AI-Driven Diagnosis Predictive Chatbot for Healthcare

Tanay Kumar[1], Ayushi Tiwari[2], Sonam Bhul[3], Prof. Dr. Mohit Kumar[4]
Email Id: [1]tanaykumar0903@gmail.com, [2]ayushitiwari.july.07@gmail.com, [3]sonam.bhule9696@gmail.com, [4]mohit.kumar@mituniversity.edu.in

Department of Information Technology, MIT Art, Design and Technology University

Abstract- An AI-Driven Predictive Diagnosis Chatbot for Healthcare is a tool which uses artificial intelligence (AI) in assistance for diagnosing medical conditions based on user inputs. This article develops a software solution from this paper, which is fueled by RAG, trying to overcome limitations of using basic LLMs, making sure three layers - retrieval layer which uses the vector database with relevant information of indexed contextualized outside sources. The augmented one enriched it with other context and at last the generation will finally blend the ability of producing proper and appropriate responses within this LLM and augmented data. This architecture is implemented using semantic chunking, hybrid search techniques, and metadata integration to enhance the retrieval accuracy. The software utilizes lightweight LLMs such as Gemini-Pro, LlaMA 2 and Mistral for efficiency and scalability. In addition, the design is multi-platform compatible, including mobile applications and integration with telemedicine workflows, which makes it especially suitable for domains such as healthcare and enterprise solutions.

Keywords- *Gemini-Pro, LlaMA 2, Large Language Model (LLM), Mistral, Retrieval-Augmented Generation (RAG)*

I. INTRODUCTION

Healthcare systems in many parts of the world are facing great challenges, such as overcrowded hospitals, a shortage of medical professionals, and an increased number of patients. These have made the capacity of healthcare providers overstretched and created obstacles for timely care. Many others, especially those living in remote or underserved areas, have limited access to medical professionals, thus not receiving necessary medical advice or diagnoses.

It becomes essential at this point to have a chance to detect health problems early. In such conditions, early detection of these problems does improve patient outcomes greatly, primarily for chronic and life-threatening diseases such as cancer, diabetes, and heart-related conditions. Therefore, to address these issues, an AI-driven predictive diagnosis chatbot can be quite significant while improving accessibility, reducing health care system burden, or speeding up medical interventions appropriately.

Our solution utilizes an AI-driven predictive diagnosis chatbot built around a streamlined three-layer pipeline: retrieval, augmentation, and generation.

1. **Retrieval Layer:** The system retrieves relevant information efficiently by utilizing a vector database that indexes external data sources. This allows for precise and rapid extraction of relevant medical data based on user queries, ensuring the chatbot can provide timely responses.
2. **Augmentation Layer:** Once the information is retrieved, the augmentation layer enriches it with additional context, ensuring that the data is not only accurate but also comprehensive. This enhances the ability of the chatbot to provide insightful and medically sound recommendations.
3. **Generation Layer:** In the final step, the enriched context is mixed with the inherent capabilities of lightweight large language models such as Gemini Pro, LlaMA 2, and Mistral to generate high relevance and accuracy in the responses that are tailored to the user's needs. This ensures that the diagnostic insights provided are actionable and reliable.

To further optimize performance, the solution incorporates advanced techniques such as semantic chunking, hybrid search strategies, and metadata integration. These enhancements improve the retrieval process, ensuring that the system can handle complex queries quickly and accurately.

Furthermore, the solution is designed to be multi-platform compatible, so it can be easily adopted in mobile applications and telemedicine platforms. This flexibility is critical in healthcare settings where accessibility, accuracy, and user-friendliness are crucial. It enhances healthcare delivery, empowers patients, and provides medical professionals with a powerful tool to support their decision-

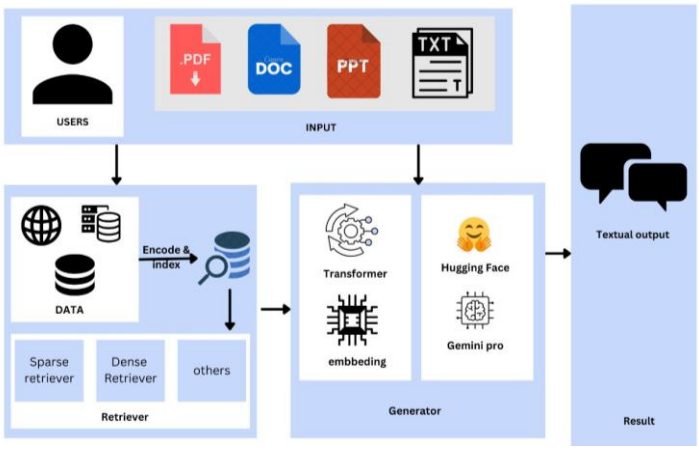
making by supporting remote workflows and seamless integration with existing healthcare systems.

