International Journal of Computer Techniques–IJCT Volume 13 Issue 2, March-April - 2026

Open Access and Peer Review Journal ISSN 2394-2231 https://ijctjournal.org/

MATERNAL HEALTH BOT USING RAG ARCHITECTURE

Ranga.Jaswanth[#1], Ootla.Eswar Sai[#2],G.Rajashekar[#3] , Ms. S.A.NEELAVANI[#4]

Email : neelavanisa.set@dsuniversity.ac.in,jaswanthranga565@gmail.com, ootlaeswar@gmail.com, rajasekhargangireddy1234@gmail.com

[123] UG Student, Department of Computer Science and Engineering, School of Engineering and Technology, Dhanalakshmi Srinivasan University, Trichy-621112-Tamilnadu.

[#4] Assistant Professor, Department of Computer Science and Engineering , Dhanalakshmi Srinivasan Institute of Technology, Trichy-621112- Tamil Nadu.

Abstract - *This study describes the deployment of a voice-activated chatbot designed to assist expectant mothers by offering trustworthy maternal health information. To provide precise and contextually aware responses, the system employs a Retrieval-Augmented Generation (RAG) technique, which combines a language model with a local knowledge store. We developed this totally with free and open-source technologies to make the solution more accessible in rural or low-resource environments. Open-source approaches, such as Whisper for speech-to-text, are used to add voice capabilities. To ensure accuracy and validity, the data is also gathered from reputable sources like the WHO and other health portals. We used PDFs from these sources for RAG and stored them in vector databases for efficient document retrieval, as this system aims to bridge the information gap.*

Keywords - Maternal health, RAG, chatbot, voice interface, NLP, vector database

I .Introduction

Having accurate and current information is vital, but it becomes much more important during pregnancy because both the mother's and the unborn child's lives are on the line. Women still struggle to get regular checkups in many low-resource settings. According to a study, pregnant women frequently face barriers to receiving healthcare services, particularly in rural areas, which can result in poor maternal outcomes and missed health alarms [1]. Furthermore, because of their hectic schedules, working women are also sometimes skipping checkups, which leads to the missed detection of critical symptoms like nutritional inadequacies. Despite eating a nutritious diet, their rigorous job life and other health issues prevent it from having the desired effect.This is why we feel there is a need for an application that would bridge this gap to create a positive impact on their lives.

We now have new methods for creating intelligent conversational bots because to the quick development of AI and language models like GPT and LLaMA. Retrieval-Augmented Generation (RAG) is one such technique [5]. By using this technique, the bots are able to search through a given collection of documents and select the most current and pertinent information to respond to, rather than only speculating. This is particularly helpful in fields where relevance and accuracy are crucial, like health. Previous research has demonstrated how RAG pipelines, which combine language creation and vector-based retrieval, enhance the contextual accuracy of chatbot responses [2].

This study investigates the effects of a voice-activated chatbot that employs free open-source tools and operates according to the RAG approach. For effective document retrieval, our system gathers PDFs from reliable and validated medical sources, converts them into vector embeddings, and stores them in vector databases. By transforming domain-specific documents into embeddings and obtaining responses using cosine similarity with tools like FAISS and Chroma, comparable RAG-based solutions have successfully applied this strategy [2]. Because of its lightweight architecture, the chatbot may run locally on any computer.

II. Related Work

These days, chatbots have gained popularity in a variety of industries, including healthcare. When it's not always feasible to see a doctor, they assist patients in accessing important health information. Afrizal et al. [1] developed a chatbot for Telegram that uses simple natural language processing to help expectant mothers identify possible pregnancy warning signals. Because the chatbot is rule-based, it can only respond to pre-defined questions; it is unable to handle alternative queries.

In a separate study, Balasooriya and the team [3] created the "Baby Bump" mobile application. Although the app had some limits, it did have several incredibly useful features, such as a pregnancy tracker and health notifications. Instead of supporting natural discourse or access to medical records for more in-depth replies, it depended on keyword matching. Additionally, it lacked voice input, which would have been more useful in scenarios where typing is challenging, as in rural areas or for people with low literacy levels.

