Intelligent Textbook Assisstance System

Dr G Bharathi Mohan
Faculty - Computer Science and Engineering
Amrita School of Computing, Chennai
Amrita Vishwa Vidyapeetham, Chennai, India
gbharathimohan@ch.amrita.edu

Akula Harshavardhan Reddy
Department of Computer Science and Engineering
Amrita School of Computing, Chennai
Amrita Vishwa Vidyapeetham, Chennai, India
akulaharsha1435@gmail.com

Adapa Hanuma Siva Sairam
Department of Computer Science and Engineering
Amrita School of Computing, Chennai
Amrita Vishwa Vidyapeetham, Chennai, India
adapasairam090904@gmail.com

Kundula Saiteja
Department of Computer Science and Engineering
Amrita School of Computing, Chennai
Amrita Vishwa Vidyapeetham, Chennai, India
bannukundula@gmail.com

***Abstract*—In a realm of natural language processing, Large lanuguage models play a vital role in creation of chatbot, text generation and summarization. In this research, We aim to create a TextBook Tutor which is trained and answers the question based on the textbook context given by user. The main objective of our project is to create a sophisticated textbook tutor which will answer the question based on textbook context,To acheive this we use natural language processing techniques, Using llama2 model which will perform various natural language processing tasks such as text generation.The research findings could greatly improve education. Which makes high quality learning resource available for people, mainly help students to develop the independent learning.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

The traditional way of learning from textbooks can be challenging for students. They may struggle to understand complex ideas, get their doubts cleared, and actively engage with the course material. As education evolves in the digital age, there is a growing need for smart systems that can provide personalized support and guidance to learners in real-time. Conversational agents, trained using advanced language models like LLAMA 2, offer a promising solution. They can transform static textbooks into dynamic and interactive learning resources.

This paper introduces an Intelligent Textbook Assistance System. This system is designed to help students by providing personalized support and insights as they study from textbooks. The system uses the LLAMA 2 framework to train conversational agents. These agents can understand the complexities of educational content and have meaningful discussions with users. By integrating AI-powered conversation features into textbooks, the system aims to improve student learning experiences, depend their understanding of the subject matter, and encourage self-directed learning.

Our Intelligent Textbook Assistance System aims to help students learn better. It provides:

Detailed explanations, examples, and clarifications to improve understanding of textbook content.

Interactive quizzes, exercises, and activities tailored to individual learning needs.

Guidance on complex topics, answering doubts, and connecting ideas across chapters and subjects.

Real-time feedback, progress tracking, and personalized study recommendations based on user interactions.

Through user-friendly interfaces, our system combines traditional textbooks with dynamic digital learning. By using advanced conversational agents, we envision students having intelligent textbook companions to enhance their learning and achieve academic success.

As you go with chatgpt ,gemini AI or which are globally used they are not trained for particular domain for information, they trained in more generalisable way. If some user wants detail description on some medical term these chatbots give a generalised answer untill unless he specifys more acuurately what actually the user looking for. Even though these models will not precicely give the answer. We build a chatbot where user can feed the domain information as pdf of any size and run the model before asking query, it will answer more accurately for the query given by the user.We used CTansformers , Huggingface embeddings and FAISS vector stores to store the embeddings generated by Sentense transformer.

In the following sections, we will explore the design, features, and potential impact of our Intelligent Textbook Assistance System on education. We'll discuss how it can promote active learning, student engagement, and knowledge retention. The combination of AI-powered conversational agents and educational resources represents an exciting advancement in educational technology, opening up new opportunities for personalized and adaptive learning experiences.

II. LITERATURE REVIEW

The paper named Design and Development of a Chatbot [1] discuss the importance of the artificial intelligence and machine learning models such as natural language processing techniques, Python NLTK decision trees and dialogue management systems to develop the chatbot and discuss about the use of chatbot in computer aided design applications.

The paper titled Automated Medical Chatbot [2] deals about the design of medical chatbot using artificial intelligence markup language, It processes the users input to detect the symtoms and shortlist possible illness. Authors evaluated their chatbot on General word percentage analysis and terminology detection tests, They acheived a accuracy of 56% on average.

The paper Towards highly adaptive Edu-Chatbot [3] discuss about the use of Camem-Bert model for recognition in edu-chatbot system. CamemBert is a language model based on a ROBERTa model, Which is a version of BERT, it shows the effective results on french dataset.

The paper An overview of chatbot technology [4] dicuss about how to develop a chatbots using algorithms such as pattern matching, artificial intelligence markup language and latent symatic analysis by understanding the user input and give responses, the process involves parsing user requests, Understanding the users intention, retrieve information, generating the responses and managing the dialogue context.

