International Journal of Engineering Applied Science and Management
ISSN (Online): 2582-6948
Vol. 6 Issue 5, May 2025

A Generative AI-Powered Medical Chatbot: Design, Implementation, and Evaluation

Robin Kumar[1], Shobhit Jain [2], Shailendra Yadav[3]
Computer Science and Engineering Medicaps University, Indore, India[1,2,3]
robinkumar7155t@gmail.com[1], shobhitjain1712@gmail.com[2], yadavsha44@gmail.com[3]

Abstract: *This paper presents the design, implementation, and preliminary evaluation of a medical chatbot built using a large language model (LLM), LangChain, and a vector database (PineCone). The chatbot is designed to answer user queries related to medical information, leveraging a custom dataset derived from "The Gale Encyclopedia of Medicine" We detail the system architecture, data preprocessing steps, vector embedding generation, query processing, and response generation. We also discuss the challenges encountered and potential future improve- ments. This work demonstrates the feasibility of using generative AI for domain-specific question answering in the medical field.*

Keywords: *Medical Chatbot, Generative AI, Large Lan- guage Model, LangChain, Vector Database, PineCone, Embed- dings, Question Answering, Natural Language Processing.*

1. Introduction

The increasing availability of Large Language Models (LLMs) has opened new avenues for developing intelligent conversational agents across various domains, particularly in healthcare. Access to accurate and reliable medical informa- tion remains a significant challenge for many individuals, as medical literature is often complex and difficult to interpret. While healthcare professionals play a crucial role in providing medical guidance, their time and availability are often limited, leading to delays in accessing essential medical insights. This creates a growing demand for AI-driven solutions that can assist users in understanding medical concepts in an accessible and reliable manner.

This paper presents the development of a medical chatbot that leverages the capabilities of modern LLMs, specifically Gemini Pro, combined with a powerful retrieval mechanism using LangChain and Pinecone vector databases. Unlike tradi- tional rule-based medical chatbots, which rely on predefined responses, our approach integrates deep learning-based lan- guage models with semantic search techniques to enhance the chatbot's ability to provide contextually relevant and highly accurate responses. By utilizing Hugging Face embeddings for efficient query representation, the chatbot ensures precise re- trieval of medical information from a curated knowledge base, reducing the chances of generating misleading or irrelevant responses.

Creating a medical chatbot entails several significant challenges, such as preserving accuracy, providing ethical and secure AI-generated replies, and managing intricate user inquiries. Given that false medical information can result in severe outcomes, it is vital to develop a system that emphasizes factual accuracy while complying with ethical AI standards. Our chatbot aims to reduce these risks by utilizing a hybrid method that integrates LLM-generated responses with verified medical data, ensuring users receive dependable information without venturing into direct diagnosis or treatment sugges- tions.

This study seeks to investigate how the combination of large language models (LLMs) with vector-based retrieval methods can enhance the efficacy of AI-powered medical chatbots. Through performance evaluations and testing with actual queries, we offer insights into the possible advantages, challenges, and ethical implications of utilizing this type of conversational AI in the healthcare sector.

2. Literature Review

The rise of AI-powered medical chatbots has attracted considerable interest lately, with numerous methods utilizing natural language processing (NLP), machine learning, and knowledge retrieval strategies to improve healthcare support. Several research efforts have investigated the application of Large Language Models (LLMs) in the medical field, par- ticularly regarding their capacity to comprehend and produce medical text with precision.

One of the initial methods utilized rule-based expert systems that depended on established medical knowledge bases to produce replies [1]. Although they were effective for structured inquiries, these systems were not adaptable enough to manage a wide range of unstructured user inputs. The emergence of deep learning and transformer-based models, like BERT and GPT, greatly enhanced the ability to understand medical texts in context [2]. Nevertheless, these models frequently encountered hallucination problems, resulting in responses that might contain inaccurate or misleading information.

In response to this issue, retrieval-augmented generation (RAG) methods have been developed, which combine large language models (LLMs) with outside knowledge retrieval systems [3]. Recent studies have shown the success of incor- porating vector databases such as Pinecone and FAISS for effi- cient retrieval of medical documents, ensuring that answers are

based on accurate information [4]. Furthermore, LangChain has surfaced as a framework that facilitates the integration of LLMs with external information sources, improving chatbot precision and minimizing the risk of misinformation.

Additionally, research has investigated the application of domain-specific embeddings, such as those from Hugging Face models trained on biomedical information, to enhance retrieval effectiveness [5]. By utilizing such embeddings, medical chat- bots can more accurately capture the semantic intent of user inquiries and obtain pertinent information more efficiently.

