CHAPTER

9

SymptoGuide: Revolutionizing digital health through retrieval augmented generation and large language models

Barnali Paul, Avhishek Nandi, Deepsubhra Guha Roy and Piyali Datta
Department of Computer Science & Engineering (Artificial Intelligence & Machine Learning), Institute of Engineering & Management, Kolkata, West Bengal, India

9.1 Introduction

With the advent of large language models (LLMs) and their widespread use in providing human-like responses, it quickly made a buzz in all fields of Technology and was adopted by many. Thanks to billions of parameters that they have been trained on, they have been able to capture intricate patterns in language and perform a wide array of language-related tasks. One field where these LLMs have been encompassed is the field of healthcare. Even if the outputs that are produced are from artificial intelligence, they are being gradually acknowledged by the health sector primarily because they unveil the competencies to process enormous amounts of medical data, be it image data or reports, which is supplementary to the doctors in improved diagnostics and also streamlining administrative tasks in hospitals. The most important factors for acceptance are its competence, precision, and data reports, which enable clinical personnel to make more informed decisions in severe settings where speedy decisions are essential. The initial skeptic has shifted thanks to explainable AI. AI systems have become part of routine workflows because they are constantly improving in reliability and support. A new generation of smart chatbots is leading the way in healthcare digitization (Nandi et al., 2025). These chatbots are commonly used for scheduling appointments, answering frequently asked questions, and managing patients through straightforward, rule-based processes. Even though they possess seamless availability with fast response to patients' interactions, using chatbots in serious medical emergencies still faces challenges. To exemplify, they can produce false information, lack clarity on how they reach conclusions, and have a limited grasp of current medical facts. We need to tackle these issues to build trust in these systems for public healthcare (Lewis et al., 2020).

The shortcomings of general language models have been addressed here, and an improved model has been developed using retrieval-augmented generation (RAG). RAG retrieves relevant documents from a knowledge base (KB) using semantic search and embedding, rather than relying only on the language model's memory. The RAG-based medical chatbot combines language models with a curated medical KB, including medical books and doctors' notes, to provide reliable, context-aware healthcare insights. Here, we start by evaluating the historical evolutions of healthcare chatbots, the current limitations of mainstream models, and the transformative role of LLMs (Huo et al., 2025). Moreover, the discussion

LLMs in Practice: Real World Applications, Challenges and Success Stories. DOI: https://doi.org/10.1016/B978-0-443-44344-2.00020-0

continues with how it is possible to implement them, and at last, how and why RAG holds the key to tackling the limitations.

9.1.1 Evolution of chatbots in healthcare

Healthcare chatbots have been introduced to us more than half a century ago, mirroring the close advancement and evolution of computer linguistics and AI. Table 9.1 portrays the timelines of the computer linguistic era and its key features.

9.1.2 Challenges in existing medical chatbots

The current systems implemented in chatbots, despite their widespread usage and buzz in the community, are prone to many errors and essentially imperfect for critical medical advice.

9.1.2.1 Clinical inaccuracy and hallucinations

LLMs are prone to giving misleading or incorrect information, known as hallucinations. LLMs are trained to generate text that is coherent and contextual rather than factually correct (Maity & Saikia, 2025). Researchers believe that hallucinations stem from the fundamental mathematical and logical structure of LLMs' architecture; therefore, impossible to eliminate them through architectural improvements, dataset enhancements, or fact-checking mechanisms. This leads to the dilemma of using a "medical chatbot" as it is unable to demonstrate resilience in high-stakes situations, where even small misjudgments can have life-threatening consequences. Survey shows that over 27% medical chatbot outputs contain factual errors, and many have failed to comply with the clinical guidelines (Busch et al., 2025).

9.1.2.2 Shallow contextual understanding

Many chatbots that use lightweight LLMs with a smaller number of parameters are unable to keep track of the context of previous questions; they cannot track multiturn dialogue, leading to fragmented or varied information in scenarios like disease follow-ups or symptom progression. This limits the chatbot to hold real-world conversations where continuity and reasoning of the conversation are required (Wang et al., 2019).

9.1.2.3 Privacy, security, and compliance

Healthcare information of a user, be it medical records or disease history, is protected under the Health Insurance Portability and Accountability Act (HIPAA) and General Data Protection Regulation (GDPR) compliance. The data cannot be stored or used by the organizations until the GDPR terms of using the data of the consumer are granted by the user. The medical chatbot has to have great security measures and Protected Health Information (PHI) protections implemented in it to first verify the identity of the user, and next, ensure compliance with these laws. There is no such robust infrastructure to pose major barriers to private data leaks (Aguzzi et al., 2025; Austin et al., 2021).

9.1.2.4 Lack of integration and explainability

The responses delivered by the LLMs are synthetic and are a result of whatever logical reasoning it is built upon to come to these responses; these are not transparent to the user, the chain-of-thought of the LLMs, which undermines the user's trust in using these systems, and medical professionals.

Table 9.1 Timelines of the computer linguistic era and their key features.

Era	Period	Key systems/tools	Technology base	Key features	Limitations
Rule-based origins	1960s–1980s	ELIZA, INTERNIST-1, and MYCIN	Pattern-matching and rule-based logic	Structured expert systems, explicit rules, and early human-computer interaction	Rigid logic, domain-limited, lacked conversation flow
Digital therapeutics and scripted tools	1990s–2000s	COGNIsoft and Beating the Blues	Scripted dialogue and CBT frameworks	Therapist-written scripts for digital CBT interventions	No learning capability, static conversations
NLP and machine learning	2010s	Ada, Babylon Health, and HealthTap	Early NLP, ML models (decision trees, and probabilistic models)	Symptom analysis, triage, intent recognition, and adaptive dialogue	Limited understanding, accuracy dependent on training data
Transformer era and LLMs	2017–present	BERT, GPT, and Woebot	Transformers and pretrained LLMs	Contextual conversation, scalable, fine-tuned clinical reasoning, and emotional tone recognition	Hallucinations, outdated knowledge, and risk in critical diagnoses
RAG opportunity	Emerging frontier	RAG-based systems	Retrieval-Augmented Generation (RAG)	Up-to-date document retrieval + response generation; more grounded and factual	Still under development, depends on the quality and freshness of the source material

Therefore, the need is clear for explainable AI, which shows exactly the reason behind the answer, more of a think-it-out-loud process. They also lack a connection to the electronic health records (EHRs) for reasons related to inadequate data security (Bhavani Peri et al., 2024; Bora & Cuayáhuitl, 2024; Chow & Li, 2025).

