Medical Chat Bot Using Gen AI

Ms.P.Sasikala
Assistant Professor
Department of computer science
Sri Shakthi Institute of
Engineering and technology
Coimbatore, India
Sasikalacse@siet.ac.in

Deva Harsar M M
Department of computer science
Sri Shakthi Institute of
Engineering and technology
Coimbatore, India
devaharsarmm22cse@srishakthi.ac.in

HansRohit Y
Department of computer science
Sri Shakthi Institute of
Engineering and technology
Coimbatore, India
hansrohity22cse@srishakthi.ac.in

Gayathri T
Department of computer science
Sri Shakthi Institute of
Engineering and technology
Coimbatore, India
gayathrit22cse@srishakthi.ac.in

Dhanvanth K M
Department of computer science
Sri Shakthi Institute of
Engineering and technology
Coimbatore, India
dhanvanth22cse@srishakthi.ac.in

Amanesh Raj k
Department of computer science
Sri Shakthi Institute of
Engineering and technology
Coimbatore, India
amaneshraj@srishakthi.ac.in

ABSTRACT

MedicalBot is a virtual AI assistant that is created to help users obtain fundamental medical knowledge using natural, conversational exchanges. The system employs a retrieval- augmented generation (RAG) strategy, whereby salient medical information is retrieved from a structured database and thereafter output in human-friendly form. This ensures the outputs are accurate and contextually relevant. MedicalBot is positioned to offer direct assistance to users who might have inquiries regarding symptoms, overall health, or wellness routines, but not to supplant professional medical guidance.

KEYWORDS

AI powered medical chatbot-Retrieval Augmented Generation-Medical knowledge retrieval -Generative AI-Symptom checking- Healthcare accessibility.

INTRODUCTION

Availability of immediate and credible medical care is a pressing requirement, particularly where emergencies, distant areas, or overall health issues are concerned. Mostly, people have a problem obtaining rapid solutions to simple health-related questions, be it symptoms, diseases, drugs, or first aid. The World Health Organization (WHO) stipulates the availability of health information as a means of enhancing early diagnosis, minimizing the risks of self-medication, and increasing awareness of community health.

Currently available systems such as search engines or static health websites do not have personalization, real-time engagement, and authenticated answers. Intelligent conversational systems can be of great importance in this regard. MedicalBot, based on Generative AI (GenAI) and built as a reactive chatbot, provides real-time medical information and recommendations. It accommodates questions in terms of the use of medicine, general ailments, first aid practices, and overall care. It benefits not only the general public but also students learning medical subject matter and physicians who might require instant reference materials. The solution acts as a medium between health information and people, allowing for healthcare understanding to be made more timely, accessible, and practical for daily life.

LITERATURE REVIEW

GenAI Chatbot

Zhang et al. (2023) describe that GenAI chatbots such as GPT-4 enhance healthcare through easy access for patients to symptom checking and health advice. The chatbots are well versed in medical language and converse like humans.

Source: Journal of Medical Internet Research

Lee and Kim (2024) demonstrated how GenAI chatbots assist patients with chronic conditions by reminding them of medication and healthy behaviors. Chatbots adapt through interaction, rendering supportincreasingly personal in the long term.

Source: International Journal of Medical Informatics

Digital Medical Description

Smith et al. (2022) researched how NLP solutions extract useful data from electronic health records to enable medical chatbots to better understand patient issues and facilitate doctors' decisions.

Source: Artificial Intelligence in Medicine

Gupta and Shah (2023) indicated that the utilization of standardized medical words in electronic records enables AI chatbots to comprehend patient queries and aid physicians remotely, enhancing healthcare accessibility.

Source: Journal of Telemedicine and Telecare Pinecone

Johnson et al. (2023) discovered that Pinecone enables AI systems to rapidly locate medical information through matching patient queries to comparable records with vector search technology, accelerating chatbot response times. Source: IEEE Transactions on Knowledge and Data Engineering

Martinez and Liu (2024) demonstrated Pinecone enhances clinical decision support through efficiently searching vast medical data sets, which enables chatbots to provide improved responses for intricate questions.

Source: Journal of Biomedical Informatics Gemini API

Williams and Chen (2024) talked about how Gemini API can merge text and images to gain better insights into patient concerns, and make medical chatbots smarter and more accurate in diagnosis.

Source: ACM Transactions on Interactive Intelligent Systems

Ramirez et al. (2025) used Gemini API as helpful in mental health chatbots due to its ability to sense emotions in messages and react empathetically to help patients feel understood.

