Retrieval-Augmented Generation based Large Language Model Chatbot for Improving Diagnosis for Physical and Mental Health

Y. Bhanu Sree
Dept. of CSE(CYS,DS)
VNRVJIET
Hyderabad, India
0000-0003-2056-9379

Addicharla Sathvik
Dept. of CSE (CYS,DS)
VNRVJIET
Hyderabad, India
addicharlasathvik@gmail.com

Damarla Sai Hema Akshit
Dept. of CSE (CYS,DS)
VNRVJIET
Hyderabad,India
akshitd371@gmail.com

Omrender Kumar
Dept. of CSE (CYS,DS)
VNRVJIET
Hyderbad, India
omrenderkumar@gmail.com

Bandaru Sai Pranav Rao
Dept. of CSE (CYS,DS)
VNRVJIET
Hyderbad, India
saipranavraob@gmail.com

2024 6th International Conference on Electrical, Control and Instrumentation Engineering (ICECIE) | 979-8-3503-8004-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICECIE63774.2024.10815693

***Abstract* - Chatbot integration in the medical domain results in improved accessibility to healthcare information and services with enhanced patient query communication and patient education. The proposed work is a novel approach to Health Care chatbot named MEDGPT. The model is developed using the Retrieval-Augmented Generation (RAG) framework integrated with external data sources such as PDF documents, CSV files, and PubMed documents related to Health Care. It combines a retriever component of the architecture to fetch relevant information from external sources and a generator component to craft contextually appropriate responses using Large Language Model (LLM's). The architecture employed tools and agents in order to generate response from multiple external sources. Tools are specialized for each data source and are designed to extract relevant information based on user queries. Agents are components within the chatbot architecture that handle different aspects of logical reasoning and decision making. When a user query, requires information from an external data source, the corresponding agent invokes the appropriate retrieval tool. Once the data is processed, it is integrated into the chatbot's knowledge base using the RAG framework. The RAG framework combines the retrieved data with the chatbot's language generation capabilities to craft contextually appropriate responses to user queries. Through extensive testing and evaluation, the chatbot achieved significant improvements in user satisfaction, response accuracy, and engagement, showcasing the potential of the RAG framework in leveraging external data sources for intelligent conversational agents in Health Care. The proposed framework enhanced the RAG based LLM chatbot's capabilities to provide relevant and accurate responses to user queries related to Health Care with respected Physical and Mental health.**

Keywords- LLM, Chatbot, RAG, Agents, Health Care.

I. Introduction

In recent years, advancements in Natural Language Processing (NLP) and Artificial Intelligence (AI) have revolutionized various sectors, including healthcare. Among these innovations, the development of medical chatbots utilizing Large Language Models (LLMs) stands out as a promising frontier in patient care and medical assistance. These chatbots, powered by sophisticated algorithms and trained on vast amounts of medical data, hold the potential to enhance healthcare accessibility, efficiency, and quality. Medical chatbots, leverage the capabilities of LLMs to understand and respond to Natural Language inputs, enabling seamless interactions between patients and healthcare providers. By simulating human-like conversations, these chatbots offer personalized medical advice, symptom assessment and even mental health support. Moreover, they serve as invaluable tools for health education, providing users with reliable information about diseases, treatments, and preventive measures. They offer immediate responses to inquiries, reducing waiting times and alleviating the burden on healthcare systems. Furthermore, the continuous learning capability of LLM-based chatbots enables them to stay updated with the latest medical research and guidelines, ensuring the delivery of accurate and evidence-based information.

The work incorporated agents into LLM based chatbots which significantly enhances their functionality, effectiveness, and user experience. This integration fundamentally changes how chatbots interact with the users, access and process information, and respond to queries in the domains of healthcare. By leveraging the Retriever Augmented Generation (RAG) framework and integrating it with specialized agents, MEDGPT chatbot is a comprehensive solution that redefines interactions and service delivery. Agents in the RAG framework serve as intelligent intermediaries that orchestrate the seamless integration of external data sources into the chatbot's knowledge base. These agents are designed to handle diverse data types and sources, including PDF documents, CSV files, YouTube videos, and website content. Each tool is tailored to a specific data source or type, agents are equipped with the capability to invoke specialized retrieval tools at the time of data retrieval. For instance, an agent makes decision according to the query, if the query needs PDF data, PDF parser tool is invoked, while another trial dealing with YouTube videos utilizes an API handler for fetching relevant content.

Furthermore, agents in the RAG framework play a crucial role in logical reasoning and decision making. They employ Natural Language Processing (NLP) techniques such as text summarization, entity recognition, and sentiment analysis to extract meaningful information from retrieved data. This processed data is then seamlessly integrated into the chatbot's knowledge base for use by the generation component of the RAG framework. The dynamic nature of agents within the

RAG framework enables the chatbot to adapt and respond contextually to user queries, leveraging external data sources to enhance the accuracy, relevance, and richness of its responses. This integration of agents with the RAG framework represents a significant advancement in chatbot development, paving the way for more sophisticated and intelligent conversational agents with a wealth of external knowledge at their disposal. The proposed work delves into the development, applications, challenges, and prospects of medical chatbots built using LLMs. By examining existing literature, case studies, and technological advancements, we aim to provide a comprehensive understanding of the role of these chatbots in transforming healthcare delivery and improving patient outcomes.