9.1.3 Role of large language models in transforming healthcare

To address the aforementioned challenges, we introduce the SymptoGuide platform in this chapter. The unification of LLMs in the chatbot architecture has catalyzed significant progress on the content and response of the chatbot in this medical domain (Kaddour et al., 2023).

9.1.3.1 Understanding complex queries

Sometimes users may not be in a position to explain the actual symptoms they are experiencing or may just want to continue a discussion they may have been having previously; with high fluency and understanding of natural language, due to the vastness of data LLMs have been trained on, LLMs can easily decipher and paraphrase complex medical queries and isolate the exact symptoms.

9.1.3.2 Personalized interaction

Memorizing the past terms of the end-user not only can train its KB to produce better answers, but it also yields user-specific responses. This leads to better compliance and acceptance among the users (Agüera Y Arcas, 2022).

9.1.3.3 Multilingual support

Among the vast datasets upon which the training of the LLMs is done, it consists of diverse documents and content of various languages all across the globe. Essentially, this means the LLMs can respond to the user's native language without any problem and even in a specific dialect (Labadze et al., 2023).

9.1.3.4 Rapid prototyping

LLMs can be fine-tuned for a specific use case quickly by focusing on that specific data. This would lead to a more holistic and better understanding of the domain by the LLMs. For example, if we fine-tune a transformer model specifically on medical data, it will be able to provide way better and more intricate insights rather than a normal LLM trained on generic data. The limitations like lack of groundedness, training cutoff, and propensity for hallucination, emphasize the need for the RAG framework (Sreeram & Sai, 2023).

9.1.4 Motivation for retrieval-augmented generation-based chatbots

RAG combines the best of two worlds,

1. *Retrieval*: In this, we have a well-structured KB that contains documents of relevant use cases and is also able to fetch real-time, accurate, and curated medical content from reliable sources.
2. *Generation*: The LLMs can synthesize personalized conversational and context-aware responses based on the prompts given to it and also the retrieved KB at its disposal to refer to for answers.

RAG is essential for medical AI for the following reasons:

- *Accuracy boost*: Linking the generated output to the verified factual authoritative medical sources reduces hallucination and improves trust in the systems.
- *Explainability*: Generated responses with citations from the sources aid in interpretability, and custom prompts to show the thinking of the model explain how the LLMs are processing the query from the user.
- *Domain adaptability*: The seamless update of the KB in the ever-changing field instantly makes the LLMs equipped with new data without the need for retraining the language model.
- *Safety and compliance*: Supporting the HIPAA-compliant workflow via regulating what kind of documents the chatbot can access to craft its responses for the user. Maintaining the privacy of medical records by not storing the sensitive information of the patients unless allowed.

In this chapter, we propose a RAG-based medical chatbot architecture that leverages large language model meta AI (LLaMA) for generation, semantic document chunking, and vector-based search using Facebook AI Similarity Search (FAISS) and custom-made prompt templates for the purpose of accurate symptom disease prediction. The architecture has been proven effective through a comparative case study against traditional search methods along with real-world test scenarios, and what kind of implementation issues we tackled (Ayers et al., 2023).

SymptoGuide is unique from other medical/health chatbots because its framework is RAG, not just a pretrained memory. Every response is rooted in a curated medical KB compiled from high-quality articles, such as the New England Journal of Medicine and the Lancet, and principal guidelines, such as the World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC). This decreases hallucinations that general bots can often have. It does not pseudo-triage, following scripted rule-based triage, when inferring abstracted, grouped symptoms that could not just be synonyms or acronyms, and providing a differential diagnosis with subsequent clear next steps after assessment.

9.2 Background and related work

9.2.1 Previous approaches in symptom analysis and disease prediction

Analyzing the problems and evaluating the solutions at hand, it can be concluded that the systems used for examining and evaluating the patients in the period spanning between the first chatbots and the modern-day ones have suffered many changes and evolved with the use of newer technologies. The default method of chatbots is the rule-based systems. These systems are resolved using a certain logic and are represented using decision trees or if-then statements to set the boundaries of the diagnosis. These systems provide a strong foundation for efficiency and accuracy for a routine question-and-answer session. Though as many benefits as these systems provide, they have their share of downfalls. They lack the ability to answer vague and unconventional descriptions of problems and need to be updated manually every time clinical guidelines change.

These limitations can be addressed by using various systems that employ machine learning methods like support vector machines (SVM), decision trees, and k-nearest neighbors (KNN). These classification algorithms are trained on labeled static datasets. They identify patterns from symptom descriptions to make predictions. Often, they work alongside natural language processing (NLP) methods. NLP helps convert unstructured user input into structured features suitable for classification. Some of these combined methods even merge decision trees with KNN models to provide multilayer reasoning. While this shift has added more flexibility and intelligence to chatbot systems, their performance continues to

rely greatly on the quality and volume of available medical databases. Every time there was some new information available, it had to be retrained (He et al., 2023).

9.2.2 Comparative analysis of rule-based versus retrieval-augmented generation-based chatbots

When comparing rule-based systems with data-driven machine learning models, several key differences become apparent. Rule-based systems operate through predefined decision trees or fixed logical rules, which makes them relatively straightforward to manage, verify, and update. However, they often struggle with unusual or complex symptom patterns and require a lot of manual effort to include new medical knowledge. Their responses tend to be uniform, offering little chance for personalization since they do not consider individual user history or context. In contrast, machine learning models can adjust better and show greater intelligence, which lets them respond more effectively to changes and new information. SVM, decision trees, and KNN are algorithms that can identify complex patterns in large amounts of medical data. These models enable more precise predictions. When combined with natural language processing, they can interpret textual symptom descriptions, making them highly valuable in medical contexts. They are capable of handling numerous user queries and adapting to new information, offering responses tailored to specific symptom combinations. However, verifying the clarity and consistency of these predictions remains challenging, as the underlying reasoning can often appear ambiguous. Regular updates, dataset integration, and retraining are therefore essential to maintain high levels of accuracy. In healthcare, this makes close monitoring and external validation essential while introducing a chatbot.

9.3 Proposed strategy for a retrieval-augmented generation-based medical chatbot

9.3.1 Architectural overview of the retrieval-augmented generation framework

SymptoGuide chatbot does not guess the answer; it follows a proper pipeline or steps to generate a proper and correct response based on the user query. This section outlines the system's architecture and workflow, which is also depicted in Fig. 9.1.

