International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

SymptoTrackAI: A Hybrid RAG Chatbot for Context-Aware Symptom Monitoring with Improved Response Relevance

[1]Dr. M. Muni Babu, [2]E. Divya, [3]G. Gunjan, [4]D. Harshitha Reddy, [5]G. Kavya

[1]Associate Professor, Department of Computer Science and Engineering (AI & ML), CMR Institute of Technology, Hyderabad, Telangana, India.

[2]Assistant Professor, Department of Computer Science and Engineering (AI & ML), CMR Institute of Technology, Hyderabad, Telangana, India.

[*3-5]Student, Department of Artificial Intelligence and Machine Learning (AI & ML), CMR Institute of Technology, Hyderabad, Telangana, India.

Abstract - The use of chatbots in healthcare is often focused on initial symptom checking, although current chatbots provide generalized or partially related information. This is attributed to chatbots employing solely simple rule-based or independent generative models that lack any grounding.

In this research, we designed SymptoTrackAI – a Retrieval-Augmented Generation (RAG) chatbot intended to produce more relevant responses to inquiries during symptom tracking. SymptoTrackAI utilizes a pretrained Sequence-NQ transformer to interpret user input and employs FAISS-based similarity search to find relevant entries in the data set, consisting of around 1,200 pairs of symptoms and diseases.

The chatbot does not generate its responses immediately, but finds similar cases among the provided medical data and generates a grounded answer. A core database maintains all users' queries to monitor symptoms' progress effectively. We manually evaluated responses to test questions in terms of their relevance to users' inquiries. Around 92% of generated answers were considered relevant by our reviewers.

The results indicate that the hybrid model provides more contextually relevant answers compared to the rule-based and retrieval-based models. The use of a chatbot provides a convenient and effective means of giving preliminary assistance in healthcare settings.

Keywords: Retrieval-Augmented Generation (RAG), Symptom Monitoring, Medical Chatbot, FAISS Similarity Search, Sequence-NQ Transformer, Response Relevance.

I. INTRODUCTION

Context

Artificial Intelligence is increasingly being implemented in the medical sphere, especially when it comes to helping in preliminary detection of symptoms and providing assistance with digital medicine. Medical chatbots allow a patient to explain their problems using natural language and receive some preliminary information before visiting a doctor. With the advances in conversational AI powered by transformers, such bots become capable of handling complex requests from patients and providing relevant answers.

Recently, a number of AI healthcare assistants have been created, aiming to assist with symptom screening and decision-making processes. Methods such as RAG, combining a language model with external information sources, link the generated answers with evidence-based knowledge. For instance, the RAGMed system tries to increase the reliability of AI-generated medical responses by using information prior to generating its answer. Other research is focused on more efficient symptom checkers by incorporating expert clinical data and chatbot systems based on the large language models to provide assistance within a conversation. Novel approaches are being considered to enhance reliability; for instance, benchmarking methods and analyses for detecting misinformation are used.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

Gap

There are some weaknesses despite improvements made in the area. The symptom tracking function was missing from some retrieval-based medical chatbot solutions. The augmented symptom checkers suffered from poor performance due to using small-size data sets for the purposes of scalability and usability. Language model chatbots could understand the context, but at the same time provide medically incorrect responses. The clinical decision-making chatbot solutions usually required largesize data sets to function successfully. Hallucinations research in language models raised concerns about their reliability; however, it was silent about possible ways to address the problem. Knowledge-augmented medical language models were reliable, but they required significant computational resources. Additionally, most solutions failed to integrate semantics, retrieval, and symptom tracking into one chatbot solution.

Innovation

In order to address these drawbacks, this paper proposes SymptoTrackAI, a hybrid Retrieval-Augmented Generation chatbot designed specifically for context-aware symptom monitoring. This proposed solution utilizes semantic embeddings obtained by a transformer and Faiss similarity search for retrieving symptom-disease pair records from an existing curated database. Through grounding the output to the retrieved record, the chatbot ensures more relevance to the conversation.