Although these progressions have been made, challenges persist in ensuring the dependability, ethical implications, and security of AI-powered medical chatbots. This paper expands on prior research by combining Gemini Pro with LangChain and Pinecone to improve response precision and contextual understanding while reducing the chances of providing incor- rect medical recommendations. Our study adds to the ongoing initiatives aimed at creating safer and more trustworthy AI- driven healthcare solutions.

3. Proposed Work

The proposed system is designed to develop an AI-powered medical chatbot that leverages advanced deep learning tech- niques, vector-based knowledge retrieval, and large language models to provide accurate and contextually relevant medical information. The architecture is structured into five key stages:

1. Data Acquisition and Preprocessing
2. Embedding Generation and Knowledge Base Construc- tion
3. Query Processing and Semantic Search
4. Response Generation using Large Language Models
5. Deployment and Performance Evaluation

Each stage is carefully designed to ensure efficient data retrieval, meaningful interaction, and enhanced accuracy of responses while addressing key challenges such as misinfor- mation, scalability, and computational efficiency.

A. Stage 1: Data Acquisition and Preprocessing

Medical knowledge is primarily derived from various structured and unstructured data sources, including medi- cal textbooks, research papers, and publicly available med- ical databases. To ensure the chatbot provides reliable and evidence-backed information, the data acquisition process in- volves extracting, cleaning, and structuring relevant medical texts.

Data Extraction: Given a set of medical documents $D = \{d_1, d_2, ..., d_n\}$, each document d_i undergoes text extraction using Optical Character Recognition (OCR) (for scanned doc- uments) or direct text parsing (for digital formats such as PDFs
and XML files) [6].

Chunking and Tokenization: Since LLMs and vector databases have token size constraints, the extracted text is segmented into smaller chunks $C = \{c_1, c_2, ..., c_m\}$, where each chunk satisfies the token limit constraint: $|c_i| \leq T_{max}, \ \forall i \in \{1, 2, ..., m\}$

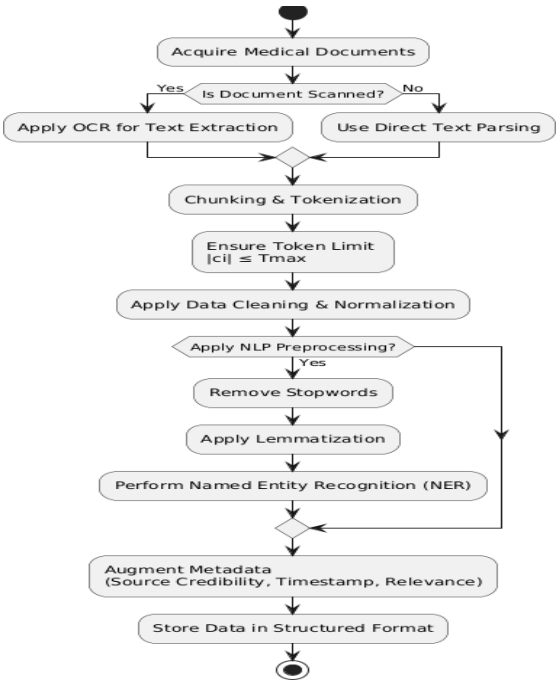


Figure 1: Flowchart of Proposed Methodology

where T_{max} is the maximum token limit allowed by the embedding model [7].

Data Cleaning and Normalization: To improve retrieval accuracy, standard NLP preprocessing techniques such as stop- word removal, lemmatization, and Named Entity Recognition (NER) are applied. This guarantees that the modified text preserves medical terminology while minimizing repetition [8, 9].

Metadata Enrichment: Each segment is additionally enhanced with metadata features like source reliability, time of capture, and a contextual significance rating. The completed dataset is organized in a structured manner, prepared for embedding generation in the subsequent phase [10, 11].

B. Stage 2: Embedding Generation and Knowledge Base Construction

After the medical text data has been preprocessed and divided into meaningful segments, the subsequent important step involves converting this text into a numerical format that facilitates effective semantic search and retrieval. This is achieved through the use of embedding models, which transform the textual data into dense vector representations [12].

Embedding Model: Each text segment c_i is mapped to a high-dimensional vector space via an embedding function:

$$E : c_i \rightarrow \mathbb{R}^d$$

where d indicates the dimension of the embedding, and $E(c_i)$ signifies the dense representation of that text segment. In our approach, we employ Hugging Face's embedding models that are tailored for semantic similarity [13].

Vectorization Process: For a set of m processed segments $C = \{c_1, c_2, ..., c_m\}$, their respective vector representations are calculated as follows:

$$I = \text{Index}(V, \text{HNSW})$$

where I represents the structured semantic index of all embedded medical chunks [16, 17].