***Step 1*:** The user's problem

It starts with the user entering a question or symptoms they are experiencing in simple English. Questions like "My throat hurts when I swallow" or "High blood pressure related to kidney issues?" SymptoGuide, just like your actual doctor during a visit, obtains this valuable information.

***Step 2*:** Question breaking down (tokenization and embedding)

SymptoGuide has the query; it will be able to comprehend it in the following manner. First, it divides the sentence into chunks of tokens and converts them to numbers with an embedding model. This process makes human language something a machine can meaningfully process.

***Step 3*:** Searching for reliable information

SymptoGuide searches with the pertinent query in its widespread, verified KB. With a fast search engine known as FAISS, it goes through a carefully investigated library of medical papers and books and identifies the most significant ones that match the request.

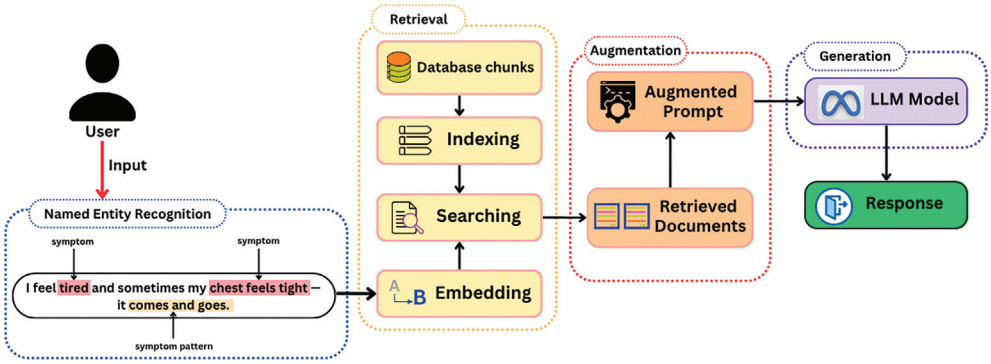

FIGURE 9.1

Working of SymtoGuide.

***Step 4*:** Building the context

The chatbot gathers top-K search results and formats them into a structured "prompt." This is hidden from the user, but what SymptoGuide uses to make sense of the query in the right context, so as to be able to give a better response. It's like giving the chatbot a cheat sheet before it answers.

***Step 5*:** Crafting the response

Lastly, this context is passed to a big language model—LLaMA—trained to answer in a coherent, medically aware, and helpful manner. It patiently weighs both your question as well as the context it has drawn from the documents, and creates a response that's conversational, accurate, and actionable.

9.3.2 Why we use the large language model meta-AI model for answers

When we build our chatbot, one of the big decisions we need to make is which large language model we choose. We choose LLaMA, a transformer-based deep learning model. LLaMa trained on huge multilingual datasets. It uses a self-attention mechanism to focus on the important parts of the input. This language model can handle ambiguous languages easily. It doesn't merely enumerate symptoms or vomit up definitions. It synthesizes, explains, and recommends in a manner that sounds natural, useful, and above all, nonintimidating. Imagine it like this: LLaMA is such a junior doctor who has read all the important medical textbooks, followed a few departments around, and now comes and sits down to chat with you. But whereas a usual doctor with not much time to waste might do so, it's always at your disposal, always waiting, and always willing to speak plainly (Yu et al., 2023).

In SymptoGuide, we further train LLaMA with medical conversations, case histories, and doctor-patient encounters.

9.3.3 Creating and curating the knowledge base–Relevance, comprehensiveness, currency, perspicuity

A language model can't give accurate answers to questions until it has access to the proper information. So, we delivered SymptoGuide with a thoroughly developed medical KB whenever it needs to respond to

a medical question (Antaki et al., 2023). We need only reliable, proven, and current sources. Therefore, we constructed our KB from peer-reviewed research articles, particularly from premier. They offer knowledge from medicine's frontiers—clinical trials, new therapies, and new risks. Reports issued by the National Institutes of Health, WHO, CDC, and other national health agencies are used as gold standards in terms of diagnosis and treatment. Data on new diseases (such as COVID-19 or monkey pox), vaccine developments, drug recalls, and treatment breakthroughs are added continuously. After collecting this data, we processed it properly to use for the chatbot. Cleaned: we eliminate distracting sections such as publisher advertisements or footers. We break up content into workable "chunks" (such as an asthma treatment section or a list of diagnostic criteria). Each piece of the knowledge chunk is tagged with metadata like disease name, type (symptom, treatment, guideline), and publication date. The system is validated for consistency and cross-checked against multiple sources whenever possible. This makes sure that what SymptoGuide brings back isn't only relevant to your question but also high-quality, specific, and up to date. In order to ensure our KB remains clinically sound, we check each entry against four guiding principles:

1. *Relevance*: Each item of information relates back to actual patient care—symptoms, diagnosis, treatment, and prevention. If it doesn't assist in answering a patient's question or inform a decision, it shouldn't be there.
2. *Comprehensiveness*: From the everyday cold to unusual autoimmune disorders, our KB covers a broad spectrum. This enables SymptoGuide to assist with day-to-day issues and less frequent medical enigmas.
3. *Currency*: Medicine changes rapidly. New therapies and new dangers are discovered frequently with continuous intensive research. That's the reason we update our KB on a regular time period, so SymptoGuide's information remains current, and the actual clinical situation is tackled easily.
4. *Clarity*: The material content should be clear and straightforward without unnecessary words. We prioritize keeping the information that is clearly explained and unambiguous, making it easier for our model. We have summaries rewritten if necessary to ensure they're appropriate for use in patient discussions.

9.3.4 Document chunking: Fixed-length versus semantic chunking

SymptoGuide's database contains thousands of pages of data from medical books and research papers. LLMs cannot consume all of it together, and it would be a really costly process to consume the whole database and give answers every time a query is hit. This is where document chunking helps, where we divide each document into smaller pieces or "chunks." This facilitates efficient storage, rapid search, and high-quality retrieval. This makes the search space concise and responses accurate. Two approaches for chunking the implemented one are the Fixed-Length chunking, which divides the documents into chunks of fixed length defined in the backend. This process is quick but can cut important or relevant text portions, causing loss of context. The second one is semantic chunking, where we use NLP techniques to find the natural breaks, like paragraph boundaries and heading changes. This automatically makes sure the created chunks are contextually relevant. Using these two types of chunking, we can optimize the process of chunking (Lee et al., 2020).

