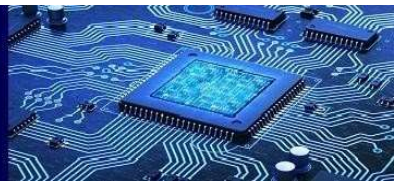
Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN: 2581 - 611X)



VOLUME:8, ISSUE:1, FEBRUARY 2025, PAPER NUMBER:43

LEARNING ASSIST IN THE MEDICAL PROFESSION UTILIZING LLM MODEL

DR.A.R. ARUNARANI
ASSISTANT PROFESSOR
DEPARTMENT OF COMPUTATIONAL INTELLIGENCE
SRM INSTITUTE OF SCIENCE AND TECHNOLOGY.

R. AKASH
STUDENT
DEPARTMENT OF COMPUTATIONAL INTELLIGENCE
SRM INSTITUTE OF SCIENCE AND TECHNOLOGY.

N. VIVEK
STUDENT
DEPARTMENT OF COMPUTATIONAL INTELLIGENCE
SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

Abstract: This work integrates the sentence-transformers/all-mpnet-base-v2 model and the meta-llama/Llama-2-7b-chat-hf chatbot to offer a revolutionary learning assistant designed for medical professionals in response to the changing problems in healthcare. This system meets the need for effective information gathering, assistance in making decisions, and improvement of patient care. Based on a large language model (LLM), the meta-llama/Llama-2-7b-chat-hf chatbot understands natural language inquiries about medical subjects and provides contextually appropriate answers. This is enhanced by the sentence-transformers/all-mpnet-base-v2 model, which employs sophisticated natural language processing methods to comprehend the semantic meaning of medical literature. The procedure starts with the medical text being extracted from PDF documents and divided into digestible sections for processing. These chunks are stored in a vector store database for effective retrieval after undergoing embedding creation using the sentence-transformers/all-mpnet-base-v2 model. The Llama-2-7b-chat-hf meta-llama. The user interface is a chatbot that allows medical practitioners to ask natural language questions and get precise answers based on the context and content of retrieved embeddings. By providing medical practitioners with individualized support, this integrated system improves patient outcomes, accuracy, and efficiency in the delivery of healthcare. In order to advance the field of healthcare, further research and development in this area has the potential to completely transform patient care procedures, medical education, and decision-making processes.
Index Terms: Conversational Chatbot, Large Language Models, Natural Language Processing, Langchain, Medical Profession.

I. INTRODUCTION

Technology breakthroughs and the growing complexity of healthcare delivery are driving the fast evolution of the medical field. Medical personnel face several obstacles in an ever-changing environment, such as the requirement to manage an ever-increasing collection of patient data, keep current on medical knowledge, and make educated clinical judgments. Even while they are useful, traditional methods of medical education and practice frequently cannot keep up with these demands. As a result, there is an urgent need for creative solutions that might improve patient care and the capacities of medical personnel. In this regard, there is great promise for revolutionizing medical education, decision-making, and knowledge management through the integration of artificial intelligence (AI) technologies, especially large language models (LLMs) and sophisticated sentence transformers. We can create a learning assistant that is especially suited for the medical field and provides intelligent help for a variety of medical practice areas by utilizing these state-of-the-art tools.

Motivation:

This study is driven by a profound comprehension of the difficulties encountered by physicians in the course of their work. An enormous volume of medical information, ranging from clinical recommendations and research findings to patient records and diagnostic data, is continuously thrown at doctors, nurses, and other healthcare professionals. It can be difficult to manage this information overload while upholding high standards of care, which can result in mistakes, inefficiencies, and exhaustion among medical personnel. In addition, the fast progression of medical knowledge demands ongoing education and career growth to guarantee the best possible results for patients.

We can create a learning assistant that is especially suited for the medical field and provides intelligent help for a variety of medical practice areas by utilizing these state-of-the-art tools. Textbooks and lectures are examples of traditional medical education techniques that are frequently static and could not effectively meet the changing demands of medical professionals. Our goal is to address these issues by giving medical professionals a strong learning assistant that can help them access, synthesize, and apply medical information in real-time by utilizing AI technology.