II. Related Work

All individuals doesn't have immediate access to healthcare services for a variety of reasons, like busy life style, staying in remote areas, negligence, financial situations, etc. So, chatbots in Health care eliminates the need for people to physically visit healthcare facilities by giving them a convenient and easily accessible way to access healthcare information, support remotely and provide assistance immediately. Large volumes of available data can be efficiently analyzed by AI tools, which can also be used to spot patterns that a clinician might overlook. New large language models (LLMs) have demonstrated impressive capabilities in various domains, including intelligent diagnostics, in the past few years. Examples of these domains include ChatGPT and Google BARD [1]. Large heterogeneous data must be ingested in order for a Large Language Model (LLM) to be trained from scratch. This process demands the mastery of complex language patterns and relationships, which calls for powerful hardware such as high-performance GPUs or TPUs to handle the enormous computational load. By fine tuning an LLM we can engineer a specialized language model capable of producing precise, context-sensitive, and safety-conscious responses in any specific domain instead of constructing from ground up [2].

Enhancing pre-trained models with smaller, task-specific datasets is a critical step in the fine-tuning process of LLMs. Fine-tuning, which acts as a crucial link between generalized pre-trained models and specialized versions, basically entails the particular needs of particular applications and general language models. This guarantees a close concurrence between the human expectations and the language model [3]. LLMs have powerful text processing capabilities and in order to utilize already learned knowledge, multimodal LLMs introduce additional modalities such as image and video to make cross-modal knowledge integration and interaction easier. multimodal LLMs do not understand the input content but more simply fit the training data distribution but Current multimodal models are still very far from understanding prompts properly [4]. It becomes essential that we combine all of the resources at our disposal and use the expertise of many sources to create CHAs(Conversational Health Agents) that provide an environment that is reliable, comprehensible, and actionable for a worldwide audience [5]. If the related legitimate concerns are proactively looked into and addressed, there may be promising uses for LLMs in health care research, education, and practice. The potential applications include improved scientific writing and versatility, utility in healthcare research, and enhanced healthcare education [6]. LLMs can be utilized in multiple languages and diverse contexts. While some will view AI chatbots as tools that could harm professions and partially replace human intervention, practitioners will see them as a chance to use them to help them in their daily work. Nutrition dietetic technicians (NDTs) and nutritionists are also utilising LLMs for feasibility [7].

Including the foundation models had opened up revolutionary possibilities, particularly in the field of healthcare- interpretation of data, prognosis and diagnosis of diseases, and care and management of patients. LLMs have recently demonstrated promising results through ChatGPT and GPT-4. These LLMs help with text-related tasks such as structured reporting and the interpretation of radiology reports. The integration of LLMs in this field can augment the interpretive skills of radiologists, facilitate patient-physician communication, and streamline workflows in clinical settings, that are usually happening in hospitals [8].

Unmatched capacity to interact with patients and provide them with different levels of medical information in an iterative fashion is the best feature. This may greatly increase the care of patients by healthcare provider . The applications of LLMs like GPT-3, Llama have to be safeguarded by mandatory algorithms assuring transparency in Decision Making [9]. LLMs are now being used in hands-on special diagnosis such Oral and maxillofacial surgery which includes Educational Tool, Patient Interaction, Surgical Planning and Decision Making, Research Assistance, Assistance in Clinical Documentation, Training Simulations, Remote Consultations, Psychological Support, Peer Collaboration, and Public Health Education. Transformers which are deep learning architecture that has encoding and decoding components are used in such applications. The input encoder transforms the words into numerical vectors in embedding space as abstract meaning [10].

Rather than fine-tuning on large amounts of domain data, LLMs are equipped with comprehensive training data and knowledge that can generate a variety of tones and symptom descriptions with suitable prompts. Appropriate prompts need to be designed for accurate response. The chatbot should be capable of inquiring key symptoms of depression, thorough questions and patient's responses [11]. A novel prompt framework can be used to create prompts that helps models interpret and extract medical information from clinical corpus. Prompt guided medicine predictions using transfer learning techniques are more accurate than a conventional pre-trained model, thereby reducing the computational requirements [12]. Different prompting strategies can be explored and analysed with unsupervised and distantly supervised emotional information on the basis of which we can experiment with LLMs for interpretable mental health analysis by guiding them to generate explanations for each of the decision. Strict human evaluations would be used to assess the accuracy of generated explanations [13].

The solution provided by prompt learning is the ability to use one model for several downstream tasks, but the prompt learning techniques used today rely on pricy templates for training. Chain of thought (COT) has demonstrated that improving how specific parts of the reasoning process are presented can boost LLM performance. AGCVT-Prompt, an auto-generated COT and verbalizer template technique, groups unlabelled texts based on their recognized topic and sentiment [14]. Recent research has shown that incorporating instruction prompts with LLMs substantially increase the efficacy of summarization tasks. However, storing the

Electronic health records (EHRs) encompassing medical diagnoses and test outcomes leads to increased performance variance and resulting in significantly distinct summaries even when prompts share similar meanings [15]. When the performance of GPT-3.5, GPT-4.0 are compared it was found that GPT-4.0 showed significantly higher accuracy and consistently performed well in all domains achieving over 50% accuracy in each domain [16].