Vakayil et al. [2] investigated the application of Retrieval-Augmented Generation (RAG) in chatbot systems in order to get over these restrictions. Their system scanned document databases and produced more precise, context-aware replies by combining the LLaMA-2 language model with programs like FAISS and LangChain. Because of this, the chatbot was noticeably smarter than conventional rule- or keyword-based systems. However, the system was still entirely text-based and didn't support voice interaction — a key feature for users who are more comfortable speaking than typing.

Ilapaka and Ghosh [4] used a similar strategy in the field of mental health. To produce more individualized and interesting interactions, their chatbot combined RAG with memory tracking and fine-tuning methods like LoRA. The techniques and architecture they employed, particularly in relation to conversation memory and document anchoring, are quite useful, even though their work was more focused on mental health than maternal care. Motivated by their efforts and those of others, our system seeks to apply those similar skills to the field of maternal health, with a particular emphasis on facilitating voice interaction to increase the system's usefulness and accessibility for users in environments with limited resources.

Feature	Existing Systems	RAG Chatbot
Information Source	Predefined responses, static FAQs	Dynamic retrieval from verified documents
Voice Interaction	Rare or unavailable	Fully voice-enabled (Speech-to-Text)
Response Accuracy	Generic and repetitive answers	Context-aware, document-grounded, and evidence-based replies
Knowledge Update Mechanism	Manual and infrequent	Easy document updates using RAG pipeline without retraining
Explainability & Transparency	Lacks source citation and explanation	Provides sources and references for every response
Offline Functionality	Often requires internet	Fully offline using lightweight local stack

Table.1. presents a side-by-side comparison of key design features between traditional chatbot systems and the proposed RAG-based maternal care chatbot.

III. System Architecture

Our voice-activated RAG chatbot's architecture is designed to be straightforward, adaptable, and simple to use with open-source tools. It provides useful and contextually relevant information about maternal health by combining language generation, document retrieval techniques, and speech processing. Voice input handling, document preparation, vector storage, response creation and

retrieval, and a user interface are the five primary components of the system.

When the user provides input, either in the form of a written or spoken question, the procedure begins. The system employs open-source speech recognition algorithms, such as Whisper, if the inquiry is voice [2]. It is delivered straight to the RAG pipeline if it is a text inquiry.

The document preparation module processes trustworthy medical content concurrently. Upon receiving this input, the system transforms it into vector embeddings, which are essentially mathematical representations of the data in n dimensions [2], [4]. A local vector database that facilitates similarity-based searches, including cosine similarity, has these embeddings. The system uses similarity-based searches to find the most pertinent similar data in the vector database based on the query. The language model, a lightweight offline model designed to keep the entire system local and accessible, receives the query after that [2]. In order to respond to the user's inquiry, it searches the vector database for the most pertinent information and retrieves it [2], [4].

The user interface now shows the completed response. Because the system is designed to function solely on a local computer, it is particularly helpful in environments with limited resources [1], [4].

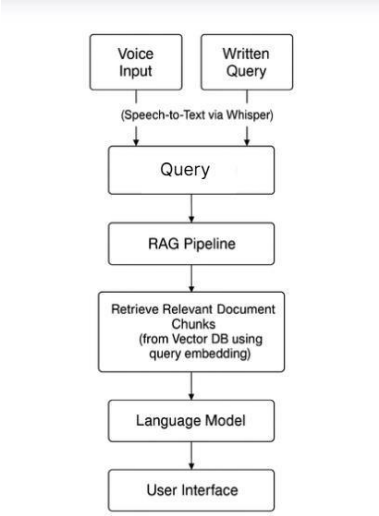

Fig.1. System Architecture

IV. Methodology

The Retrieval-Augmented Generation (RAG) technique [2], [5] is the foundation of this architecture, which aims to create a dependable chatbot that provides accurate and current responses about maternal health. Because of its lightweight and modular design, the system can run fully on local computers with free and open-source software. The functioning prototype was created using the following methodology:

A. Vectorization and Document Processing

The first step in the procedure is gathering the pertinent documentation from reputable sources, like the World Health Organization and various national health portals. We can make sure that this system can handle additional file types, such as Word and Excel, by making minor adjustments. All of these PDFs are uploaded into a folder. After that, a text-splitting approach is used to analyze the folder and split the documents into smaller pieces, or coherent text segments [4]. The primary goal is to produce efficiently searchable, semantically meaningful chunks. After that, pre-trained open-source models are used to incorporate each fragment into a numerical vector [4]. Following that, these embeddings are kept in a local vector database such as ChromaDB [4], which facilitates retrieval based on similarity using techniques like cosine similarity.The purpose of dividing the text into chunks rather than keeping the entire thing as one is to skip duplicates, which makes the knowledge base clean, structured, and less redundant.