II. RELATED WORK

This section examines the current body of research regarding dental chatbots and the incorporation of Large Language Models (LLMs) alongside Retrieval-Augmented Generation (RAG) techniques within the healthcare sector. The investigations concentrate on enhancing the performance, accessibility, and efficiency of chatbots, particularly within the realms of dental and healthcare applications.

Year	Title	Author	Key Features
2023	A systematic review and meta-analysis on ChatGPT and its utilization in medical and dental research	Hiroj Bagde, Mohammad Khursheed Alam	Integration of LangChain, RAG enhancements, and optimized LLMs to improve productivity, accuracy, and customization in dental health chatbot interactions.
2024	Chatbots and Large Language Models in Radiology: A Practical Primer for Clinical and Research Applications	Rajesh Bhayana	Exploration of LLMs in radiology, highlighting their superior performance in benchmarks while cautioning against potential inaccuracies in clinical applications.
2024	Health-LLM: Personalized Retrieval-Augmented Disease Prediction Model	Mingyu Jin, Qinkai Yu, Chong Zhang, Dong Shu	Fine-tuning of Health-LLMs with LLaMA-2 for disease prediction, achieving 83.3% accuracy, but facing challenges with performance on varied patient data outside the training scope.
2024	Integrating RAG with LLMs in Nephrology: Advancing Practical Applications	Jing Miao, Charat Thongprayoon, Supawadee Suppadungsuk, Oscar A. Garcia Valencia, Wisit Cheungpasitporn	LLMs combined with RAG techniques in nephrology to improve accuracy and relevance in medical advice for chronic kidney disease, while addressing issues like inaccuracies and hallucinations in AI-generated information.
2024	Leveraging LLM: Implementing an Advanced AI Chatbot for Healthcare	Ajinkya Mhatre, Sandeep R. Warhade, Sayali Kokate, Omkar Pawar, Samyak Jain	LLM-powered healthcare chatbots, achieving 61% accuracy in answering health queries, while highlighting issues of accuracy, bias, and ethics in AI-driven medical advice.
2024	Building Certified Medical Chatbots: Overcoming Unstructured Data Limitations with Modular RAG	Leonardo Sanna, Patrizio Bellan, Simone Magnolini, Marina Segala, Saba Ghanbari Haez, Monica Consolandi, Mauro Dragoni	Use of modular RAG in certified medical chatbots to achieve 85% retrieval accuracy for medical files, while addressing challenges in topic modeling and consistency in unstructured text.
2024	A Medical Chatbot: Your Healthcare Assistance	Harsh Jain	Development of a medical chatbot using the LLaMA 2 model to provide accurate information and emotional support
2024	Efficiency-Driven Custom Chatbot Development: Unleashing LangChain, RAG, and Performance-Optimized LLM Fusion	S. Vidivelli, Manikandan Ramachandran, A. Dharunbalaji	Development of a highly efficient medical chatbot using LangChain, RAG, and fine-tuned LLMs like LoRA and QLoRA

III. RAG ARCHITECTURE DIAGRAM



The diagram depicts a RAG approach. The retrieval-augmented generation enhances the large language models, by retrieving relevant information in the knowledge base to answer queries provided by the user. In this process, a query is converted into an embedding, and then searching a knowledge base for embeddings of similar queries; thereafter combining the retrieved information with that of the query is the input to the LLM to generate a more informative response.

IV. PROPOSED METHOD AND WORKFLOW

To develop a reliable and efficient dental chatbot. The proposed model uses advanced AI technologies like LangChain, Retrieval-Augmented Generation (RAG), and Performance-Optimized Large Language Models (LLMs) to design an intelligent and responsive conversational agent specifically for dental healthcare. The methodology is broken down into two main parts: computational methodology, which focuses on the technical processes and problem-solving strategies, and experimental methodology, which outlines the practical steps and testing procedures used to bring the model to life.

