

Digital Health Partner: AI for Customized Treatments with RAG and LLM

Dr. K. Narsimhulu¹; S J Musharraf Ali²

¹Associate Professor; ²M.Tech Student

¹Department of Computer Science Engineering,
Rajeev Gandhi Memorial College of Engineering & Technology, Andhra Pradesh, India

²Department of Computer Science Engineering,
Rajeev Gandhi Memorial College of Engineering & Technology, Andhra Pradesh, India

Publication Date: 2025/06/13

Abstract: The healthcare system has had a difficult time keeping up with the demands of this age. People are experiencing many medical problems as a result of the rapid population growth, which has made it difficult for the medical system to handle and treat. To solve the problem, we employ technology for the health care revolution, which enables us to receive prompt and precise assistance. In this project, we are developing an AI chatbot with the use of artificial intelligence. As a virtual assistant, an AI chatbot assists us in giving accurate information about the problems, prescribes if the problem is minor, and directs users to see a doctor if the situation is serious. We use natural language processing (NLP) technologies in this chatbot to make the conversation sound human. The conversations are often upgraded. Additionally, it has unique features like multilingual support and the ability to schedule doctor appointments for convenient times. Additionally, it looks up the closest dates for the doctor's appointment and assists users in taking their medications on time. To make our model run effectively, we have employed techniques like RAG, Streamlit, LLM, and FAISS in addition to NLP. To organize and analyze PDF files, we have also introduced data intake tools such as PyPDF2. According to the HIPAA Act, which states that maintaining the use of protected data is the most crucial function, the healthcare system must guarantee that data privacy and security are among the most crucial elements. AI technology can be used to develop a chatbot that will transform the healthcare system and enable us to efficiently receive the right therapy at the right moment.

Keywords: Natural Language Processing (NLP), RAG, Streamlit, LLM, FAISS, Virtual Assistant, AI, ML, Python, Data Ingestion, PyPDF2, Chatbot.

How to Cite: Dr. K. Narsimhulu; S J Musharraf Ali (2025) Digital Health Partner: AI for Customized Treatments with RAG and LLM. *International Journal of Innovative Science and Research Technology*, 10(6), 419-425.
<https://doi.org/10.38124/ijisrt/25jun282>

I. INTRODUCTION

Naturally, a lot of people are experiencing health problems as a result of rapid changes in the environment and abnormal behavior. These include the use of hazardous gases in various locations, the consumption of inappropriate and unbalanced meals, and many other factors. The healthcare department must handle all of the patients and provide appropriate treatment at the appropriate time, among other issues, as a result of the growing human concerns. An appropriate prescription is still another factor to take into account. When we look at our hospitals, we find that most of them are overcrowded, which makes it difficult to treat patients in emergencies since they have to wait to see a doctor, putting their lives in danger. There are moments when things get really bad. And occasionally, people ignore the disease's serious signs because they believe they are just a typical issue.

Therefore, using AI technology, we suggested a model as a chatbot to overcome all of these difficulties. Chatbots are useful because software intended to mimic human-like communication with consumers. Using the information in the dataset, this chatbot attempts to answer users' questions. To improve efficiency, the model also learns and refines its functionality based on user feedback.

Artificial intelligence (AI) chatbots may be employed for offering services such as medical guarantee, exercise suggestion, health care maintenance tips, drug reminders, solutions to simple questions regarding issues, and scheduling appointments at suitable times according to user requirements. Our chatbot's operation is enhanced with the support of NLP technology. For instance, it can converse in any language. By considering the symptoms, chatbots might be able to diagnose the illness immediately and give the user the best solution.

Through a variety of services, this digital health partner assists patients in preventing health problems before they arise. Our system makes use of approaches like RAG, FAISS, and LLM. By using these approaches, our system has evolved to the point where it will offer recommendations and safety measures to be followed by the symptoms that patients have provided. This system is convenient to use because it has been built in a digital version that can be accessed through the internet on any electronic device.

By adding new features like voice chat augmentation, it has a better chance of succeeding in the future. Because the current system does not allow the uninformed to write in symptoms, this expansion is helpful to everyone. The main purpose of this expansion is to help the people who can't unable to express their symptoms in terms of chat, but they can express in words, so we have a thought of introducing a voice chat to take the input from the users through the voice chat. This will reduce the patient's time to express their symptoms. On the other hand, it will also help the system to understand the patient's input and produce effective output.