The paper Chatbot developments in business world [5] discusses about the developments and implementation of chatbots in the business world mainly focusses on enhance the customer service, it deals about benefits of chatbots in business world and says the importance of natural language processing and Artificial Intelligence in chatbot developments and also presents the methods for sentimental analysis based on words to positive, negeative and neutral.

In the paper titled ChatDoctor:A Medical Chat Model Fine-Tuned on LLAMA Model using Medical Domain Knowledge [6], Authors developed the chatdoctor model which is specialized language model fine-tuned on llama model using medical domain knowledge, The authors aim to address the limitations of prevelant large language models by giving accurate results and also discusses the limitations of the large language models such as lack of contexual understanding.

The paper Financial News Analytics Using Fine-Tuned Llama 2 GPT Model [7] discusses how the Llama 2 language model was customized for financial datasets. This allowed it to perform various tasks, such as analyzing financial market text, summarizing text, and identifying named entities with sentiments. The study explores efficient fine-tuning methods like PEFT/LoRA, which can adapt pre-trained language models for specific applications without high costs. It also examines the use of advanced techniques like Low-Rank Adaptation (LoRA) to optimize GPU usage and achieve performance comparable to full fine-tuning.

The paper LLaMAntino: LLaMA 2 Models for Effective Text Generation in Italian Language" discusses the development of LLaMAntino [8] , a set of Italian language models based on the LLaMA 2 models. The authors focused on adapting the LLaMA 2 models to better handle the Italian language. They fine-tuned the models using the UltraChat dataset, which had been translated into Italian. This aimed to improve the models' understanding and generation capabilities for Italian-specific applications. The adaptation involved supervised fine-tuning training and the creation of model adapters. The resulting LLaMAntino models demonstrate strong linguistic abilities for various Italian natural language processing tasks, such as text generation, sentiment analysis, and question answering.

The paper Fine-Tuning LLaMA for Multi-Stage Text Retrieval [9] explores using the LLaMA model to improve text retrieval. It looks at using LLaMA as both a dense retriever (RepLLaMA) and a pointwise reranker (RankLLaMA). The study tested this on MS MARCO datasets and found that large language models work better than smaller ones. The RepLLaMA-RankLLaMA pipeline shows strong performance without special training, beating existing methods in both passage and document retrieval. The models can handle longer text without needing to break it up, which helps. Overall, the research demonstrates the potential of large language models to enhance retrieval tasks.

The paper A Proposed Academic Chatbot System using NLP Techniques [10] suggests developing a chatbot system that uses Natural Language Processing (NLP) techniques. This chatbot would answer academic and non-academic questions from users visiting a college's website. The goal is for the chatbot to provide instant responses without needing human involvement, reducing the workload for university staff. The system aims to give users relevant information efficiently, making it hard to distinguish the chatbot from a human. The chatbot uses NLP techniques to understand and process user queries, allowing it to provide accurate and appropriate responses.

The paper LLama 2: Open Foundation and Fine-Tuned Chat Models [11] introduces Llama 2, a set of pre-trained and fine-tuned large language models (LLMs) optimized for dialogue use cases. The models have been fine-tuned and show improved performance compared to open-source chat models on various benchmarks. Human evaluations for helpfulness and safety indicate that Llama 2-Chat models may be a suitable substitute for closed-source models. The aim is to allow the community to build on this work and contribute to the responsible development of large language models.

The paper Chatbot for Healthcare System Using Artificial Intelligence [12] focuses on developing a chatbot for healthcare systems using AI. The chatbot uses natural language processing techniques to interact with users and provide information. It employs keyword ranking, sentence similarity calculation, and database storage to understand user queries and generate appropriate responses. Methods like n-gram, TF-IDF, and cosine similarity are used for keyword ranking and sentence similarity calculation. The chatbot retrieves the most similar sentences from the database to answer user queries.

The paper A Chatbot System for Education NLP Using Deep Learning [13] focuses on designing intuitive and natural interaction between humans and computers, particularly in chatbot systems. The proposed work aims to enhance the capabilities of chatbots by intelligently identifying and gathering missing data from users to generate better responses. The study primarily examines chatbots that can serve the needs of small to medium-sized organizations, with the goal of creating interactive, easy-to-maintain, and cost-effective systems. The chatbot system combines an external knowledge base, a modified AIML system, and a relational database management system (RDBMS) within an integrated big data framework.

The paper Chatbot Using NLP [14] describes a chatbot system designed for college inquiries. This system uses Natural Language Processing (NLP) to understand user questions and provide relevant responses about the college's facilities and courses. The goal is to create an interactive interface where users can communicate with the chatbot through text or text-to-speech. The system analyzes user input to generate suitable replies, fostering a connection between humans and the machine.