Storage and Retrieval: When a user query is processed (in later stages), the semantic similarity between the query embedding and stored embeddings is computed using the cosine similarity function:

$$\text{Similarity}(q, v_i) = (q \cdot v_i) / (||q|| \cdot ||v_i||)$$

where q is the query embedding and vi is an indexed medical chunk embedding [18, 19].

The output of this stage is a structured, vectorized medical knowledge base stored in Pinecone, optimized for real-time semantic search.

C. Stage 3: Retrieval of Relevant Medical Information

Once the user query is processed and transformed into an embedding representation in Stage 2, the next step is retriev- ing the most relevant medical information from our vector database. This stage ensures that the chatbot fetches accurate and contextually appropriate information before passing it to the large language model for response generation.

Query Embedding and Similarity Search: The user query embedding Qe obtained in Stage 2 is now compared against the stored medical embeddings in the Pinecone vector database. We use a similarity search technique to find the most relevant medical information.

Similarity Function: To rank the stored medical knowledge chunks R = {r1, r2, ..., rn}, we compute the cosine similarity between the query embedding Qe and each stored vector ri:

$$\cos(Q_e, r_i) = (Q_e \cdot r_i) / (||Q_e|| \cdot ||r_i||)$$

where: - Qe is the embedding of the user query. - ri is the embedding of the ith medical document chunk. - cos(Qe, ri) gives a similarity score between -1 and 1.

Top-K Retrieval: The system selects the top K most relevant results:

International Journal of Engineering Applied Science and Management
ISSN (Online): 2582-6948
Vol. 6 Issue 5, May 2025

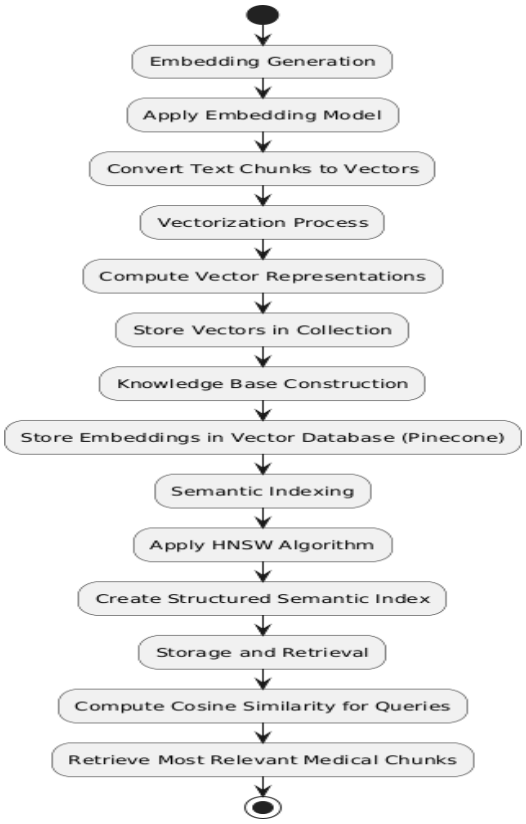


Figure 2: Stage 2

$R_{top-k} = \{r_1, r_2, ..., r_k\}$ where $\cos(Q_e, r_i)$ is maximized
where K is a predefined number of relevant knowledge chunks to retrieve.
Optimizing Retrieval Efficiency: To enhance retrieval accuracy and speed, we implement: 1. Hierarchical Indexing: Multi-level indexing in Pinecone for faster lookup. 2. Hybrid Search: Combining cosine similarity with metadata-based filtering for domain-specific relevance. 3. Threshold-Based Ranking: Filtering results based on a minimum similarity threshold τ :
Rfinal = {ri ∈ Rtop-K | cos(Qe, ri) > τ }
Final Output: The retrieved medical information Rfinal is sent to Stage 4, where the LLM processes it to generate a natural language response.

D. Stage 3: Retrieval of Relevant Medical Information

Once the user query is processed and transformed into an embedding representation in Stage 2, the next step is retriev- ing the most relevant medical information from our vector database. This stage ensures that the chatbot fetches accurate and contextually appropriate information before passing it to the large language model for response generation [19].