II. LITERATURE SURVEY

➤ *Existing Models:*

- **Healthcare Chatbots Based on Rule-Based Systems:** The majority of early healthcare chatbots were rule-based, using decision trees and pre-written if-then logic. By associating keywords with particular answers, these systems could manage basic and commonly asked questions.
- **Machine Learning-Based Healthcare Chatbots:** Rule-based systems started to give way to machine learning (ML) models as AI advanced. Large datasets of patient inquiries or medical interactions were used to train these chatbots.
- **Retrieval-Based Chatbots for Medical Queries:** Retrieval-based solutions were developed to overcome the drawbacks of static machine learning chatbots. With this method, when a user asks a question, the chatbot finds the most pertinent material in a pre-established knowledge base.

➤ *Disadvantages of Existing Models*

Even with significant advancements, current healthcare chatbots continue to have several drawbacks:

Limited Knowledge Coverage: The quickly expanding field of medical knowledge cannot be included in static databases.

- **Lack of Context Awareness:** A lot of models are unable to preserve the context of a conversation when a user switches between turns.

- **Poor customisation:** Depending on the user's background or particular medical circumstances, responses frequently lack customisation.
- **Outdated Information:** Retraining, which is resource-intensive, is frequently necessary to update the knowledge base in classical machine learning models.

These drawbacks highlight the necessity of a more intelligent, flexible, and dynamic system architecture for healthcare support.

➤ *Proposed Model:*

We suggest a cutting-edge AI chatbot system built on Retrieval-Augmented Generation (RAG) in conjunction with a potent Large Language Model (LLM) to overcome the difficulties encountered by earlier systems. This architecture is intended to provide consumers with logical, individualized, and reliable health advice by retrieving the mostcurrent, pertinent medical data from an organized knowledge base.

The following are the main goals of the suggested approach:

- Easily incorporate validated medical data into chatbot dialogues.
- Improve the model's capacity to produce answers based on trustworthy sources.
- Preserve contextual awareness and conversational fluency.
- Use a web interface to offer lightweight, real-time interactivity.

➤ *RAG and LLM-Based Personalized Healthcare Chatbot:*

The system workflow is organized into the following steps:

- Data Collection and Knowledge Base Creation
- Text Embedding and Indexing
- Query Handling Using RAG Architecture
- Deployment through Streamlit

III. METHODOLOGY

In developing the system, we have used a dataset that includes a patient's symptoms and conditions related to those symptoms. We have gathered the dataset by studying various reports collected from distinct hospitals. Using this dataset, we have trained our system, which will generate the patient's condition based on the symptoms provided by the patient. In this system, we have also used RAG, which will produce the result not only based on the training of the model, but also will retrieve the result/answer through facts found in various reports/documents.

