National College of Ireland

Medical Assistant Utilizing Large Language Models with Retrieval Augmented Generation and Vector Search

MSc Research Project
Master of Science in Artificial Intelligence

Divyasree Harikrishnan
Student ID: 23291508

School of Computing
National College of Ireland

Supervisor: Professor Lavish Thomas

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Divyasree Harikrishnan
Student ID:	23291508
Programme:	MSc in Artificial Intelligence **Year:** 2024-25
Module:	MSc (Research) Practicum
Supervisor:	Professor Lavish Thomas
Submission Due Date:	11-Aug-2025
Project Title:	Medical Assistant Utilizing Large Language Models with Retrieval Augmented Generation and Vector Search
Word Count:	5077 **Page Count** 16

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other authors written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Divyasree Harikrishnan

Date: 11-Aug-2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Medical Assistant Utilizing Large Language Models with Retrieval Augmented Generation and Vector Search

Divyasree Harikrishnan
Student ID: 23291508

Abstract

A trending strategy in the development of domain-specific assistants is to integrate the concepts of the Retrieval-Augmented Generation (RAG) with Large Language Models (LLMs) to build efficient high-quality assistants. The given research will develop a chatbot medical assistant offering credible data on health conditions, diseases, treatments, and preventative measures. The knowledge base is a pre-processed knowledge base based on "The Gale Encyclopedia of Medicine (Second Edition)." A custom set of domain-specific embeddings is produced with the PubMedBERT model and embedded representations are stored in the Qdrant cloud vector database, serving to allow rapid retrieval. The use of cosine similarity means that it uses the vector search which is provided to a RAG pipeline supported by two LLMs, LLaMA-3 70B through Groq API and Google MedGemma 4B through Hugging Face API. I compared the PubMedQA datasets to conduct an evaluation on their performance with metrics being ROUGE-L, BLEU scores, and response latency. The preliminary results show that the LLaMA-3 with the help of PubMedBERT embeds is considerably ahead of MedGemma, reaching the ROUGE-L score of 0.1741 and BLEU score of 0.0229 with lower latency. The chatbot was implemented on Streamlit framework which supports session-based memory to create continuity in conversation. The proposed research proves that the combination of LLMs, vector search, and RAG pipelines is effective to create medical-specific assistants.

Key Words: Medical Assistant, Large Language Models, Retrieval-Augmented Generation, Vector Search, PubMedBERT, LLaMA-3, MedGemma

1 Introduction

Artificial Intelligence (AI) and large language models (LLMs) (Ozmen and Mathur, 2025) have been promising in healthcare and especially its capability to improve patient education, clinical decision support, and dissemination of general health information. Although recent LLMs, such as GPT-4 and LLaMA, have shown impressive results in achieving an understanding of natural language and generating natural language, their uses in the sensitive medical field are questionable. Among the greatest risks of these are the possibilities of a hallucination, having no domain specificity, and failure by the models to provide valid sources when issuing medical advice. There is a significant risk involved to these disadvantages in the practical healthcare settings where accuracy and credibility is required.
Retrieval-Augmented Generation (RAG) (Benavent, Venerito and Michelena, 2025) has become a possible solution to overcome this inadequacy, through supplementing LLM-

produced results with information retrieved through a well-trusted and curated knowledge base. This hybrid system will take into consideration the factual anchoring by outside sources with the generative capability by LLMs. In addition to RAG, vector search implemented with embedding models will allow performing efficient semantic search, guaranteeing that the most relevant information is returned to the language model so that it could be used to generate responding. Collectively, RAG and vector search structures present a satisfying direction toward developing a formidable, sound, and situationally fit medical assistant.

The study shows a specific medical chatbots framework in the healthcare domain based on large language models and a RAG pipeline and a vector search system that specializes in healthcare information. This knowledge base is built upon The Gale Encyclopedia of Medicine (Second Edition) (Longe, Blanchfield and Gale Research Company, 2002), a medical text of knowledge and authority. The structure includes PubMedBERT embedding to produce high precision semantic representations of medical articles, which are stored and indexed in the Qdrant vector database (qdrant.tech, n.d.). In the case of the generative backbone, LLaMA 3 70B through Groq and Google MedGemma 4B IT through Hugging Face models will be tested according to their accuracy in giving medically correct and context-sensitive answers.

The significant impact of the presented research is that it created and tested a medical assistant that effectively answers health-related questions but avoids false information to the extent. A side contribution is a comparative study of two state-of-the-art encoder based LLMs in a RAG environment, which runs on the medical QA task (such as PubMedQA) (Github.io, 2019) and a benchmark base on ROUGE, BLEU, and latency. In Section 2 this paper describes related work that is followed by methodology, design specifications, implementation details, evaluation each in its respective Sections, 3, 4, 5 respectively and culminated by future work directions in Section 7.

2 Related Work

Large Language Models (LLMs) with Retrieval-Augmented Generation (RAG) and vector search are starting to become popular in the development of medical chatbots and medical assistants, which has opened a promising possibility of bringing the accessibility of healthcare and expertise of clinical decision support to everyone. Existing medical chatbots have utilized pre-trained language models for natural language understanding. It has shown great results on a BERT based medical chatbot (Babu and Sekhar Babu Boddu, 2024) which utilizes Bidirectional Encoder Representations from Transformers to aid healthcare communication and achieved 98% accuracy, 97% precision and 98% F1 score much better than its LSTM, SVM, BI LSTM counterparts. The ability of bidirectional contextual understanding is proven to achieve proper and advanced interpretation of medical terms. The approach, however, has high computational demand and the possibility that the training data may be biased, and may not be able to handle exceedingly uncommon medical cases due to lack of training data.