The paper Developing a Chatbot [15] focuses on creating task-based chatbots using English textbooks in Dialogflow. This is for teaching language skills in the national English curriculum. The approach involves a two-layer chatbot system. It starts with a basic rule-based chatbot and then trains a sequence-to-sequence recurrent neural network (RNN) model on a public dataset. Another development is a chatbot-based academic system that uses natural language processing (NLP) techniques. This system can answer both academic and non-academic queries on a college website. Additionally, the researchers have created and released Llama 2, a collection of pre-trained and fine-tuned large language models (LLMs). These LLMs are optimized for dialogue use cases and outperform open-source chat models on various benchmarks.

III. OVERVIEW

Dataset is any textbook of particular domain in which you need to train the model in pdf fomat.We took the book named ”71763-gale-encyclopedia-of-medicine.-vol.-1.-2nd-ed”,which has 637 pages, where it has the information regaurding entire medical domain.

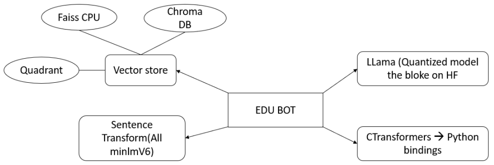
Faiss CPU
Chroma DB
Quadrant
Vector store
LLama (Quantized model the bloke on HF
EDU BOT
Sentence Transform(All minlmV6)
CTransformers → Python bindings

Fig. 1. Mind Map of Model

Dataset is given to the llm based chat bot model. Figure. 1. shows the mind map for the model to build. In this paper we have build abmodel using langchain, llama 2 as large language model, with sentence transformer to create word embeddings and Vector stores to store the embeddings, Retrival QA chain to retrive the ans finally with the help of chainlit we build a user interface.

A. Sentence Transformers

To represent text in vectors of fixed size by encoding sentence transformers are trained, these fixed size vectors are also referred as embeddings. Typically sentence transformers are designed to encode paragraph or sentences into embeddings, but they can also take each word as a sentence and generate word embeddings.Steps follwed by sentence transformers: 1. Tokenization: Input text is splitted into individual words. 2. Embedding Generation: Enbedding representation is obtained by passing each word to sentence transformer. 3.Aggregation: To have a single vector it can aggregate the word embeddings for the input text. Contextual content cannot be captured effectively by Sentence transformers to capture those one can go with Word2Vec, Glove or FastText. But if out-of-vocabulary words are requires from embeddings and want to go with pre-trained model sentence transformers is best. The sentence transformer which we used in our work was ”All mini lm V6”.

B. Vector DB/Stores

To manage and store the vector embeddings an efficient database or a system is designed which is termed as vectore DB or vector stores. In these vectorestores embeddings of different types can be stores such as embeddings of images, text, structured and un structured data. For vector embeddings An efficient way of storing is done.These are designed in an optimized way such that Based on queries embeddings are retrived fast. To enable quick search operations vectorstores incorporates indexing techniques like approximate nearest neighbor search algorithm etc. Based on users requirements retrieval parameters and storage can customize.

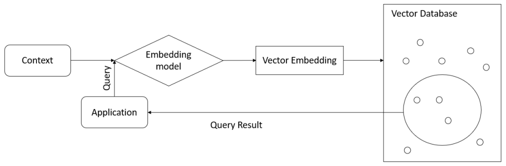

Fig. 2. Workflow of Sentence transformer and Vector DB

Figure. 2. Describes how sentence transformers and vector DB work together to retrieve the answer / Result of query given by the user. AS shown in flow diagram the context is stored in vector db as vector embeddings created by embedding model which here we take it as sentence transformer and when query given the information is retrieved from stored vectore data base. In our work we used FAISS DB.

C. LLAMA2

Llamas are gentle, woolly creatures from South America. They typically stand 5.5 to 6 feet tall at the shoulder and weigh around 280 to 450 pounds. These herbivores mainly eat grasses and other plants. Llamas have long, banana-shaped ears and can live up to 20 years when cared for properly. Valued for their strength, sure-footedness, and ability to carry heavy loads over rough terrain, llamas are popular pack animals in their native Andes region. They are also increasingly used as therapy animals and in animal-assisted interventions worldwide.