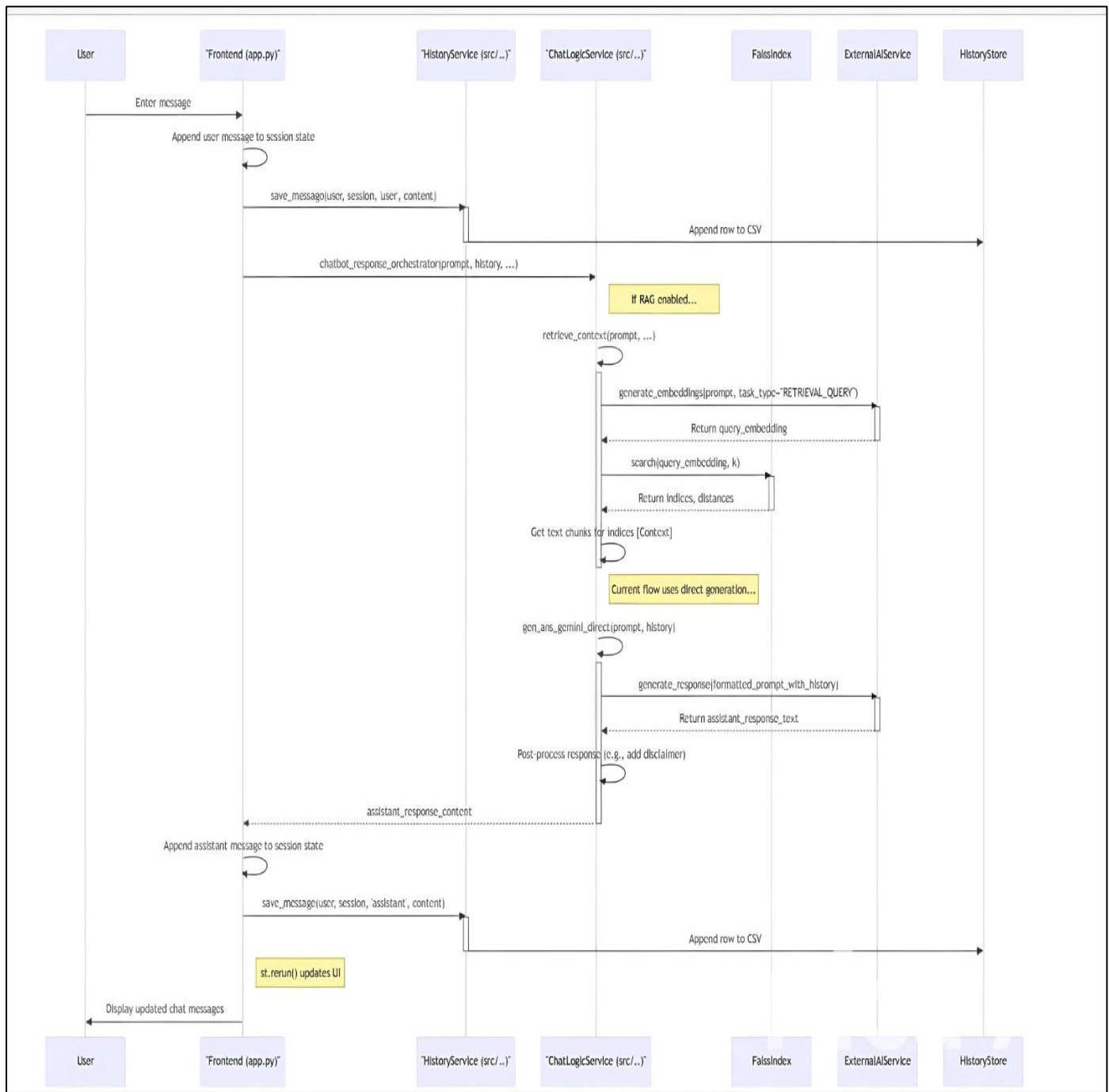


Fig 1 Architecture

➤ Storage of Data:

CSV files can be used to store user information and chat history while prototyping; however, there are a few disadvantages, including:

- Security: Passwords and other sensitive information stored in CSV files will be at risk. It is necessary to ensure the passwords.
- Adaptability: CSV files will work well for handling a small number of users, but they lose their effectiveness when dealing with large numbers of people. No SQL

database can have the upper hand in managing operations in such circumstances.

- Synchronicity: There is a risk of significant data loss when we attempt to submit repeated requests in CSV files. We can use databases to manage synchronization.

➤ RAG Implementation:

The model essentially employs a simple chat system that, by using previous exchanges, makes common conversations quicker and easier; nevertheless, it does not analyze the files to obtain better outcomes. The model can function more effectively and produce better results simply by switching to the `gen_ans_rag` method.

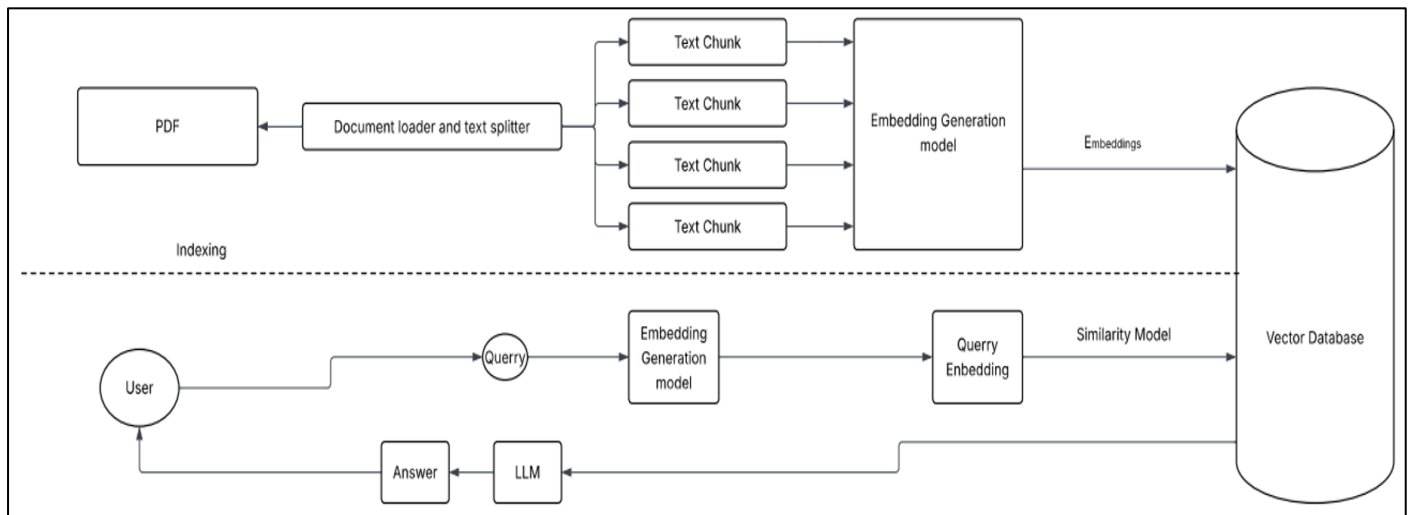


Fig 2 RAG Flow chart

➤ Fault Management:

The basic try and except block is used by the present system to identify faults and provide the user with an error message. We may strengthen it by using well-structured logs that show the important errors and give consumers the appropriate error details without revealing important details.