GPT-4 exhibits significant promise across multiple scientific fields, such as biology, drug discovery, computational chemistry, materials science, and partial differential equations. GPT-4 shows potential for scientific research, but it also has some drawbacks. To make the most of GPT-4, researchers should exercise caution, double-check the model's outputs, try out various prompts, and combine its powers with specialized AI models or computational tools to guarantee accurate findings and top performance in their respective fields of study [11]. It is challenging to evaluate texts produced by Natural Language Generation (NLG) systems automatically. It has been demonstrated that traditional reference-based metrics like BLEU and ROUGE have comparatively minimal association with human evaluations, particularly for jobs requiring diversity and creativity, because they can be applied to new tasks without human references, large language models (LLMs) are recommended as reference-free metrics for NLG evaluation in recent studies [17]. Healthcare domain is a critical field which involves medicine and medical research, so unknown bias in LLMs may have detrimental effects on patient outcomes. Because LLMs generate text that reflects the training data they use, they may reinforce prejudices related to language, culture, gender, and race [18]. LLMs like GPT-3.5, GPT-4.0 and BARD show limited proficiency in Inductive, mathematical and Multi-hop Reasoning tasks which can be improved by fine-tuning and Data Augmentation [19].

III. Methodology

The proposed work had employed the Retrieval Augmented Generation (RAG) framework for seamless integration of external data into the chatbot development process. Initially, relevant data is gathered from sources based on the domain and objectives of the chatbot. Once the data is prepared, the RAG framework is employed to incorporate external data into the chatbot's knowledge base. Specialized retrievers are created for multiple data sources and are integrated into chatbot agents, enabling dynamic tool invocation for most relevant data retrieval and processing using the RAG framework.

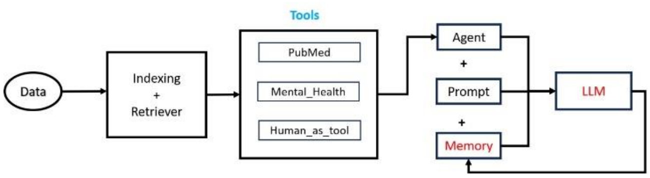


Fig-1: Work Flow of Diag-GPT

A. Data/Resources

Large Language Models have already been trained on large corpus but it's not been updated with the existing data. Even when a personalized model is to be created, data needs to be feed constantly. There are various data sources which can be integrated with the LLMs dynamically. The proposed methodology includes data sources like Conversational data of CSV format, PDF data. PDF data is used to fetch data from a pdf that user uploads. The main data source is PUBMED, by "The National Center for Biotechnology Information, National Library". This database - comprises more than 35 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full text content from PubMed Central and publisher web sites.

B. Retrieval Augmented Generation (RAG)

RAG is a Framework that enables an LLM to understand the external context passed as PDF, text file, videos, etc. and use this private knowledge for doing specific to task. RAG is a method for adding new data to enhance LLM knowledge. Although LLMs are capable of reasoning about a wide range of subjects, their expertise is restricted to the public data available to them during their training period. To develop AI applications that can make decisions about confidential information or data added after a model's cutoff date, you must add the precise information the model requires to expand its knowledge. Retrieval Augmented Generation (RAG) is the process of bringing the relevant data and inserting it into the model prompt. A RAG application contains two components, i.e., indexing and Retrieval and generation.

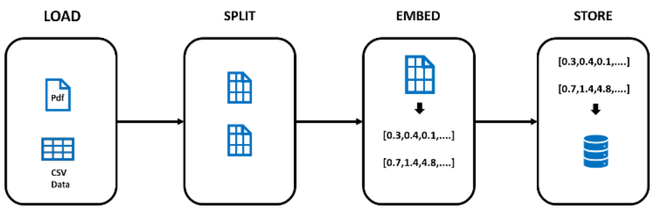


Fig-2: Architecture of indexing component of RAG

C. Indexing

Indexing includes collecting data from the resource, making some necessary modifications like splitting into chunks. These chunks are converted into embedding's and those embedding's are stored in vector database for further retrieval of information for a specific task.

Firstly, Diagnostic-GPT defined the specific data sources relevant to the chatbot's domain and used document loaders to load data from a source as Document's. A Document is a piece of text and associated metadata. For instance, there are document loaders for loading a simple .txt file. They can be used for loading the text contents of any web page, or even for loading a transcript of a YouTube video. Then splitting is done on the documents by dividing a large document into smaller, more manageable sections or segments. It is a crucial step in organizing and analyzing language data effectively. By breaking down a document into smaller units, it becomes easier to extract meaningful insights, identify patterns, and perform various language analysis tasks. The chunks of textual data are transformed into numerical representations using embedding models that can be processed by machine learning algorithms. These embeddings are used in various Natural Language Processing (NLP) tasks, such as understanding text, analyzing sentiments, and translating languages. In Lang Chain, these models can generate embeddings for both queries and documents. When a query is embedded, the text string is converted into an array of numbers, each representing a dimension in the embedding space. In order to store and search over unstructured data, it is embedded and stored the resulting embedding vectors. Then at query time, the embedded unstructured query are used to retrieve the embedded vectors, that are 'most similar' to the embedded query. Vector store is used for storing embedded data and perform vector search in the vector storage.

