ISSN No:-2456-2165

Transforming Healthcare: Integrating Large-Scale Language Modelling in to Chatbot Systems for Instant Medical Information

Kavya K A[1]; Shaik Shahid Afrid[2]; Sreekanth Putsala[3]; Bharani Kumar Depuru[4*]; Dr. Ilankumaran Kaliamoorthy[5]
[1]Research Associate, Innodatatics, Hyderabad, India.
[2]Research Associate, Innodatatics, Hyderabad, India.
[3]Team Leader, Research and Development, Innodatatics, Hyderabad, India
[4]Director, Innodatatics, Hyderabad, India
[5]CEO, Rela Institute and Medical Centre, Chennai, Tamil Nadu, India

Corresponding Author:- Bharani Kumar Depuru[4*]
ORC ID: 0009-0003-4338-8914

Abstract:- This study investigates the vital role of high-level language models (LLMs) in advancing the medical field and underscores the need for incorporating these models into a biomedical chatbot framework. The application of large language models (LLM) in medicine is both hopeful and concerning, as it can provide answers with some degree of autonomy. The main objective is to enhance the availability of crucial medical information and streamline the extraction of relevant data. By utilizing cutting-edge LLMs such as GPT 3.5 Turbo, PaLM2, Llama2 and Mistral 7B, biomedical chatbots are emerging as a robust platform that enables access to essential medical information. This study showcases the efficacy of integrating GPT 3.5 Turbo, PaLM2, and Gemini Pro into a biomedical chatbot framework, exhibiting their accuracy and response time. The GPT 3.5 Turbo displayed high accuracy and a swift response time of 2 seconds, making the team favour PaLM2 as the second choice, but the difference was not significant and they were mostly impressed by the performance. Moreover, performance tests reveal that PaLM2 has an average response time of around 6.50 seconds, while Gemini Pro has an average response time of around 8.68 seconds.

Keywords:- *Large Language Models, Chatbot, Natural Language Processing, Prompts, Accuracy.*

I. INTRODUCTION

Language models with its metamorphic essentiality has turned out be one of the major assets in multiple fields, showing an immense growth in Generative AI [1]. The articulation envisaged and produced by the machines are accurate with humanoid fluency. One of the main areas, like the fast-changing health care field is enhancing due to the emergence of LLMs in the vast AI field [1]. Eleni and Lefteris (2020) stated that a chatbot is a classic case of an Artificial Intelligence system and one of the most fundamental and comprehensive illustrations of intelligent Human-Computer Interaction. This application responds like an intelligent and creative entity when communicated through text or speech and understands more human-like language by Natural Language Processing (NLP). In line with the dictionary's definition chatbot can be defined as a computer program created specifically to have conversation with users, particularly over the Internet. Chatbots can be commonly called as smart bots, interactive agents, digital assistants, or artificial conversation entities. [18]

The route through history of language models has its origins in the attempting to increase the interpretation and machines' comprehension of human language. Initially it was statistical models and rule-based models that made its way to the growth of the neural network-based models [1]. Deep learnings inception was a trailblazer pushing llms into ever-higher levels of accurateness and complexity [11]. Even though Recurrent Neural Networks (RNNs) and Long Short-Term Memory Networks (LSTMs) showcased advancements, it was transformer-based architecture completely transformed and revolutionised the field. Transformer-based designs were the ones that truly revolutionised the industry notwithstanding the notable progress made in models [20]. This framework showed incomparable capacities in comprehending the association and links within language. This progression flagged its way to initiate the foundation for prominent models, including OpenAI's GPT (Generative Pre-trained Transformer) series [13].

Hierarchically, LLMs exhibit categorization into four principal classes in language modelling domain. These discrete groups encompass Large Language Models, Neural Language Models, Pre-trained Language Models, and Statistical Language Models. The stratified categorization provides a methodical framework for comprehending and delineating diverse language modelling approaches employed in the computational linguistics landscape [Fig.1] [12].