The merger of RAG with LLMs offers a promising way of dealing with the shortcomings of standalone language models, especially in healthcare applications. Integrating external validated medical literature helped improve response accuracy and eliminate hallucinations of

RAG models implemented for improving clinical decision support in plastic surgery (Ozmen and Mathur, 2025). RAG architectures (Benavent, Venerito and Michelena, 2025) applied to rheumatology applications demonstrated that RAG-augmented LLMs also achieve a substantially higher accuracy in medical tasks. The main contribution of RAG is the ability to base responses on trusted clinical evidence and supply a clear reasoning path with the aid of explicit citations. RAG (Kresevic et al., 2024) proved effective in clinical guideline interpretation in the context of managing chronic Hepatitis C Virus with GPT-4 Turbo, improving accuracy from 43 % baseline to 99 % due to the importance of data formatting and the reformatting of guidelines as a structured entity. Multimodal capabilities could not overcome the innate problems that LLMs have parsing non text sources such as medical tables.

Research into locally deployed solutions has been due to privacy and data security concerns. In (Matteo Magnini, Gianluca Aguzzi and Montagna, 2025), we see that we can preserve privacy through small language models (SLMs) with under 4 billion parameters for medical chatbots as they allow the desired architecture to be private in design. Evaluation demonstrated that SLMs have comparable semantic quality for generic medical questions, but there is a big performance gap in handling complex tasks such that intent recognition accuracy is 77% versus 91% for cloud-based models such as Gemini Pro 1.5. The study suggests that SLMs deployed locally can perform sensitive medical tasks but show that they need to be fine-tuned for better utility and need to be integrated with RAG.

Almanac (Zakka et al., 2024) a comprehensive RAG framework for clinical medicine heavily boosts the safety and reliability of LLMs in the healthcare field. On the ClinicalQA dataset the factuality was increased by 18%, 95% safety against adversarial prompts compared to ChatGPT. Contrary to ChatGPT, the framework answered all clinical calculation scenarios and had the right result. While physicians preferred ChatGPT's outputs 57% of the time. Furthermore, a systematic analysis (Bora and Cuayáhuitl, 2024) of the RAG-based LLMs for use cases of medical chatbots in resource limited environments along with an open-source model [Flan-T5- Large, LLaMa-2-7B, Mistral-7B] comparing study revealed that fine tuning on the RAG achieves the optimal performance with Mistral-7B + RAG + Fine tuning achieving 57% exact match accuracy of multiple-choice questions. Through the study, they concluded that RAG and fine tuning are both essential to better performance in medical question answering tasks.

There have been several (though limited) studies which investigate RAG use in medical domains. The task was addressed using a RAG framework (Hung et al., 2024) based on GPT-4 that performed recommendation of clinical trials in head and neck oncology with an F1 score of 0.77 and 100% recall and 63% precision, substantially outperforming the baseline non-RAG and GPT 4. However, moderate precision, variation from cancer type to cancer type and limitations from small sample size and single institution focus were exhibited by the system. Similarly, SentimentCareBot (Nayinzira and Mehdi Adda, 2024) which uses sentiment analysis and RAG for mental health support showed that sentiment analysis can help in making the document relevant, the Multiquery RAG approach with MistralAI had a considerable performance improvement. The RAG framework for surgical decision support, SurgeryLLM (Ong et al., 2024), incorporated external knowledge from evidence based surgical guidelines. The advantages in finding abnormal labs, finding missing investigations,

and in giving guideline-based management recommendations were well performed by the model in comparison to unmodified LLMs.

To conclude, the state of the art proves that the combination of Retrieval-Augmented Generation (RAG) and Large Language Model (LLM) like GPT-4, Mistral-7B, and LLaMA-2-7B has enhanced the performance and stability of medical chatbots to a large degree. There are, however, several limitations that still exist among them being the high computational requirements, biased training data, moderate precision in clinical recommendations and difficulties in dealing with structured medical information such as clinical tables and diagnostic reports. Studies also reveal that the use of locally deployed Small Language Models (SLMs) has privacy benefits with a decrease in performance when dealing with intricate medical queries as opposed to that of cloud-based models. Besides, although the current instances of knowledge bases heavily depend on utilizing textual resources, what is needed is more elaborate and structured forms of databases that would serve as a medical base to be able to have specific insights and distinction amid comparable diseases or forms of treatment.

To overcome these disadvantages, a domain-specific medical chatbot framework, utilizing RAG pipeline and vector search is proposed in this research. Knowledge base will be built on The Gale Encyclopedia of Medicine (Longe, Blanchfield and Gale Research Company, 2002) and hence credible information on medical information will be retrieved. PubMedBERT embeddings (Huggingface.co, 2025b) grant high-quality vector representation that stores in Qdrant database (qdrant.tech, n.d.). One of the comparative analyses of MedGemma 4B IT (Huggingface.co, 2025a) and LLaMA 3 70B via Groq (Groq.com, 2025) and Hugging Face is made to determine the best LLM backbone. The assessment is done on PubMedQA datasets (Github.io, 2019) regarding accuracy, the fact correctness, latency, and model size.