9.3.5 Retrieval mechanism: How SymptoGuide gets the correct info

SymptoGuide gives the factually correct answer from its enormous KB, divided into chunks, through the following process,

- *Translating questions into vector*: When the user types a question like "What are the symptoms for typhoid?" The system translates that sentence into an embedded vector, which is a set of numbers that captures the meaning of your question.
- *Semantic search with FAISS*: FAISS is used to find similar chunks by comparing the question/query vector with every document vector present and returning the closest matches.
- *Cosine similarity scoring*: The K closest matches are ranked based on how well those matches with the query, implementing cosine similarity, a mathematical method that measures vector similarity.
- *Feeding to LLaMA for answer generation*: The best-ranked information chunks are sent to the LLM model, which reads them and generates a contextually relevant and specific response. In a sense, SymptoGuide behaves like a medical intern who refers to well-trusted notes before responding in milliseconds.

9.3.6 Response generation and contextual reasoning

After the pieces of the question document are entered into SymptoGuide, it begins to generate a response. The real magic occurs when it takes these pieces and turns them into useful, actionable insights for the user. SymptoGuide uses transformers and attention mechanisms that allow it to focus on what is most important in the input. Whether it's a specific symptom, a new treatment guideline, or an unusual patient history, the model weighs what matters most when producing its output. It provides safe and smart advice. The main goal is to be medically accurate. This means it follows clinical standards and relies on trustworthy sources. It avoids using complex medical terms, making it easy for users to understand. It not only provides information, it also delivers beneficial advice, like recommending whom to contact or what type of care should be taken. SymptoGuide remembers the earlier conversations such that it can understand the follow-up questions and confirm that the conversation gives you the correct information (Adamopoulou & Moussiades, 2020).

9.3.7 Customized prompt templates for disease diagnosis

We applied prompt engineering here to reduce the hallucination and response quality. Every query contains keywords and symptoms that are extracted from the user's query and the patient's history.

The prompt contains the instructions that the model should follow with the query and the extra information fed to it. For example:

"*From what is presented here in the way of history and symptoms, make a probable diagnosis and what to do next in straightforward easy to understand language*." These tailored questions are like a GPS which guide the model away from vague or off-sentence answers and maintain on-topic, informative, and precise conversation.

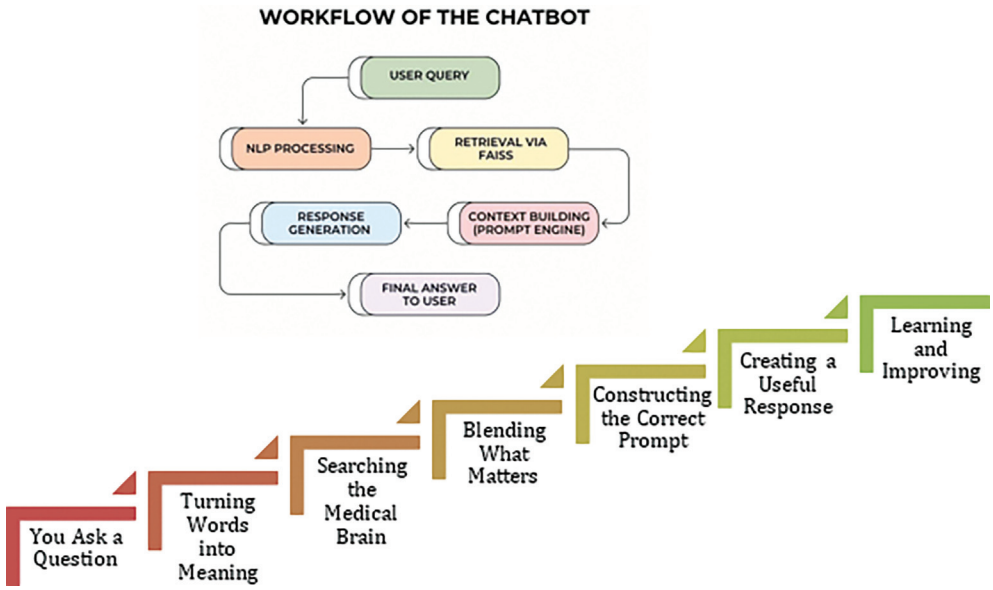

FIGURE 9.2

Flowchart of SymptoGuide.

9.4 Workflow of the chatbot: How SymptoGuide works—From your question to a reliable answer

In this chapter, we go through step by step how SymptoGuide works behind the scenes, as described in Fig. 9.2. Here, we explain how we implemented the whole technology in one place and created the SymptoGuide chatbot. When a user types a question, every part is designed for delivering clear, accurate output. We'll also show how SymptoGuide is different from a typical web search, and how it handles vague or unclear symptom descriptions.

9.4.1 Step-by-step query-to-response pipeline

SymptoGuide has a clever and systematic seven-step process in order to learn your question, identify the correct information, and provide a useful response.

Step 1:*You ask a question*

When a user types something such as, "I've fatigue and chest pain." No formatting or special keywords required—just write naturally.

Step 2:*Turning words into meaning*

SymptoGuide converts your query to a numerical form (a vector) with a model such as Bidirectional Encoder Representations from Transformers (BERT). This assists the system in understanding not only the words, but also what they mean. For instance, it is aware that "tired" and "fatigued" are equivalent.

Step 3:*Searching the medical brain*

Your question vector is matched with a vast library of medical material—research articles, guidelines, and textbooks—by a lightning-fast mechanism called FAISS. SymptoGuide extracts the most salient pieces of information.

Step 4:*Blending what matters*

SymptoGuide now has the best relevant information, which it converts into its own and blends with the given question. Now, this is ready for the language model to be comprehended in full context.

Step 5:*Constructing the correct prompt*

A prompt is now constructed with the key information from the given query (e.g., symptoms, duration, severity, etc.).

Step 6:*Creating a useful response*

The LLaMA model now interprets the prompt, considers the given symptoms, and delivers an answer.

Step 7:*Learning and improving*

After providing the output, SymptoGuide now logs feedback. This helps the system learn over time to give more personalized responses and improve accuracy in the future.

9.4.2 Case study: SymptoGuide versus web search for symptom analysis

While feeling excessively tired with a sudden pressure in the chest, the user enters the symptom into a search engine, he/she can be bombarded immediately with entries like blogs, advertisements, and forums with frightening headlines. These may provide the user with various possibilities, such as stress or a heart attack, to name a few. You're more confused—and more frightened—than when you started. SymptoGuide does better. It doesn't simply search; it knows the common language. You don't have to know fancy medical jargon. Just give input simply and clearly, and SymptoGuide will understand. It identifies important symptoms from your message and looks only at trusted, expert sources such as clinical guidelines and medical literature. SymptoGuide gives clear and concise answers without attaching an enormous number of links. Just provides a straightforward response. Unlike a search engine, SymptoGuide does the difficult work of filtering, comprehending, and explaining—so you don't have to. It provides health answers you can rely upon, comprehend, and use.