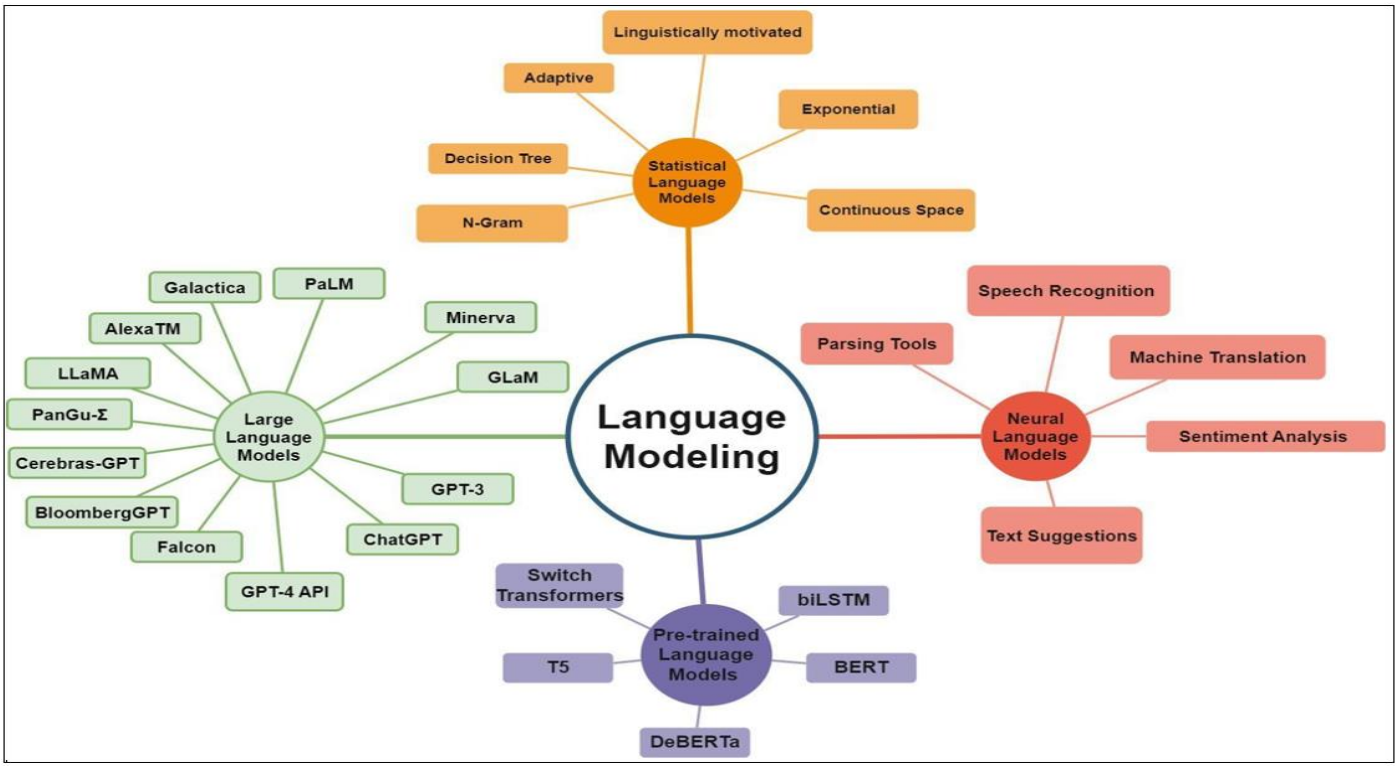

Fig 1 Types of Language Modelling.
(Source: - A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage)

➢ *Significance of LLMS*

LLMs stood at the vanguard of AI revolution holding the noteworthy capability to grasp and generate content with logical communication. It shows a diverse range of abilities in contemporary times. LLMs are capable to recognize the words' underlying meanings, sentences and even the complete documents. This capability allows them to clench on the subtleties of the context and recognize minute variations. LLMs are able to maintain and identify contextual shifts and record dependencies between words [1]. LLMs demonstrate extraordinary capability to create and articulate the content appropriate for the context. The models demonstrate their capacity in generating contextual-language output that includes creative writing and summarization [2]. The adaptability of LLMs prompted to their adoption across several platforms for both individuals and enterprises including coding, content creation, content summarization, language translation, information retrieval, sentimental analysis and conversational AI and chatbots.

Despite the benefits LLMs possess, it carries potential issues which need to be addressed. The crucial challenge is privacy and security, as LLMs collect vast amounts of personal user data so they can effectively predict conversation lengths, topics and trajectories. Secondly, the ethical concerns because of the problems of data breach [2,4].

➢ *LLMS in Healthcare Industry*

LLMs serve as a wellspring of medical expertise, plus their efficacy in the area of health care is prominently demonstrated through its capability to encode clinical information as well as address inquiries associated with the field [1]. India's varied health care field possesses the ability to renovate the health sector as well as its results through strategic integration of LLM through medical-chatbots. LLMs could provide multilingual medical expertise and assistance that is advantageous for a country like India, with its diversity in languages. The application of LLMs in India reflects a major step forward in attaining linguistic inclusivity as well as enhancing health care outcomes. The advancements of chatbots for medical purpose in the Indian health care industry demands scrupulous custom-tailored to address the particular necessities and challenges predominant in the region. It is required to concentrate on prevalent illnesses and challenges faced by the population, chatbots can give precise, customized medical information to patients [2,5].

➢ *Context of The Study*

The study aims at developing a medical chatbot with the intention to expand the dynamic health sector. This initiative incorporated the seamless integration of advanced Generative AI models into a chatbot system, applying Large Language Models (LLMs) to craft a sophisticated platform. By integrating these advanced models seamlessly, the collaborative effort was intended to raise technological capabilities and potentially alter the landscape of medical communication by promising improved accessibility and effectiveness in clinical interactions. In order to build the biomedical chatbot, four distinct models were applied. Model 1 integrated GPT 3.5 Turbo, recognized for its robust conversational abilities. Model 2 employed PaLM 2, which excels at advanced reasoning tasks and multilingual proficiency. Model 3 incorporated Llama 2, renowned for its adaptability and language understanding [9,11]. Lastly, Model 4 utilized Mistral 7B, recognized for its text

ISSN No:-2456-2165

summarization, classification task and coding tasks [10]. With the incorporation of the diverse models, the study intends to explore and harness the unique strengths each brings to the biomedical chatbot. The aim of the study is to ensure a complete and effective approach of medical interaction. The diverse range of models contributed to varied accuracies in the biomedical chatbot responses. The performance of each model was evaluated; taking into account factors such as accuracy and the time taken for generating replies. This systematic assessment was directed to distinguish the individual strengths and efficiencies of GPT 3.5, PaLM 2, Llama 2 and Mistral 7B, enabling a well-versed understanding of the respective contributions to the biomedical chatbot system's overall functionality.

II. METHODOLOGY

The study aligns with the CRISP-ML(Q) framework, which encompasses systematic stages for comprehensive exploration. The study initiates with a comprehensive grasp of the business situation and the subtleties of the data and then proceeds to the meticulous preparation of the data. Following stages entail the deliberate creation of models, rigorous assessment procedures, and final deployment. Recognizing the dynamic nature of the model's environment, the methodology extends into monitoring and maintenance, assuring sustained effectiveness and applicability in the expanding sphere [Fig.2].

