2025 International Conference on Computing and Communication Technologies (ICCCT) | 979-8-3315-3757-9/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICCCT63501.2025.11019682

QuickAid: A Hybrid RAG and LLM Framework for Improving Chatbot Accuracy and Relevance in First-Aid Guidance

1st Divy Shikha
CSE Department
Amity University Uttar Pradesh
Noida, India
divyshikha1890@gmail.com

2nd Sanya Bansal
CSE Department
Amity University Uttar Pradesh
Noida, India
sanyabansal10005@gmail.com

3rd Aakriti Raman
CSE Department
Amity University Uttar Pradesh
Noida, India
ramanaakriti661@gmail.com

4th Seema Sharma
CSE Department
Amity University Uttar Pradesh
Noida, India
ssharma26@amity.edu

***Abstract*—India is facing significant healthcare issues, including a doctor-patient ratio of 1:836, delayed emergency response times, and terribly limited medical infrastructure in rural areas where 65% of the population reside, while healthcare facilities number only 25%. With overcrowded hospitals dealing with routine non-emergency cases, proper first aid counsel could help in such cases. Also, in case of emergencies step by step first aid guidance help patients to save their life by managing critical situations, like cardiac arrest, choking, or severe bleeding before reaching hospitals. To bridge the healthcare gap, there is a need of medical chatbot that provides immediate, cost effective and multimodal emergency guidance especially in rural areas . The rapid advancements in artificial intelligence(AI) have significantly transformed healthcare applications, particularly in the development of medical chatbots. To create an intelligent First Aid Medical Chatbot, this study proposes QuickAid, a hybrid approach that integrates Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs). RAG improves the ability of the chatbot to provide accurate, real-time first aid advice by retrieving relevant medical knowledge from authoritative and evidence-based sources and combining it with generative AI capabilities. This ensures that users receive accurate, contextually relevant, and timely emergency support. The dataset is designed to cover essential first aid requirements, including choking, headaches, and other medical emergencies. The chatbot facilitates multimodal interaction through both text and audio for better accessibility. Its hybrid strategy helps to lower the risk of errors by employing structured retrieval. QuickAid's effectiveness is supported by a BLEU score of 0.87 and ROUGE scores of 0.92 (ROUGE-1), 0.85 (ROUGE-2), and 0.91 (ROUGE-L), reflecting its high accuracy and fluency.**

***Index Terms*—Retrieval-Augmented Generation (RAG), Large Language Models (LLMs), FAISS-based Similarity Search, SentenceTransformer Embeddings, Groq LLaMA3 Model**

I. Introduction

In emergencies, moments are crucial. Imagine a victim of intense chest pain at midnight, with no doctor in sight. At such times, one reliable source of first-aid guidance may pull you through. Conventional health care systems have to deal with a lot of delay in consultation, which makes it impossible in some cases for patients not to get attention at the last-minute hour of need. This is where QuickAid: AI-powered medical chatbot comes into play, offering instant support and bridging the gap between patients and healthcare providers. This study enhances medical chatbot capabilities by integrating Retrieval-Augmented Generation (RAG) with Large Language Models (LLMs) to improve response accuracy and contextual relevance. Traditional NLP models often struggle with complex medical terminology and diverse patient inquiries. To address this, our chatbot utilizes SentenceTransformer embeddings and FAISS-based similarity search, ensuring precise information retrieval from structured medical knowledge bases. Additionally, stringent authentication measures safeguard user data, reinforcing privacy and security. By overcoming limitations in existing medical chatbots. The aim of this study is to develop a system that provides a more reliable, responsive, and context-aware healthcare solution.