9.4.3 Exact match versus multiple symptom matching: How SymptoGuide deals with both simple and complicated health questions

Not all health issues have an easy solution. Sometimes symptoms directly indicate the illness, but many times symptoms are vague and indicate many different things. SymptoGuide can handle both scenarios.

1. *When the cause is clear*: If you say something like you have a burning sensation when you urinate. That's a pretty precise symptom, and SymptoGuide is likely to immediately identify it as a urinary tract infection. In these instances, SymptoGuide doesn't simply label the condition. It tells you what it is, how to recognize the signs, what treatment may be useful, and what kind of doctor to see—such as a general physician or a urologist. It sets you on a clear course of action rather than leaving you in the dark.

2. *When it is more complex*: Suppose now you tell them you've been tired recently, or that you frequently get headaches. These are generic symptoms that can appear in so many different illnesses, ranging from as benign as dehydration or sleep deprivation to more complex conditions like anemia, thyroid disorders, or even chronic fatigue syndrome. In this scenario, SymptoGuide doesn't assume. Rather, it provides you with a couple of potential explanations and explains how they contrast.

   It could say, for instance:

   "*Fatigue can result from low iron, sleep, or thyroid. If it's been occurring for some time, a simple blood test should be able to refine it.*"

It can also softly recommend what type of doctor or tests might assist you in having a better understanding.

9.4.4 Comparative evaluation with existing solutions

Old-fashioned tools such as rule-based chatbots or basic symptom testers tend to come across as robotic. They might demand precise keywords, provide generic responses, or not learn to respond to how you word your query. Web searches tend to lead you down a rabbit hole of contradictory or unverified information. SymptoGuide is designed to come across as more human and more useful. It knows the context and remembers previous components of the conversation. Accepts natural language—just speak the way you would under any circumstances. Cites actual sources when offering medical guidance. Don't freak out when your question isn't black-and-white. In a world of scattered, untrustworthy health information, SymptoGuide provides a down-to-earth, compassionate, and smart guide to navigate your worries.

9.5 Case studies and user scenarios

We now look at the impact on patient engagement. To test the strength and flexibility of the model in real-life situations, we created extensive simulated user scenarios and test cases. These case studies mimic common questions that a patient might ask during a telehealth or clinical visit. They will help us show and measure the system's ability to handle unclear conversation flows, identify diseases with overlapping symptoms, and then provide clear and understandable steps for the user to take after the conversation. Additionally, we are examining how well the chatbot manages specific cases and how closely it matches a real doctor's diagnosis (Huo et al., 2025). The following discussions illustrate how the chatbot operates in practice.

9.5.1 Scenario 1: Ambiguous symptoms—Chest discomfort

User Input: "*I feel tired sometimes also my chest feels tight. It comes and goes.*"

- *Symptom extraction:* Using Named Entity Recognition and semantic parsing, the chatbot successfully identifies the distinct symptoms. They are - Fatigue, Chest tightness, and an intermittent pattern of occurrence

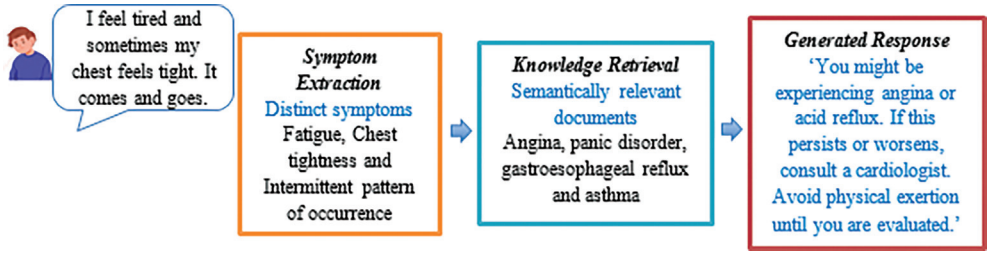

FIGURE 9.3

SymptoGuide example test case for scenario 1.

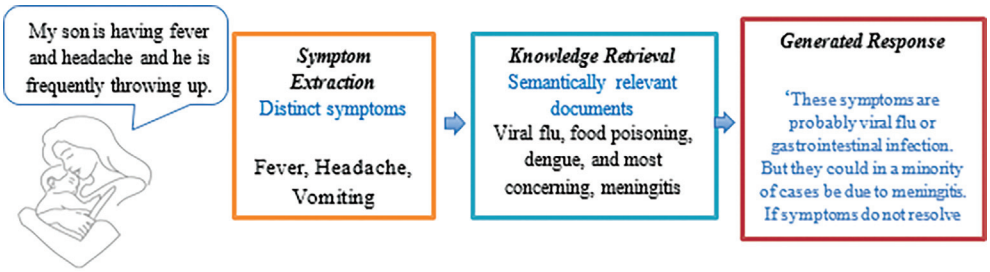

FIGURE 9.4

SymptoGuide example test case for scenario 2.

- *Knowledge retrieval:* The RAG pipeline from the KB, which was indexed, finds semantically relevant documents. The retrieved documents include medical literature associated with conditions like angina, panic disorder, gastro-esophageal reflux, and asthma.
- *Generated response:* Upon consolidating these documents, the language model determines that immediate attention is not required, as the user input showed no such indication. Therefore, it generates a precautionary guided response "You might be experiencing angina or acid reflux. If this persists or worsens, consult a cardiologist. Avoid physical exertion until you are evaluated." The differential diagnosis mindset presents common yet clinically relevant possibilities and asks to seek consultation at the earliest (Fig. 9.3).

9.5.2 Scenario 2: Overlapping symptoms in multiple diseases

User Input: "My son is having fever and headache and he is frequently throwing up."

- *System behavior:* The chatbot identifies the different symptoms. It searches for related documents based on context from the KB. The retrieved information includes various diagnoses such as viral flu, food poisoning, dengue, and, most concerning, meningitis (as depicted in Fig. 9.4). It concludes that the main symptoms are fever, headache, and vomiting. This combination may indicate several children's illnesses. Based on this information, the system provides a balanced and informative response: "These symptoms are probably viral flu or a gastrointestinal infection. However, in a few

cases, they could be due to meningitis. If symptoms do not improve or if there is neck stiffness, seek emergency care." This response shows how the model can prioritize common diagnoses while still considering serious ones. Its concise format highlights urgent symptoms and offers clear, actionable steps for further care.