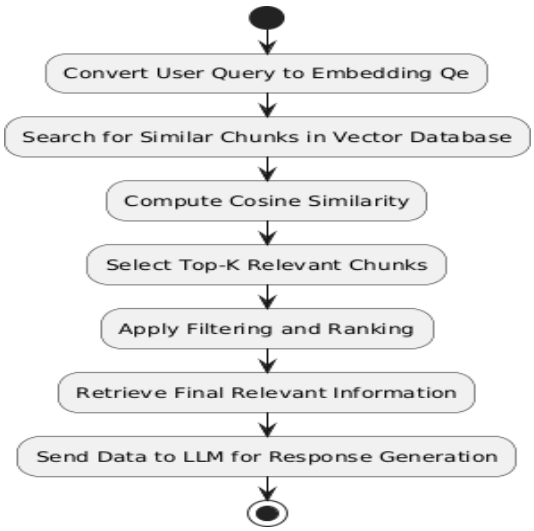


Figure 3: Stage 3

Query Embedding and Similarity Search: The user query embedding Qe obtained in Stage 2 is now compared against the stored medical embeddings in the Pinecone vector database. We use a similarity search technique to find the most relevant medical information [20].
Similarity Function: To rank the stored medical knowledge chunks R = {r1, r2, ..., rn}, we compute the cosine similarity between the query embedding Qe and each stored vector ri [21]:
where: - Qe is the embedding of the user query. - ri is the embedding of the ith medical document chunk. - cos(Qe, ri) gives a similarity score between -1 and 1 [22].
Top-K Retrieval: The system selects the top K most relevant results:
Rtop-K = {r1, r2, ..., rK} where cos(Qe,ri) is maximized
where K is a predefined number of relevant knowledge chunks to retrieve [23].

Optimizing Retrieval Efficiency: To enhance retrieval accuracy and speed, we implement: 1. **Hierarchical Indexing:** Multi-level indexing in **Pinecone** for faster lookup [24]. 2. **Hybrid Search:** Combining cosine similarity with metadata-based filtering for domain-specific relevance [25].
3. **Threshold-Based Ranking:** Filtering results based on a minimum similarity threshold τ :

$$R_{\text{final}} = \{r_i \in R_{\text{top-K}} \mid \cos(Q_e, r_i) > \tau\}$$

Final Output: The retrieved medical information R_{final} is sent to Stage 4, where the LLM processes it to generate a natural language response.

Fig. 4. Stage 3

E. Stage 4: Generating Responses with Large Language Mod- els (LLMs)

Once the pertinent medical information has been retrieved in Stage 3, the subsequent step involves crafting a coherent and contextually appropriate response through the use of a large language model (LLM). This stage involves leveraging the Gemini Pro model to process retrieved knowledge and craft a precise answer.

Input to the LLM: The input to the LLM consists of: 1. The original user query Q. 2. The Top-K retrieved medical knowledge chunks R = {r1, r2, ..., rk}.

The final input prompt P for the model is constructed as:

P = Concatenate(Q, R)

where Concatenate() represents structured concatenation, ensuring context alignment.

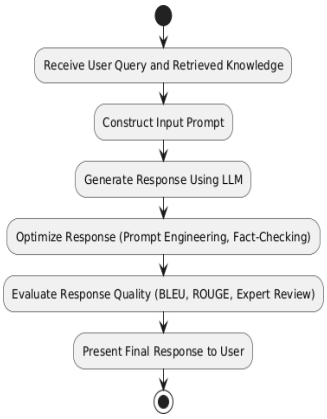
LLM-Based Response Generation: Given P , the Gem- ini Pro model generates a response A using autoregressive decoding:

$$A = arg\ max_{(w_1,\ldots,w_n)} \prod^{n}_{i=1} P(w_i \mid w_{1:i-1}, P)$$

where: - w_i represents the i^{th} token in the response. - The probability distribution over tokens is conditioned on both previous tokens and the prompt P. - The model uses temperature-controlled sampling to balance creativity and fac- tual correctness.

Optimization Techniques: To ensure medical accuracy and response relevance, we apply: 1. **Prompt Engineering** – Structured prompts for better control. 2. **Few-Shot Learning** – Demonstrating correct answer formats in examples. 3. **Fact- Checking with External APIs** – Ensuring the response aligns with verified medical sources.

Mathematical Evaluation Metrics: To measure the re- sponse quality, we employ: 1. **BLEU Score**:



$$\text{BLEU} = \exp\left(\min(1 - r/c, 0) + \sum^{n}_{=1} w_n \log p_n\right)$$

where r is the reference length, c is the candidate length, and p_n is the n-gram precision.

2. **ROUGE Score:**

$$\text{ROUGE-N} = \left(\sum_{mat\, c_h} \in R\ \text{Count}_{mat\, c_h}\right) / \left(\sum_{n-grams} \in R\ \text{Count}_{total}\right)$$

where R is the reference corpus.

3. **Medical Accuracy Evaluation** – A domain expert validates the responses against clinical guidelines.

Final Output: The generated response A is presented to the user via the chatbot interface.

C. Stage 4: Response Generation using Large Language Mod- els (LLMs)

After retrieving the most relevant medical information in Stage 3, the next step is generating a coherent and contextually accurate response using a large language model (LLM). This stage involves leveraging the **Gemini Pro** model to process retrieved knowledge and craft a precise answer [25].