II. Related Work

Numerous studies emphasize both the potential and existing challenges of AI-driven, NLP-based medical chatbots. S. Biradar and S. shastri 2024 [1] developed a chatbot which uses ML Techniques like SVM for predicting infectious diseases, and achieved an 97.4% accuracy, but it struggled with complex queries that went beyond predefined symptoms. B. Suvarnamukhi, et al., 2025 [2] proposed an EHR-integrated chatbot that faced difficulties with maintaining conversational flow. R. Jegadeesan, et al., 2023 [3] developed a BERT-based model that enhanced language comprehension but required significant domain-specific fine-tuning and encountered issues

979-8-3315-3757-9/25/$31.00 ©2025 IEEE

with ambiguous medical Terminology. Moreover, author utilized retrieval-based NLP with KNN to classify diseases by severity and achieved 75% accuracy but it lacked a system for updating data dynamically. M. Mittal et al., 2021 [4] indicated that well-known chatbots like HOLMeS and IBM Watson have inadequate word prediction and pattern recognition capabilities, which restricts their ability to produce innovative medical insights. In the study conducted by C. Bulla, et al., 2020 [5] highlighted that speech-enabled chatbots show potential for improving accessibility but often struggle with real-time symptom assessment. H. Mendapara et al. (2021) [6] highlighted that these systems struggle with generalization and require continuous data updates and retraining. J. Lee, et al., 2020 [7] emphasized the need for architectures integrating medical knowledge bases with LLMs. Consequently, this study highlights the necessity of a hybrid approach that combines optimized LLMs and strict security protocols to improve accuracy, contextal understanding, and adherence to healthcare data standards. LLM-based models like BioBERT and ClinicalBERT [8] employed Bidirectional Encoder Representations from Transformers (BERT) [9] to enhance medical language comprehension but sometimes struggle with contextual data.

III. Methodology

Figure 1 described the steps taken to implement the proposed chatbot. It follows a structured process to ensure accurate and accessible first-aid assistance. The Key steps include architecture design, data processing, multilingual integration, semantic embedding, model training, and evaluation, ensuring reliability and usability.

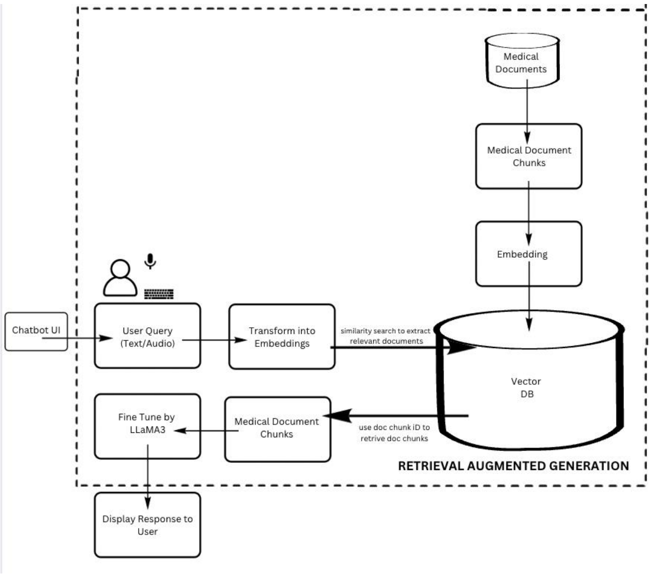


Fig. 1. Architecture of proposed QuickAid Chatbot

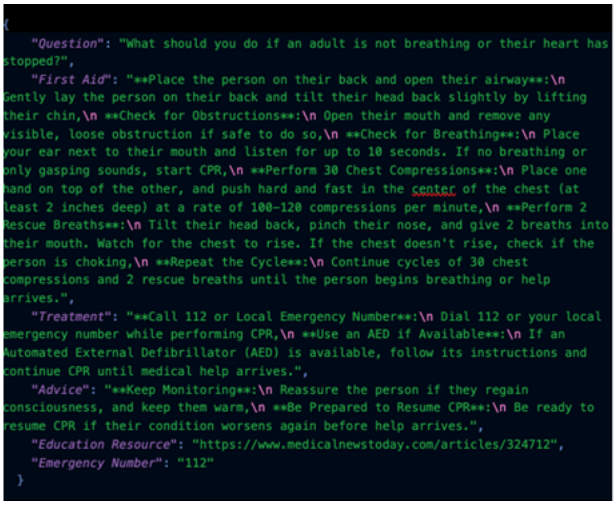
A. Proposed Architecture

The chatbot takes care of the text or audio queries through a Flask backend responsible for authentication, session tracking, and routing. The queries are embedded by a transformer model, and the matches are stored in a FAISS index for a quick first-aid retrieval. FAISS (Facebook AI Similarity Search) is a powerful library developed by Facebook AI Research that enables rapid similarity searches and clustering of dense vectors. It is crucial for efficient and precise retrieval in extensive datasets [10]. Stored solutions handle high-confidence matches, while low matches use a larger language model.