The RAG model ensured that the response covered a range of possibilities so that no potential conditions would be overlooked. The LLM's response followed custom prompts designed to be cautious without causing alarm. Recognizing the intermittent nature of the symptoms, the model did not recommend over the counter medications but instead advised seeking medical consultation after outlining the potential issues.

9.5.3 Practical insights from realistic test cases

To determine the performance of SymptoGuide in more general conditions, we curated a list of 50 anonymized user inquiries that duplicated the symptoms, such as cough, dizziness, rashes, backache, and childhood fever. Each question runs through the system in order to measure its staging with a range of medical specialties. The metrics concentrated on three main measures: at par with the opinion of the medical professional, the accuracy of the top three diseases identified, and response time. SymptoGuide responses show alignment with expert suggestions in 76 % of the cases, while attaining the top three diagnosis accuracy of 84%, indicating its capacity to retrieve clinically applicable conditions consistently. Additionally, the system had an average response time of less than 2 seconds, including both retrieval and generation phases. The model performed really well in a department that contains a vast literature covered in the KB, such as general medicine, pediatrics, and internal medicine. These findings highlight SymptoGuide's promise as a trustworthy clinical decision support system, with some areas to be improved in the future, notably KB coverage and dealing with imprecise or underspecified symptom inputs.

9.6 Implementation details

9.6.1 Tools, frameworks, and libraries

SymptoGuide is developed using tools and algorithms that are popular and used in all the big projects of the industry in the field of machine learning and NLP. It also uses cloud computing to support scalability. The backend is made using Python and FastAPI, which facilitates the use of a RESTful API, helping in routing and asynchronous processing of user queries. For the brain, SymptoGuide uses the LLaMA model, which gives high performance in generating responses to instructions given as custom prompts. The medical document of the KB and the user input are processed using BioBERT. It is used as an embedding model, which converts the text into vectors that can be later used for semantic similarity search. We use BioBERT as it is already trained on biomedical text and optimized for healthcare-related tasks. FAISS search engine is a high-speed similarity search engine that is implemented during the time of identifying and getting the contextually relevant chunks into the LLMs. It uses libraries like SpaCy and LangChain to help in preprocessing the big document chunks into logical chunks. Information about the user is stored in PostgreSQL for scalability and ad-hoc queries. The vector representations are stored in Pinecone, which offers both semantic search and scalable retrieval options. The deployment

is managed by Docker images and Amazon Web Services, which allow features like auto scaling, GPU usage, etc. This makes the tech stack of SymptoGuide modular and maintainable, and deployment-ready in the healthcare sector.

9.6.2 Managing latency and cost optimization in real-time systems

The users generally expect a quick and precise response from the chatbot at the time of emergency. The response time of SymptoGuide is less than two seconds, even for complex and elaborate operations such as making personalized recommendations based on past sessions. It uses FastAPI to serve multiple user requests concurrently without lagging. It also implements caching by storing the answers to the most frequent questions asked by the user. SymptoGuide retrieval makes sure only the relevant part, which is what the model actually needs to answer the query, reaches the LLM. It prefetches answers and removes the unwanted and unnecessary materials upfront. This results in a system that is intuitive enough to engage with the user but runs at low cost, providing access to wider healthcare institutions.

9.6.3 Data privacy and ethical considerations

Healthcare data comes with certain obligations toward securing the private data of the user and maintaining the anonymity of the user. SymptoGuide aligns its methodology with compliance with the strict global healthcare data privacy of HIPAA in America and GDPR in Europe. SymptoGuide never knows the actual user of the query, and it never saves any personal records in the database. It doesn't recall your name, your query, or where you're from. All is done anonymously, and general usage trends alone are stored to aid in system enhancement—never your individual data. To secure your data while in transit, SymptoGuide employs secure, encrypted connections. The same kind of protection as used with online banking. In addition to that, all answers the system provides are logged with anonymous IDs so that developers can examine performance or correct things, never knowing who asked what.

Notably, SymptoGuide is up front about what it is and isn't. It's not a physician, and it doesn't claim to be. Each answer is accompanied by a clear notice that the information is for informational purposes only and not a substitute for actual medical advice. In fact, one of its biggest strengths is knowing when to back away—if someone complains of chest pain or difficulty breathing, the system is designed to immediately tell people to get emergency treatment.

Wherever possible, SymptoGuide also indicates where it gets its answers from—quoting sources or documents—so that users know it's not fabricating but instead drawing on actual, vetted medical information. Overall, SymptoGuide is meant to function as a reliable guide, making sense of users' symptoms without invading their privacy while also promoting professional care when most needed.

9.7 Results and evaluation

To gain a real understanding of how well SymptoGuide performs, we must look beyond the simple fact of how it was built. In this chapter, we take a close-up look at just how accurate, consistent, and useful the system is in use. Here we test the ability of the model on how accurately it can suggest diagnosis and follow clinical guidelines, and by how much it can outperform other alternative options like search engines or normal language models.

9.7.1 Disease prediction and recommendation accuracy

SymptoGuide is tested on an amalgamation of real-world and synthetic datasets, which include more than 100 queries regarding health issues and general queries that we expect to see the user to ask the chatbot. It contained both simple and complex vague symptom queries to test the chatbot. With around 76% accuracy, SymptoGuide precisely identifies the underlying disease with the highest recommendation based on the symptoms provided. With the broaden scope of the top three matched diseases, the accuracy jumped to 84%. In comparison to clinical recommendations, its options matched those of an actual physician 81% of the time. The system had a low false positive rate of 6.5% and a false negative rate of 9.2%, where it could not correctly identify the potential disease. These statistics showcase that our chatbot is based on strong, solid, verified medical knowledge, which grounds the response which makes the real difference. In contrast, the language models normally hallucinate and provide irrelevant answers, which creates unnecessary panic in the mind of the user.

9.7.2 User experience and trust

We asked 40 users to try out the system with various kinds of queries covering different kinds of health situations. The users gave positive feedback on the responses of SymptoGuide. The average grade was 4.2 out of 5 based on the user feedback, with comments on how easy and intuitive it was to use the chatbot as a whole. The survey also showed us that the user relied on the response of the chatbot and loved the fact that when it was unsure, it gave out a list of 3 prone conditions and overall advice on what to do next. Users also liked the fact that the response's tone was calming, so that the user doesn't go into a state of anxiety or panic. They also noted that the chatbot's transparency, like citing its information sources and showing step-by-step reasoning, helped build confidence in its answers.