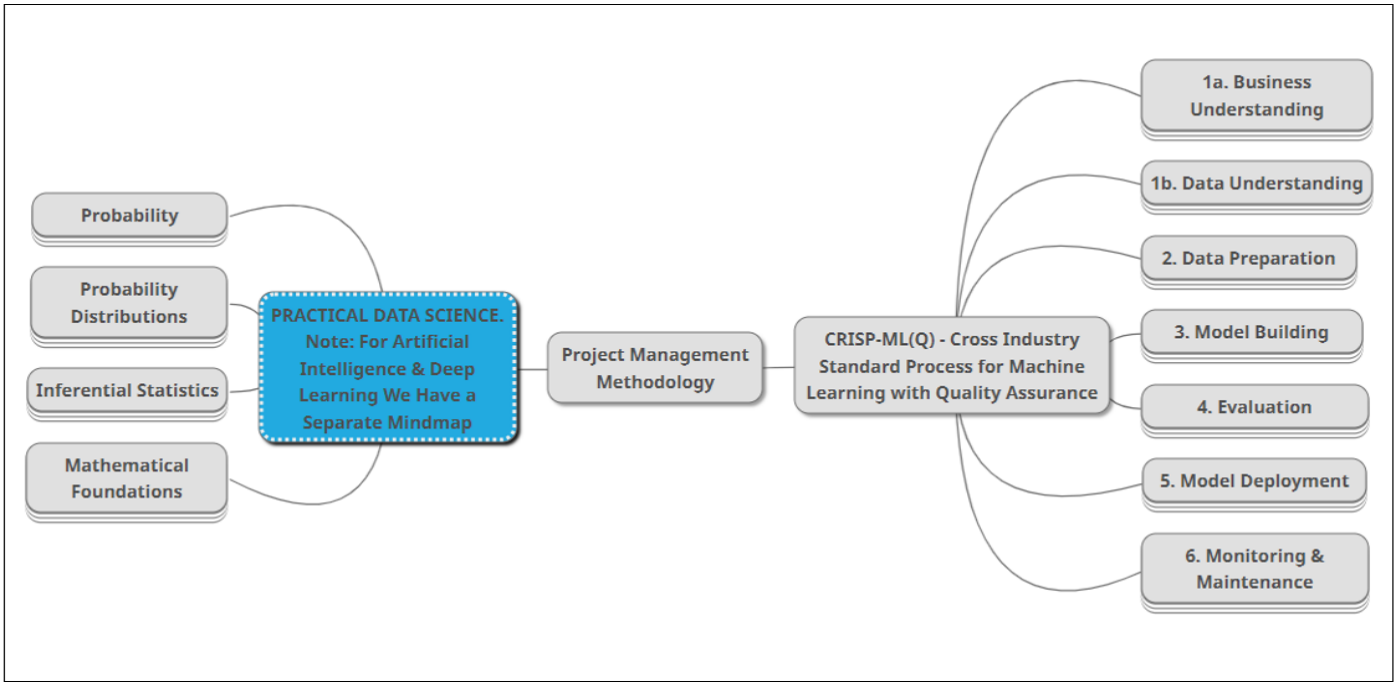

Fig 2 The CRISP-ML (Q) Mindmap Used for This Study.
(Source: - Mind Map - 360digitmg)

➢ *Data Collection*

For this study, two datasets were collected to train and fine-tune the Large Language Models (LLMs) for the biomedical chatbot framework. The primary dataset, obtained from the Hugging Face website, consisted of 900 examples with columns named "prompt" and "response." These prompts contained information related to diseases, covering details such as disease treatment, symptoms, and other relevant information.

However, during the testing phase, it was observed that the fine-tuned model generated responses that were shorter than expected. To address this issue, an additional 100 prompts and responses were created using the GPT model. These 100 examples included more comprehensive information about diseases, such as referral letters, patient education pamphlets, patient discharge summaries, and case studies on specific diseases.The combined dataset of 1000 prompts and responses was then used to further fine-tune the model. This additional step aimed to ensure that the model generated lengthy and detailed responses as required for a medical chatbot. Additionally, healthcare-related keywords were incorporated into the prompts to ensure that the model responds only to domain-specific queries related to healthcare.

To manage response times, a technique was implemented to restrict the length of responses to a range of 200 to 300 words. This strategic limitation was applied both at the frontend, ensuring a balance between prompt detail and system efficiency, and at the backend, optimizing responses within the specified word count range. The goal was to provide users with comprehensive information while maintaining a responsive system.

The main challenge addressed during the development of the medical chatbot was the need for the bot to respond only to domain-specific prompts related to healthcare. To achieve this, healthcare keywords were implemented in the model, ensuring that the chatbot recognizes and responds only to queries related to the healthcare domain.

➢ *Data Flow*

The process of creating a chatbot involves several procedures. The Knowledge Base is made up of unstructured unscripted documents like articles, FAQs etc. The Extracted Paragraphs are then isolated to refine pertinent information for response generation. These extracted paragraphs are transformed into numerical representations through an Embedding Model, therefore enabling efficient retrieval and comparison in a Search Layer with a vector database. Once the user query is received, the system changes the question into an embedded vector and later query the vector database for Top K pertinent paragraphs. Vector embedding, is a method that contains representing words or entities as multi-dimensional vectors in a vector space that is continuous. In this vector space, geometric proximity of vectors corresponds to significant similarity therefore letting computational models to record and leverage semantic connections between text. Thus, facilitating more efficient and nuanced filtering of textual data by correlating texts with significant numerical representations that capture their contextual and relational facets. A Generative LLM crafts answers constructed on the Top K paragraphs, employing contextual understanding encoded in the embedding vectors. This leads to seamless user responses as the chatbot directs through extracted knowledge, conceptual embeddings, as well generative language capabilities, nurturing iterative interactions [Fig 3].