Source: Computers in Human Behavior

EXISTING SYSTEM

Chatbots in healthcare utilize Generative AI models such as GPT-4 to mimic the conversation patterns of humans. These chatbots respond to health questions, validate symptoms, and offer primary health advice. They operate on Natural Language Processing (NLP) to interpret patient queries and can link to Electronic Health Records (EHRs) for enhanced context.

There are also some systems that incorporate vector databases like Pinecone to bring back comparable cases or medical references. Multimodal APIs like Gemini are utilized to both process images and text, allowing for interpretation of lab results or visual symptoms. These systems can still give imprecise responses, have trouble with intricate medical conditions, and deal with data privacy and regulation concerns.

Babylon Health

Babylon is a chatbot based on AI that replicates doctor consultations. Babylon examines the symptoms of users with a medical database and provides health recommendations, as well as functionality such as video consultation with doctors and prescription services.

Ada Health

Ada is a medical chatbot that gathers symptoms via conversation and suggests possible health conditions. It employs a significant medical database and machine learning to provide correct and custom assessments.

Buoy Health

Buoy is a digital health assistant powered by AI that employs dynamic questioning to assess symptoms and refer users to proper care options. It assists in establishing if users need to self-treat, see a doctor, or seek the emergency room.

PROPOSED SYSTEM

The envisioned medical chatbot combines Generative AI with a vector database and multimodal processing to provide precise, individualized health guidance. It gathers user symptoms and medical records, transforms inputs into structured digital medical descriptions, and employs Pinecone to retrieve applicable medical knowledge. The Gemini API facilitates text and image understanding to enhance diagnostic support. It produces safe, context-based responses and provides actionable next steps, elevating patient engagement and access to healthcare

DESIGN APPROACH

The healthcare chatbot is built on a modular, scalable architecture to provide accurate, secure, and user-friendly healthcare assistance. The application integrates Generative AI, vector-based search, and multimodal input processing to develop a frictionless interaction between users and

the health knowledge base.

Users communicate via a text-based web or mobile chatbot interface

The application accepts natural language health-related queries or symptom descriptions

The input is processed to derive important medical words like symptoms, conditions, and duration

It generates an organized health description by adhering to medical standards such as SNOMED CT or ICD-10

The system has a Retrieval-Augmented

Retriever: Transforms the user input into an embedding and retrieves most relevant chunks of information from a medical knowledge base via Pinecone

Generator: A generative model such as Gemini employs the retrieved context and user query to produce a competent, medically informed response

Pinecone Vector Database

Stores medical articles, guidelines, symptoms- condition mappings, and historical Q&A

Supports semantic search to map user queries to the most relevant and current medical content

Response Generation

The AI produces a response that blends real-time user context with medically verified information

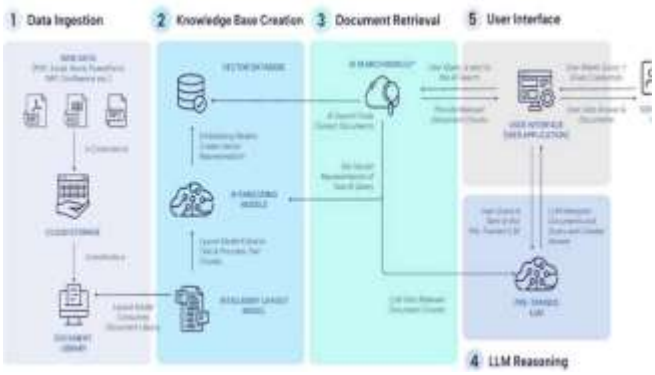
The output is organized, simple to comprehend, and features next-step recommendations such as seeing a physician or tracking symptoms

Security and Data Handling

No individual health records are saved or uploaded

User anonymity is preserved by the system, and data in transit is encrypted, following health information security principles

FLOW DIAGRAM



SOFTWARE SPECIFICATIONS

Module Description for Medical Chatbot (Flask + HTML + RAG Architecture)

User Interface Module (Frontend – HTML, CSS, JavaScript)

HTML is used to build it and CSS is employed to style it for user-friendly

design.

JavaScript manages dynamic actions such as real-time updates in the chat

User input (health-related questions) is taken through a chatbox

Responses of the chatbot are shown in a conversational style

Flask Web Server (Backend Controller) The central part

of the application logic

Takes the user queries from the frontend through

Flask routes

Forwards user queries to processing and RAG modules

Returns the resulting responses for display on frontend

Input Processing Module (Python)Parsing and cleaning of user input

Entity recognition (symptoms, duration,condition)

Conversion of input to embedding-acceptable format for retrieval

Retrieval Module (Using Pinecone Vector DB)

Conversion of input to a vector by sentence embedding model (e.g., SentenceTransformer).