9.7.3 Comparison with traditional search and other artificial intelligence models

When people are sick and look for help online, they use search engines or generic AI assistants. These aids, however, were not designed for medicine—and it is evident.

SymptoGuide is distinct in the following aspects:

It employs verifiable medical information instead of broad internet material, greatly improving the reliability of its suggestions. It has a low hallucination rate of about 4% to 6%, while typical language models go astray in 15% to 30% of cases. It deals with a broad scope of symptom types and question styles— "My stomach hurts after eating" to "Should I worry about this rash?"—and it knows how to answer. Perhaps most importantly, SymptoGuide tells the user what to do next, whether that's consulting a specialist or monitoring for some warning signs.

9.7.4 Quantitative metrics: Precision, recall, and response time

SymptoGuide performance metric showcased a precision of 0.81, which means the suggested diagnosis as per the query was indeed correct. Its recall was 0.84, suggesting its high rate of all possible correct diagnoses. Its overall F1-score was 0.825, which reflected high overall accuracy. The response time

took nearly 2 seconds, which was ok for a real-time application. Every answer was short and readable—usually between 55 and 110 words long—and found a good balance between detail and readability. The retrieval system operated using three to five chunks of information from the KB per answer, maintaining the context compact and manageable. Cumulatively, these measures reveal that SymptoGuide is not just accurate and reliable but also effective, hence suitable for deployment in real-life use in clinics, community health applications, or personal health tools.

9.8 Discussion

9.8.1 Strengths and limitations of the proposed approach

SymptoGuide focuses on safety, responsiveness, and user trust. Its strength lies in providing verified medical literature from the KB, which led to the grounding of the response given by the LLM (Kim et al., 2023) and reducing the hallucination of the model. This retrieval of medical documents ensures consistent clinical advice and is able to interpret everyday layman's terms to understand the symptoms clearly, just like a doctor.

There are certain limitations of SymptoGuide, such as the accuracy of the model solely relying on the comprehensive information available in the KB. Also if there are gaps in the treatment, they can degrade the response quality (Misischia et al., 2022). It is not multilingual yet, leading to a limited reach among people. When the symptoms of certain diseases overlap, it gives broader possibilities that require assessment and supervision from a medical professional.

9.8.2 Scalability in clinical practice

SymptoGuide needs to incorporate the rigid data protection laws of healthcare data, therefore integrate with EHR systems for the security of the data. Also to obtain regulatory approval from the FDA (Neha et al., 2025; Paneru et al., 2024). It also needs to adapt and adjust the responses according to the regional disease and how medical professionals practice there. There is another need for proper traceability of the document that is being fed to the LLM from the KB, so as to remain accountable.

9.8.3 Minimizing risks of misdiagnosis and hallucinations

SymptoGuide relies on trusted sources and, with the always updated KB, provides a much comprehensive outline of the symptoms rather than assumptions or misinformation. Unclear or severe symptoms are generally referred to the professionals with disclaimers. It can work alongside doctors to provide aid in high-intensity situations, and it keeps on improving using user feedback.

9.9 Future directions

SymptoGuide demonstrates a great promise here that every individual will receive healthcare information through AI. If we can connect it with any useful device, it has a high potential in the future. This

chapter is all about how we can improve SymptoGuide in the future, smarter, more personal, and more helpful for patients and physicians alike.

9.9.1 Integration with electronic health records

Right now, SymptoGuide gives helpful, accurate answers based on the symptoms users type in. But imagine if it also knew your medical history—your past illnesses, medications, allergies, and lab results. That's what could happen if SymptoGuide is connected to a hospital's EHRs. By integrating with EHR systems, it can personalize its advice, taking into account your health background. Physicians can receive assistance in recording cases, since SymptoGuide might summarize notes or complete medical forms. It speeds up time by filling patient records automatically before or after a consultation. Naturally, this would need rigorous security, consent of the patient, and adaptation to healthcare requirements such as Health Level 7 (HL7) and Fast Healthcare Interoperability Resources (FHIR)—technological frameworks that allow medical software to speak with one another securely (Adamopoulou and Moussiades, 2020).

9.9.2 Real-time telemedicine applications

Telemedicine is increasingly a standard method for receiving medical assistance from home. Yet physicians continue to spend time taking history from patients and condensing notes, which means time taken away from listening and counseling. SymptoGuide can be a virtual telemedicine assistant by taking symptoms ahead of the appointment and providing physicians with a clean summary. Providing live support on the call, for example, by proposing potential diagnoses or red flags for what the patient is reporting. Following up with patient messages, such as reminders or health education hints. This would offload the burden on physicians and make patients feel more supported—even beyond the appointment.

9.9.3 Multimodal extensions (voice and wearable sensor data)

Most individuals prefer to speak rather than type. And with health trackers and smart watches becoming the norm, there's an opportunity to make SymptoGuide even more interactive. In future updates, SymptoGuide could speak and listen, employing speech recognition to process voice questions such as, "Why am I dizzy after my run?" Employ wearable data, like heart rate or oxygen saturation, to provide better feedback in real time. Examine pictures, such as verifying a picture of a rash on the skin or even scanning X-rays (with medical supervision), due to advanced computer vision algorithms. Each of these brings a new dimension to the discussion—making it feel more natural and helpful, like conversing with an actual, well-informed medical aide.

9.9.4 Continuous learning from new medical knowledge

Medical science continues to change. New diseases, treatments, and research articles emerge on a daily basis. If SymptoGuide does not continue to learn, it will become outdated. SymptoGuide has the feature of applying a net search and scan verified and acclaimed medical websites and journals for learning

new treatment guidelines published. This could lead to high-quality responses in the future with the updated KB.

9.10 Conclusion and discussions

SymptoGuide is an embodiment of the advancement in the field of healthcare using LLMs and RAG architecture. By merging the generative capability of the LLaMA model and an updated medical KB, SymptoGuide offers responses that are not only contextually appropriate but also easy to understand for nonmedical persons. RAG makes sure the drawbacks of the LLMs, like hallucinations, are evaded and provides reliability in the responses it provides. SymptoGuide factors a stable and flexible design centered on medical use, unlike general AI large language models or rule-based symptom checkers. It provides routes to trace back the response to its original medical document, which helps foster user trust in the model and supports ethical use of AI. Simulation with large datasets like those faced in real-world tests the pipeline used by SymptoGuide and showcases its capability as a powerful tool in the space of AI in healthcare. The scalability makes it possible for the model to be integrated with telemedicine platforms, public health advisory services, and outpatient handling systems. With improved retrieval scope, domain-specific accuracy, and model efficiency, it can become a vital digital resource for both patients and healthcare professionals.