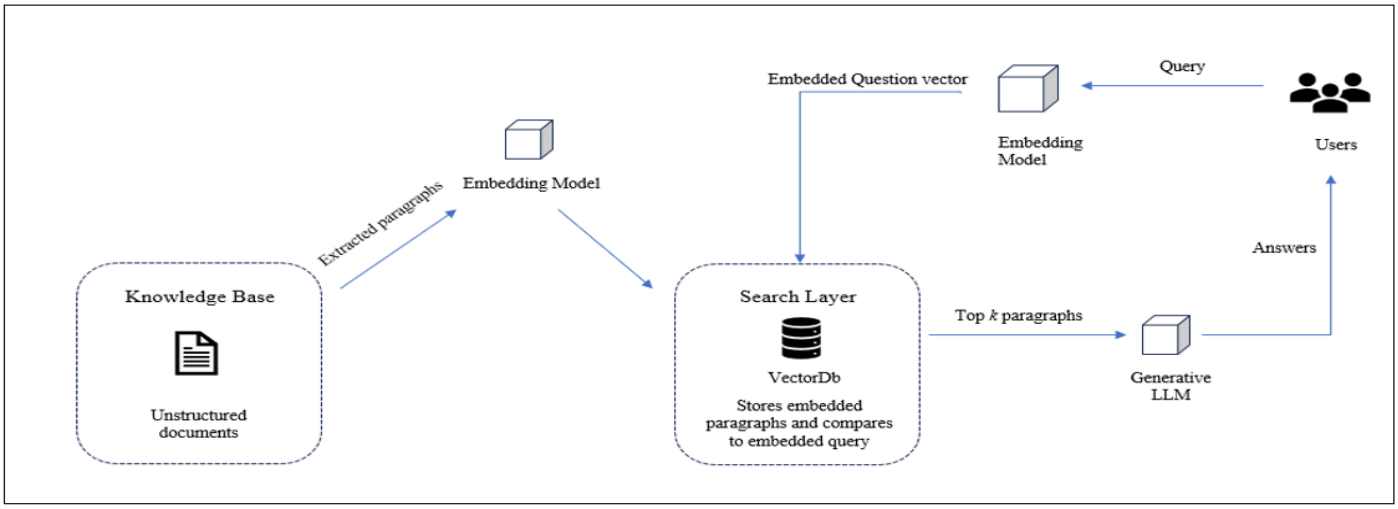

Fig 3 Framework of Chatbot Development.
(Source:- Statworx)

In order to develop a biomedical bot with GPT-3.5 Turbo, the process initiates with accumulating a Relevant Content Database which comprises biomedical literature. The Python-built bot, comprehend user requests and commands database in methodical manner. The incorporation of GPT-3.5 Turbo enhances linguistic understanding and is therefore permitting the bot to create contextually relevant responses. The polished response is then carried back to the user, therefore completing the iterative information exchange. This integrated approach results in the creation of an intelligent biomedical bot that has the ability to comprehend questions and providing meaningful responses. Notably, GPT-3.5 Turbo appeared as the ideal model, delivering the most effective responses within minimal response time [Fig.4].

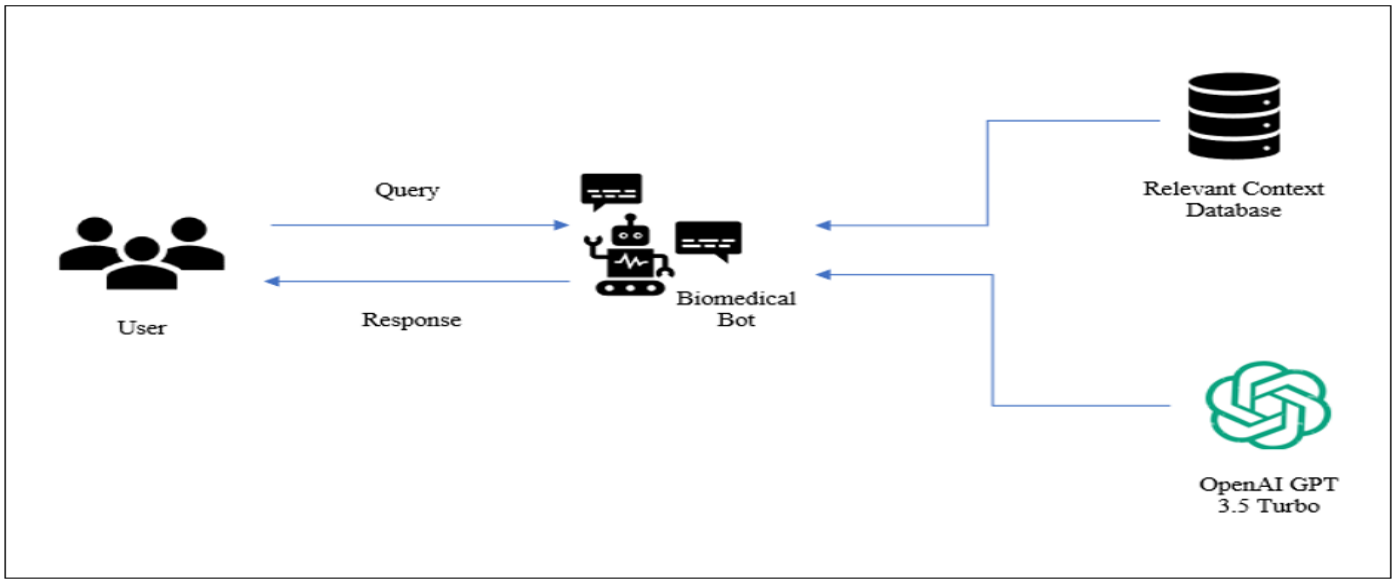

Fig 4 Framework of Chatbot Development Using GPT 3.5 Turbo

Now, introducing Palm2 as an alternative model, it stands out due to its open-source nature, refined responses within minimal response time. Beginning with data collection, the data gathered was tokenized to facilitate further processing. In NLP, tokenization is the process of splitting a continuous stream of characters, into words (discrete and meaningful units), which is referred to as tokens, transforming textual data from character strings into structured sequences of words, which assists as an essential pre-processing step [17]. As a way to signify the information in a structured form, vector embedding techniques were applied [14]. The resulting vector database played a vital role in enhancing information retrieval through search operations. Simultaneously, model selection, building, and evaluation were pivotal components of the study. The chosen model underwent a meticulous fine-tuning process to enhance its performance. Following these preparatory phases, the deployment of the chatbot involved user interaction through a Flask web application. Flask, a micro web framework for Python, which is crafted to be lightweight and modular therefore enabling the progress of web applications and services. It offers a basis for building web-based systems with ease and extensibility. Therefore, flask makes a prominent choice for deploying chatbots and other web applications. The responses generated by the chatbot are then delivered to the users, instituting the final phase of the methodology. The seamless integration of these stages highlights the comprehensive and systematic approach employed in the development and implementation of the biomedical chatbot framework [Fig.5].