3 Research Methodology

The research methodology consists of a structured pipeline with five important phases of Data Gathering, Data Preparation, Knowledge Base Construction, Model Development and Retrieval, and Evaluation and Deployment. Such methodical process guaranteed the creation of a flexible and precise medical question answering chatbot. The working process in general is represented in Figure 1.

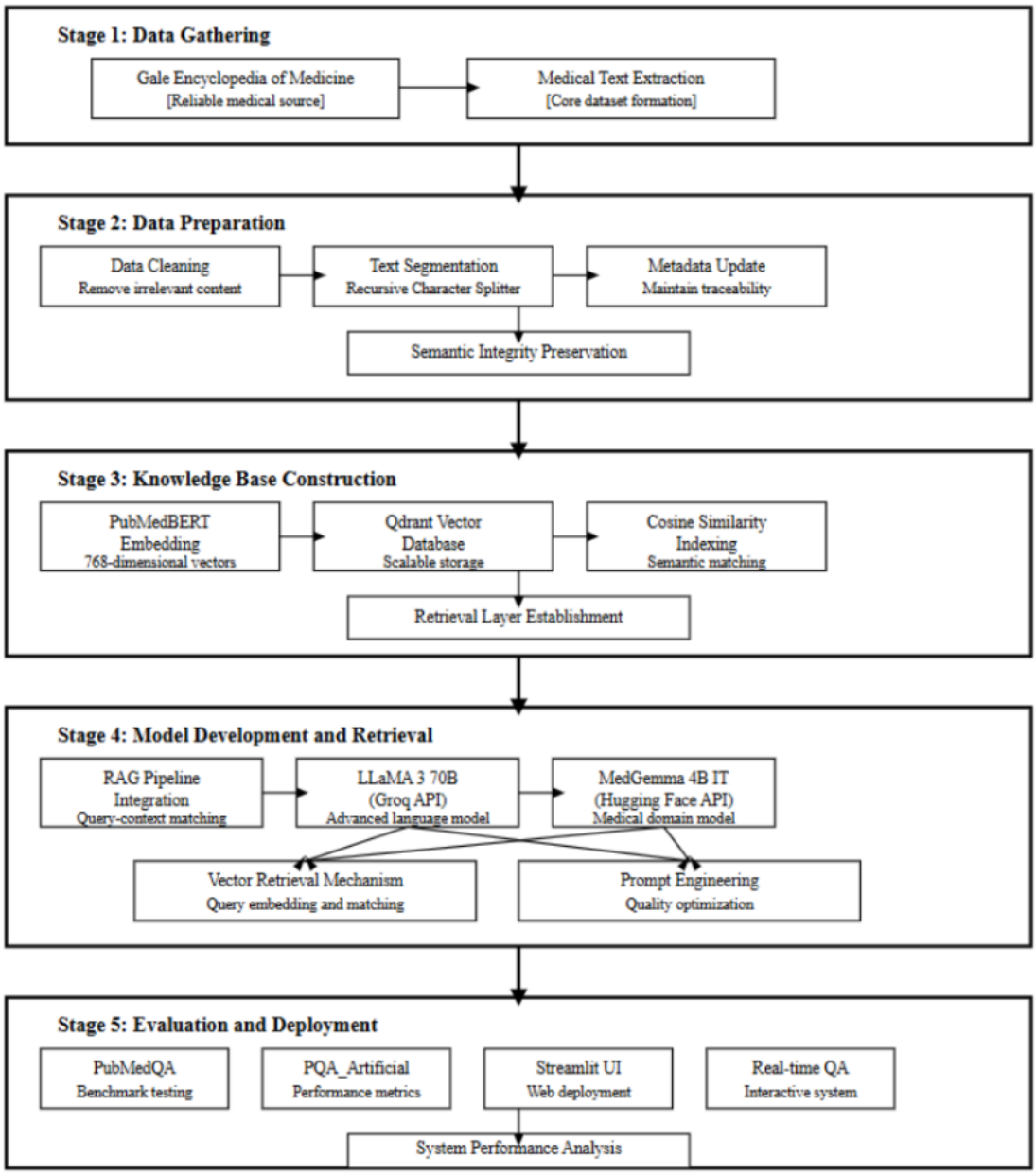

Figure 1: Research Methodology

The first step, Data Gathering entailed the acquisition of quality medical text in The Gale Encyclopedia of Medicine. The resource has been chosen due to its high medical relevance and a medically reviewed resource. The textual content that was extracted was placed as the core data set which formed the basis of knowledge in the retrieval system of the chatbot.

The second step, Data Preparation phase, Raw data was prepared where raw data was cleaned and arranged towards further processing. The noise included the mix of irrelevant information, headers, links of the navigation and advertisements was manually deleted. This made sure that unimportant and inapplicable information was eliminated. Metadata was refreshed to remain transparent and traceable. The cleaned text was split into smaller portions through Recursive Character Text Splitter to allow the transformer-based language models to

process them well. This has done two things, one is that it maintained the semantic integrity of the content, and the other is that the token lengths were kept at a reasonable range.