In our work LLama 2 model is used. The LLAMA 2 algorithm is an improved version of the original LLAMA

(Locality Aware Memory Access) cache replacement algorithm. LLAMA 2 is designed to enhance the efficiency of cache management in modern computer systems, especially when memory access patterns are not uniform. A key advantage of LLAMA 2 is its ability to dynamically adapt to changing access patterns and workload characteristics. Unlike the original LLAMA, which focused mainly on spatial and temporal locality, LLAMA 2 incorporates additional features such as frequency-based promotion and demotion policies. These policies prioritize cache entries based on their access frequencies, allowing LLAMA 2 to effectively mitigate cache pollution and improve overall cache hit rates. This, in turn, leads to enhanced system performance and reduced memory latency. Furthermore, LLAMA 2 introduces mechanisms to better handle skewed access patterns and irregular memory access behavior, making it a more versatile and effective cache replacement algorithm compared to its predecessor.

IV. Methodology

As in the Figure. 3. we implemented the model and created a chatbot by giving ”The GALE ENCYCLOPEDIA of MEDICINE SECOND EDITION” as input pdf and get the answers for quires using llm and QA retrival chain.

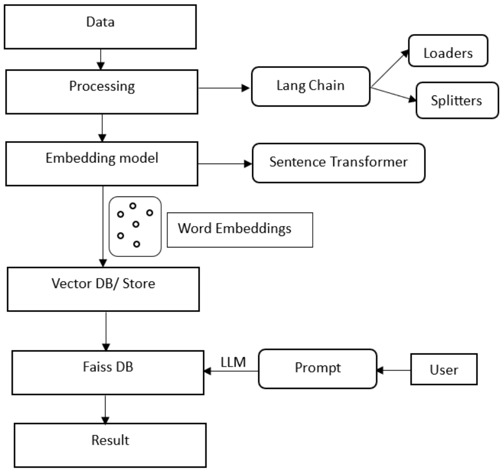

Fig. 3. Proposed Architecture

Using langchain loaders and splitters the large pdf is loaded with the help of splitters it splits the text into chunks with chunk size of 500 and having an overlap of 50. These chunks are sent to sentence transformer to get word embeddings. These take words or paragraphs as input and generates vector of numericals which contain the meaning, here ALL mini lm v6 is the transformer we used. THis transformer is trained with text and code of huge dataset. v6 indicates version 6.

After creating word embeddings these are need to be stored for that vector db/stores are helpful. All the generated word embeddings are get stored in vector stores while storing we use FAISS. FAISS is a library that interacts with datastore which helps to search the embeddings efficiently. It assign index for the vectors stored in database. For datasets which may not fit to memory this FAISS optimizes, it follow searching nearest neighbour that are based on query vector. It can also supports various distance metrics and search algorithms.

Creates a custom prompt template which defines as to use the following pieces of information to answer the users question , If don’t know the answer just say that I don’t know don’t try to make up an answer with context and question it returns prompt. When user give a query llm is getting loaded which llama2 is our llm model which was imported from TheBloke Hugging face and the model was GGML where it is not the direct model launched by meta AI as we are creating the bot to run it on cpu we require a quantised mode which can be loaded with transformers.The model released by METAAI requires GPU and Tensorflow to run so, we proceed with ”Llama-2-7B-chat-GGML” quantised model and used CTransforme to load the model.

Next to develop user interface Chainlit provides easy and good end user experience. It allows to interact with chatbot and your query. Chainlit has an ability to integrate well with many libraries such as FAISS to retrieve answer from vectoe based embeddings. user can also perform pre and post processing to refine user query and to adapt the content given as output for better flow. We intialized chainlit with content Starting bot.... Hi welcome to Bot what is your Query and after answer to query the msg variable is get update with the retrieved answer.

A. Retrieving Mechanism

When the query is given by the used it is getting converted to vector of embeddings and loads LLM to understand the context of question and Retrival chain starts to find most similar embedings that matches to the query vector in vector stores it initalises FAISS which search for the relevant embedding vector with nearest index where it finds the relevent content in the embedings stored in vector db the information is retrieved along with the source page no. The information get from different source pages are given to llama2 this analyse the query, context given by retrivel chain and generate response and the response sent to chainlit to display over the interface along with the source page no and what information it retieved from the text book given as input.

In our study, we meticulously curated a dataset comprising ten distinct medical questions. Our methodology involved presenting theseu questions to our custom-built chatbot and comparing its responses with those generated by ChatGPT. Furthermore, to enrich our analysis, we included original answers sourced from Quora. This rigorous approach enabled us to conduct a comprehensive evaluation of our chatbot’s performance against a state-of-the-art language model, providing valuable insights into its effectiveness and potential areas for improvement in the medical domain.

The rough score function calculates the similarity between two text responses using the cosine similarity metric.creates a count vectorizer object and fits it to the reference and user responses. This process converts the text into a matrix of token counts.