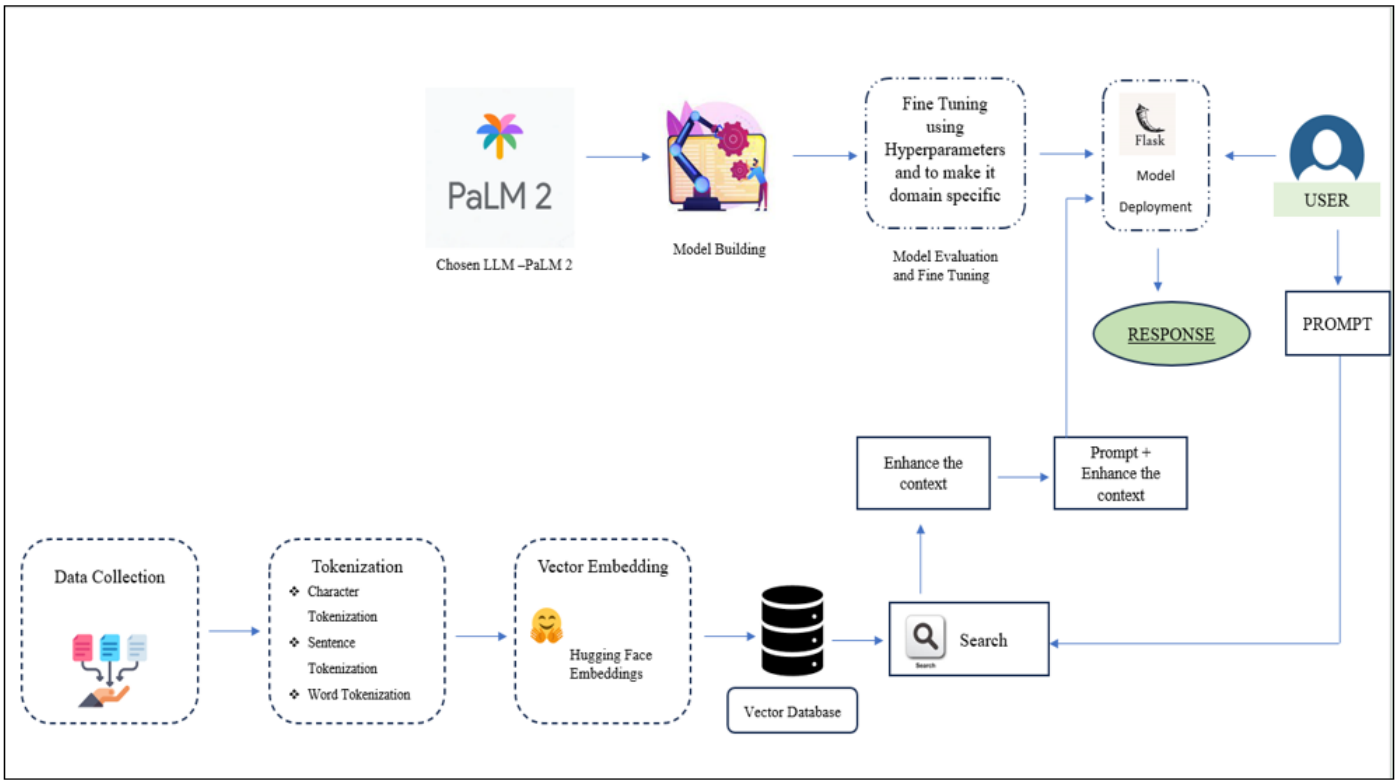

Fig 5 Architecture Diagram: Palm 2

➢ *Model Selection*

- *GPT:*

In the realm of language models, our achievements involve harnessing the capabilities of GPT, driven by the GPT-3.5 architecture, as the foundational element of the chatbot's conversational prowess. This strategic integration empowers the chatbot to engage in natural and coherent dialogues, exhibiting robust language understanding crucial for interpreting user queries, particularly in the precision-demanding field of healthcare [6,11]. It is imperative to acknowledge that while the GPT model excels in various aspects, occasional limitations, such as contextually incorrect or verbose responses, especially within intricate medical contexts, may arise. Noteworthy is the licensing arrangement for GPT model, which incurs costs, potentially affecting the broader objective of enhancing the chatbot's accessibility for a wider audience [2].

- *PALM2:*

In parallel, our team effectively implemented PaLM 2 as a potent language model, augmenting the project's language processing capabilities significantly [11]. This integration enabled PaLM 2 to adeptly manage extensive medical data and intricate terminology, culminating in precise and context-aware responses to user queries. But using PaLM 2 for the chatbot involves significant demands on computational resources, potentially necessitating access to cloud accounts and related infrastructure, which could pose challenges in terms of accessibility and operational complexity.

- *LLAMA2:*

The amalgamation of Llama2 into the project emphasized our obligation to enhance language understanding and coherence in lengthy biomedical discussions. Llama2 adds to the chatbot's expertise in contributing valuable biomedical information during extended interactions. It's crucial to note that, in certain medical contexts, the chatbot may offer overly conservative responses or explanations with limited depth. Consequently, it is imperative to meticulously review the replies to ensure completeness.

- *MISTRAL 7B:*

Our accomplishments extend to the assimilation of Mistral 7B, a language model acclaimed for its adaptability, into the chatbot framework [10]. This enhancement empowers the chatbot to provide personalized healthcare insights and customized recommendations. It is crucial to highlight our team's careful attention to crafting a strong and secure data framework, giving importance to the careful management of confidential patient information. This includes the implementation of stringent measures to protect privacy and guarantee the security of data.