The third step, Knowledge Base Construction each segment of the processed texts converted to 768-dimensional vectors with PubMedBERT, biomedical language model that was proposed on PubMed literature. During this embedding, natural language was turned into numeric vector representations fit to be used with similarity-based retrieval. The resultant vectors were saved to the Qdrant vector database whereby scalable and fast searching of vectors is enabled. The vectors were compared with the help of cosine similarity, as this method proved to identify the semantically similar entries. In this step the retrieval layer was created, allowing the chatbot to find matching content to the user queries.

The fourth step, Model Development and Retrieval phase in which the vector retrieval mechanism was translated into a Retrieval-Augmented Generation (RAG) pipeline. As a user posts a query, this query is embedded into that stored in vectors and searched against to retrieve appropriate chunks of text. These bits of information are then fed to large language models as context to produce informed response. There were two pretrained models through the Groq API and Hugging Face API uses LLaMA 3 70B and MedGemma 4B IT. These models were chosen due to their superior language skills and their knowledge in the sphere of medicine accordingly. Instead of using extra training and model compression methods, they were reduced to the timely prompt engineering of the generation quality and the relevance of the context.

The fifth step, Evaluation and Deployment was in terms of gauging the system with the help of benchmark data utility sets like PubMedQA Dataset. Performance was measured in terms of accuracy, relevance of response and latency. Based on these evaluations, a few prompt and retrieval refinements were accomplished. The chatbot is deployed into a web-based UI using Python Streamlit which is a lightweight responsive web application. One would be able to connect to the system in real-time and post a medical question and get a proper context-sensitive response. This configuration allows it to easily scale as well as to integrate it in the future with external medical information systems.

4 Design Specification

The medical question-answering system is designed in a way in which a retrieval-augmented generation (RAG) structure is employed which combines a knowledge base constructed using biomedical texts with the state-of-art language models to provide real-time answers to the given queries. The system architecture mainly focuses on two fundamental parts illustrated in Figure 2. the retrieval-based knowledge pipeline and the language model interaction module followed by a final Evaluation and frontend deployment pipeline.

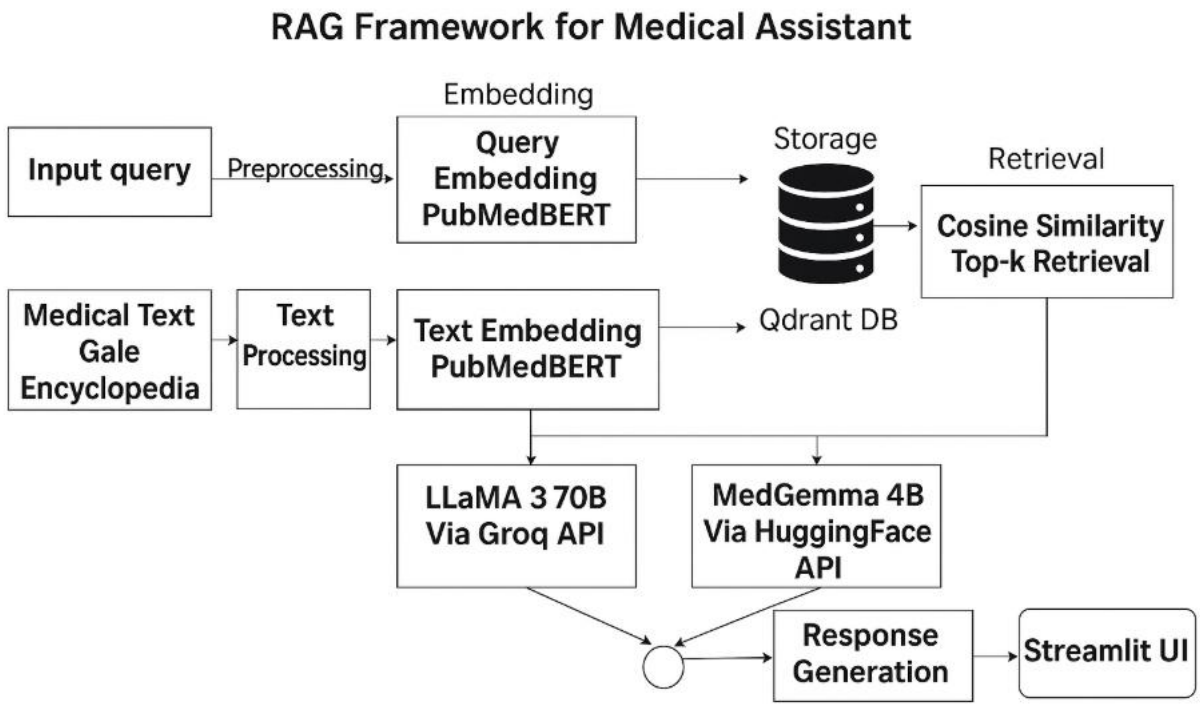

Figure 2: RAG Framework for Medical Assistant

4.1 Retrieval-Based Knowledge Pipeline

This element performs data ingestion, pre-processing, semantic embedding, and similarity-based retrieving. The structured text contained in the pipeline was written based on the "The Gale Encyclopedia of Medicine" which is a medically authoritative resource. The text data is extensively processed such as updating metadata, removal of irrelevant information and recursive segmentation to retain a meaningful text fragment as tokens.