D. Retrieval and Generation

The objective is to develop a chatbot that takes a user query as input and is embedded into vectors and does semantic search in the vector store. The query then retrieves the similar context and gives as an input to LLM which generates response. LangChain defines a Retriever interface which wraps an index that can return relevant Documents given a string query. The most common type of Retriever is the Vector Store Retriever, which uses the similarity search capabilities of a vector store to facilitate retrieval.

In the context of RAG, the tool is invoked by the retriever to process the retrieved information and extract relevant data or insights. The tool is a specific function that an agent can use to communicate with the environment and access data. A tool acts as a connection point between an agent and different data sources or functions. Each tool is designed to perform a specific task. For example, a tool might be designed to query Wikipedia for information, translate text from one language to another, or access data from a CSV file. CSV, PDF, PubMed tools are employed to generate the most appropriate response. PubMed tool is to provide access to PubMed records and provides an easy approach to search data within it.

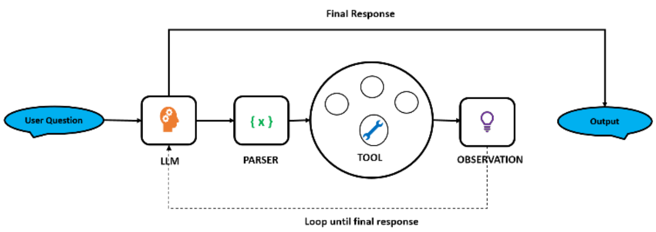


Fig-3: Architecture of Retrieval component of RAG

The agent is created as mentioned in Fig 3. The tools in RAG that promoted a modular, flexible, and adaptable chatbot architecture are used. This approach simplifies development, improves maintainability, and ultimately leads to more capable and informative chatbots. Agents are chains that has a unique ability of function calling, as many times as it needed in a loop. Agent is used as executor that can execute multiple tools in parallel, leveraging the computing resources efficiently and reducing processing time. This parallelism improves overall system throughput and responsiveness. Prompt template is created that provides a standardized format for querying the retriever component. This ensures consistency in how queries are formulated, making it easier to retrieve relevant information from external sources. By using a prompt template, the quality of queries is enhanced and is sent to the retriever. The template includes keywords, context, or structured information that help the retriever understand the user's intent and retrieve more accurate and relevant data. Prompt templates guides the retriever in selecting appropriate sources or data subsets for retrieval. This context can be crucial for retrieving information that is specific to the user's query or situation.

IV. RESULTS

The Fig-4 and Fig-5 showcase the responses generated by our chatbot utilizing the Retrieval-Augmented Generation (RAG) framework integrated with external data sources including PDF documents, CSV files and PUBMED. These responses highlight the chatbot's ability to retrieve and process information from diverse data types, providing contextually relevant and informative answers to user queries.

A. CASE STUDIES

CASE-1:

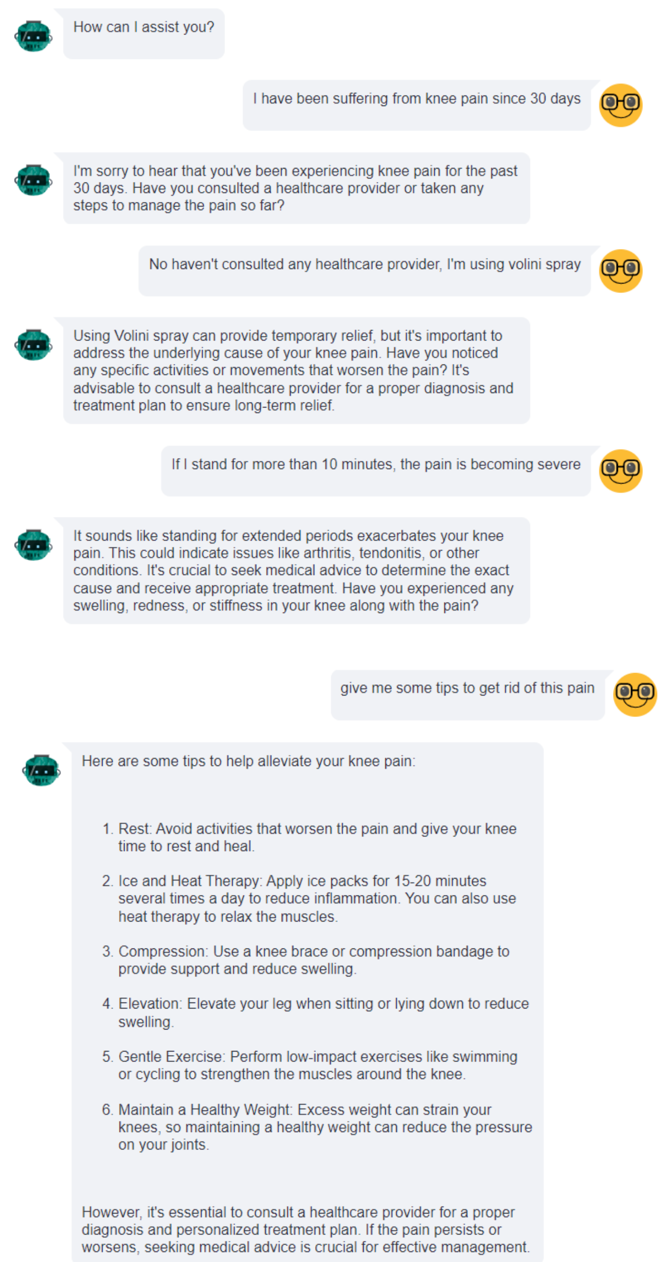
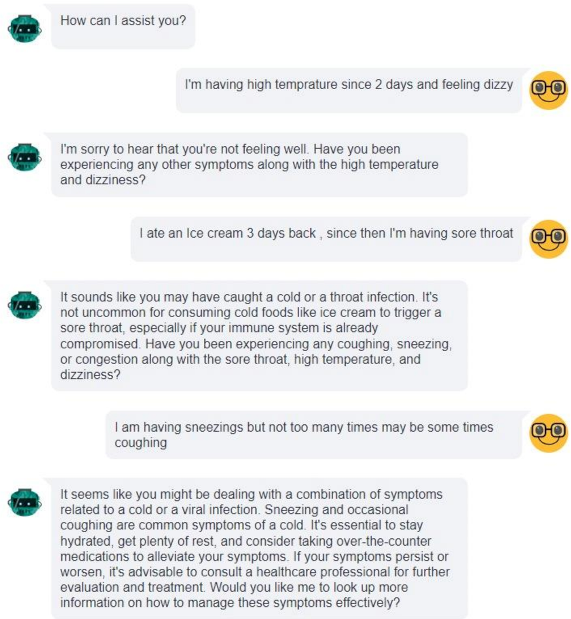