B. Data Collection

The knowledge base of the proposed chatbot is built from a curated dataset obtained from medical databases, emergency guidelines, and WHO manuals. The dataset consists of structured questions and answers and pre-processed with the help of NLP techniques, namely text cleaning, tokenization, and vectorization, and audio transcription.

For the proposed work, the dataset includes medical emergencies like choking, cardiac arrest, burns, and allergies. The data is stored in JSON format shown in figure 2 which is designed to provide first aid guidance by matching queries with the most relevant emergency response information. It includes a ”Question” field to identify the emergency scenario, ”First Aid” for immediate steps, ”Treatment” for medical actions, ”Advice” for additional precautions, ”Education Resource” for further learning, and ”Emergency Number” for contacting emergency services. It covers various situations, such as choking, headaches, burns, and CPR, to assist in first aid response.



```
{
    "Question": "What should you do if an adult is not breathing or their heart has
stopped?",
    "First Aid": "**Place the person on their back and open their airway**:\n
Gently lay the person on their back and tilt their head back slightly by lifting
their chin,\n **Check for Obstructions**:\n Open their mouth and remove any
visible, loose obstruction if safe to do so,\n **Check for Breathing**:\n Place
your ear next to their mouth and listen for up to 10 seconds. If no breathing or
only gasping sounds, start CPR,\n **Perform 30 Chest Compressions**:\n Place one
hand on top of the other, and push hard and fast in the center of the chest (at
least 2 inches deep) at a rate of 100-120 compressions per minute,\n **Perform 2
Rescue Breaths**:\n Tilt their head back, pinch their nose, and give 2 breaths into
their mouth. Watch for the chest to rise. If the chest doesn't rise, check if the
person is choking,\n **Repeat the Cycle**:\n Continue cycles of 30 chest
compressions and 2 rescue breaths until the person begins breathing or help
arrives.",
    "Treatment": "**Call 112 or Local Emergency Number**:\n Dial 112 or your local
emergency number while performing CPR,\n **Use an AED if Available**:\n If an
Automated External Defibrillator (AED) is available, follow its instructions and
continue CPR until medical help arrives.",
    "Advice": "**Keep Monitoring**:\n Reassure the person if they regain
consciousness, and keep them warm,\n **Be Prepared to Resume CPR**:\n Be ready to
resume CPR if their condition worsens again before help arrives.",
    "Education Resource": "https://www.medicalnewstoday.com/articles/324712",
    "Emergency Number": "112"
  }
```

Fig. 2. Sample JSON Record

C. Data Preprocessing

The raw data collected is preproccessed to ensure transformation into a structured format suitable for analysis. This ensures a standardized dataset for further processing. The steps taken for data preprocessing for QuickAid is as follows:

I. Text Normalization In the first step, the dataset was cleaned and normalized by removing unwanted characters like

@, #, %, &, *, (), [], {} and other punctuations and symbols. The text was then converted into lowercase to avoid case-sensitive mismatches using a function (1).

$$X' = \text{lower}(X) \tag{1}$$

Here, X' stands for the actual text.

II. Tokenization In the second step of preprocessing, the normalized text is broken down into words. This process converts raw text into meaningful units, such as words (”first aid”) or sub words (”first-aid”).

III. Vectorization In the third step, the text is converted into numerical representations using SentenceTransformer embeddings, mapping user queries and knowledge base into numerical representations that capture their meaning and semantic relationships. The FAISS (Facebook AI Similarity Search) indexing technique is then applied to store and retrieve these embeddings quickly, by running a similarity search on the database to get coherent responses.