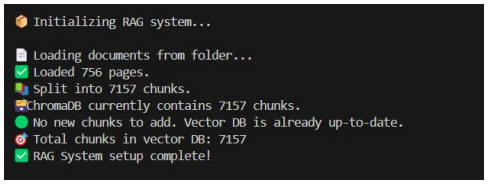

Fig.2. RAG system initialization output displaying document loading, chunking process, and vector database population with duplicate handling capability.

B. Query Input and Context Retrieval

Both voice-activated and text-based interactions are supported by the chatbot. Through a user interface, users may immediately enter inquiries. Upon submission, a query is transformed into a vector embedding, which is then compared to the embeddings in the local vector database to identify the most pertinent and comparable data points. Based on semantic similarity, the retrieval module chooses the most similar chunks or the best

matching segments. The response generator receives these chunks as input once they have been merged with the user query [2], [4].

C. Response Generation Using LLM

Here, the user query and the context that was retrieved are entered into the LLM. The final response is generated by large language models using frameworks such as Ollama [2], [4]. Because these models are CPU-optimized, they are completely internet-independent.

external APIs or connection. Reliability is increased and hallucinations, the primary issue with present systems, are lessened thanks to the output from the retrieved context.

D. User Interface and Voice Feature

Although the prototype is still text-based, it will be able to accept inquiries and display answers in an easy-to-use format after the user interface is enhanced and the voice module is integrated utilizing open-source modules such as Whisper [2]. We can incorporate the ability to upload documents into the UI with even a minor improvement. In the same way, the user can upload their own document and receive responses from it.

Fig.3. Pregnancy support chatbot interface with RAG system integration, showing successful initialization and sample query input for pregnancy-related medical information retrieval.

V. Results

We tested the chatbot using various questions that people might ask during pregnancy in order to simulate real-world scenarios and assess its performance. The purpose of this is to evaluate the chatbot's dependability and response to these queries.

We experimented with questions about diet, symptoms, cord care, and other topics. The algorithm was able to look through the stored documents and find pertinent information that was both on topic and pertinent.

Sample-query:

Fig.4. User interface showing the chatbot responding to a cord care query with evidence-based information retrieved from WHO guidelines.

Instead of hallucinating, the chatbot in this case just used the confirmed and trustworthy content that was retrieved. One of the main advantages of the RAG approach is that it maintains responses pertinent and based on the real materials. Without requiring internet access or sophisticated gear, this version functions flawlessly on a standard computer and is completely offline.

Phase	Feature	Description	Tools/Methods
Phase 1	Multilingual Support	Enable interaction in Indian regional languages	Indic NLP, Translate API
Phase 2	Agentic PDF Collect	Auto-fetch PDFs and ingest using LangChain agent	LangChain + RPA
Phase 3	User Feedback loop	Collect & retrain on user	Prompt injection, RLHF

		corrections	
Phase 4	Mobile App Interface	Build user-friendly mobile app	Flutter/React Native

Table 2. Development roadmap outlining the planned enhancements for the RAG-based maternal care chatbot, including feature goals and associated tools for each phase.

VI. Future Scope

Later on, we can develop and enhance this project. To improve its functionality, we may add an agent that searches the web to retrieve information from trustworthy sources [6]. This agent might look for pertinent information or PDFs, download them, and then feed the content into the RAG pipeline so that it can be divided into manageable chunks. After that, these fragments might be saved in the vector database as vector embeddings. This would eliminate the requirement for manual labor. This functionality might make it easier for the chatbot to remain current at all times.
The addition of language support is another significant enhancement. This makes the chatbot easier to use by allowing users to use it in the language of their choice. We could use the Google Translate API for this update, or we could use tools like the Indic NLP Library to help incorporate local languages.
We can even include user feedback loops. With these characteristics, the system can remain accurate and intelligent while also being easier to use and more user-friendly.