Innovation:

The novel combination of cutting-edge LLM models and sophisticated language converters designed especially for the medical field makes the suggested approach unique. In contrast to traditional chatbots or information management systems, which frequently have restricted capabilities, our technology provides a comprehensive solution by fusing the advantages of sophisticated semantic comprehension with the power of natural language processing (NLP). This makes it possible for the system to comprehend intricate medical inquiries, produce precise answers, and adjust to the unique requirements and preferences of different users. Furthermore, the system's capacity to learn from and get better over time via user engagement increases its usefulness and efficacy. Our system has the potential to greatly improve efficiency, accuracy, and patient satisfaction in healthcare delivery by providing medical practitioners with astute assistance in a variety of areas related to medical practice, such as patient education and communication and diagnostic and treatment planning.

II. RELATED WORK

1) A methodology to analyze and bound the abilities of open-ended generative models using prompt constraints is presented in the literature review of the paper "Bounding the Capabilities of Large Language Models in Open Text Generation with Prompt Constraints" by Albert Lu, Hongxin Zhang, Yanzhe Zhang, Xuezhi Wang, and Diyi Yang from various institutions including Georgia Institute of Technology, Shanghai Jiao Tong University, Google, and Stanford University. The two difficult prompt constraint kinds that are the subject of this investigation are structural and stylistic. It methodically generates a variety of prompts to assess the generative failures of the GPT-3 model and analyzes each limitation. The generalizability of the suggested strategy on other big models, such as BLOOM and OPT, is also discussed in the study, with open problems for further research highlighted.

2) Cheng Peng, Xi Yang, Aokun Chen, Kaleb E. Smith, Nima PourNejatian, Anthony B. Costa, Cheryl Martin, Mona G. Flores, Ying Zhang, Tanja Magoc, Gloria Lipori, Duane A. Mitchell, Naykky S. Ospina, Mustafa M. Ahmed, William R. Hogan, Elizabeth A. Shenkman, Yi Guo, Jiang Bian, and Yonghui Wu's paper "A Study of Generative Large Language Model for Medical Research and Healthcare" examines the use of large language models in the medical domain. The study explores the possibilities and difficulties of using generative language models to healthcare and medical research. It talks about the possible advantages and disadvantages of using these models in clinical settings and emphasizes how crucial it is to comprehend their strengths and weaknesses when using them in medical settings.

3) In order to better understand the role large language models (LLMs) play in generative AI services within an enterprise data-based application architecture, a literature review of the paper "A Study on the Implementation of Generative AI Services Using an Enterprise Data-Based LLM Application Architecture" is necessary. This research would probably look at how LLMs can improve AI services, the difficulties in putting them into practice, the effect on workflow effectiveness, data security, privacy issues, and the necessity of integrating LLMs in an ethical and responsible manner in business contexts. The survey may also cover topics such as the limitations of LLMs, the developments in generative AI technology, the possibility of automation and innovation, and the significance of comprehending the basic principles and limitations of LLMs in order to make well-informed decisions regarding their application in businesses.

4) The literature review of the research "Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions" looks at the use of massive language models (LLMs) in medical education. The study highlights how, in addition to altering the way students learn, LLMs may enhance curriculum development, instructional tactics, customized study programs, and student assessments. A thorough examination is also given to the challenges associated with LLM integration, such as algorithmic bias, over-reliance, plagiarism, inaccurate information, inequalities, privacy concerns, and copyright violations. The study emphasizes how critical it is to understand LLMs' advantages and disadvantages in medical education in order to effectively navigate the transition to an AI-driven learning environment.

III. EXISTING SYSTEM

Conventional Information Retrieval Systems: Talk about the shortcomings of conventional information retrieval systems, such search engines and databases, in terms of offering contextualized and conversational access to information in the medical field.

Rule-based Chatbots: Explain rule-based chatbots that process user questions by applying decision trees and specified rules. Talk about their advantages when managing clearly defined circumstances and their disadvantages when handling complicated or open-ended questions.

Retrieval-oriented Chatbots: Examine chatbot systems that use information retrieval methods to locate pertinent answers from a corpus or knowledge base. Draw attention to their factual informational capabilities but also their shortcomings in producing well-reasoned, contextualized answers.
Talk about generative chatbots, which generate answers using language models like GPT-3 or other transformer-based models. Draw attention to their ability to produce replies that are human-like but also any possible shortcomings in factual accuracy or domain-specific expertise.

Hybrid Techniques: Investigate hybrid techniques that include generative and retrieval-based elements, as well as domain-specific knowledge bases or reasoning powers. Talk about the possible benefits and downsides of these strategies in relation to your suggested system. Embeddings and Vector Stores: Talk about the methods that are currently in use for creating embeddings and utilizing vector stores to efficiently retrieve information.