D. Implementation

I. Retrieval Augmented Generation (RAG)

To improve accuracy and contextually appropriate responses, retrieval augmented generation, combines up-to-date information sources with large language models. Without RAG, LLMs rely solely on their pre-trained knowledge which can become outdated and may lack critical, real time medical information. To supplement the limited knowledge of LLM, RAG retrieves up-to-date and relevant information from the authorised databases when the query is generated by the patient and ensures accurate context-aware and evidence based responses. RAG extends the powerful capabilities of LLM without the need to retrain the model. To generate responses to the patient queries, RAG allows LLMs to access information outside of their training data that allows them to produce more refined responses without extensive training or fine tuning. The proposed solution uses RAG to add evidence-based first-aid documents from the authorised sources to the chatbot’s knowledge base. RAG supports continuous knowledge updates for the chatbot without the need for costly LLM retraining that enhances the chatbot’s utility. Combining RAG with the LLM augments the model’s reasoning capabilities and provides continuous access to updated knowledge. Consequently, the chatbot can provide strong, evidence-driven explanations that adapt to the latest available information.

II. Multimodal Interaction, Multilingual and Location-Based Services

- Multimodal Interaction: The proposed work features multimodal interaction which enable patients to interact with QuickAid through multiple input formats i.e., text and speech. The chatbot incorporates speech-recognition functionality which improves user experience for patients who are incapable to type in unsuitable times. This work integrated Groq API (Whisper model) for audio-to-text processing. The conversion from speech-to-text is (2):

$$\mathbf{X} = \mathit{Whisper}(J, F) \tag{2}$$

where J represents the input audio file, F represents the Whisper model used for transcription and X is the output transcribed text.

- Multilingual Speech Processing: Furthermore, QuickAid supports multilingual interactions, making it accessible for patients with language barriers. When the patient records an audio, the Whisper model detects the language and converts it into text. The function for multilingual speech-to-text transcription is (3) :

$$X = \mathit{Whisper}(J, F, L) \tag{3}$$

where J represents the input audio file, F represents the Whisper model used for transcription, L denotes the target language code, and X is the output transcribed text in the specified language.

- Location-Based Services
  The hospital location feature will aid patients to find nearby hospitals by processing their queries. The chatbot integrates LLaMA API with geolocation data to provide real-time hospital advisories. For example, in ”I’m feeling heart pain, please get me the list of hospitals,” the system recognizes the medical urgency and automatically detects the patient’s location bsased on their current position. The nearby hospitals are identified and retrieved as follows:
  The first step is to process the query to identify keywords related to symptoms and determine the patient’s current location. Utilising this information, the chatbot queries the LLaMA API to search the hospital database. Based on the results, the chatbot recommends hospitals, providing names, addresses, and contact details. Finally, patients can reach out to the nearest facility.

E. Model Training

The chatbot went under model training to ensure that it can accurately understand and respond to diverse medical queries. If the chatbot is not trained, it would struggle to interpret variations in user input, recognise medical terminology , and provide accurate first-aid advice. By integrating RAG with a custom-built first-aid dataset, the chatbot effectively retrieves relevant information and has acquired an understanding of semantic relationships between symptoms and treatments, allowing it to generate contextually and medically accurate advice. The embeddings in the database are indexed with FAISS (Facebook AI Similarity Search) for easy retrieval. The training process minimizes binary cross-entropy (BCE) loss, which is given by (4):

$$J = -\sum_{k=1}^{M} [t_k \log \hat{p_k} + (1 - t_k) \log(1 - \hat{p_k})] \tag{4}$$

where t_k is the ground truth label, $\hat{p_k}$ is the predicted probability, and J represents the binary cross-entropy loss.

F. Model Evaluation

The RAG model was assessed and compared using BLEU and ROUGE scores to evaluate the correctness, coherence, and relevance of various chatbot responses against the constructed first-aid instructions. The achieved scores for the BLEU and ROGUE metrics for the proposed chatbot are presented in Table I.

1) BLEU Score: A high BLEU score of 0.87 suggests that the responses generated by the chatbot closely align with the reference answers.