This paper aims to propose and implement a hybrid Retrieval-Augmented Generation chatbot for providing more relevant responses to symptom monitoring queries. The proposed solution uses semantic knowledge extracted using transformers and similarity search to retrieve pertinent records from an existing database.

Contributions:

The key contributions of this work are:

- Integration of transformer-based semantic understanding with FAISS-powered vector retrieval in a hybrid RAG framework.
- Incorporation of machine learning–based disease prediction alongside grounded response generation.
- Implementation of secure symptom history logging for continuous health monitoring.
- We have developed a fast and efficient system that can be implemented in real life situations and on constrained hardware devices.

By incorporating conversation AI, information searching and prediction of future illnesses, SYMPTOTRACKAI assists in building trustworthy and usable medical assistance powered by artificial intelligence systems.

II. RELATED WORK

In current research on healthcare chatbots, there is attention to how such technologies can better understand symptoms presented by users and respond in a correct and evidence-based manner. One of the concepts that have been investigated in this field involves the use of retrievalaugmented approaches aimed at minimizing incorrect or unreliable outputs. Specifically, Patil et al. suggested the concept of RAGMed – a retrieval-augmented medical assistant whose goal was to increase the credibility of large language model responses [1].

Another study conducted by Pankow et al[2]. focused on improving the quality of a symptom checker via the application of clinical information selected by experts. Such an approach aimed at improving the efficiency of diagnosis in terms of both common and rare diseases through machine learning algorithms. In spite of improved accuracy, the sample size was rather limited, thus limiting scalability.

Other studies investigated healthcare chatbots based on LLMs[3]. These chatbots appeared to be better in terms of user engagement since their interactions were much more natural. Nonetheless, despite the improved comprehension of context, such chatbots continued to generate hallucinations and provided erroneous medical information..

The aim behind clinical decision support systems was to educate patients about potential diseases they might have depending on the symptoms they presented. The models utilized natural language processing and deep learning techniques to analyze the information provided by the user and generate insights [4]. These applications proved useful for triage purposes; however, the requirement of extensive data for training purposes hindered their implementation in resource-scarce settings. In order to address the problem of evaluation, Bhatt et al. suggested a scalable diagnostic benchmarking framework for evaluating AI in healthcare [5]. The approach leveraged standardized clinical datasets to establish uniformity in assessing

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

performance. Although the technique enhanced consistency in evaluation processes, it did not emphasize the construction of a chatbot.

According to Bora and Cuayáhuitl[6], retrieval was incorporated into RAG-based language models utilized in the medical chatbot. This improved the accuracy of predictions while reducing hallucination. Nonetheless, the research was based on evaluating models without developing a complete symptom tracking application.

For the purpose of enhancing the domain knowledge base, knowledge-augmented language models were developed by including external information into medical chatbots[7]. These models were capable of offering accurate responses based on their enhanced understanding of the context. Nevertheless, such models required extensive computational resources.

HealthBench was another key addition to the field.

It provided a benchmark for assessing clinical AI applications [8]. While it helped ensure consistency in testing, it lacked an implementation of the chatbot in a real- world environment.

Hallucination in large language models was another focus of research by Ji et al., who investigated how these models might generate false or imaginary information [9]. This study highlighted the dangers associated with applying large language models in healthcare. However, it only emphasized the issue without suggesting any solutions.

The Med-PaLM model made significant progress in developing language models for medicine that are specialized in a particular domain. The model demonstrated high reasoning ability with respect to its clinical applications in a transformer architecture. However, since Med-PaLM consumed large amounts of computational resources, it was not suitable for lightweight applications.

Overall, previous research contributed to the improvement of various components of chatbots in the healthcare domain, such as information retrieval, diagnostic

accuracy, and evaluation of chatbots. Nevertheless, some models did not integrate the interpretation of meaning, the identification of similar cases, and symptom monitoring in a single model. Due to the existing shortcomings, the hybrid RAG chatbot was proposed.