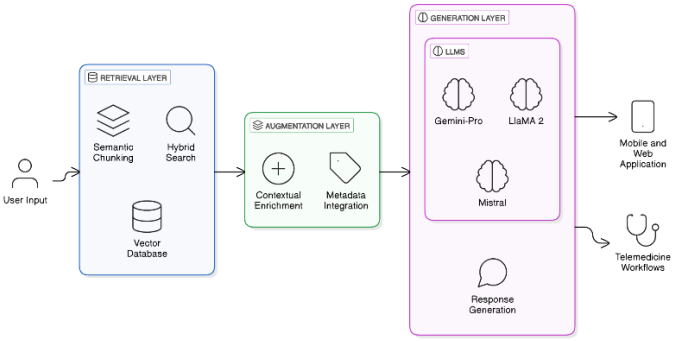
The primary challenge addressed in this study is the development of a dental chatbot capable of providing accurate, relevant, and personalized responses while overcoming the limitations of traditional AI systems, such as:

- Inaccurate and hallucinated outputs from LLMs.
- Limited accessibility to structured dental knowledge.
- Performance inefficiencies in real-time query processing.
- Integration with unstructured and semi-structured data sources.

To address these issues, the proposed solution focuses on:

- Integrating LangChain for customization and modular design.
- Using RAG to retrieve real-time data and mitigate hallucinations.
- Incorporating fine-tuned LLMs like LLaMA 2, LoRA, and QLoRA to enhance response generation and accuracy.

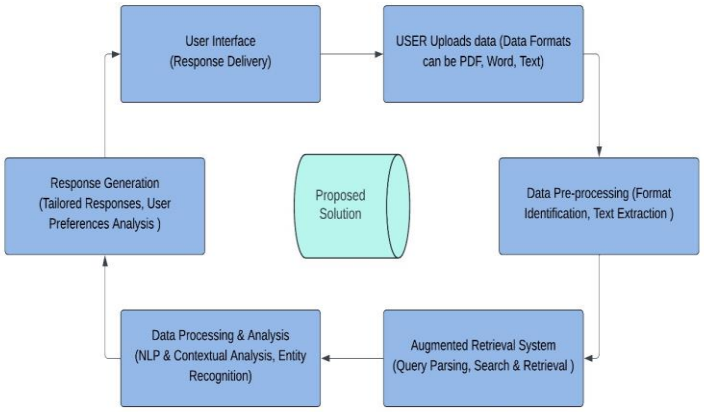
3.1. ARCHITECTURE



The chatbot uses a three-layer architecture:

- Retrieval Layer: Leverages a vector database to retrieve relevant information using RAG.
- Augmentation Layer: Enhances the retrieved information with contextual details.
- Generation Layer: Uses fine-tuned LLMs to generate precise, context-aware responses.

3.2. BLOCK DIAGRAM OF THE PROPOSED MODEL



The diagram describes an intelligent system that would accept user-uploaded data in one of the formats mentioned: PDF, Word, or plain text. The data goes through a pre-

processing where the formats are identified and extracted. The process continues by an augmented retrieval system with functionalities such as query parsing, search, and retrieval so that the right information will be acquired. The data obtained is then analyzed through NLP, contextual techniques, and entity recognition to get meaningful insights. Finally, the response is generated based on the user's preference and sent through a user-friendly interface so that communication is accurate and efficient.

3.3. WORKFLOW OF MODEL

The proposed method employs Google's Generative AI together with the LangChain framework to collect, process, and analyze document content from uploaded PDF files. It allows users to ask questions about documents with the goal of obtaining valid insights. The workflow involved can be divided into following steps:

1. It allows users to import PDF files. The program deals with an array of files and aggregates all information for further processing.

2. Text Content Extraction Files are uploaded which are processed with PdfReader to extract text contents. Scanned or Image-based PDF files require an addition of OCR tool (like Tesseract) to improve accuracy.

3. Content partitioning: The content is broken up into chunks for consumption using the RecursiveCharacterTextSplitter; this is both for indexing and retrieval, in preserving sections 'meaning.

4. Vectorization and Storage: The text segments are turned into vectors using the embedding from Google's Generative AI (GoogleGenerativeAIEmbeddings) and then stored in a vector database through FAISS for fast similarity queries.

5. The interface submits an inquiry to which the query is parsed to find the most relevant text fragments that capture the semantic similarity.