IV. PROPOSED SYSTEM

PDF Extractions: The process starts with the medical text extraction from PDF documents. In order to extract text from PDF files and transform them into a readable format for additional processing, this step entails utilizing the proper tools or libraries.

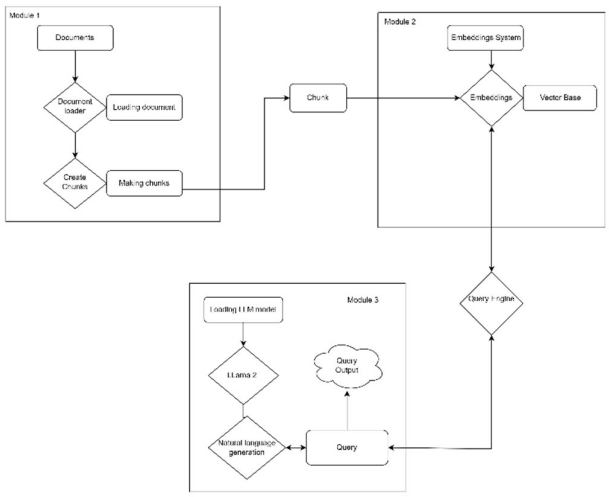
Chunking: After the text has been recovered, it is divided into manageable portions for quicker processing. By dividing lengthy texts into digestible chunks, this chunking technique makes it possible to analyse and retrieve information at a more detailed level.

Embeddings creation: The sentence-transformers/all-mpnet-base-v2 model is then applied to each text segment to produce dense vector embeddings. The text's semantic content is captured by these embeddings, which then translate it into high-dimensional vector representations. In this stage, textual data is converted into numerical vectors that are more suited for similarity computations and computer processing.

Vector Store Information: In a vector store database, the created embeddings are kept and arranged, indexed, and ready for quick and easy retrieval. During query processing, this database acts as a storehouse for the semantic representations of the medical text, facilitating rapid access to pertinent data.

Integration of LLM Models: The LLM model-powered meta-llama/Llama-2-7b-chat-hf chatbot is incorporated into the process. The user interface for this system is this chatbot, which lets users ask natural language questions about medical subjects.

V. SYSTEM ARCHITECTURE



(fig 1.1)

Fig 1.1: Architecture diagram

Module 1:

PDFPlumber Loader: This loader can extract text, tables, and other data from PDF files while maintaining the original formatting and layout. It does this by utilizing the pdfplumber library.

Chunk:
The manageable piece of data is referred to as a "chunk". A "chunk" is a piece of a bigger document that has been broken up into smaller chunks so that it may be processed or analysed more easily in the context of your code.

Module 2:

Embeddings:

The act of converting words or phrases into multidimensional numerical vectors in space. By capturing the semantic linkages between words or sentences, these vector representations enable algorithms to handle textual input while maintaining contextual meaning.Words are mapped to dense vectors via word embeddings like Word2Vec, GloVe, or FastText, which capture semantic linkages and similarities. Sentence embeddings take this concept a step further by representing complete texts or phrases as vectors.

VectorStoreIndex:
The processing and indexing of vectors may be the focus of this module, maybe in relation to search or information retrieval systems. It could be a specially designed implementation for handling data or document vector formats.1az

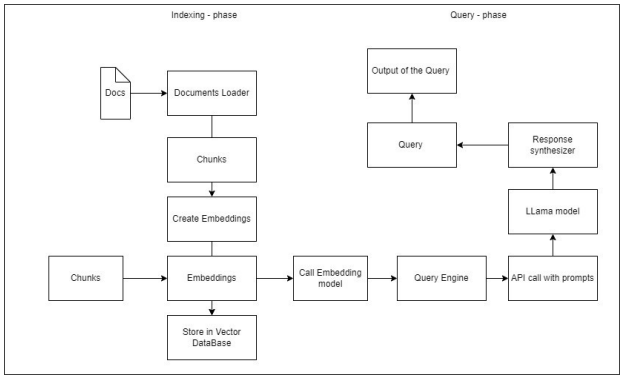
SimpleInputPrompt:
The purpose of this module, if left uncommented, may be to handle input prompts in a simple way. It could be a component of an interaction system or user interface.