➤ Layered Design:

The paradigm for organizing and safeguarding the code seems simple due to the layered architecture, which separates the responsibilities into many modules.

➤ Streamlit:

Streamlit is the frontend technology that works better for designing applications, but when the public uses the application, it gets affected by high load. The backend management of resources like LLM and index building may make it difficult to control the system's speed when traffic is heavy.

➤ Efficient Index Storage:

The FAISS index is now cached using @st.cache_resource and is constructed into memory. Although this works well for small programs, it becomes less effective when dealing with large indexes or often restarted apps. We can serialize the index to disk and restore it at startup rather than recreating it each time we turn on. Because they provide capabilities like automated persisting, scalability, and robust filtering for more complex requirements, specialized vector databases are better suited for production situations.

➤ System Algorithm:

• Phase 1: Input

The model should receive the patient's symptoms as input.

• Phase 2: Text Preprocessing

The following techniques are used to preprocess the provided data:

- ✓ Tokenisation: The input data will be in the form of sentences, which are broken down into words.
- ✓ Lemmatization: The provided data is lemmatized, which eliminates prefixes and suffixes from the keywords and vectorizes (matches) the data. Duplicate words are then identified and eliminated. The resulting data will be distinct as a result.
- ✓ Text cleaning: removing extraneous information such as punctuation, special characters, and spaces.

• Phase 3: Vectorization

The preprocessed data is mapped onto the dataset that was used to train the model. The highest number of user-provided symptoms that match the symptoms of any ailment in the dataset is found via vectorization.

• Phase 4: Prediction

The patient's condition is predicted after the symptoms have been mapped to the dataset.

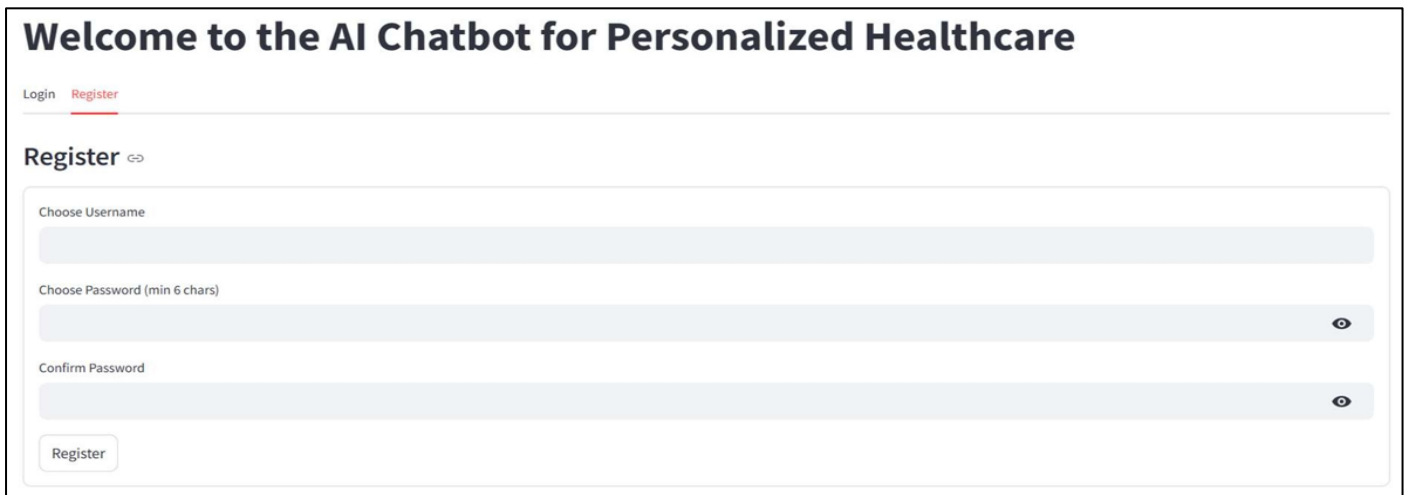
• Phase 5: Response Generation

The predicted condition of the patient is displayed as the output along with the precautions to be taken.