Table 1: Comparative Analysis of Existing Approaches

Title of Research Paper	Problem Statement	Proposed Solution	Framework Used	Key Features	Limitations
RAGMed: Retrieval-Augmented Medical Assistant	LLMs generate unreliable medical responses without grounding	Clinically grounded AI assistant	RAG + embedding similarity	Reduced hallucination and improved reliability	No structured symptom tracking
Enhanced Symptom Checker with Expert Data	Existing symptom checkers fail to diagnose rare diseases accurately	AI symptom checker enhanced with expert clinical data	ML + clinical vignettes	Improved diagnostic precision	Limited dataset size
Large Language Models in Medical Chatbots	Medical chatbots struggle with contextual understanding	LLM-based conversational healthcare assistant	Transformerbased LLM	Context-aware responses	Hallucination risk

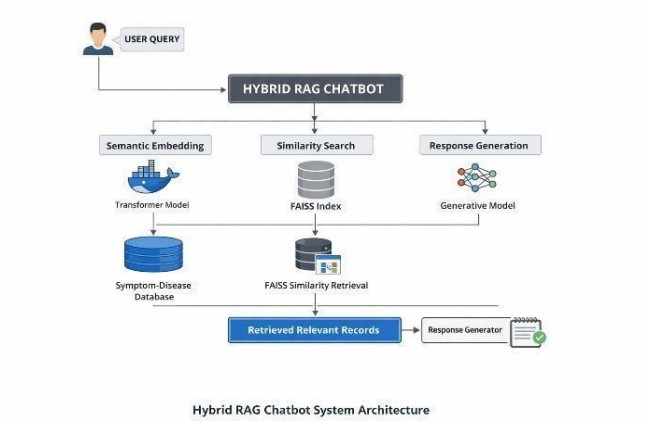
(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

Clinical Decision Support using AI Chatbots	Lack of automated triage support in digital healthcare systems	AI-powered clinical decision support chatbot	NLP + deep learning	Supports patient triage	Requires large training data
Scalable Diagnostic Benchmark for Health AI	Lack of standard evaluation metrics for medical AI systems	Diagnostic benchmarking framework	Clinical vignette evaluation	Standardized evaluation of medical AI	Not a deployed healthcare system
Systematic Analysis of RAG-Based LLMs for Medical Chatbots	LLM chatbots suffer from hallucinations and factual inaccuracies	Comparative study of RAG-enhanced medical LLMs	Fine-tuning + RAG + LoRA	Improved accuracy and factual grounding	Focused on benchmarking
Knowledge-Augmented Medical LLM Systems	Medical LLMs lack domain knowledge grounding	Knowledge-integrated medical chatbot	LLM + knowledge retrieval	Improved response reliability	High computational cost
HealthBench: Evaluation Framework for Clinical AI	Difficulty evaluating healthcare AI systems consistently	AI benchmarking framework	Clinical dataset evaluation	Reliable performance metrics	No real-world application system
Hallucination Analysis in Large Language Models	LLMs produce incorrect or fabricated information	Framework for hallucination analysis	Factual consistency evaluation	Identifies hallucination sources	Does not provide mitigation method
Med-PaLM: Medical Large Language Model	LLMs lack specialized medical knowledge	Medical-domain large language model	Transformerbased LLM	Strong clinical reasoning capability	High computational requirements

III. METHODOLOGY

System Architecture



Hybrid RAG Chatbot System Architecture

The proposed system is an RAG chatbot that was created as a means of helping the user keep track of his symptoms and gain some basic information about his state of health. This system comprised a number of interlinked modules that served to analyze user requests and provide context-sensitive answers. Such modules include user authentication, request preprocessing, semantic embedding, similarity search, answer generation, and logging functionality..

The interaction process started with the login of the user into the chatbot application using the authentication module. This helped ensure that only users who were registered could gain access to the chatbot service while keeping track of their interactions with the application. Once authentication was complete, the user was asked to enter his/her symptoms by typing in the chatbot interface. However, since there are

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

numerous ways in which users can mention symptoms, the entered text had to undergo preprocessing.