Asks Pinecone vector database to return top relevant documents/chunks

Returns these relevant contexts to be used for response generation

Generation Module (RAG - GPT/Gemini API)

Combines user input + content retrieved

Passes combined input to generative AI model through Gemini API

Gets and formats context-aware, medically competent response

Response Renderer (Frontend Integration)

Accepts output from Flask backend and dynamically refreshes the chat

window

Displays the chatbot's response in a readable and clear manner

Ensures user-friendly rendering for both desktop and mobile users

Security and Privacy Layer

Flask manages sessions securely without preserving individualized health

information

HTTPS guarantees encrypted data transportation User

input isn't stored or logged to ensure privacy

CONCLUSION

The developed Medical Chatbot system demonstrates an effective application of Retrieval- Augmented Generation (RAG) architecture to provide users with

accessible, reliable, and real-time medical information. By integrating Python Flask for backend processing and a clean HTML interface for interaction, the chatbot ensures an engaging user experience. The system bridges the gap between healthcare awareness and accessibility, especially for individuals in remote or underserved regions. Though not a substitute for professional medical consultation, it empowers users with initial health guidance and improves their ability to make informed decisions.

FUTURE WORKS

1. **Multilingual Support** – Incorporate language translation models to cater to non- English-speaking users
2. **Voice-Based Input and Output** – Integrate speech recognition and text-to- speech for more inclusive accessibility
3. **Document Upload Feature** – Allow users to upload prescriptions or reports for enhanced analysis
4. **Context Memory** – Improve the chatbot's ability to handle follow-up questions and maintain conversation flow
5. **Integration with IoT Devices** – Connect with health tracking devices (like fitness bands) for personalized suggestions
6. **Security Enhancements** – Implement end-to-end encryption and compliance with healthcare data standards (e.g., HIPAA)

REFERENCES

[1] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345-1359, Oct. 2010.

[2] J. Devlin, M. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint

[3] A. Vaswani et al., "Attention is All You Need," Advances in Neural Information Processing Systems (NeurIPS), pp. 5998-6008, 2017.

[4] S. Thorne et al., "The Fact Extraction and VERification (FEVER) Shared Task," Proceedings of NAACL-HLT, pp. 809–819, 2018.

[5] M. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," arXiv preprint arXiv:2005.11401, 2020.

[6] Y. Zhang et al., "DocChat: An Information Retrieval-Based Healthcare Chatbot System," Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM), pp. 2617–2625, 2019.

[7] A. Rajpurkar et al., "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," arXiv preprint arXiv:1711.05225, 2017.

[8] L. Deng and X. Li, "Machine Learning Paradigms for Speech Recognition: An Overview," IEEE Transactions on Audio

[9] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," EMNLP 2020: Systems Demonstrations, pp. 38–45, 2020.

[10] A. Sahu, M. Routh and A. Sinha, "An Intelligent Medical Chatbot Using NLP and Machine Learning," International Journal of Computer Sciences and Engineering, vol. 7, no. 4, pp. 498–502, 2019.

[11] S. Soni and S. Dey, "Healthcare Chatbot System Using NLP and Deep Learning," Journal of Web Engineering, vol. 21, no. 1, pp. 59–78, 2022.

[12] Pinecone, "Pinecone Vector Database for Scalable Semantic Search,"

[13] Google Cloud, "Gemini API Overview – Generative AI for Developers,"

[14] Flask Documentation, "Flask Web Framework for Python

[15] OpenAI, "GPT-4 Technical Report," OpenAI, 2023.

[16] S. M. Kamath and M. R. Bendre, "Design and Implementation of a Medical Chatbot," International Research Journal of Engineering and Technology (IRJET), vol. 7, no. 8, pp. 3824–3827, Aug. 2020.

[17] S. R. Kang, "AI in Healthcare: A Review of NLP-Based Medical Chatbots," Healthcare Informatics Research, vol. 27, no. 3, pp. 192–201, July 2021.

[18] L. Tanwar and A. Agrawal, "Digital Health Assistants Using AI for Primary Care," International Journal of Health Sciences, vol. 6, no. 1, pp. 101-108, 2022.

[19] H. Alam and R. Vohra, "Development of AI-Powered Web Application for Health Consultation," International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 8, no. 10, pp. 5112–5115, 2019.

[20] R. B. Mishra and D. S. Naik, "Intelligent Chatbot for Healthcare with AI Integration," Journal of King Saud University - Computer and Information Sciences, 2023.