IV. PERFORMANCE EVALUATION

➤ Step 1: Register Page

When we run the code successfully, we can see the register page on the screen. We need to create the credentials to use in the future.

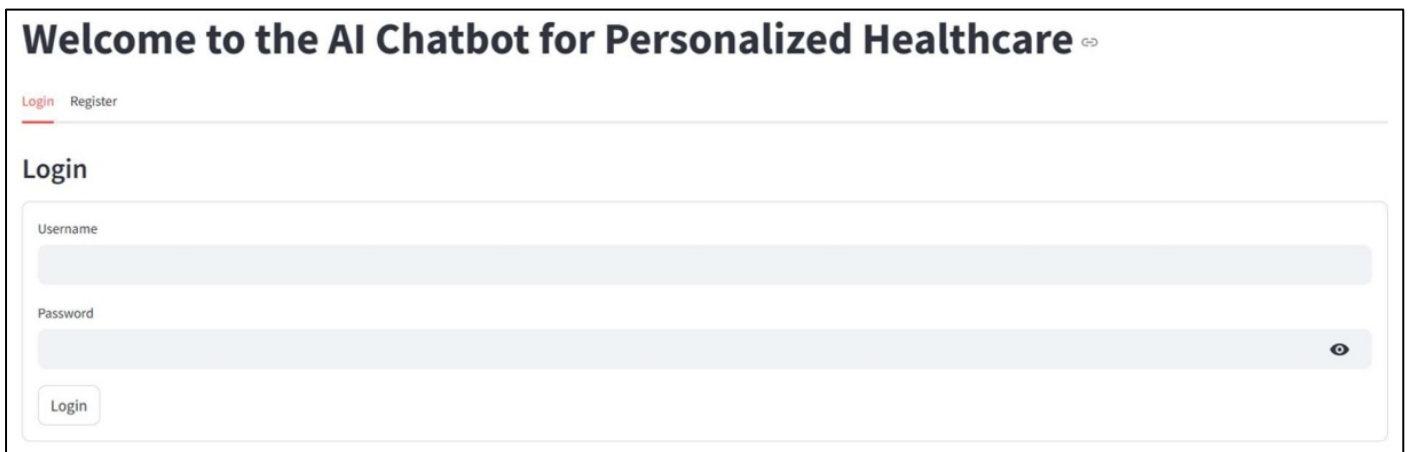


The screenshot shows the 'Register' page of the 'Welcome to the AI Chatbot for Personalized Healthcare' application. At the top, there are links for 'Login' and 'Register', with 'Register' being the active link. Below the header, the title 'Register' is followed by a small icon. The form contains three input fields: 'Choose Username', 'Choose Password (min 6 chars)', and 'Confirm Password'. Each field has a corresponding password visibility icon (an eye) on the right. At the bottom left of the form is a 'Register' button.

Fig 3 Register Page

➤ *Step 2: Login Page*

After the successful registration, the page jumps to the login page. Here, we have to provide the details we entered on the registration page.