Fig-4: Response of MedGPT for health issue regarding knee

The conversation built on the health issue of knee pain between user and chatbot is depicted in the Fig3. The response is generated using LLM integrated with multiples tools such as csv, pdf and PubMed which retrieve data from external sources specialized for each data source, enabling the chatbot to fetch relevant information based on user queries. In the proposed system, when a user query requires information from an external data source, the corresponding agent invokes the appropriate retrieval tool. For example, if the query is related to a PDF document, the agent responsible for PDF data invokes the PDF retrieval tool to extract relevant information. The conversation started with enquiring the user regarding his query, followed by query of the user. While retrieving the response for knee pain, agent invoked csv data to check for the personalized. Physical health data of the user and further asked questions about health history. In the same way, when the user asked about the remedies, agent invoked the PubMed which retrieved the most relevant citation paper and generated response.

CASE-2:



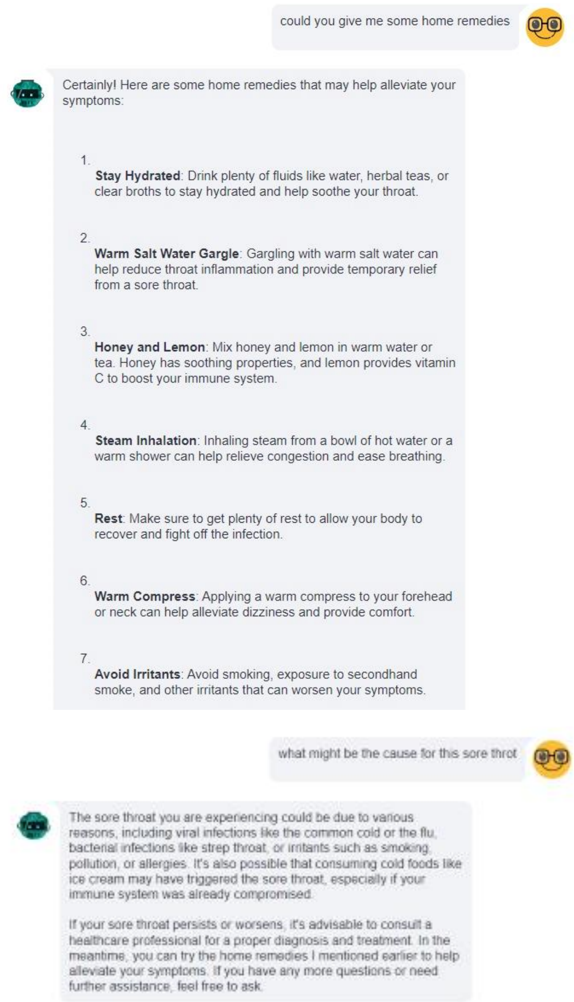


Fig-5: Response of MedGPT for health issue regarding Fever

The Conversational exchange between the user and chatbot on issue of fever and dizziness is depicted in Fig-4. Initially the chatbot responded by inquiring about the health history of the user and followed by analysis of the health history, response given by the chatbot addressing the query is quite accurate. It generated multiple possible remedies to relieve from illness with enhanced accuracy and empathy.

In one of the case study, the proposed architecture had used to address the patients question. The real time execution was done using an Interface created using streamlit. Patients queried the DIAG-GPT and found the descent results performed by the LLM.