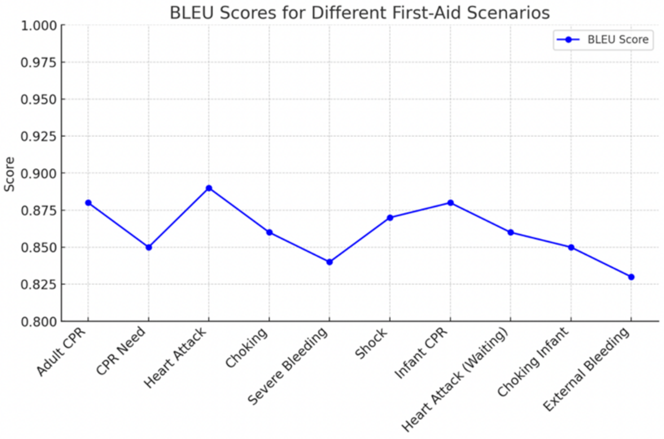


Fig. 3. BLEU scores across first-aid scenarios

2) ROUGE SCORE: This metric evaluates word and phrase overlaps, resulting in ROUGE (Recall-Oriented Understudy for Gisting Evaluation):

- ROUGE-1: 0.92 – Indicates excellent word similarity.
- ROUGE-2: 0.85 – Reflects highly coherent phrase matching.
- ROUGE-L: 0.91 – Demonstrates high fluency and structural accuracy.

This confirms that the RAG model can retrieve and generate high-quality medical responses to provide real-time accuracy.

TABLE I
BLEU AND ROUGE SCORES FOR FIRST-AID MEDICAL CHATBOT

METRIC	SCORE	DESCRIPTION
BLEU	0.87	Indicates high similarity to reference answers
ROUGE-1	0.92	Shows excellent word similarity
ROUGE-2	0.85	Represents strong phrase coherence
ROUGE-L	0.91	Signifies fluency and structural accuracy

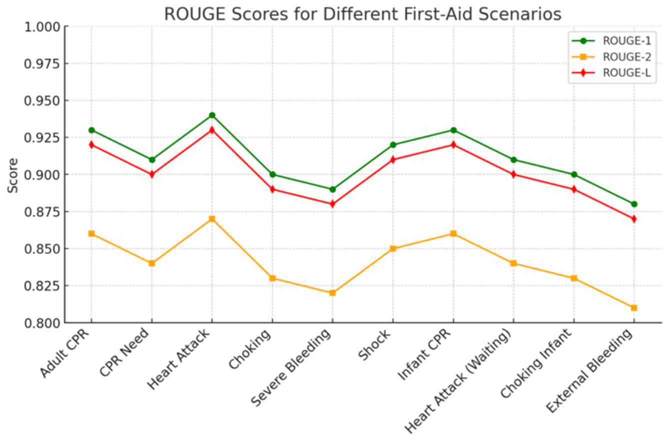
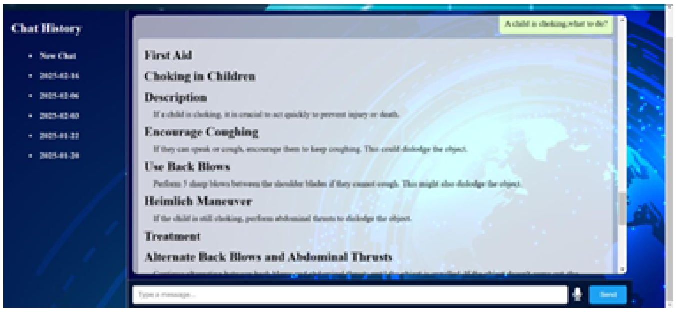