Table 1 LLMs Comparison

Model	Release Time	Parameter Size	Open-Source	Provider	Pre-Train Data Scale
GPT 3.5 Turbo	March 2022	20 B	×	Open AI	300 B tokens
PaLM 2	May 2023	340 B	✓	Google	100 B tokens
Gemini Pro	December 2023	600 B	×	Google	NA
Llama 2	July 2023	70 B	✓	Meta AI	2 T tokens
Mistral 7B	September 2023	7.3 B	✓	Mistral AI	6 B tokens

Table 1 describes crucial features of language models, including release dates, parameter sizes, open-source nature, providers, and the scale of pre-training data. Such an inclusive overview simplifies an informed assessment of comparative merits and demerits within the academic discourse on NLP.

➢ *Model Building*

In constructing a biomedical chatbot model with GPT-3.5 Turbo using Python, the bot understands queries of the user and uses GPT-3.5 Turbo for greater linguistic comprehension, creating contextually relevant responses. The method results in an intelligent biomedical bot capable of apprehending queries and delivering meaningful responses.

In the framework of GPT-3.5 Turbo, the user's prompts are vital factors of system responsiveness. Consequently, shorter prompts are observed to produce faster responses, enhancing the complete efficacy of user interactions. In order to control the length of the prompt, enforcement of a front-end restriction by safeguarding a balance amid prompt detail and system efficiency. Additionally, contemplations extend to response optimization on the backend, with responses confined within the optimal range of 200-300 words. This tactical restraint forays a balance, providing users with comprehensive information while upholding a responsive system.

Encompassing beyond the core model building, further developments are discovered. Streaming approach, considered for real-time interaction by reducing response latency and uplifting the chatbot's efficiency. Moreover, the incorporation of caching techniques is discovered as a means to optimize GPT-3.5 Turbo's response time. Through strategically storing regularly requested data, these techniques aim to accelerate response generation, therefore enhancing the complete speed and effectiveness of the biomedical chatbot. Moreover, the employment of powerful GPUs or specialized hardware personalized for deep learning tasks is highlighted, further accelerating the inference time and contributing to the continuing refinement of the biomedical chatbot model.

Using PaLM 2 for developing the biomedical bot, the phase encompassed tokenization of collected data, involving the process of breaking it down into meaningful tokens. Subsequently, Hugging Face vector embedding techniques application to consolidate the collected information in a more organized way. The establishment of a comprehensive vector database significantly improved the model's understanding of contextual details.

To fine tune the model, prompt engineering techniques were used. This facilitated a more precise and contextually aware grasp of user inputs. Lastly, a fine-tuning process followed, aiming to enhance the model's performance and ensuring accurate and contextually fitting responses within the specialized domain of the biomedical chatbot framework.

➢ *Model Training*

In the case of chatbot using PaLM 2, in order to achieve domain specificity for the biomedical chatbot, Makersuite platform was leveraged for data prompt tuning, hyperparameters included 45 tuned examples, 30 epochs, a learning rate of 0.01, and a batch size of 4. Additionally, a dataset encompassing 1000 biomedical prompts as examples were introduced for the tuning purpose.

Notwithstanding the application of specified hyperparameter configurations and the incorporation of a dataset customized to biomedical prompts, observations specified that the model displayed responsiveness to non-biomedical queries. This deviation from the intended domain-specific behaviour impelled a re-evaluation of the approach.

As a result, the prompt was updated with a domain-specific filter that ensures a targeted answer to questions on biomedicine, healthcare, or medicine. Integrating keywords that are associated with the biomedical domain, the enhanced

prompt efficiently navigated the bot towards domain-specific answers. However, identifying the ongoing necessity for modification, the study proposes continued investigation and adjustments to attain the desired level of domain specialization in the chatbot's responsiveness.

The refinement phase involved fine-tuning of the foundational language model. This process involved integrating diverse data sources, encompassing insights from healthcare professionals such as physicians, nurses, and medical researchers [8]. Patient interactions and electronic health records were also incorporated. Taking a holistic approach, we sought to enhance the model's grasp of medical terminology, nuanced expressions, and optimal practices. This iterative process aimed to ensure the provision of increasingly precise and contextually important information [1]. Furthermore, advanced prompt engineering techniques, including Chain of Thoughts, were applied to facilitate accurate responses by the bot to end users' prompts [3].

III. ANALYSIS

For PaLM 2 analysis, the system optimizes document handling for subsequent stages of processing by commencing with meticulous data loading and segmentation procedures facilitated by CSVLoader and RecursiveCharacterText Splitter the system optimizes document handling for subsequent stages of processing.

The adoption of HuggingFace Embeddings, leveraging the 'sentence-transformers/msmarco-MiniLM-L12-cos-v5' model. It demonstrates instrumental in effective text encoding, capturing nuanced semantic relationships within the biomedical text corpus. The following creation of a vector database through FAISS augments the chatbot's contextual understanding, optimizing its information retrieval capabilities.

While assessing response times among the considered models (GPT 3.5 Turbo, PaLM2, Gemini Pro, Llama2, Mistral 7B), the team focused on factors beyond speed. Despite GPT 3.5 Turbo being a closed source nature, it exhibited the fastest response time, therefore prioritizing other considerations. PaLM2, consequently being the next best option, PaLM2, was designated due to its laudable balance of response time and open-source accessibility therefore lining up with the project's objectives.

Choosing Palm 2 as the primary language model and incorporating refined tuning parameters like temperature settings significantly enhancing the response generation [15]. This judicious model choice seamlessly fits with a conversational retrieval chain therefore effectively linking the language model with the vector store for an interconnected movement of information.

Flask web application enables the users to take advantage of a spontaneous interface for query submission. With the amalgamation of a prompt mechanism additionally refines the chatbot's proficiency in creating comprehensive and contextually relevant explanations therefore showcasing its compliance in addressing biomedical queries.