6. It will generate answers with the integration of the retrieved text segments with a personalized question-answering framework and prompt template in order to achieve contextually correct answers. Response Delivery: The response generated is delivered in real-time to the users, giving them proper insights from their documents.

3.4. TECHNOLOGIES USED

1. Streamlit:
Provides the user interface for uploading files, asking questions, and viewing results. It creates an interactive and responsive web application.

2. PyPDF2:
Extracts text from uploaded PDF documents. It works for text-based PDFs, though OCR tools may be needed for image-based files.

3. LangChain Framework:
Manages the integration of text processing, embeddings, and AI models. It simplifies the workflow by connecting different components seamlessly.

4. Google Generative AI (Gemini-Pro):
Used for generating text embeddings and answering questions. Embeddings enable efficient document searches, while the chat model generates accurate answers.

5. FAISS:
A library for storing and retrieving vectorized text chunks, enabling fast and accurate similarity searches for relevant document content.

6. RecursiveCharacterTextSplitter:
Splits large documents into smaller chunks, preserving context for better search and retrieval.

7. Google API Key:
Authenticates access to Google Generative AI services, powering the embedding and question-answering features.

3.5. EQUATIONS AND MATHEMATICAL FRAMEWORK

The model integrates the mathematical foundation of self-attention (used in LLMs) and semantic similarity metrics for retrieval in RAG.

- Dot-Product Attention:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where,
Q (Query): The current token is being considered.
K (Key): The tokens it's being compared to.
V (Value): The information to extract.
d_k: Scaling factor

- Cosine Similarity:

$$\text{Similarity}(a, b) = \frac{a \cdot b}{\|a\|\|b\|}$$

Used for ranking retrieved document in the vector database.

3.6. FORMULATION AND SPECIFICATIONS

- LangChain Framework: Implements modular components for dynamic query routing and task specific execution.
- RAG Integration: Combines real-time retrieval with generative capabilities for grounded response generation.

3.7. SOLVING PROCEDURE

- User queries are processed into embeddings using an embedding model (e.g., Google embedding or Sentence Transformers).
- The query embedding is matched against stored document embeddings in a vector database (e.g., Chroma).
- Relevant chunks are retrieved, augmented, and passed to the LLM for final response generation.

3.8. EXPERIMENTAL METHODOLOGY

The experimental approach combines LangChain for framework design, real-time retrieval via RAG, and fine-tuned LLMs to generate response accuracy. Every component, through iterative evaluation, aims to be as optimal and accurate as possible in performance.

3.9. ALGORITHMS

- Data Retrieval Algorithm: Implements semantic search in a vector database to fetch relevant chunks of information.
- Response Generation Algorithm: Combines RAG outputs with fine-tuned LLM responses for accuracy.
- Chunking Algorithm: Divides large datasets into manageable chunks using recursive or semantic chunking techniques.

Algorithm 1: Algorithm for an LLM application using RAG

Indexing Process

1. embeddings ← load("google_embedding_model")
2. docs ← load("data_sources")
3. chunks ← chunker.split(docs)
4. chunk_embeddings ← embeddings.embed(chunks)
5. database ← index(google_chunk_embeddings)

Query Process

1. INITIALIZE llm_model ← load ("performance_optimized_llm")
2. while TRUE do
3. query ← get.user_input()
4. query_embedding ← embeddings.embed(query)
5. relevant_chunks ← database.search (query_embedding,similarity_search)
6. retrieved_context ← merge(relevant_chunks)
7. final_prompt ← create_prompt(sys_prompt, query, retrieved_context)
8. response ← llm_model.generate(final_prompt)
9. return response to user
10. end while

3.10. EVALUATION METRICS

- Accuracy: Measuring response correctness against ground truth.
- Response Time: Time taken to generate responses.
- User Satisfaction: Feedback from simulated patient interactions.
- Error Rate: Incidence of hallucinated or irrelevant responses.

V. FUTURE SCOPE AND LIMITATIONS

- **Limitations:**

The current version of the tool faces challenges with the accuracy of the extraction of text, especially if this is being extracted from a scanned or image-based PDF that can result in partial or even incorrect content. The users cannot access the system without having a Google API key, and the system has poor performance when dealing with large or complex documents. User privacy-related concerns also arise based on managing uploaded documents as concerns with the security of the information involved. Scalability also is a challenge especially in growing user bases to lead to performance bottlenecks and increase processing times.

