victor, Neha Jindal,
A comprehensive solution to retrieval-based chatbot construction,
Computer Speech & Language, Volume 83,2024,101522,ISS 0885-2308, https://doi.org/10.1016/j.csl.2023.101522.
[9] Caldarini, Guendalina & Jaf, Sardar & McGarry, Kenneth. (2022). A Literature Survey of Recent Advances in Chatbots. 13. 41. 10.3390/info13010041.
[10] Skrebeca, Julija & Kalniete, Paula & Goldbergs, Janis & Pitkevica, Liene & Tihomirova, Darja & Romanovs, Andrejs. (2021). Modern Development Trends of Chatbots Using Artificial Intelligence (AI). 1-6. 10.1109/ITMS52826.2021.9615258.

[11] Kooli, Chokri. (2023). Chatbots in Education and Research: A Critical Examination of Ethical Implications and Solutions. Sustainability. 15. 5614. 10.3390/su15075614.