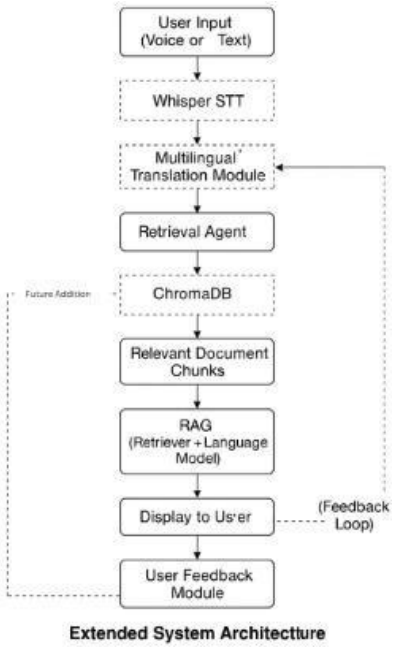

Fig.5. System Architecture with integrated Future Scope

VII. ETHICAL AND PRIVACY CONSIDERATIONS

Protecting personal information is essential in healthcare, particularly in maternal care. Sensitive information may be included in user inquiries about medical issues that are handled by the suggested chatbot. The system is built to run completely offline, removing the exposure of cloud-based data in order to reduce risk.

Additionally, the chatbot interface has to include explicit disclaimers that emphasize that this technology is not a replacement for expert medical advice. Although responses are based on trustworthy sources, clinical judgment should always come first.

Future generations of the chatbot should be checked for bias in order to maintain fairness, particularly in multilingual use scenarios where inaccurate translations may result in inaccurate information or

responses that are culturally offensive. By integrating user corrections to increase transparency and trust, feedback mechanisms also conform to ethical AI requirements.

VIII. Conclusion

In order to provide accurate and dependable information about maternity care utilizing open-source and free resources, the study introduces a voice-activated chatbot based on RAG architecture. For pregnant women in low-resource locations, the solution helps to close the information gap by integrating a lightweight, offline-friendly architecture.

Reducing hallucinations and improving accuracy, the chatbot gets verified information from reliable sources such as WHO and other national health portals, converts it into vector embeddings, and then responds.

Automation with agentic RAG and enhanced interactions with multilingual support are potential avenues for future growth. The project demonstrates how effective AI-powered solutions can be, even when implemented with little funding.

References

[1] S. H. Afrizal, N. Hakiem, A. Erna Permanasari, H. Syaifullah Albab, G. Yoki Sanjaya and L. Lazuardi, "A User-Centered Design of Natural Language Processing for Maternal Monitoring Chatbot System," *2022 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, Jakarta, Indonesia, 2022, pp. 244-248,doi:10.1109/ICIMCIS56303.2022.10017517.

[2] S. Vakayil, D. S. Juliet, A. J and S. Vakayil, "RAG-Based LLM Chatbot Using Llama-2," 2024 7th International Conference on Devices, Circuits and Systems (ICDCS), Coimbatore, India, 2024, pp. 1-5, doi: 10.1109/ICDCS59278.2024.10561020.

[3] D. Balasooriya *et al*., "Baby Bump: A Monitoring System for Pregnant Mothers and Babies," *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, Pune, India, 2024, pp. 1-6, doi: 10.1109/I2CT61223.2024.10544017

[4] A. Ilapaka and R. Ghosh, "A Comprehensive RAG-Based LLM for AI-Driven Mental Health Chatbot," *2025 7th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA)*, Ankara, Turkiye, 2025, pp. 1-5, doi: 10.1109/ICHORA65333.2025.11017017.

[5] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Schwenk, D. Kiela, and S. Riedel, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474. [Online]. Available: https://arxiv.org/abs/2005.11401

[6] A. Singh, N. Gupta, and S. Das, "Agentic-RAG: A Survey on Agentic Retrieval-Augmented Generation," *arXiv preprint*, arXiv:2501.09136, 2025. [Online]. Available: https://arxiv.org/abs/2501.09136