- **Future Scope:**

The project can be extended to support multiple formats besides PDFs, including Word and PowerPoint documents, and integrate Optical Character Recognition (OCR) for

image-based PDFs. These improvements would make the tool more versatile and accessible to users who work with a wide variety of document types. Moreover, the addition of advanced search features, such as fuzzy search and Boolean queries, would greatly enhance the accuracy and flexibility of user queries. It would allow users to upload documents directly from their cloud accounts, making it easier. Automatic document summarization would also be added for quick insights from long texts. Finally, user profiles with authentication would enable saving of document histories and tailoring the experience to make it more personalized and efficient.

VI. CONCLUSION

In conclusion, the proposed AI-driven predictive diagnosis chatbot presents a transformative solution to the significant challenges faced by healthcare systems around the world, including overcrowded hospitals, limited access to care, and the necessity for early disease detection. This system, by employing a streamlined three-layer pipeline—retrieval, augmentation, and generation—guarantees efficient and accurate responses to user inquiries, supported by advanced AI methodologies such as semantic chunking, hybrid search strategies, and metadata integration. The use of lightweight and scalable LLMs greatly increases the flexibility and performance of the chatbot, hence making it applicable to multiple healthcare applications, such as telemedicine and mobile settings. This solution has great potential to enhance accessibility, accuracy, and efficiency in healthcare, especially among underserved populations, due to its ability to process sophisticated queries and provide accurate insights that are context-specific. Future advancements could further expand the functionalities of the system, thus potentially making it an indispensable utility in improving healthcare delivery and enterprise applications in varied environments.

REFERENCES

[1] A systematic review and meta-analysis on ChatGPT and its utilization in medical and dental research, Hiroj Bagde, Mohammad Khursheed Alam, December 2023

[2] Chatbots and Large Language Models in Radiology: A Practical Primer for Clinical and Research Applications, Rajesh Bhayana, January 2024

[3] Health-LLM: Personalized Retrieval-Augmented Disease Prediction Model, Mingyu Jin, Qinkai Yu, Chong Zhang, Dong Shu, February 2024

[4] Integrating RAG with LLMs in Nephrology: Advancing Practical Applications, Jing Miao, Charat Thongprayoon, Supawadee Suppadungsuk, Oscar A. Garcia Valencia, Wisit Cheungpasitporn, March 2024

[5] Context based LLM chatbot using RAG, Rahull Borana, March 2024

[6] New Technologies for Spoken Dialogue Systems: LLMs, RAG and the GenAI Stack, Graham Wilcock, March 2024

[7] Leveraging LLM: Implementing an Advanced AI Chatbot for Healthcare, Ajinkya Mhatre, Sandeep R. Warhade, Sayali Kokate, Omkar Pawar, Samyak Jain, May 2024

[8] Use of large language model-based chatbots in managing the rehabilitation concerns and education needs of outpatient stroke survivors and caregivers, Jin Rui Edmund Neo, Joon Sin Ser, San San Tay, May 2024

[9] Building Certified Medical Chatbots: Overcoming Unstructured Data Limitations with Modular RAG, Leonardo Sanna, Patrizio Bellan, Simone Magnolini, Marina Segala, Saba Ghanbari Haez, Monica Consolandi, Mauro Dragoni, May 2024

[10] A Medical Chatbot: Your Healthcare Assistance, Harsh Jain, June 2024

[11] Efficiency-Driven Custom Chatbot Development: Unleashing LangChain, S. Vidivelli, Manikandan Ramachandran, A. Dharunbalaji, August 2024

[12] Efficiency-Driven Custom Chatbot Development: Unleashing LangChain, RAG, and Performance-Optimized LLM Fusion, S. Vidivelli, Manikandan Ramachandran, A. Dharunbalaji, August 2024

AUTHORS

First Author – Tanay Kumar, Final Year, MIT Art, Design and Technology University, tanaykumar0903@gmail.com

Second Author – Ayushi Tiwari, Final Year, MIT Art, Design and Technology University, ayushitiwari.july.07@gmail.com

Third Author – Sonam Bhul, Final Year, MIT Art, Design and Technology University, sonam.bhule9696@gmail.com

Fourth Author – Prof. Dr. Mohit Kumar, Professor, MIT Art, Design and Technology University, mohit.kumar@mituniversity.edu.in