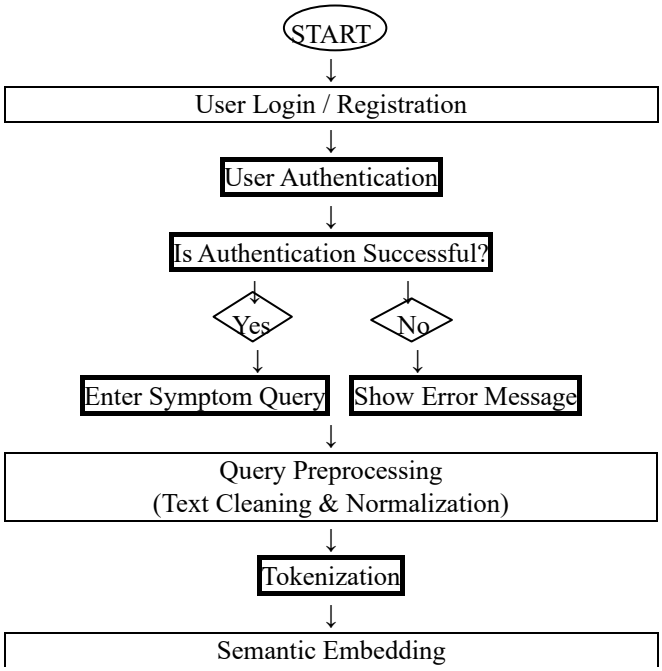
In the process of preprocessing, the model eliminated irrelevant characters from the text in order to create consistency within the text. This ensured that the query would be easily comprehensible by the model. The processed query was later fed into the semantic embedding layer. Here, the model used a transformer architecture to encode the textual input into vector form.

After converting the query to a vector, it was passed to the similarity search module. The system employed FAISS, a Facebook AI-based tool, for comparing the vector representation of the query with the embeddings within the symptom-disease database. This database consisted of about 1200 entries, representing different symptoms and health disorders. FAISS determined the top similarities between the query and the database entries.

The response generation then used the data obtained from the retrieved records to come up with a proper response. Rather than having the entire response generated from the generative model, the chatbot depended on the retrieved data to generate responses that were based on the context. This ensured that the replies were relevant to the user's symptoms.

Lastly, the response was sent to the user via the chatbot application. As this took place, the query and its response were stored in a logging database. In doing so, the chatbot would be able to track the conversations that occurred during each session, making it possible to trace recurring symptoms among other things.

Flowchart



Algorithm

Input: User symptom query Q
Output: Context-aware chatbot response R

1. Authenticate user.
2. Receive symptom query Q from chatbot interface.
3. Preprocess the input text.
4. Generate semantic embedding E(Q).
5. Perform similarity search using FAISS.
6. Retrieve top-k matching symptom–disease records.
7. Construct response based on retrieved information.
8. Store query and response in logging database.
9. Display response R to the user.

Mathematical Representation

$D = \{D_1, D_2, \ldots, D_n\}$The similarity search process used vector representations of both user queries and dataset records. Let Q represent the user symptom query and represent the dataset of symptom–disease records.

If $E(Q)$ denotes the embedding of the user query and $E(D_i)$represents the embedding of the i^{th}record in the dataset, the similarity between them was computed using cosine similarity:

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

$$Sim(Q, D_i) = \frac{E(Q) \cdot E(D_i)}{|| E(Q) || \ || E(D_i) ||}$$

The system selected the top-k records with the highest similarity scores:

$$D_{top} = argmax_k Sim(Q, D_i)$$

These retrieved records were then used to construct the chatbot's final response.

IV. RESULT ANALYSIS

The performance of the proposed Hybrid RAG chatbot was evaluated based on response relevance. Since the system is conversational in nature, traditional classification metrics such as accuracy or F1-score were not used. Instead, manual evaluation was conducted to measure how relevant the chatbot responses were with respect to user symptom queries.