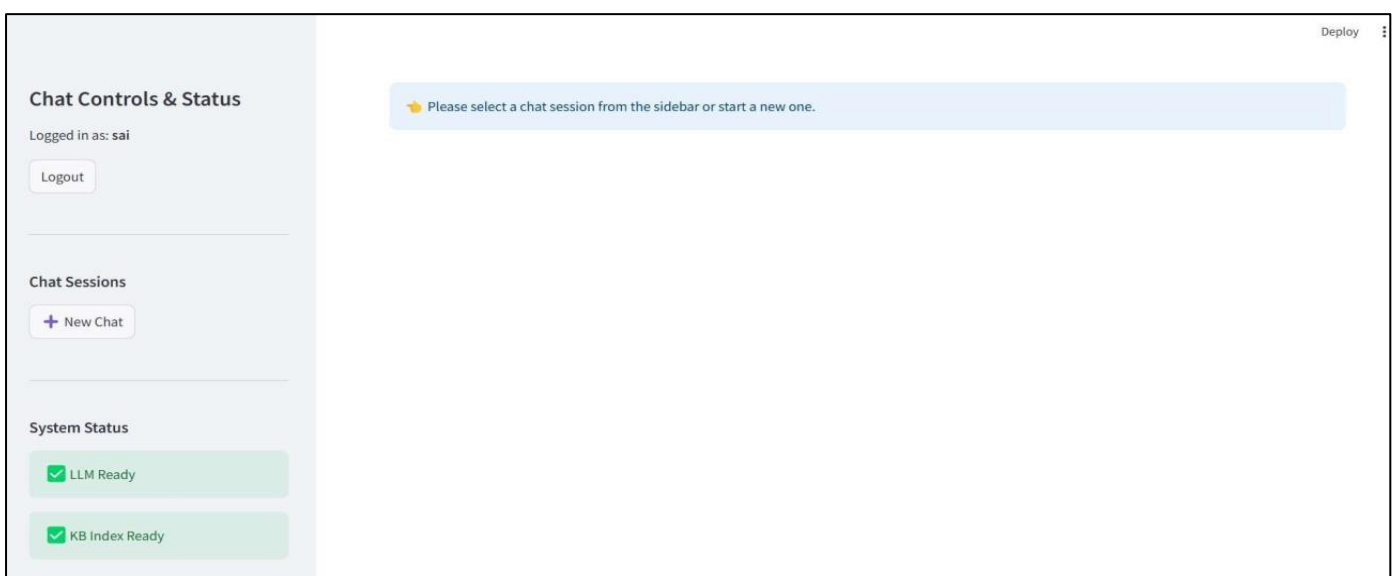


The screenshot shows the 'Login' page of the 'Welcome to the AI Chatbot for Personalized Healthcare' application. At the top, there are links for 'Login' and 'Register', with 'Login' being the active link. Below the header, the title 'Login' is followed by a small icon. The form contains two input fields: 'Username' and 'Password'. The 'Password' field has a password visibility icon (an eye) on the right. At the bottom left of the form is a 'Login' button.

Fig 4 Login Page

➤ *Step 3: Chatbot Interface*

After entering the credentials, the chatbot interface will be displayed, and it will ask the users to start a new chat.



The screenshot shows the 'Chatbot Interface' of the application. On the left side, there is a sidebar with three sections: 'Chat Controls & Status' (showing 'Logged in as: sai' and a 'Logout' button), 'Chat Sessions' (with a '+ New Chat' button), and 'System Status' (showing 'LLM Ready' and 'KB Index Ready' with green checkmarks). The main area on the right has a light blue header with a yellow star icon and the text 'Please select a chat session from the sidebar or start a new one.' In the top right corner of the main area is a 'Deploy' button.

Fig 5 Chatbot Interface

➤ *Step 4: Conversational Page*

When the user clicks on the new chat, the page will be directed to the conversational page where the conversation starts.