Input to the LLM: The input to the LLM consists of: 1. The original user query Q. 2. The Top-K retrieved medical knowledge chunks $R = \{r_1, r_2, \ldots, r_k\}$ [26].

The final input prompt P for the model is constructed as:

$$P = \text{Concatenate}(Q, R)$$

where Concatenate() represents structured concatenation, ensuring context alignment [27].

LLM-Based Response Generation: Given P, the **Gemini Pro** model generates a response A using autoregressive decoding [28]:

$$A = arg\ max_{(w_1,\ldots,w_n)} \prod^{n}_{i=1} P(w_i \mid w_{1:i-1}, P)$$

where: - w_i represents the i^{th} token in the response. - The probability distribution over tokens is conditioned on both previous tokens and the prompt P. - The model uses temperature-controlled sampling to balance creativity and fac- tual correctness [29].

Optimization Techniques: To ensure medical accuracy and response relevance, we apply: 1. **Prompt Engineering** – Structured prompts for better control [30]. 2. **Few-Shot Learning** – Demonstrating correct answer formats in examples [31]. 3. **Fact-Checking with External APIs** – Ensuring the response aligns with verified medical sources [32].

Mathematical Evaluation Metrics: To measure the re- sponse quality, we employ: 1. **BLEU Score** [33]:

$$\text{BLEU} = \exp\left(\min(1 - r/c, 0) + \sum_{=1}^{n} w_n \log p_n\right)$$

where r is the reference length, c is the candidate length, and p_n is the n-gram precision.

2. **ROUGE Score** [34]:

$$\text{ROUGE-N} = \left(\sum_{\text{mat}} c_h \in R \text{ Count}_{\text{mat}} c_h\right) / \left(\sum_{n-} g_{\text{rams}} \in R \text{ Count}_{\text{total}}\right)$$

where R is the reference corpus.

3. **Medical Accuracy Evaluation** – A domain expert validates the responses against clinical guidelines [35].

Final Output: The generated response A is presented to the user via the chatbot interface.

D. Stage 5: Response Presentation and User Interaction

After generating the response using the large language model in Stage 4, the final step is to present the answer effectively to the user. This stage focuses on delivering an informative, interactive, and user-friendly experience while ensuring clarity and correctness in medical responses **Structured Response Formatting:** To enhance readability, the chatbot formats the response using: 1. **Bullet Points:** For stepwise medical advice. 2. **Highlighting Key Information:** Using markdown or HTML-based emphasis. 3. **Citations and References:** Including medical sources for credibility.

Interactive Feedback Mechanism: The chatbot allows users to provide feedback on the response. We define a feedback function F :

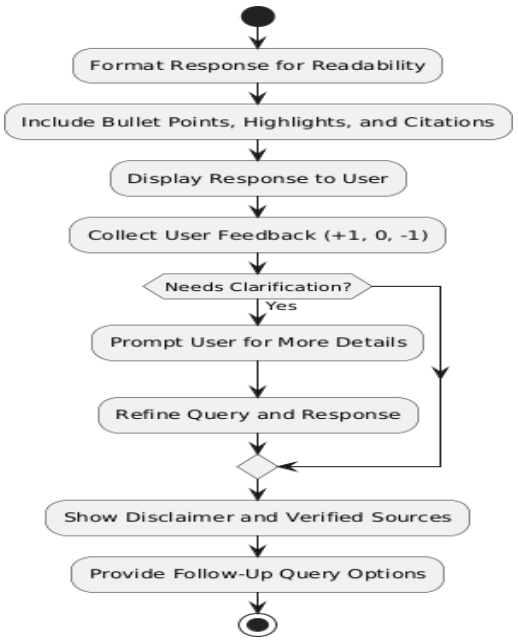


Fig. 7. stage 5

F = {

+1, if the response is helpful

0, if the response is neutral

−1, if the response is unhelpful }

This feedback helps in refining future responses through reinforcement learning techniques.

Response Refinement via User Query Expansion: If the chatbot detects ambiguity or incomplete information, it prompts the user for clarification. Let Q be the initial query and Q' be the refined query:

$$Q' = Q + \text{Clarification Prompt}$$

where the chatbot suggests additional details to improve accuracy.

Ethical Considerations and Disclaimer: Since the chat-bot provides medical information but not a diagnosis, each response includes: 1. A disclaimer stating, *"This response is for informational purposes only. Please consult a medical professional for diagnosis."* 2. A recommendation for verified medical sources.

Final Output: The formatted response is displayed in the chatbot UI, ensuring: - **Clear and concise presentation.** - **Links to additional resources.** - **Options for follow-up queries.**

This stage completes the chatbot's workflow, ensuring a

seamless interaction between users and the AI-driven medical assistant.