Every part of the text chunk is processed to a 768-dimensional vector using a PubMedBERT based transformer pre-trained embedding model on biomedical literature. All these vector embeddings can be found in the Qdrant vector database that enables real-time and scalable searches of the vectors similarity. Referring to the text chunks corresponding to the semantic closeness of the user query is accomplished by cosine similarity. This mechanism of retrieving the information makes sure only the most appropriate context is conveyed through to the language generation module and increases preciseness and factual content.

4.2 Language Model Interaction Module

This module takes the queries made by the user and it embeds the queries and uses Qdrant to find the relevant documents. These documents are subsequently attached to the inquiry as a structure prompt. The prompt is transferred to pretrained language models that are called using Groq API, namely, LLaMA 3 70B and via Hugging Face API for MedGemma 4B IT.

The reasons why these models were selected are that they mix superior general understanding of language and domain-specific medical knowledge.

The models execute a context-sensitive generation where their retrieved medical chunks are used to respond to the queries made by the users. These models are not fine-tuned or quantized thus; they do not lose their initial performance parameters. Rather, there is prompt engineering of positioning the information retrieved, whenever contextual prompt is used, in the effort to persuade the models to head the right way toward the better, factual reply. Latency inference parameters are kept to a minimum through high-throughput potential of Groq LPU.

4.3 Evaluation and Frontend Integration

The effectiveness of the systems is established with the help of standard QA resources, such as PubMedQA. The metrics of evaluation are accuracy, relevance, and the response time. The generated texts are also evaluated based on ROUGE-L and BLEU Score which penalises heavily for the generated text. Such benchmarks ensure that chatbot provides high reliability and performance in different medical queries.

The system is run on a simplistic Python Streamlit interface. The frontend provides the user with the opportunity to type a query, receive answers and monitor conversation history in real-time. This UI should be lightweight and easy to use not only by ordinary users but also by medical professionals for a quick lookup of the medical information

5 Implementation

The medical question-answering system was designed as an application available over the web built with the help of the Streamlit framework to develop frontend and using Python to develop backend operations. Its execution is in modular architecture with the RAG pipeline combined with capabilities of a vector search and inference of a language model. Knowledge base construction phase implemented preprocessing standardization of the text information contained in the medical book, which is called The Gale Encyclopedia of Medicine (Second Edition). This medical data is pre-processed to remove unnecessary data and limiting only the required data for processing. Irrelevant metadata is also removed to provide more clarity. The cleaned data was then split into text chunks with a size of 500 characters and with 20 characters of an overlap to conserve semantic integrity and to have the optimal token length to fit token models. PubMedBERT model was used to create vector embeddings for the medical text using hugging face transformers library. The vectors of each text segment were generated using a model prepared with the due, which was trained on biomedical literature, namely, the model of NeuML/pubmedbert-base-embeddings. These embeddings were then saved in Qdrant cloud vector database with the use of cosine similarity as distance metric on semantic retrieval.

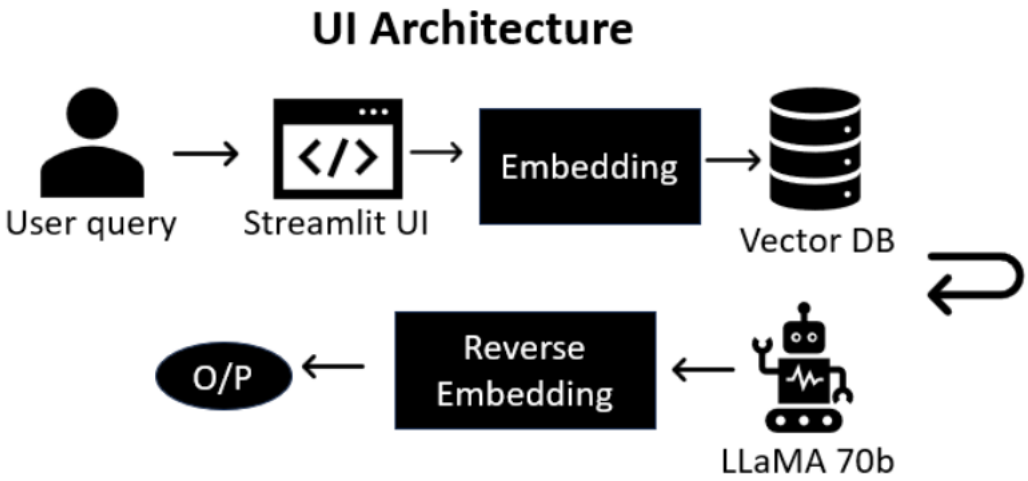

Figure 3: UI Implementation

The similarity search functionality available in Qdrant with a retrieval limit of 3 most relevant documents per query was used as the mechanism of the retrieval. The same embedding process is applied to user queries and knowledge base content and thus have similar representations in a vector database. To ensure that irrelevant material is excluded and the quality of the responses remained high, similarity threshold was put to 0.6. Two language frameworks were incorporated by using APIs. LLaMA-3 70B was accessed through Groq API with optimized inference settings to provide consistent and focused answer. This Google MedGemma 4B IT was trained with Hugging Face API via the transformers pipeline comparable parameter setups. The two models have effectively engineered prompts with retrieved contexts and instructions on how particular questions are to be answered with medical significance. From the two models, based on benchmarks evaluated best model is chosen to build a user interface. From the evaluation LLaMA was found to perform well and went on to build the UI. Figure 3. shows the high-level architectural diagram to implement the UI. To support the conversation context and conversation state within the conversation flow, streamlit application uses the session state management feature. Its interface has a primary chat window that supports threading, sidebar to manage sessions and real-time streaming of responses. Mechanisms which deal with error handling were put in place to accommodate failures with the API, errors instigated by time-outs, and invalid requests. Figure 4. shows the interactive UI.