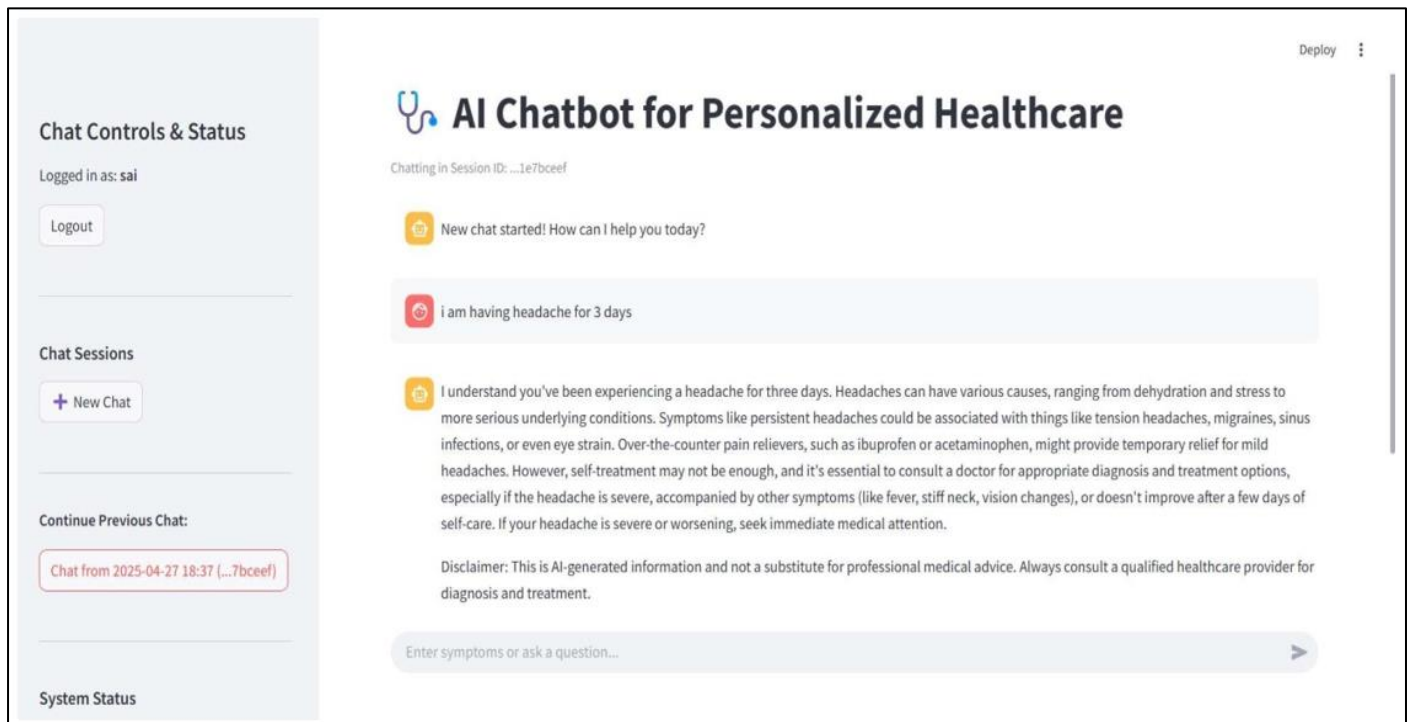


Fig 6 Conversational Page

➤ *Step 5: Final Disease Prediction/Output*

After the conversation with the user, the model predicts the disease by considering the symptoms.

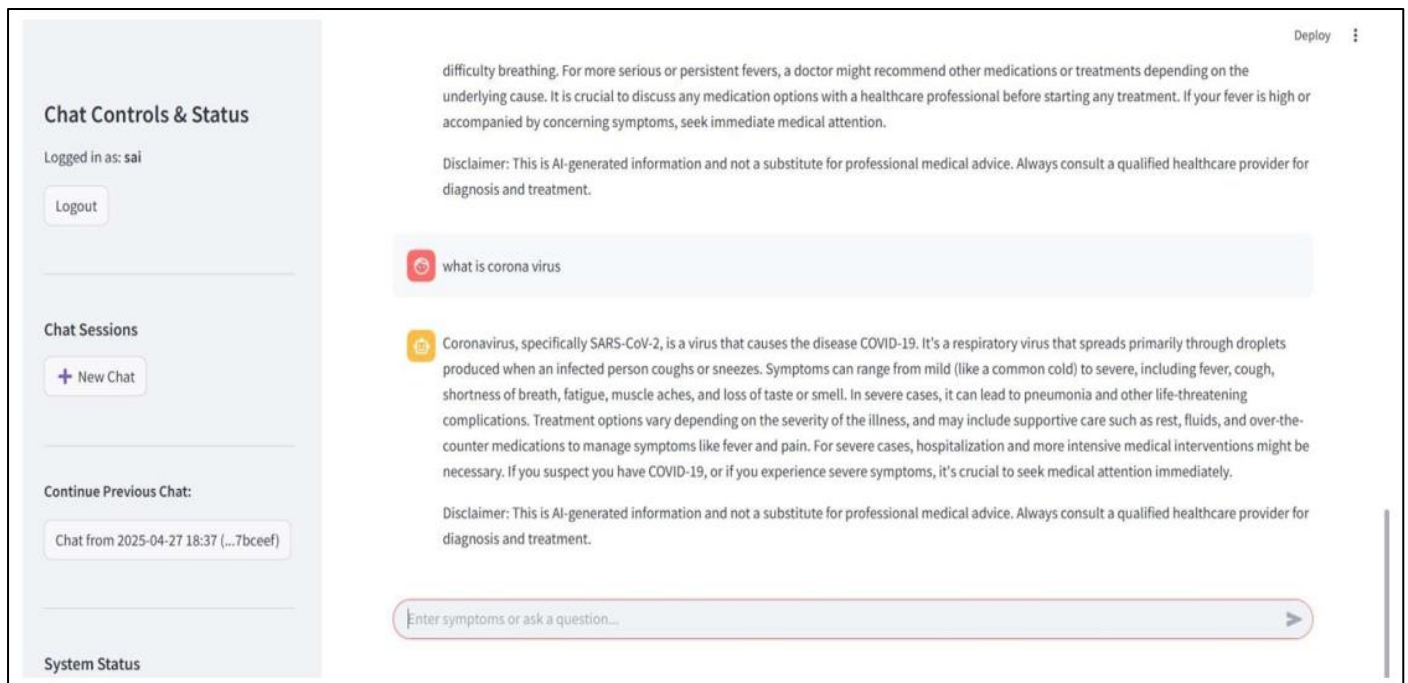


Fig 7 Output

V. CONCLUSION

The primary goal of the project is to create a model that will help users with common health-related questions, human-like discussion, what to do, and prescription

reminders. The chatbot created for this project is crucial to the healthcare system's future. To enhance system performance, methods such as Streamlit, RAG, LLM, and FAISS indexing are employed. This combination of technologies allows for efficient retrieval of information and

ensures that users receive accurate and timely responses to their inquiries. By integrating these advanced techniques, the chatbot can frequently improve its understanding of user needs and provide a more personalized experience.