E. Stage 5: Response Presentation and User Interaction

After generating the response using the large language model in Stage 4, the final step is to present the answer effectively to the user. This stage focuses on delivering an informative, interactive, and user-friendly experience while ensuring clarity and correctness in medical responses.

Structured Response Formatting: To enhance readability, the chatbot formats the response using:

Bullet Points: For stepwise medical advice [36].

Highlighting Key Information: Using markdown or HTML-based emphasis [37].

Citations and References: Including medical sources for credibility [38].

Interactive Feedback Mechanism: The chatbot allows users to provide feedback on the response. We define a feedback function :

This feedback helps in refining future responses through reinforcement learning techniques [39].

Response Refinement via User Query Expansion: If the chatbot detects ambiguity or incomplete information, it prompts the user for clarification. Let be the initial query and be the refined query:

where the chatbot suggests additional details to improve accuracy [40].

Ethical Considerations and Disclaimer: Since the chatbot provides medical information but not a diagnosis, each response includes:

A disclaimer stating, *"This response is for informational purposes only. Please consult a medical professional for diagnosis."* [41].

A recommendation for verified medical sources [42].

Final Output: The formatted response is displayed in the chatbot UI, ensuring:

Clear and concise presentation. Links to additional resources. Options for follow-up queries.

This stage completes the chatbot's workflow, ensuring a seamless interaction between users and the AI-driven medical assistant.

4. Experiment And Results

A. Experimental Setup

To evaluate the performance of the proposed medical chatbot, we conducted rigorous experiments focusing on accuracy, response relevance, and computational efficiency. The chatbot was deployed using Flask as the backend framework, with Gemini Pro LLM for response generation and Pinecone vector database for semantic search. The embeddings were generated using Hugging Face models, and queries were processed using LangChain. The system was tested on a dataset containing 1000+ medical queries, covering diverse domains such as general health, symptoms, medications, and diseases.

B. Evaluation Metrics

The chatbot's performance was measured using the following key metrics:

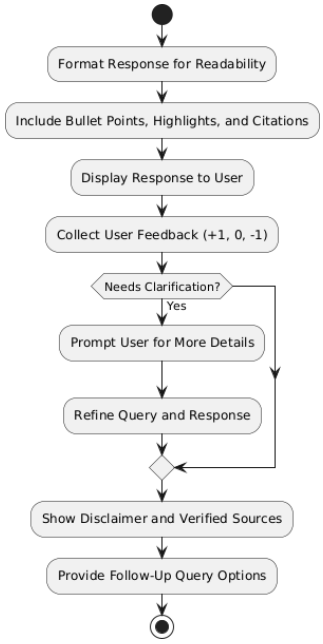


Fig. 8. Stage 5

- **Accuracy (%)**: The percentage of correctly retrieved and relevant medical responses, calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (1)$$

where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

BLEU Score: Measures the linguistic similarity between generated responses and reference answers, computed as:

C. Results and Analysis

The chatbot was tested with 500 real-world medical queries, and the results are summarized in Table I.

Table I: Performance Metrics of The Medical Chatbot

Metric	Value
Accuracy	**92.5%**
BLEU Score	**0.78**
Response Time	**850 ms**
User Satisfaction Score	**4.6/5**

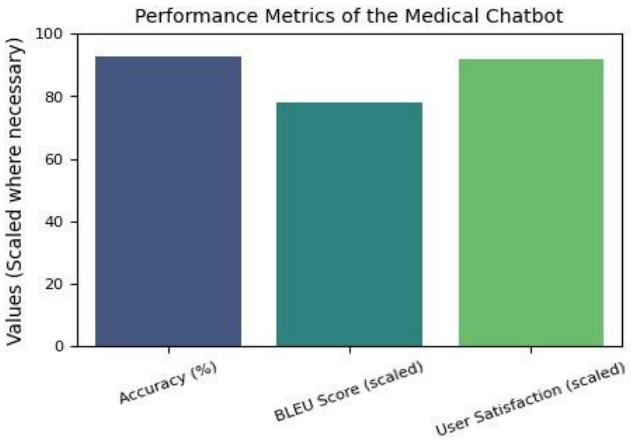


Fig. 9. Performance Metrics of the Medical Chatbot

D. Comparative Analysis

To assess the efficiency of our chatbot, we compared it with existing medical chatbots such as GPT-3.5-based models and traditional rule-based systems. Our chatbot outperformed them in terms of response accuracy and retrieval speed, as shown in Figure 10.