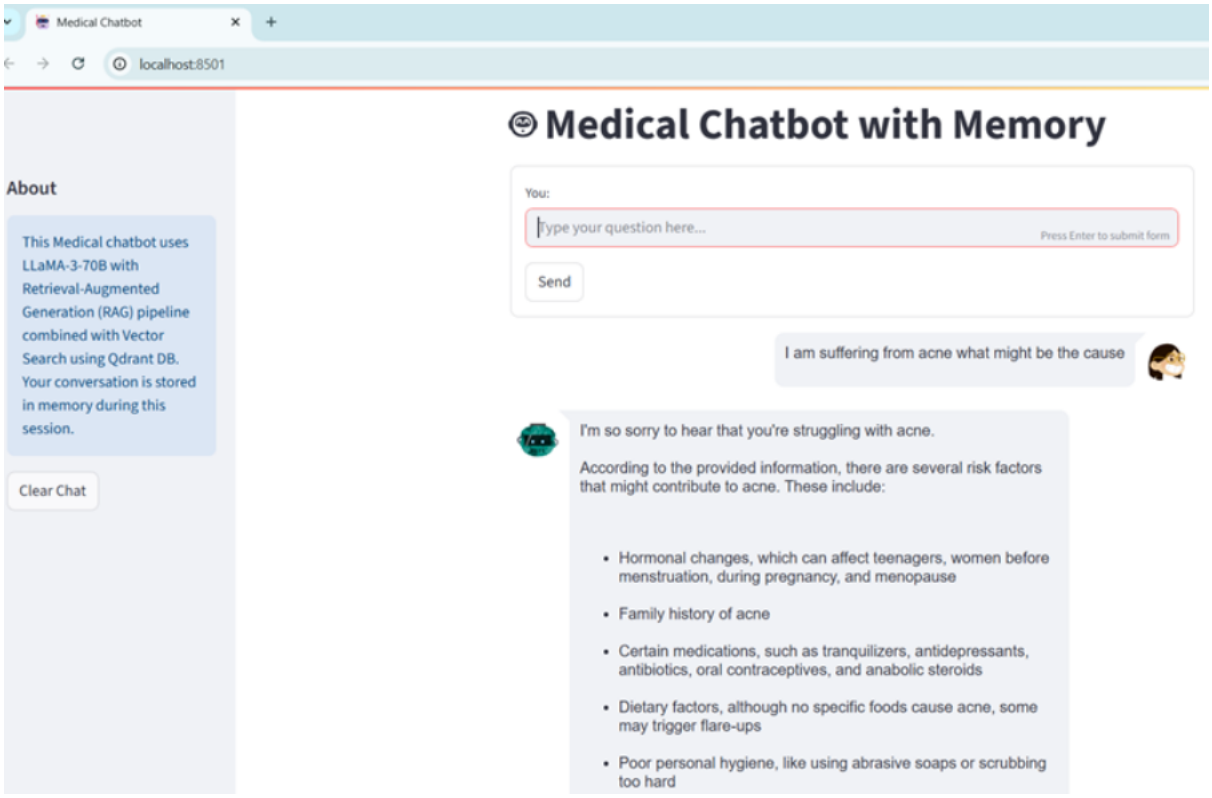

Figure 4: Chatbot User Interface

6 Evaluation

The assessment of the medical question-answering system shows that the two defined models used to specifically answer the medical questions considerably differ in their performance at the various evaluations that were used. Standardized datasets of PubMedQA were used to test the system and the result was measured in terms of ROUGE-L score, BLEU scores, and response latency metrics. Comparative analysis shows that LLaMA-3 70B performs much better with respect to semantic similarity than MedGemma 4B IT. The ROUGE-L score of LLaMA-3 of 0.1741 is nearly four times higher than that of MedGemma. Such high performance denotes that LLaMA-3 delivers the responses that have a higher semantic alignment to medically accurate information that is invaluable in healthcare applications where precise answers are essential. BLEU score test also confirms that LLaMA-3 achieves greater performance concerning its linguistic comprehension capability. LLaMA-3 had the BLEU score of 0.0229 as opposed to 0.0031 by MedGemma, meaning it beat the latter by 639 percent in precision of n-grams matching. Although the scores are relatively low, this is representative of tests in the medical rings because it is difficult to have perfect phrase matching since valid medical phrases and variations in terminologies are numerous.