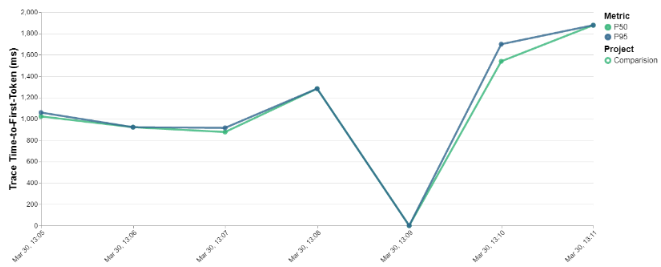


Fig-6: Tracing First Token per Minute

The graph in Fig-5 measures the time it takes for the LLM to generate the first token (TTFT) in its response, which is a metric for evaluating the model's responsiveness. A blue line trends slightly upwards over time, showing the TTFT measured at each second. These labels along the blue line refer to the 50th percentile (P50) and 95th percentile (P95) of the TTFT measurements. The P50 represents the median TTFT, and the P95 represents the TTFT that is slower than 95% of the measured responses.

Fig-7: LLM Latency per Minute

The graph shows how long it takes the LLM system to respond to a prompt at various points in time. The P50 and P95 values along the line indicate the median and the 95th percentile response times. Ideally, you would want these values to be as low as possible for a more responsive LLM. A Trace is a collection of runs that are related to a single operation in Fig-6. The scope of the LLM looks like it starts from the 1000ms for every token and that reaches at most of 1900ms there are positions where the graphs represents that matrices deviate from one another.

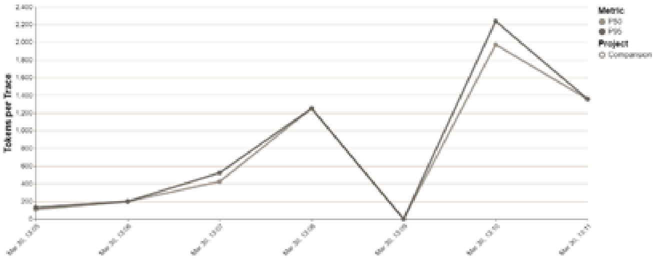


Fig-8: Tokens per Trace

LLM Latency is the crucial evaluation metric for LLMs. The graphs ranges from 200ms to 2000ms, with tick marks every 200ms. The P95 line shows the TTFT at each second, and it fluctuates between approximately 400ms and 1400ms over the course of the measured period. shown in Fig-7

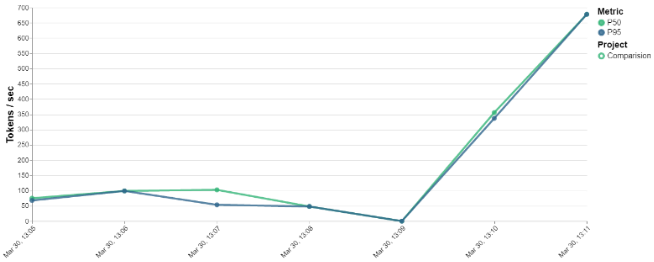


Fig-9: Tokens per Second

When evaluating LLMs, metrics often focus on aspects like response quality, accuracy, fluency, and response time**.** Tokens per Trace refers to the average number of tokens generated in a single LLM trace in Fig-8. This could be helpful to understand the model's verbosity or tendency to generate longer outputs. The P95 line shows the TTFT at each second, and it fluctuates between approximately 400ms and 1600ms over the course of the measured period.

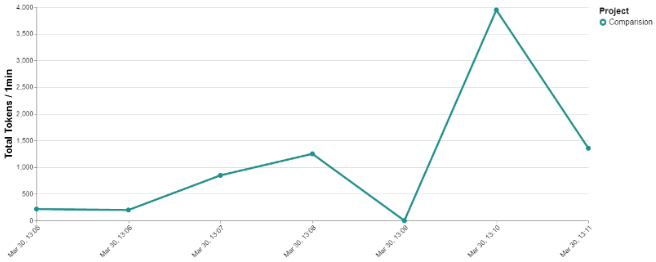


Fig-10: Total Tokens Retrieved per Minute

However, the TTFT is estimated which varies by at least 1200ms based on the scale of the graph in Fig-9. Half of the responses took less than 1000 milliseconds to generate the first token in the metric of P50 as in P95 95% of the responses took less than 1600 milliseconds to generate the first token.

The number of tokens generated per second for some duration where conversation existed between the chatbot and patient represents the token generation capability with in a second and the model performs by generating the 50 tokens for the first second and eventually grows according to the user query. The highest average number of tokens observed over 95% of the generation process was around 680 tokens.