References

Adamopoulou, E., & Moussiades, L. (2020). An overview of chatbot technology. *IFIP Advances in Information and Communication Technology, 584*, 373–383. https://doi.org/10.1007/978-3-030-49186-4_31.

Agüera Y Arcas, B. (2022). Do large language models understand us? *Daedalus, 151*(2), 183–197. https://doi.org/10.1162/daed_a_01909.

Aguzzi, G., Magnini, M., Farahmand, A., Ferretti, S., Pengo, M. F., & Montagna, S. (2025). RAG-enhanced open SLMs for hypertension management chatbots. *Journal of Medical Systems, 49*(1). https://doi.org/10.1007/s10916-025-02297-7.

Antaki, F., Touma, S., Milad, D., El-Khoury, J., & Duval, R. (2023). Evaluating the performance of ChatGPT in ophthalmology: An analysis of its successes and shortcomings. *medRxiv*. https://doi.org/10.1101/2023.01.22.23284882.

Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C., Terry, M., Le, Q., Sutton, C., Program synthesis with large language models. arXiv. (2021), Reterived from: https://arxiv.org.

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine, 183*(6), 589–596. https://doi.org/10.1001/jamainternmed.2023.1838.

Bhavani Peri, S. D., Santhanalakshmi, & Radha, R. (2024). Chatbot to chat with medical books using retrieval-augmented generation model. In *Proceedings of NKCon 2024 - 3rd edition of IEEE NKSS's flagship international conference: Digital transformation: Unleashing the power of information* https://doi.org/10.1109/NKCon62728.2024.10774900.

Bora, A., & Cuayáhuitl, H. (2024). Systematic analysis of retrieval-augmented generation-based LLMs for medical chatbot applications. *Machine Learning and Knowledge Extraction, 6*(4), 2355–2374. https://doi.org/10.3390/make6040116.

Busch, F., Hoffmann, L., Rueger, C., van Dijk, E. H. C., Kader, R., Ortiz-Prado, E., Makowski, M. R., Saba, L., Hadamitzky, M., Kather, J. N., Truhn, D., Cuocolo, R., Adams, L. C., & Bressem, K. K. (2025). Current applications and challenges in large language models for patient care: A systematic review. *Communications Medicine, 5*(1). https://doi.org/10.1038/s43856-024-00717-2.

Chow, J. C. L., & Li, K. (2025). Large language models in medical chatbots: Opportunities, challenges, and the need to address AI risks. *Information (Switzerland), 16*(7). https://doi.org/10.3390/info16070549.

He, K., Mao, R., Lin, Q., Ruan, Y., Lan, X., Feng, M., Cambria, E., A survey of large language models for healthcare: From data, technology, and applications to accountability and ethics. arXiv. (2023), https://doi.org/10.48550/arXiv.2310.05694.

Huo, B., Boyle, A., Marfo, N., Tangamornsuksan, W., Steen, J. P., McKechnie, T., Lee, Y., Mayol, J., Antoniou, S. A., Thirunavukarasu, A. J., Sanger, S., Ramji, K., & Guyatt, G. (2025). Large language models for chatbot health advice studies: A systematic review. *JAMA Network Open, 8*(2). https://doi.org/10.1001/jamanetworkopen.2024.57879.

Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., McHardy, R., Challenges and applications of large language models. arXiv. (2023), Reterived from: https://arxiv.org.

Kim, J. K., Chua, M., Rickard, M., & Lorenzo, A. (2023). ChatGPT and large language model (LLM) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine. *Journal of Pediatric Urology, 19*(5), 598–604. doi:10.1016/j.jpurol.2023.05.018.

Labadze, L., Grigolia, M., & Machaidze, L. (2023). Role of AI chatbots in education: Systematic literature review. *International Journal of Educational Technology in Higher Education, 20*(1). https://doi.org/10.1186/s41239-023-00426-1.

Lee, M. C., Chiang, S. Y., Yeh, S. C., & Wen, T. F. (2020). Study on emotion recognition and companion chatbot using deep neural network. *Multimedia Tools and Applications, 79*(27-28), 19629–19657. https://doi.org/10.1007/s11042-020-08841-6.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W. T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in neural information processing systems*. United Kingdom: Neural Information Processing Systems Foundation.

Maity, S., & Saikia, M. J. (2025). Large language models in healthcare and medical applications: A review. *Bioengineering, 12*(6), 631. https://doi.org/10.3390/bioengineering12060631.

Misischia, C. V., Poecze, F., & Strauss, C. (2022). Chatbots in customer service: Their relevance and impact on service quality. *Procedia Computer Science, 201*, 421–428. https://doi.org/10.1016/j.procs.2022.03.055.

Nandi, A., Paul, B., Datta, P., & Roy, D. G. (2025). MedMate: A contextual approach for disease diagnosis using retrieval-augmented generation. *Lecture Notes in Networks and Systems, 1380*, 429–441. https://doi.org/10.1007/978-981-96-5822-0_34.

Neha, F., Bhati, D., & Kumar Shukla, D. (2025). Retrieval-augmented generation (RAG) in healthcare: A comprehensive review. *AI, 6*(9). https://doi.org/10.3390/ai6090226.

Paneru, B., Thapa, B., & Paneru, B. (2024). Leveraging AI in ayurvedic agriculture: A RAG chatbot for comprehensive medicinal plant insights using hybrid deep learning approaches. *Telematics and Informatics Reports, 16*. https://doi.org/10.1016/j.teler.2024.100181.

Sreeram, A. A. S., & Sai, P. J. (2023). An effective query system using LLMs and LangChain. *International Journal of Engineering Research & Technology, 12*(6). https://doi.org/10.17577/IJERTV12IS060161.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*. United States: Neural Information Processing Systems Foundation.

Yu, P., Xu, H., Hu, X., & Deng, C. (2023). Leveraging generative AI and large language models: A comprehensive roadmap for healthcare integration. *Healthcare., 11*(20), 2776. https://doi.org/10.3390/healthcare11202776.