Fig. 4. ROUGE Score Graph: ROUGE scores across first-aid scenarios

IV. DISCUSSION AND RESULT

Our experiments validate that with the aid of the RAG model, QuickAid gives accurate, context-based first-aid instruction. Using a filtered dataset derived from both medical databases and WHO manuals, QuickAid handles queries under FAISS indexing and retrieves pertinent document segments. These are supplemented by conversational context and input into Groq's LLaMA3 to produce coherent responses. For instance, when a user asks about treatments for viral fever, FAISS retrieves very quickly relevant information, and LLaMA3 provides a well-structured and medical-grade response. Performance evaluation metrics have confirmed the capabilities of QuickAid to achieve BLEU score of 0.87 and ROUGE scores of 0.92, 0.85, and 0.91 for ROUGE-1, ROUGE-2, and ROUGE-L respectively. In addition, QuickAid enables speech-to-text and language detection with its multilingual support through Whisper API, allowing users the convenience of communication in their language of preference. This LLaMA3 LLM enabled geo-location service makes QuickAid even more effective by allowing in real-time hospital recommendations based on the user's searched or asked location. These functionalities work together to make QuickAid a reliable, quick, and efficient first-aid assistant for giving medical directions at any moment and any place.



A child is choking,what to do?

First Aid

Choking in Children

Description

If a child is choking, it is crucial to act quickly to prevent injury or death.

Encourage Coughing

If they can speak or cough, encourage them to keep coughing. This could dislodge the object.

Use Back Blows

Perform 5 sharp blows between the shoulder blades if they cannot cough. This might also dislodge the object.

Heimlich Maneuver

If the child is still choking, perform abdominal thrusts to dislodge the object.

Treatment

Alternate Back Blows and Abdominal Thrusts

Fig. 5. Chatbot Response to user query

V. Conclusion

QuickAid stands as a potential AI-powered first-aid assistant through retrieving-augmented generation (RAG), FAISS indexing, and LLaMA3 LLMs to provide context-aware medical guidance. Its multimodal nature, including speech-to-text transcription via the Whisper API and hospital recommendations based on geolocation, increases its accessibility and response in emergencies. Quantitative performance metrics, including the BLEU and ROUGE scores, reinforce that QuickAid can generate relevant and structured responses along with multilingual queries with voice input while considering location-based assistance, making it deployable in real-world usage situations. In the future, the focus will be placed on incorporating new medical literature into the dataset so that QuickAid's knowledge and responsiveness towards varying health conditions is improved. Increasing the chatbot's multilingual and location-based services will broaden the accessibility of the chatbot to different languages and regions. Apart from integrating OCR based analysis on medical attachments, real-time clinical data, AI-based chronic disease risk assessment, and model pruning to reduce the burden on the system's resources will also be integrated. All of these advances are aimed at improving QuickAid's first-aid assistance accuracy as well as making it more effortless, enabling users to receive accurate healthcare information when needed.

References

[1] S. Biradar and S. Shastri, "Medical chatbot: Ai based infectious disease prediction model," *Journal of Scientific Research and Technology*, pp. 1–12, 2024.

[2] B. Suvarnamukhi, A. Praveena, C. R. C. Prakash, and K. N. Babu, "Automated chatbot for healthcare using nlp and ml," *International Journal of Information Technology and Computer Engineering*, vol. 13, no. 1, pp. 1–10, 2025.

[3] R. Jegadeesan, D. Srinivas, N. Umapathi, G. Karthick, and N. Venkateswaran, "Personal healthcare chatbot for medical suggestions using artificial intelligence and machine learning," *European Chemical Bulletin*, vol. 12, no. 3, pp. 6004–6012, 2023.

[4] M. Mittal, G. Battineni, D. Singh, T. Nagarwal, and P. Yadav, "Web-based chatbot for frequently asked queries (faq) in hospitals," *Journal of Taibah University Medical Sciences*, vol. 16, no. 5, pp. 740–746, 2021.

[5] C. Bulla, C. Parushetti, A. Teli, S. Aski, and S. Koppad, "A review of ai based medical assistant chatbot," *Res Appl Web Dev Des*, vol. 3, no. 2, pp. 1–14, 2020.

[6] H. Mendapara, S. Digole, M. Thakur, and A. Dange, "Ai based health-care chatbot system by using natural language processing," *International Journal of Scientific Research and Engineering Development*, vol. 4, no. 2, 2021.

[7] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[8] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical bert embeddings," *arXiv preprint arXiv:1904.03323*, 2019.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

[10] B. L. AN *et al.*, "Docbot: Integrating llms for medical assistance," 2024.