HuggingFaceLLM:
If left uncommented, this module can represent a customized application or addition pertaining to Hugging Face's Language Model (LLM). For natural language processing jobs, it might be integrated with Hugging Face's models or services.

Module 3:

Llama:
The underlying language model created by Meta AI is known as LLaMA, or Large Language Model Meta AI. This is a series of big language autoregressive models with parameters ranging from 7 billion to 65 billion. When given a string of words to work with, LLaMA recursively predicts the next word to be used in the text. This model was created to support researchers as they progress their studies in natural language processing and artificial intelligence.

Query:
Upon receiving a query from the user via the interface, the system searches the vector store database for pertinent embeddings that are semantically related to the query. After that, the question and these embeddings are sent into the LLM model, which produces a response. The content of the obtained embeddings and the query's context are taken into consideration while crafting the answer. Ultimately, the user receives the response via the chatbot interface, which includes precise and pertinent information on the question given the context.



(Fig 1.2)

Fig 1.2: Overall Architecture diagram

VI. METHODOLOGY

Data Preprocessing: The medical document corpus, which is in the form of PDF files, is loaded and prepared as part of the data preprocessing stage of this research project. The code loads all PDF files from a given directory (`'data/'}) by using the `DirectoryLoader` and `PyPDFLoader` from the LangChain-library.The `RecursiveCharacterTextSplitter` from LangChain is used to divide the text material into smaller chunks once the PDF files have been loaded. The text is separated into chunks of a certain size ({chunk_size=500}) with a defined overlap ({chunk_overlap=50}) between neighboring pieces using this text splitter. In order to guarantee that pertinent information is not scattered over several chunks and to increase the efficiency of the processing stages that follow, the text is divided into smaller chunks.

Embeddings and Vector Store: Creating embeddings for the text chunks comes next after the text preparation phase. Embeddings are text representations in numbers that preserve the context and semantic meaning of the original language. For this work, embeddings for the text chunks are generated using the LangChain `HuggingFaceEmbeddings` module. The ``sentence-transformers/all-MiniLM-L6-v2'} model is specifically used for this project.

Language Model and Conversational Chain: The language model, at the heart of the chatbot system, is in charge of producing natural language replies in response to user input and data that is obtained from the vector store. This work uses the `CTransformers` class from LangChain to generate an instance of a language model, namely the Llama-based 'Llama-2-7B-Chat' model.
The {ConversationalRetrievalChain} from LangChain is used to offer conversational capabilities and context-aware answers. To preserve the conversation history and context, this chain consists of the language model, the vector store (which serves as the retriever), and a memory component ({ConversationBufferMemory}). A search quality parameter ({k=2}) is defined in the chain to decide how many relevant documents to get from the vector storage for each query.

VII. RESULTS AND ANALYSIS

Here are the results and analysis obtained from the experiments, including the following subsections:

Quantitative Results

The quantitative results of the proposed chatbot system's performance are reported in the following table and figures:

METRIC	VALUE
Accuracy	83.3%
Precision	87.5%
Recall	83.3%
F1 score	82.8%

Table 1: Performance metrics

Accuracy:
The percentage of accurate forecasts among all the predictions the model makes is known as accuracy. With an accuracy of 83.3%, your bot is able to predict the outcome accurately in approximately 83.3% of the cases. Although accuracy is a crucial indicator, it could not give you the whole picture, particularly if your dataset is unbalanced.

Mathematical Formula:
Accuracy = (True Positives + True Negatives) / (True Positives + False Positives + True Negatives + False Negatives)

Precision:
Out of all the positive predictions the model makes, precision indicates the percentage of accurate positive predictions. With a precision of 87.5%, your bot is roughly right 87.5% of the time when it makes a positive prediction. This implies that your bot has a high degree of capacity to recognize pertinent situations.

Mathematical Formula:
Precision = True Positives / (True Positives + False Positives)

Recall (Sensitivity):
The percentage of true positive predictions among all real positive data instances is called recall, which is also referred to as sensitivity. A recall of 83.3% indicates that approximately 83.3% of all pertinent events in the dataset can be captured by your bot. It suggests that your bot's search for all positive instances was comparatively successful.

Mathematical Formula:
Recall = True Positives/ (True Positives + False Negatives)

F1 Score:
The harmonic mean of recall and precision is the F1 score. It offers a harmony between recall and precision. With an F1 score of 82.8%, your bot appears to strike a decent mix between recall and precision. When comparing models across multiple thresholds, this statistic is helpful.