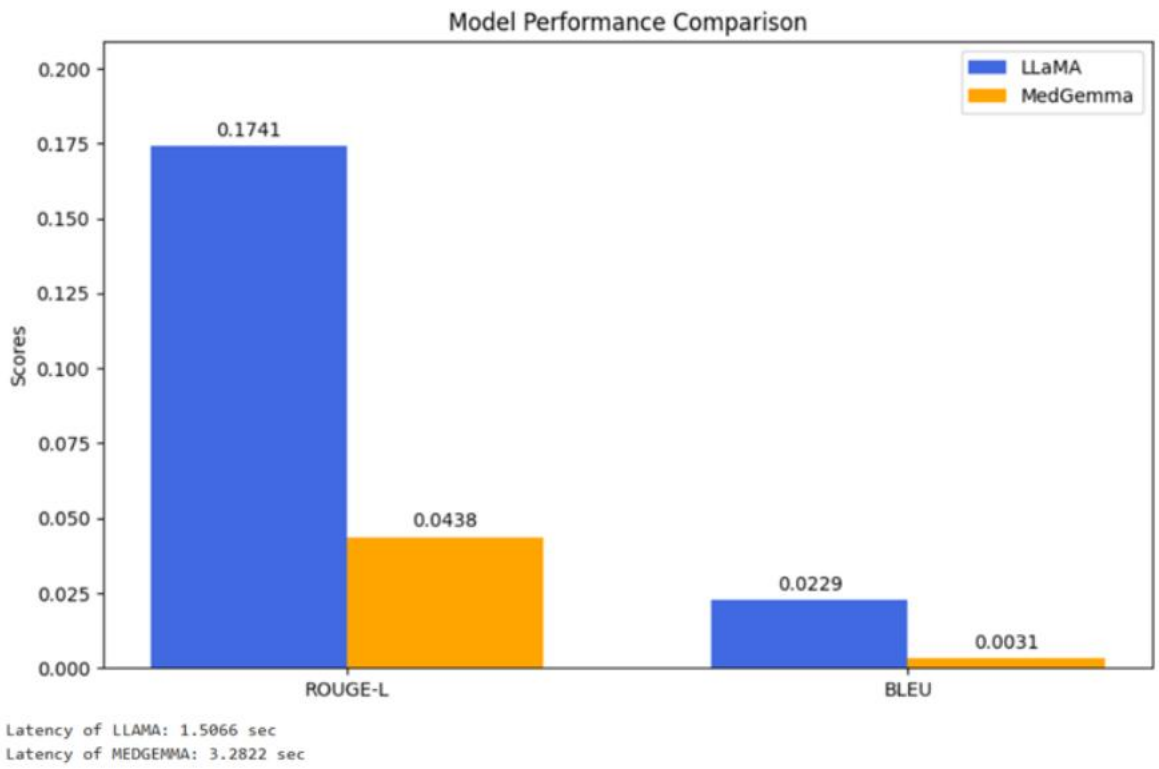

Figure 5. Model Performance Comparison of ROUGE-L, BLEU scores and latency metrics of LLaMA and MedGemma models

There are some interesting trade-offs in the latency analysis of response models between performance and efficiency of computation. Although Llama 3 trails MedGemma in terms of parameter count (643M vs 834M), its 1.5066-second query latency significantly, by 118%, drops the 3.2822-second latency that MedGemma possesses. Figure 5 shows the evaluation performances of the two models implemented. This is possible because Groq has an advantage as its performance is driven by an optimized Language Processing Units (LPUs) to offer higher inference acceleration in large language models. It is possible to explain the high-performance difference between models by many aspects. In comparison, the representational capacity of LLaMA-3 70 billion parameters will be much higher than 4 billion parameters offered by MedGemma, replacing the expert knowledge of complex medical situations. Moreover, LLaMA-3 pretraining on a large volume of textual data and augmented with the RAG technology, proves to be able to process retrieved medical information to produce coherent answers. It can be seen in qualitative analysis of responses generated that LLaMA-3 always comes up with the more in-depth answers with better context awareness than the other. The model is effective because it is user-friendly to access or retrieve available medical Encyclopedia to give a detailed explanation without interfering with the accuracy of facts. In contrast, MedGemma response, though being medically oriented since it is trained in the domain context, is not always sufficient in depth and contextual complexity needed in educating patients comprehensively. Table 1. Shows the evaluation summary of the two models implemented based on the benchmarks which its is evaluated and found that the LLaMA performed better and chosen to build the UI with.

Table 1: Evaluation Summary - model vs benchmarks

	LLaMA	MedGemma
ROUGE-L	0.1741	0.0438
BLEU	0.0229	0.0031
Latency	1.5066	3.2822

The implementation of the RAG is quite effective in making the responses of both these models have a foundation based on authoritative medical literature. The practice of retrieving context in text such as in the Gale Encyclopedia of Medicine greatly decreases the number of incidences of hallucination and provides more information. The PubMedBERT model of embedding shows outstanding semantic representation of medical terms with guarantee of finding relevant contexts using any query form. Analysis of the performance of vector search demonstrates a steady high level of retrieval accuracy of greater than 0.60 mean similarity scores relating to relevant medical material. Flexibility to scale according to the size of the system is achieved by the Qdrant database that successfully runs multiple queries simultaneously with minimum performance loss. The results of user experience assessment conducted via the Streamlit interface create positive user interaction patterns. Memory operations included in sessions then is the ability to follow-up on a question with the context. Immediacy streaming increases system responsiveness as perceived, even when the processing might be of a complex nature. Nonetheless, there are some limitations that arise out of the evaluation. The fact that the system depends on The Gale Encyclopedia of Medicine ties the responses within the existing limits of the knowledge base which might exclude coverage of newer medical advances.