RESULTS

Table.1 Rough Score(RS)

Question	RS (our Model)	RS(Chat gpt)
Explain the causes and clinical manifestations of Parkinson's disease	0.678	0.593
What is the role of antibiotics in treating bacterial infections, and why is antibiotic resistance a concern?	0.806	0.485
How does the renin-angiotensin-aldosterone system (RAAS) regulate blood pressure and electrolyte balance in the body?	0.772	0.754
What are the risk factors, symptoms, and treatment options for coronary artery disease?	0.862	0.551
How can we predict disease susceptibility and prevent diseases before they manifest?	0.642	0.488
What are the primary risk factors for cardiovascular diseases, and how can they be mitigated or managed effectively?	0.755	0.684
What are the environmental factors contributing to the rise in respiratory diseases like asthma and COPD, and how can public health policies address these factors?	0.609	0.604
What are the socio-economic determinants influencing the prevalence and outcomes of infectious and chronic diseases, and how can healthcare disparities be addressed?	0.805	0.626
What are the emerging infectious diseases, and what strategies should be employed to prevent and control their spread?	0.557	0.493
What are the challenges in combating antimicrobial resistance, and what strategies are being developed to preserve the effectiveness of antibiotics?	0.884	0.437
What are the most common types of cancer worldwide, and what are their risk factors?	0.909	0.390

Average Rough score of our mode: 0.755

Average Rough Score of chatgpt: 0.55

Table.1 shows the Rough score for Each question by comparing the response given by our model and Chat gpt model with standerd answers which are collected from standard textbooks and google

I. Conclusion

The Intelligent Textbook Assistance System is a significant advancement in educational technology. It uses AI-powered conversational agents to enhance student learning. The system integrates personalized support, interactive features, and real-time feedback into traditional textbooks. This empowers students to better understand and actively engage with course material. By combining LLAMA 2, Sentence Transformers, and Vector Stores, the system efficiently retrieves accurate information from textbooks and provides tailored assistance to users. This innovative approach has the potential to promote active learning, improve student engagement, and foster knowledge retention in the digital age.

References

[1] Tamrakar, Rohit, and Niraj Wani. ”Design and development of CHATBOT: A review.” ResearchGate, Apr (2021).

[2] Srivastava, Prakhar, and Nishant Singh. ”Automatized medical chatbot (medibot).” In 2020 International Conference on Power Electronics and IoT Applications in Renewable Energy and its Control (PARC), pp. 351-354. IEEE, 2020.

[3] Tarek, A. I. T., Mohamed El Hajji, ES-SAADY Youssef, and Hammou Fadili. ”Towards highly adaptive edu-chatbot.” Procedia Computer Science 198 (2022): 397-403.

[4] Adamopoulou, Eleni, and Lefteris Moussiades. ”An overview of chatbot technology.” In IFIP international conference on artificial intelligence applications and innovations, pp. 373-383. Springer, Cham, 2020.

[5] Sari, Azani Cempaka, Natashia Virnilia, Jasmine Tanti Susanto, Kent Anderson Phiedono, and Thea Kevin Hartono. ”Chatbot developments in the business world.” Advances in Science, Technology and Engineering Systems Journal 5, no. 6 (2020): 627-635.

[6] Li, Yunxiang, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. ”Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge.” Cureus 15, no. 6 (2023).

[7] arXiv:2308.13032 [cs.CL]

[8] Basile, Pierpaolo, Elio Musacchio, Marco Polignano, Lucia Siciliani, Giuseppe Fiameni, and Giovanni Semeraro. ”LLaMAntino: LLaMA 2 models for effective text generation in Italian language.” arXiv preprint arXiv:2312.09993 (2023).

[9] arXiv:2310.08319 [cs.IR]

[10] N. Rakesh, N. Ravi, O. N. Daivajna and S. Ramesh, ”A Proposed Academic Chatbot System using NLP Techniques,” 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2022, pp. 1300-1304, doi: 10.1109/ICOEI53556.2022.9777231.

[11] https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/

[12] L. Athota, V. K. Shukla, N. Pandey and A. Rana, ”Chatbot for Healthcare System Using Artificial Intelligence,” 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2020, pp. 619-622, doi: 10.1109/ICRITO48877.2020.9197833.

[13] C. Kavitha and K. P. Kavitha, ”A Chatbot System for Education NLP Using Deep Learning,” 2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), Chennai, India, 2023, pp. 1-7, doi: 10.1109/ICONSTEM56934.2023.10142830.

[14]

[15] Goyal, Palash, et al. ”Developing a chatbot.” Deep Learning for Natural Language Processing: Creating Neural Networks with Python (2018): 169-229.