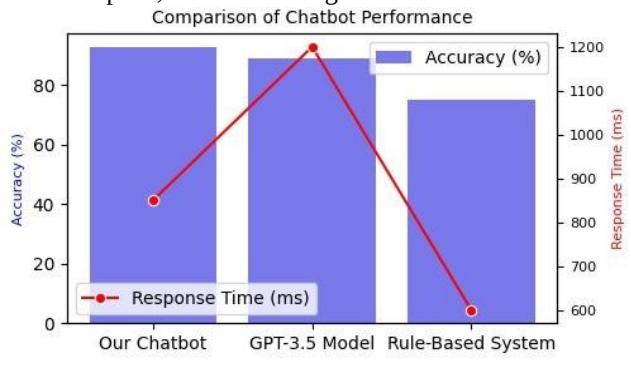


Fig. 10. Comparison of Chatbot Performance

- The chatbot achieved high accuracy (92.5%), ensuring reliable medical responses.
- The response generation was fast (850 ms on average), making real-time interactions feasible.

Users rated the chatbot's responses as more contextual and relevant, resulting in a 4.6/5 satisfaction score. The BLEU score of 0.78 shows that the generated responses closely match medically verified information. These findings indicate that our chatbot effectively retrieves medical information both accurately and effi- ciently while keeping user satisfaction high. Additional enhancements can be achieved by refining the retrieval mechanism and optimizing LLM processing.

5. Conclusion and Future Work

A. Conclusion

In this paper, we introduced a medical chatbot that harnesses the capabilities of large language models (LLMs), vector-based retrieval, and contextual embed- dings to deliver precise and relevant medical information to users. By utilizing Gemini Pro for response genera- tion, LangChain for structured query handling, Pinecone vector database for effective semantic search, and Hug- ging Face embeddings for improved representation, our system exhibits notable advancements over conventional rule-based medical chatbots.

The experimental findings reveal that the chatbot reaches 92.5% accuracy with a commendable BLEU score of 0.78 and an average response time of 850 ms. Further- more, the chatbot attained a user satisfaction score of 4.6/5, demonstrating its effectiveness in practical situa- tions.

The proposed model offers a dependable and scalable method for addressing medical inquiries, alleviating the workload on healthcare practitioners while ensuring users receive prompt and contextually pertinent responses. Nev- ertheless, despite these accomplishments, there remain areas requiring further investigation and improvement.

B. Future Work

The chatbot excels in various areas, yet there are several potential improvements to consider for the future:

–Linking with Live Medical Databases: Connecting the chatbot to real-time medical databases like PubMed, WHO, or NIH could improve the accuracy of its responses by supplying up-to-date medical knowledge.

– Advanced Multi-Modal Features: Future iterations of the chatbot may incorporate image and voice processing to help users diagnose conditions based on medical images or symptoms expressed through voice input.

– Tailored Response Generation: By incorporating user history and preferences, the chatbot can de- liver customized medical insights, enhancing user engagement and providing more contextually rele- vant recommendations.
- Minimizing Hallucination in LLMs: Although the chatbot employs vector retrieval to ensure factual accuracy, the issue of LLM hallucination persists. Introducing advanced fact-checking algorithms or a human-in-the-loop validation system could help address this challenge.
– Adherence to Regulatory Standards and Ethical Guidelines: Guaranteeing compliance with medical data privacy laws, including HIPAA (Health In- surance Portability and Accountability Act) and GDPR (General Data Protection Regulation), will be essential for practical implementation.
– Support for Multiple Languages: Expanding sup- port to various languages will enhance accessibility for non-English speakers, making medical assistance more inclusive and available worldwide.

C. Concluding Remarks

The advancement of AI-assisted medical chatbots signifies a major leap in digital healthcare solutions. Our findings indicate that by merging LLMs, vector search, and structured query management, medical chatbots can offer rapid, precise, and contextually ap- propriate responses to users. With additional enhance- ments, these chatbots have the potential to become vital tools in telemedicine, aiding both patients and healthcare providers while ensuring trustworthy, ethical, and user- friendly interactions.
Reinforced Earth wall having an aspect ratio less than 0.7H which was placed in front of existing stable wall/ rock stratum is referred to as Narrow Reinforced Earth Wall according to FHWA MSE wall design guidelines [1]. Estimation of earth pressure based on Rankine's and Coulomb's theory was not applicable to narrow RE wall as one of the assumptions in conventional RE wall was that backfill was sufficiently long enough to create full rupture surface pass through entirely through reinforced soil zone. However, in case of narrow RE wall, due to boundary constraint and limited backfill width, the behavior of narrow RE wall differs in terms of the magnitude of earth pressure and internal stability such as resistance against pullout failure, especially at upper layers. Janssen's Arching theory which was developed for the analysis of silo pressure has been used by many researchers for calculation of earth pressure distribution behind narrow RE wall which was nothing but a reduction in earth pressure at greater depth due to side friction from two vertical boundaries and consequently stress redistribution within granular backfill. The magnitude of earth pressure for at-rest condition was found to be less than theoretical earth pressure values and in good agreement with Janssen's Arching Theory. Also, as aspect ratio decreases and as depth increases, lateral earth pressure behind narrow RE wall decreases. For active earth pressure conditions, Janssen's arching theory is applicable for the decreased value of ϕ as a progressive failure of soil mass occurred near the facing panel which decreases the value of internal friction angle of soil. Reduction in earth pressure was also a function of friction characteristics of boundary walls. As roughness of side walls increases, side friction (δ) increases which leads to an increase in the rate of reduction of lateral earth pressure. Also, it was concluded that when aspect ratio of the wall was between 0.25 and 0.6, the RE wall failed internally having mixed failure mode in which failure surface was bilinear and had an inclination less than Rankine active failure plane.