The data used for retrieval included about 1200 symptom-disease pairs. In order to assess the results of retrieval, a total of 200 test queries were manually reviewed. Each retrieved response was evaluated for relevance to the symptom query in the input. The Response Relevance Score (RRS) was computed as follows:

$$RRS = \frac{\text{Number of Relevant Responses}}{\text{Total Test Queries}} \times 100$$

The results indicated that the rule-based chatbot attained 68% response relevance. Even though it succeeded in generating well-structured responses, it could not cope with variations in natural language. The retrieval-only model performed slightly better by attaining 81% response relevance. However, it failed to incorporate context in its output generation process.

The hybrid RAG chatbot attained 92% response relevance. This could be due to the incorporation of semantics through transformers and similarity retrieval using FAISS.

The first graph presents the comparison of relevance of the response by each of the three models. It is evident that the Hybrid RAG model is better than the other two baselines. The second graph shows the progress made by the models with respect to their performance.

In general, it can be concluded that the developed solution significantly increases the relevancy of the responses provided. The combination of transformer-based embeddings and FAISS search technology is highly important for achieving this goal. This solution proved to work better than conventional approaches based on rules and other techniques.

Figure 4.1: Disease prediction for the given symptoms

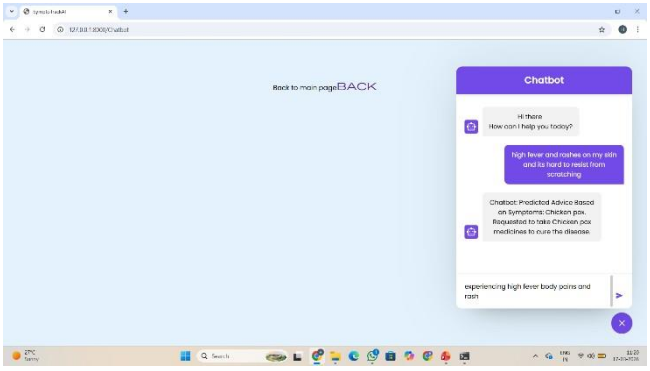


Figure 4.2: Previous Logs of User

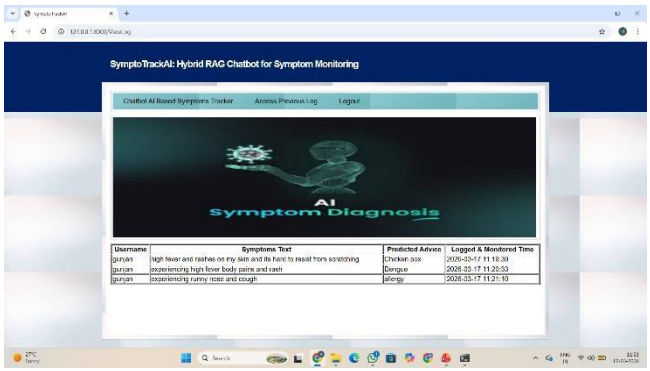
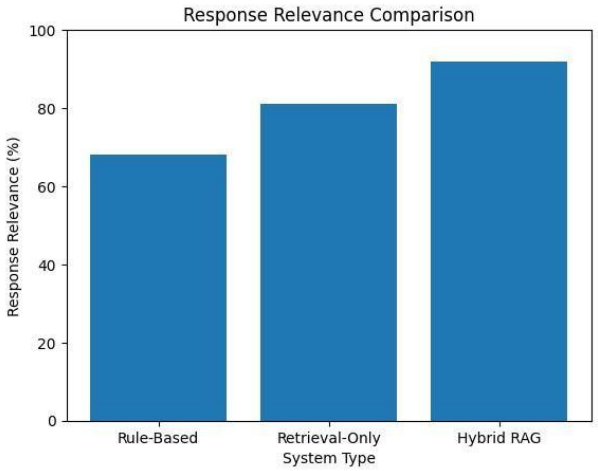