REFERENCES

- [1]. Lekha Athota, Vinod Kumar Shukla, Nitin Pandey, Ajay Rana, "Chatbot for Healthcare System Using Artificial Intelligence," 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Amity University, Noida, India. June 4-5, 2020.
- [2]. Rohit Binu Mathew, Sandra Varghese, Sera Elsa Joy, Swanthana Susan Alex, "Chatbot for Disease Prediction and Treatment Recommendation using Machine Learning", Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019). IEEE.
- [3]. Srivastava, P., & Singh, N (2020, February). Automated medical chatbot (medibot). In 2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC) (pp.351-354). IEEE.
- [4]. Bharti, U., Bajaj, D., Batra, H., Lalit, S., Lalit, S., & Gangwani, A. (2020, June). Medbot: Conversational artificial intelligence-powered chatbot for delivering tele-health after COVID-19. In 2020 5th International Conference on Communication and Electronics Systems (ICES) (pp. 870-875). IEEE.
- [5]. Gentner, T., Neitzel, T., Schulze, J., & Buettner, R. (2020, July). A Systematic literature review of medical chatbot research from a behavior change perspective. In 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC) (pp. 735-740). IEEE.
- [6]. S. Divya, Indumathi, S. Ishwarya, M. Priyasankari, S. Kalpanadevi | A Self-Diagnosis Medical Chatbot Using Artificial Intelligence | Institute of Electrical and Electronics Engineers June 2019 Softić, A., Husić, J. B., Softić, A., & Baraković, S. (2021, March). Health Chatbot: Design, Implementation, Acceptance, and Usage Motivation. In 2021 20th International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-6). IEEE.
- [7]. Badlani, S., Aditya, T., Dave, M., & Chaudhari, S. (2021, May). Multilingual Healthcare Chatbot Using Machine Learning. In 2021 2nd International Conference for Emerging Technology (INCET) (pp. 1-6). IEEE.
- [8]. Madhu, D., Jain, C. N., Sebastain, E., Shaji, S., & Ajayakumar, A. (2017, March). A novel approach for medical assistance using a trained chatbot. In 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 243-246). IEEE.
- [9]. Sunny, A. D., Kulshreshtha, S., Singh, S., Srinabh, B. M., & Sarojadevi, H. (2018). Disease diagnosis system by exploring machine learning algorithms. *Int. J. Innov. Eng. Technol*, 10(2), 14-21.
- [10]. Kandpal, P., Jasnani, K., Raut, R., & Bhorge, S. (2020, July). Contextual Chatbot for healthcare purposes (using deep learning). In 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4) (pp. 625-634). IEEE.
- [11]. Hwang, T. H., Lee, J., Hyun, S. M., & Lee, K. (2020, October). Implementation of an interactive healthcare advisor model using a chatbot and visualization. In 2020
- [12]. International Conference on Information and Communication Technology Convergence (ICTC) (pp. 452-455). IEEE
- [13]. Avila, C. V. S., Franco, W., Venceslau, A. D., Rolim, T. V., & MP, V. (2021). MediBot: An Ontology-Based Chatbot to Retrieve Drug Information and Compare Its Prices.
- [14]. Mathew, R. B., Varghese, S., Joy, S. E., & Alex, S. S. (2019, April). Chatbot for disease prediction and treatment recommendation using machine learning. In 2019, 3rd International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 851- 856). IEEE.
- [15]. Rahman, M. M., Amin, R., Liton, M. N. K., & Hossain, N. (2019, December). Disha: An implementation of a machine learning based Bangla healthcare Chatbot. In 2019 22nd International Conference on Computer and Information Technology (ICCIT) (pp. 1-6). IEEE.
- [16]. Ayanouz, S., Abdelhakim, B. A., & Benhmed, M. (2020, March). A smart chatbot architecture based on NLP and machine learning for health care assistance. In Proceedings of the 3rd International Conference on Networking, Information Systems & Security (pp.1-6).
- [17]. Kandpal, P., Jasnani, K., Raut, R., & Bhorge, S. (2020, July). Contextual Chatbot for healthcare purposes (using deep learning). In 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4) (pp. 625-634). IEEE.
- [18]. Athota, L., Shukla, V. K., Pandey, N., & Rana, A. (2020, June). Chatbot for Healthcare System Using Artificial Intelligence. In 2020, 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO) (pp. 619-622). IEEE.