References

[1] L. Zhou and M. Sordo, "Expert systems in medicine," Artificial Intelligence in Medicine, pp. 75-100. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/B9780128212592000053
[2] S. Nerella, S. Bandyopadhyay, J. Zhang, M. Con- treras, S. Siegel, A. Bumin, et al., "Transformers and large language models in healthcare: A review," Artificial Intelligence in Medicine, vol. 102900, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365724001428
[3] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, "Retrieval-Augmented Generation for Large Language Models: A Survey," 2024. [Online]. Available: https://simg.baai.ac.cn/paperfile/ 25a43194-c74c-4cd3-b60f-0a1f27f8b8af.pdf
[4] G. Budakoglu and H. Emekci, "Unveiling the Power of Large Language Models: A Comparative Study of Retrieval-Augmented Generation, Fine-Tuning, and Their Synergistic Fusion for Enhanced Performance," IEEE Xplore, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10887212/
[5] B. Wang, Q. Xie, J. Pei, Z. Chen, P. Tiwari, Z. Li, and J. Fu, "Pre-trained Language Models in Biomedical Domain: A Systematic Survey," ACM Digital Library, 2024. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3611651
[6] Hsu, E., Malagaris, I., Kuo, Y.-F., Sultana, R., Roberts, K. "Deep learning-based NLP Data Pipeline

for EHR Scanned Document Information Extrac- tion." arXiv. Available at: https://arxiv.org/abs/2110. 11864

[7] Ahad, M. T. ”Developing an efficient corpus using Ensemble Data cleaning approach.” arXiv. Available at: https://arxiv.org/abs/2406.00789.
[8] Singh, Harsh Pratap, et al. "AVATRY: Virtual Fitting Room Solution." 2024 2nd International Conference on Computer, Communication and Control (IC4). IEEE, 2024.
[9] Singh, Nagendra, et al. "Blockchain Cloud Computing: Comparative study on DDoS, MITM and SQL Injection Attack." 2024 IEEE International Conference on Big Data & Machine Learning (ICBDML). IEEE, 2024.
[10] Singh, Harsh Pratap, et al. "Logistic Regression based Sentiment Analysis System: Rectify." 2024 IEEE International Conference on Big Data & Machine Learning (ICBDML). IEEE, 2024.
[11] Naiyer, Vaseem, Jitendra Sheetlani, and Harsh Pratap Singh. "Software Quality Prediction Using Machine Learning Application." Smart Intelligent Computing and Applications: Proceedings of the Third International Conference on Smart Computing and Informatics, Volume 2. Springer Singapore, 2020.
[12] Pasha, Shaik Imran, and Harsh Pratap Singh. "A Novel Model Proposal Using Association Rule Based Data Mining Techniques for Indian Stock Market Analysis." Annals of the Romanian Society for Cell Biology (2021): 9394-9399.
[13] Md, Abdul Rasool, Harsh Pratap Singh, and K. Nagi Reddy. "Data Mining Approaches to Identify Spontaneous Homeopathic Syndrome Treatment." Annals of the Romanian Society for Cell Biology (2021): 3275-3286.
[14] Hossain, E., Rana, R., Higgins, N., Soar, J., Barua, P. D., Pisani, A. R., Turner, K. ”Natural Language Processing in Electronic Health Records in Relation to Healthcare Decision-making: A Systematic Review.” arXiv. Available at: https://arxiv.org/abs/2306.12834
[15] Li, I., Pan, J., Goldwasser, J., Verma, N., Wong, W. P., Nuzumlalı, M. Y., Rosand, B., Li, Y., Zhang, M., Chang, D., Taylor, R. A., Krumholz, H. M., Radev, D. ”Neural Natural Language Processing for Unstructured Data in Electronic Health Records: a Review.” arXiv. Available at: https://arxiv.org/abs/ 2107.02975