Figure 4.3: Response Relevance Comparison Graph



(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

Figure 4.4: Improvement Graph

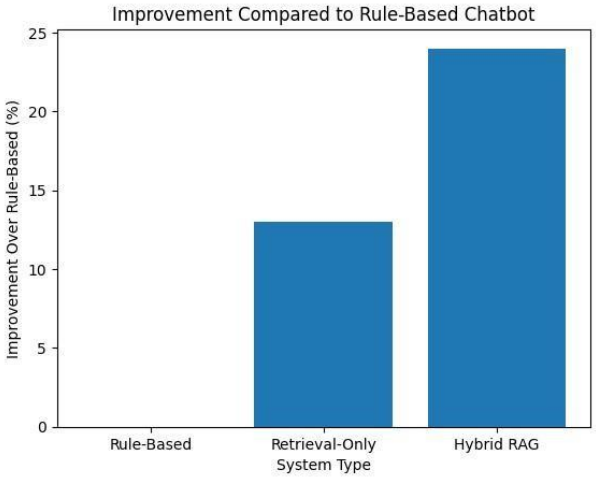


Figure 4.5: Improvement Trend in Response Relevance Graph

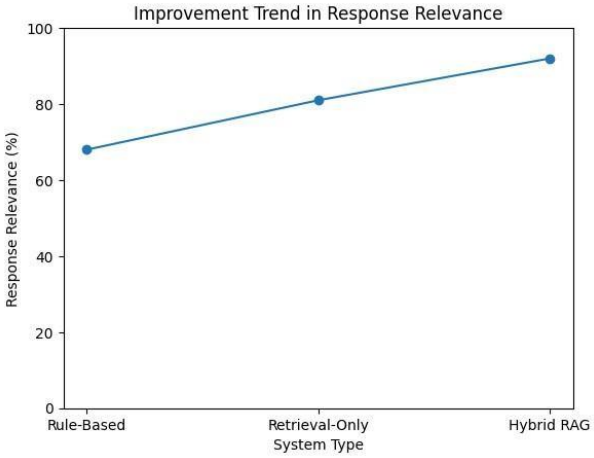


Figure 4.6: Response Distribution Chart

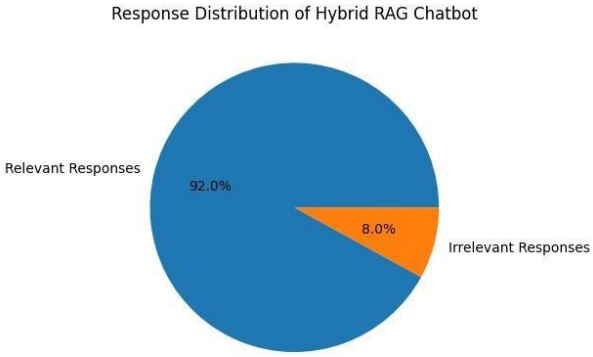


Table 2: Master Result Analysis Table

System Type	Response Relevance (%)	Improvement Over Rule-Based (%)
Rule-Based Chatbot	68%	-
Retrieval-Only System	81%	13%
Proposed Hybrid RAG Chatbot	92%	24%

In the table above, the rule-based chatbot has the response relevance score of 68%, which was considered the base model. Incorporation of the similarity search increased the performance by 13% to become 81%. On the other hand, the proposed hybrid RAG chatbot had the highest response relevance score of 92%, which was an improvement of 24% from the rule-based approach..

V. DISCUSSION

The purpose of this research is to enhance the relevance of responses in AI-driven symptom tracking. In the suggested Hybrid RAG chatbot, semantic embedding and similarity-based search are used to guarantee that the responses are relevant to medical records. The disadvantages associated with the rule-based system and the retrieval model alone have been minimized using this method. The logging process is an additional advantage that helps track the patient's symptom history. It can be noted from the findings that the use of semantic encoding and retrieval grounding leads to enhanced context comprehension and consistency.