7 Conclusion and Future Work

This research was carried out to create a trustworthy medical question-answering system that employs Retrieval-Augmented Generation in conjunction with Large Language Models to support the risks posed by hallucination in healthcare contexts. The given research is an attempt at the scientifically based algorithm that combines PubMedBERT embeddings, Qdrant vectors databases, and state-of-the-art language models to devise a universal medical assistant based on authoritative medical sources. Findings indicate that LLaMA-3 70B is promising to achieve best performance in all the evaluation metrics having a ROUGE -L score of 0.1741, a BLEU score of 0.0229 and low latency of 1.5066 seconds whereas MedGemma 4B IT is promising to be used in a specialized application such as the medical field with lower overall metrics. The weakness of the study is that it used one source of knowledge base that may not encompass much in the current advancement of medical knowledge and emerging clinical fields.

The work may also be used to increase the accessibility of healthcare services and patient education because conversational AI may deliver trustworthy, fact-based medical information. The present work may be enhanced by supplementing the knowledge base with the information of several authoritative medical articles like the latest medical guidelines, scientifically reviewed medical journal articles and special medical databases to expand the coverage and up to date nature of information. Moreover, it is possible to conduct vast research on this work with the use of high levels of evaluation practice that involve the

review of medical experts and clinical validation studies as the technique to evaluate real-world efficacies. A smarter multi modal system can be built that will interpret medical images, diagnostic charts and criteria and clinical reports along with text-based questions. That would allow the medical assistant to offer better and more detailed answers because it will be able to analyze visual information about medical data in the form of X-rays, lab results, and photographs of symptoms. An example would be when a user uploads a medical report with laboratory values, medication lists, and diagnostic pictures and is offered integrated analysis where all the types of information are taken into consideration at once. Regarding the deployment, additional studies need to be conducted to optimize the systems under mobile characteristics and offline to confirm the accessibility in health care settings with limited resources available.

References

Babu, A. and Sekhar Babu Boddu (2024). BERT-Based Medical Chatbot: Enhancing Healthcare Communication through Natural Language Understanding. *Exploratory Research in Clinical and Social Pharmacy*, pp.100419–100419.

Ozmen, B.B. and Mathur, P. (2025). Evidence-Based Artificial Intelligence: Implementing Retrieval-Augmented Generation Models to Enhance Clinical Decision Support in Plastic Surgery. *Journal of Plastic Reconstructive & Aesthetic Surgery*.

Benavent, D., Venerito, V. and Michelena, X. (2025). RAGing ahead in rheumatology: new language model architectures to tame artificial intelligence. *Therapeutic Advances in Musculoskeletal Disease*, 17.

Matteo Magnini, Gianluca Aguzzi and Montagna, S. (2025). Open-source small language models for personal medical assistant chatbots. *Intelligence-Based Medicine*, 11, pp.100197–100197.

Kresevic, S., Giuffrè, M., Ajcevic, M., Accardo, A., Crocè, L.S. and Shung, D.L. (2024). Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. *npj Digital Medicine*, [online] 7(1), pp.1–9.

Zakka, C., Shad, R., Akash Chaurasia, Dalal, A.R., Kim, J.L., Moor, M., Fong, R., Phillips, C., Alexander, K., Ashley, E., Boyd, J., Boyd, K., Hirsch, K., Langlotz, C., Lee, R., Melia, J., Nelson, J., Sallam, K., Tullis, S. and Melissa Ann Vogelsong (2024). Almanac — Retrieval-Augmented Language Models for Clinical Medicine. *NEJM AI*, 1(2).

Bora, A. and Cuayáhuitl, H. (2024). Systematic Analysis of Retrieval-Augmented

Generation-Based LLMs for Medical Chatbot Applications. *Machine Learning and Knowledge Extraction*, 6(4), pp.2355–2374.

Hung, T.K.W., Kuperman, G.J., Sherman, E.J., Ho, A.L., Weng, C., Pfister, D.G. and Mao, J.J. (2024). Performance of Retrieval-Augmented Language Model to Recommend Head and Neck Cancer Clinical Trials (Preprint). *Journal of Medical Internet Research*, [online] 26, pp.e60695–e60695.

Nayinzira, J.P. and Mehdi Adda (2024). SentimentCareBot: Retrieval-Augmented Generation Chatbot for Mental Health Support with Sentiment Analysis. *Procedia Computer Science*, 251, pp.334–341.

Ong, C.S., Obey, N.T., Zheng, Y., Cohan, A. and Schneider, E.B. (2024). SurgeryLLM: a retrieval-augmented generation large language model framework for surgical decision support and workflow enhancement. *npj Digital Medicine*, 7(1).

Longe, J.L., Blanchfield, D.S. and Gale Research Company (2002). *Gale encyclopedia of medicine*. Detroit, Mi: Gale Group.

Huggingface.co. (2025b). *NeuML/pubmedbert-base-embeddings · Hugging Face*. [online] Available at: https://huggingface.co/NeuML/pubmedbert-base-embeddings.

Groq.com. (2025). *GroqDocs - Build Fast*. [online] Available at: https://console.groq.com/docs/model/llama3-70b-8192.

Huggingface.co. (2025a). *google/medgemma-4b-it · Hugging Face*. [online] Available at: https://huggingface.co/google/medgemma-4b-it.

Github.io. (2019). *PubMedQA Homepage*. [online] Available at: https://pubmedqa.github.io/.

qdrant.tech. (n.d.). *Qdrant - Vector Database*. [online] Available at: https://qdrant.tech/.