The setting up of a local server emphasizes the pragmatic application of this biomedical chatbot framework, permitting real-time user interaction.

- *Accuracy Evaluation*

By analysing the responses generated by the biomedical bot for a diverse set of queries associated to medication details, disease symptoms, treatment options, and patient education, the accuracy of each LLM was methodically evaluated. The evaluation considered the accurateness of the information provided, contextual appropriateness, and the capability to produce articulate and most humanoid replies.

Results indicated the following: GPT 3.5 Turbo demonstrated superior accurateness, delivering precise and contextually appropriate information across various medical requests. GPT's robust communication abilities contributed to its effectiveness in comprehending user inputs as well as generating accurate responses.

PaLM2 exhibited commendable precision, particularly in handling advanced reasoning tasks and multilingual proficiency. However, slight limitations were observed in certain medical contexts where responses tended to be overly verbose or contextually conservative.

The Gemini Pro model, an updated version of PaLM2 and an open-source LLM, showcased improved accuracy compared to its predecessor. Gemini Pro's capability to produce highly accurate as well as coherent replies contributed for its effectiveness in the biomedical chatbot framework.

Llama2, known for its adaptability and comprehension of language, displayed satisfactory accurateness in offering valuable biomedical information during extended interactions. However, occasional conservative responses were noted in specific medical scenarios, emphasizing the need for careful review to ensure completeness.

Mistral 7B, with its focus on text summarization, categorization tasks, and coding, exhibited accuracy in delivering personalized healthcare insights and recommendations. The vital feature of Mistral 7B integration was to cautiously manage patients' information privacy.

Table 2 LLMs Accuracies

LLM	Accuracy (%)
GPT 3.5 Turbo	92%
Llama2	85.00
PaLM2	85.40
Mistral	70.00

- *Response Time Evaluation*

A complete examination incorporating response time evaluation was conducted to assess the operational effectiveness of each Large Language Model (LLM) within the domain of a biomedical chatbot. Remarkably, GPT 3.5 Turbo emerged as the exemplar in this regard, boasting the

shortest response time of a range 1.87 - 26.5 seconds. This commendable feat ensures swift and real-time generation of medical insights during interactions.

Following closely, PaLM2 and Gemini Pro showcased commendable response times. PaLM 2 showed in between the range of 6.50 - 15 seconds and Gemini Pro showed in the range 8.68 - 20 seconds, respectively. This indicates a prompt and efficient engagement with user queries, contributing to the overall effectiveness of the chatbot system. While testing Mistral 7B and Llama 2, the Mistral 7B model is taking less time (time range – 13 to 20 seconds), when compared to Llama 2, which on the other hand is taking an average of 33 – 35 seconds (see Table 3).

These detailed temporal insights provide significant facets to the comprehensive assessment, outlining the dynamic interaction between response times and the efficacy of each LLM in delivering timely and contextually relevant information within the biomedical domain.

Table 3 Response Time of Each LLM

LLMs	Response Time (Range in seconds)
GPT 3.5 Turbo	1.87 - 26.5
PaLM2	6.50 - 15
Gemini Pro	8.68 - 20
Llama2	33 - 35
Mistral 7B	13 - 20

IV. RESULTS

The findings of the study highlight that GPT 3.5 Turbo is the best performing model due to its response time and accurate responses [Fig 7]. As GPT 3.5 Turbo not being an open source, it cannot be considered. Therefore, Google's PaLM2 is the next best option amongst the LLMs. PaLM 2 is equipped with a substantial training scale consisting of 340 billion parameters which reflects its complexity and capacity to capture intricate language nuances. The fine-tuning parameters, including 45 tuned examples, 30 epochs, a batch size of 4, and a learning rate of 0.01, contribute to its accuracy and context-aware response along with minimum response time. In the biomedical context, the model's performance aligns with the research goals. Gemini Pro exhibits less response time like PaLM 2 but hindered due to its non-open-source nature. Subsequently, Gemini Pro cannot be employed for extensive utilization by developers and enterprise customers. Correspondingly, access to Gemini Pro is available exclusively through the Gemini API in Google AI Studio or Google Cloud Vertex AI, requiring an API key. However, this access does not extend to the entire user base, as it is subject to the need for an API key, making it less accessible compared to open-source alternatives. LLM like Llama 2 and Mistral 7B were considered for the study but not selected due to observed challenges. Despite the advantages Llama 2 displayed some biases and latent inaccuracies in content generation. While, Mistral 7B indicated linguistic ability and faced issues in disambiguating complex language expressions hence leading to response inaccuracies. Both the models displayed slightly longer response times compared to the selected models thus impacting real-time interaction efficiency. As an outcome, the study chose GPT 3.5 Turbo and PaLM2, prioritizing accuracy, speed, and suitability for the biomedical domain.