VI. CONCLUSION

In the current study, the development of a Hybrid Retrieval-Augmented Generation Chatbot was conducted for intelligent symptom surveillance. This involved the use of semantic embeddings in combination with similarity searches through the use of FAISS. From the comparison analysis, there is evidence of improved response relevance in contrast to the rule-based and purely retrieval-based models. Additionally, the chatbot has a centralized logging feature to log user interactions. Overall, this shows that combining retrieval with semantic understanding makes conversations more reliable.

VII. FUTURE WORK

Despite the Hybrid RAG chatbot proving its efficacy by responding more relevantly and accurately, there is still room for improvement in enhancing this system. In

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

the future, efforts will concentrate on increasing system functionality, scalability, and clinical reliability.

1. The symptom disease data set would be improved by adding more information from credible medical resources.s
2. The system will have a classification of risk to help recognize cases that require an urgent response.
3. The chatbot will provide multilingual capability to ensure that patients can interact with the system regardless of their language.
4. It will have a cloud-based system architecture for greater ease of scalability and data storage.
5. Studies would be carried out by healthcare professionals on real-world applicability.

IX. REFERENCES

[1] V. Patil, A. Mehta, and S. Kulkarni, "RAGMed: A Retrieval-Augmented Medical AI Assistant," IEEE Journal of Biomedical and Health Informatics, vol. 29, no. 2, pp. 890–902, 2025.

[2] T. Pankow, L. Richter, and K. Weber, "Enhanced Symptom Checker Using Expert-Curated Clinical Data," Computers in Biology and Medicine, vol. 176, 2025.

[3] M. Kumar and S. Singh, "Conversational AI for Clinical Decision Support in Healthcare," Healthcare Informatics Research, vol. 31, no. 1, pp. 45–56, 2025.

[4] L. Wang et al., "Mitigating Hallucination in Large Language Models for Healthcare Applications," *IEEE Transactions on Artificial Intelligence*, 2025.

[5] R. Patel and K. Shah, "AI-Based Symptom Monitoring Systems for Early Disease Detection," *Biomedical Signal Processing and Control*, vol. 84, 2025.

[6] Bora and H. Cuayáhuitl, "Systematic Analysis of Retrieval-Augmented Generation-Based LLMs for Medical Chatbot Applications," Machine Learning and Knowledge Extraction, vol. 6, pp. 2355–2374, 2024.

[7] M. Bhatt, S. Iyer, and D. Rao, "A Scalable Diagnostic Benchmark for Health AI Systems," Nature Digital Medicine, vol. 7, no. 1, 2024.

[8] Y. Zhang and L. Wang, "Knowledge-Augmented Large Language Models for Medical Conversational Systems," IEEE Access, vol. 12, pp. 102345–102358, 2024.

[9] S. K. Sharma and R. Gupta, "Machine Learning-Based Disease Prediction Using Symptom Data," *Expert Systems with Applications*, vol. 228, 2024.

[10] H. Lee and J. Kim, "Context-Aware Conversational Systems Using Hybrid AI Architectures," *Applied Artificial Intelligence*, vol. 38, no. 2, 2024.

[11] R. Chen, Y. Zhang, and L. Wu, "Semantic Similarity Search in Medical Text Using Deep Embeddings," *Information Processing & Management*, vol. 61, no. 2, 2024.

[12] Z. Ji, N. Lee, R. Frieske, et al., "Survey of Hallucination in Natural Language Generation," ACM Computing Surveys, vol. 55, no. 12, 2023.

[13] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Advances in Neural Information Processing Systems (NeurIPS), 2023.

[14] Johnson, M. Douze, and H. Jégou, "Billion-Scale Similarity Search with FAISS," IEEE Transactions on Big Data, 2023.

[15] D. Gupta and N. Arora, "Vector-Based Information Retrieval Using Deep Learning for Healthcare Data," *Pattern Recognition Letters*, vol. 173, 2023.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)