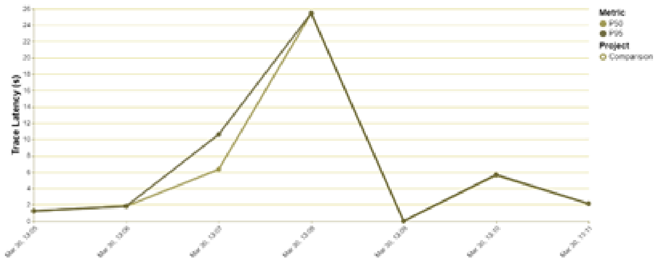


Fig-11: Trace Latency per Second

B. COMPARISON WITH EXISTING LLMs

The Latency distribution between the models were calculated. Latency refers to the time delay for data to travel from one point to another in a system. Counting the number of times the latency occurred and observed that an outlier suggests some of the query took too long to extract answer from the LLM that is around 495 seconds. The proposed framework has a steady and consistent latency output. The other models had a huge fluctuation starting from 1.14 to 495 seconds. The latency depends on the query that user give some questions take time to generate answer according to the question. The comparisons are shown in Table-1

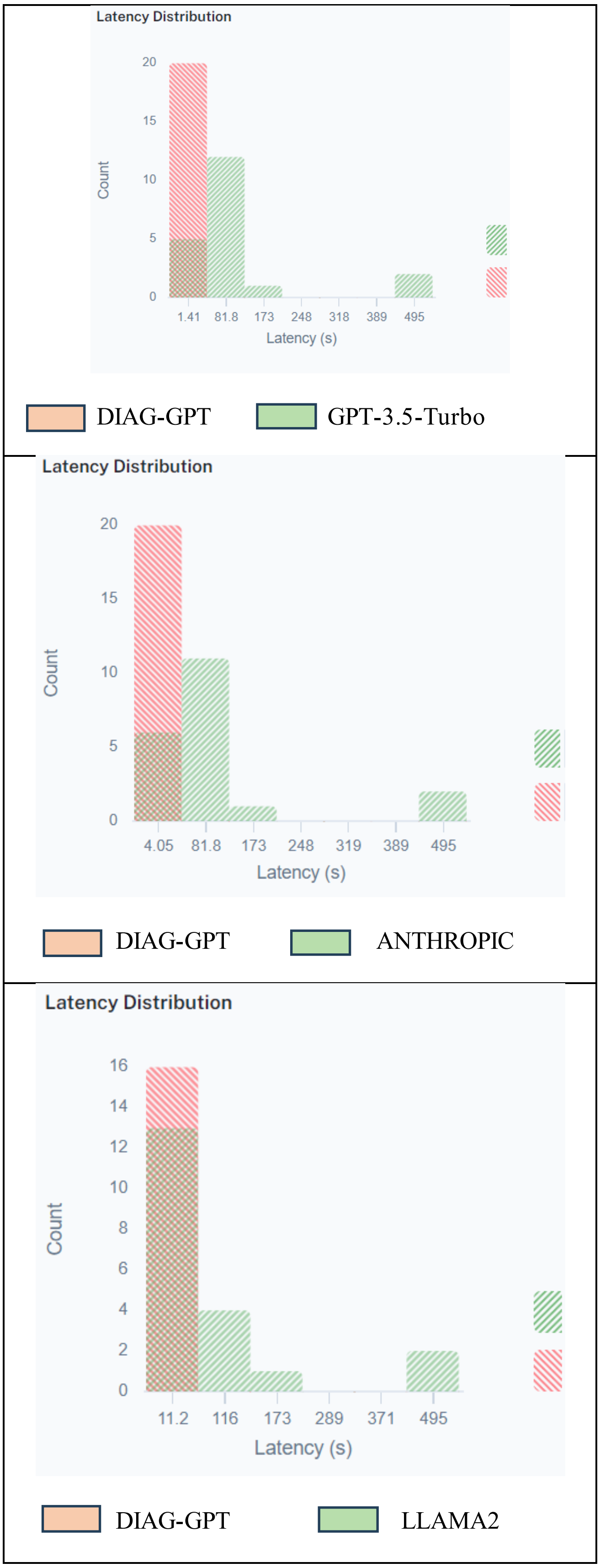


Table-1: Comparison of MED-GPT with Other LLMs

Feedback on various aspects are extracted for different LLMs. The LLMs considered are GPT-3.5-TURBO, LLAMA2, ANTHROPIC and DIAG-GPT. The Conciseness is the most important aspect that promotes the real world inspection of the problem. It improves readability and clarity. It keeps the audience engaged and ensures message is understood. COT contextual accuracy aims to bridge the gap between the raw capabilities of LLMs and human-like reasoning, making them more reliable and versatile tools for various tasks. It's important to consider the potential for harmfulness when developing or using new technologies, products, or ideas. By being mindful of potential risks, which can take steps to mitigate them and promote safety and well-being. Helpfulness is a valuable quality that can benefit both the giver and the receiver. By being helpful, chatbots can Strengthen relationships, create a positive environment and Boost well-being. Focusing on relevant information saves time and effort by directing your attention to what truly matters. It allows for better decision-making by ensuring you consider all the crucial factors. In communication, relevance helps keep conversations focused and prevents tangents.

V. Discussion & Conclusion

The research explores the potential of medical chatbots utilizing Large Language Models (LLMs) in healthcare, these chatbots simulate human-like interactions, offering personalized medical advice and information. It explores the advancement, applications, challenges, and future prospects of such chatbots, emphasizing their part in upgrading healthcare openness and quality Moreover it introduces an innovative method called the Retrieval-Augmented Generation (RAG) framework, which harnesses the power of external data sources, such as PDFs, CSVs, YouTube videos, and websites, to greatly enhance the richness and accuracy of information provided by chatbots. By seamlessly integrating external data sources like PDFs, CSVs, YouTube videos, and websites, the chatbot can provide contextually relevant responses, showcasing its versatility and adaptability. Customized chatbots enable efficient retrieval and processing of data, ensuring that user queries receive precise and timely responses. The chatbot's capability to engage users effectively and provide accurate information indicates an overall positive user experience.