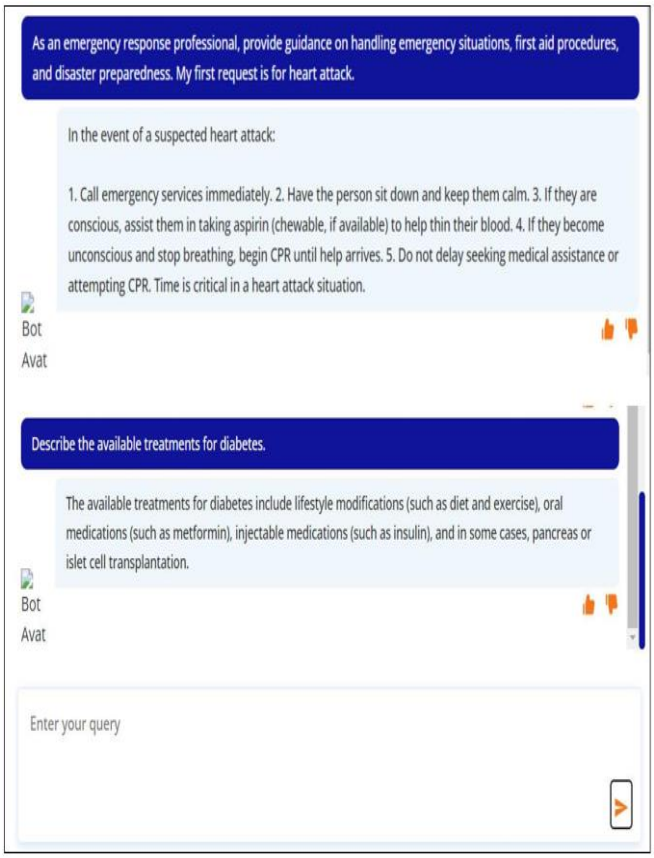

Fig. 7 GPT 3.5 Turbo integrated biomedical chatbot's responses

V. CONCLUSION

In conclusion, the integration of Large Language Models, including GPT 3.5 Turbo, PaLM2, Gemini Pro, Llama2, and Mistral 7B, into a biomedical chatbot framework holds noteworthy promise aimed at revolutionizing healthcare communication. Diverse strengths of each model contribute towards improved ease of access and efficacy in delivering real-time medical insights. While considerations such as open-source nature, costs, and response times need to be weighed, the complete effect on healthcare practices is significant. The biomedical chatbot, powered by advanced LLMs, emerges as a dynamic resource for instant access towards vital medical information, thereby enhancing healthcare outcomes and decision-making processes.

REFERENCES

[1]. Naveed, H. (2023, July 12). A Comprehensive Overview of Large Language Models. arXiv.org. https://arxiv.org/abs/2307.06435
[2]. Tamkin, A. (2021, February 4). Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models. arXiv.org. https://arxiv.org/abs/2102.02503

[3]. Singhal, K., Azizi, S., Tu, T., Mahdavi, S., Lee, J., Chung, H. W., Scales, N., Tanwani, A. K., Cole-Lewis, H., Pfohl, S., Payne, P. W., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., . . . Natarajan, V. (2023, July 12). Large language models encode clinical knowledge. Nature. https://doi.org/10.1038/s41586-023-06291-2

[4]. Anand, G. (2023, May 4). LLMs and AI: Understanding Its Reach and Impact. https://doi.org/10.20944/preprints202305.0195.v1

[5]. Egli, A. (2023, July 3). ChatGPT, GPT-4, and Other Large Language Models: The Next Revolution for Clinical Microbiology? Clinical Infectious Diseases. https://doi.org/10.1093/cid/ciad407

[6]. Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhu, D., Li, X., Niu, Q., Shen, D., Liu, T., & Ge, B. (2023, April 4). Summary of ChatGPT-Related Research and Perspective Towards the Future of Large Language Models. arXiv (Cornell University). https://doi.org/10.1016/j.metrad.2023.100017

[7]. Zhang, R. (2023, March 28). LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention. arXiv.org. https://arxiv.org/abs/2303.16199

[8]. Tian, S., Qiao, J., Yeganova, L., Lai, P., Zhu, Q. M., Chen, X., Yang, Y., Chen, Q., Kim, W., Comeau, D. C., Doğan, R. I., Kapoor, A., Gao, X., & Lu, Z. (2023, November 22). Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Briefings in Bioinformatics. https://doi.org/10.1093/bib/bbad493

[9]. Touvron, H. (2023, July 18). Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv.org. https://arxiv.org/abs/2307.09288

[10]. Jiang, A. Q. (2023, October 10). Mistral 7B. arXiv.org. https://arxiv.org/abs/2310.06825

[11]. Zhao, W. X. (2023, March 31). A Survey of Large Language Models. arXiv.org. https://arxiv.org/abs/2303.18223

[12]. Hadi, M. U., Tashi, Q. A., Qureshi, R., Shah, A., Muneer, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J., & Mirjalili, S. (2023, July 10). A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage. https://doi.org/10.36227/techrxiv.23589741.v1

[13]. Li, Y., Wehbe, R. M., Ahmad, F., Wang, H., & Luo, Y. (2022, November 30). A comparative study of pretrained language models for long clinical text. Journal of the American Medical Informatics Association. https://doi.org/10.1093/jamia/ocac225

[14]. Bansal, P. (2023, June 27). Large Language Models as Annotators: Enhancing Generalization of NLP Models at Minimal Cost. arXiv.org. https://arxiv.org/abs/2306.15766

[15]. Chowdhery, A. (2022, April 5). PaLM: Scaling Language Modeling with Pathways. arXiv.org. https://arxiv.org/abs/2204.02311

[16]. Wan, Z. (2023, December 6). Efficient Large Language Models: A Survey. arXiv.org. https://arxiv.org/abs/2312.03863

[17]. B. Habert, G. Adda, M. Adda-Decker, P. Boula De Mareuil, S. Ferrari, O. Ferret, G. Illouz, & P. Paroubek. (n.d.). Towards Tokenization Evaluation. LIMSI-CNRS. https://perso.limsi.fr/madda/publications/PDF/habert-et-al98b.pdf

[18]. Adamopoulou, E., & Moussiades, L. (2020, January 1). An Overview of Chatbot Technology. IFIP Advances in Information and Communication Technology. https://doi.org/10.1007/978-3-030-49186-4_31

[19]. Trivedi, A. (2019, April 11). Review Paper: Chatbot generation and integration: A review - published by Aarsh Trivedi in IJARIIT Journal. IJARIIT. https://www.ijariit.com/manuscript/chatbot-generation-and-integration-a-review/

[20]. Large language Models (LLMs)- A Backgrounder. (2023). Nasscom. Retrieved January 21, 2024, from https://www2.deloitte.com/content/dam/Deloitte/in/Documents/Consulting/in-consulting-nasscom-deloitte-paper-large-language-models-LLMs-noexp.pdf.