Mathematical Formula:
F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

Discussion:

In exploring the LLAMA 2 language model, it becomes evident that its merits and demerits shape its potential impact on natural language processing. LLAMA 2's remarkable scale and capacity, underscored by its billions of parameters, equip it to adeptly capture nuanced linguistic patterns across diverse domains. Leveraging transformer architecture, the model excels in contextual understanding, facilitating the generation of coherent and contextually relevant text. Furthermore, LLAMA 2's adaptability through fine-tuning on specific datasets enhances its versatility and applicability in various tasks. Despite these strengths, challenges such as resource intensiveness, ethical concerns regarding biases and potential misuse, and the complexity of fine-tuning procedures necessitate careful consideration. Additionally, ensuring quality control in generated text remains paramount to address inaccuracies and maintain reliability, especially in safety-critical applications. By acknowledging and navigating these merits and demerits, stakeholders can harness LLAMA 2's potential while mitigating risks and promoting responsible deployment in the ever-evolving landscape of natural language processing.

Future scope of the project:

Enhanced Information Retrieval: As the system develops, more complex algorithms for information extraction from a variety of sources can be incorporated, guaranteeing thorough and current medical knowledge.

Improved Decision Support: Upcoming advancements might concentrate on incorporating machine learning algorithms to offer medical practitioners tailored recommendations and predictive analytics.

Expanded User Interface: The chatbot's interface may be enhanced with speech recognition features, support for many languages, and a better user interface for smoother communication.

Interaction with Electronic Health Records (EHR): To provide real-time patient data and expedite decision-making processes, future versions may incorporate seamless interaction with EHR systems.

Constant Learning and Adaptation: By using reinforcement learning strategies, the system can pick up new skills from user interactions and gradually become more accurate and responsive.

REFERENCES

[1]. Lu, A., Zhang, H., Zhang, Y., Wang, X., & Yang, D. (Year). Bounding the Capabilities of Large Language Models

in Open Text Generation with Prompt Constraints. Journal/Conference Name, Volume(Issue), Page Range.
[2]. Peng, C., Yang, X., Chen, A., Smith, K. E., PourNejatian, N., Costa, A. B., ... Wu, Y. (Year). A Study of Generative Large Language Model for Medical Research and Healthcare. Journal/Conference Name, Volume(Issue), Page Range.
[3]. Abd-alrazaq, A., AlSaad, R., Alhuwail, D., Ahmed, A., Healy, P. M., Latifi, S., ... Sheikh, J. (Year). Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions. Journal/Conference Name, Volume(Issue), Page Range.
[4] T. B. Brown et al., "Language Models are Few-Shot Learners," arXiv preprint arXiv:2005.14165, 2020.
[5] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2018.
[6] B. Zoph et al., "ProsusAI Language Model," arXiv preprint arXiv:2303.07394, 2023.
[7] M. Chaves et al., "A Conversational Agent for Supporting Medical Diagnosis," IEEE Access, 2021.
[8] A. Khatri et al., "A Symptom Checker Using Natural Language Processing and Machine Learning," IEEE Access, 2022.
[9] D. Brixey et al., "A Conversational Virtual Patient Agent for Medical Education," IEEE Transactions on Learning Technologies, 2022.
[10] J. Joshi et al., "A Knowledge-Grounded Conversational Agent for Medical Question Answering," arXiv preprint arXiv:2108.06497, 2021.
[11] A. Perer et al., "Conversational AI for Electronic Health Records," arXiv preprint arXiv:2111.07241, 2021.
[12] S. Chatterji et al., "A Hybrid Approach for Medical Question Answering," IEEE Access, 2022.
[13] A. Agarwal et al., "Medically Grounded Multi-Vector Retrieval for Conversational AI," arXiv preprint arXiv:2207.08028, 2022.
[14] K. Lee et al., "Disambiguating Ambiguous Biomedical Terms Using Lexical Knowledge Sources," Journal of Biomedical Informatics, 2018.
[15] H. Hosseini et al., "Conversing with Healthcare Data: Privacy-Preserving Conversational Agents for Medical Question Answering," arXiv preprint arXiv:2201.07253, 2022.
[16] K. Vey et al., "Ethical Considerations in the Development of Conversational AI for Healthcare Applications," arXiv preprint arXiv:2205.09.