Future research may involve further investigation into incorporating additional data sources, enhancing the accuracy of information retrieval, and integrating more sophisticated natural language generation techniques. Overall, the integration of external data sources into the chatbot's knowledge base through the Retrieval-Augmented Generation (RAG) framework holds huge potential for advancing the capabilities of conversational agents across diverse domains, especially in the healthcare sector.

VI. References

[1] Loredana Caruccio a , Stefano Cirillo a,∗ , Giuseppe Polese a , Giandomenico Solimando a , Shanmugam Sundaramurthy b , Genoveffa Tortora a. Expert Systems With Applications.

[2] Ou Zheng,PhD,* Dongdong Wang, Chenzhu Wang, Shengxuan Ding and Mohamed Abdel-Aty. TrafficSafetyGPT: Tuning a Pre-trained Large Language Model to a Domain-Specific Expert in Transportation Safety

[3] Konstantinos I. Roumeliotis a , Nikolaos D. Tselikas a,∗ , Dimitrios K. Nasiopoulos b. LLMs in e-commerce: A comparative analysis of GPT and LLaMA models in product review evaluation

[4] Shuhan Qi a,b,d , Zhengying Cao a , Jun Rao a,∗ , Lei Wang c , Jing Xiao c , Xuan Wang a,b What is the limitation of multimodal LLMs? A deeper look into multimodal LLMs through prompt probing

[5] Mahyar Abbasian, M.Sc.1*, Iman Azimi, Ph.D.1 , Amir M. Rahmani, Ph.D.1 , and Ramesh Jain, Ph.D.1. Conversational Health Agents: A Personalized LLM-Powered Agent Framework

[6] Malik Sallam 1,2. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns

[7] Angeline Chatelan, PhD, MSc, RD; Aurélien Clerc, MSc, RD; Pierre-Alexandre Fonta, MSc Eng. “ChatGPT and Future Artificial Intelligence Chatbots: What may be the Influence on Credentialed Nutrition and Dietetics Practitioners?

[8] Jiancheng Yang a,* , Hongwei Bran Li b,c , Donglai Wei d. The impact of ChatGPT and LLMs on medical imaging stakeholders: Perspectives and use cases,

[9] Julian Varghese1, *, Julius Chapiro2. ChatGPT: The transformative influence of generative AI on science and healthcare.

[10] B. Puladi, C. Gsaxner, J. Kleesiek, F. Hölzle, R. Röhrig, J. Egger: The impact and opportunities of large language models like ChatGPT in oral and maxillofacial surgery: a narrative review. Int. J. Oral Maxillofac. Surg. 2024; 53: 78–88..

[11] Siyuan Chen1 , Mengyue Wu2∗ , Kenny Q. Zhu3∗ , Kunyao Lan4 , Zhiling Zhang5 , Lyuchun Cui6 1,2,3,4,5Shanghai Jiao Tong University, Shanghai, China 6Shanghai Mental Health Center, Shanghai, China LLM-empowered Chatbots for Psychiatrist and Patient Simulation: Application and Evaluation

[12] Awais Ahmed a , Xiaoyang Zeng a , Rui Xi a , Mengshu Hou a,b,∗ , Syed Attique Shah c MED-Prompt: A novel prompt engineering framework for medicine prediction on free-text clinical notes

[13] Kailai Yang 1 Shaoxiong Ji ∗ 2 Tianlin Zhang ∗ 1 Qianqian Xie †1 Ziyan Kuang 4 Sophia Ananiadou 1,3. Towards Interpretable Mental Health Analysis with Large Language Models.

[14] Xu Gu a , Xiaoliang Chen a,b,∗ , Peng Lu b , Zonggen Li c , Yajun Du a , Xianyong Li a. AGCVT-prompt for sentiment classification: Automatically generating chain of thought and verbalizer in prompt learning.

[15] Yu-Neng Chuang1 , Ruixiang Tang1 , Xiaoqian Jiang2 , PhD, and Xia Hu1 , PhD 1 Rice University, Houston, TX; 2 University of Texas Health Science Center, Houston, TX SPeC: A Soft Prompt-Based Calibration on Performance Variability of Large Language Model in Clinical Notes Summarization

[16] Tongyao Wang a , Juan Mu b , Jialing Chen a , Chia-Chin Lin a,*, Comparing ChatGPT and clinical nurses’ performances on tracheostomy care: A cross-sectional study

[17] Yang Liu Dan Iter Yichong Xu Shuohang Wang Ruochen Xu Chenguang Zhu G-EVAL: NLG Evaluation using GPT-4 with Better Human Alignment

[18] Ethics of large language models in medicine and medical research,

[19] Jessica López Espejel, El Hassane Ettifouri ∗ , Mahaman Sanoussi Yahaya Alassan, El Mehdi Chouham, Walid Dahhane GPT-3.5, GPT-4, or BARD? Evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts.