A Context-Aware Medical Support System Using Retrieval-Augmented Generation with LangChain, Hugging Face, Mistral LLM, and FAISS

A PROJECT REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF

MASTER OF TECHNOLOGY
IN
ARTIFICIAL INTELLIGENCE

Submitted by

SUMIT KUMAR MISHRA (23/AFI/29)

Under the supervision of

DR. MINNI JAIN

COMPUTER SCIENCE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042

JUNE, 2025

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, **Sumit Kumar Mishra**, Roll No's – **23/AFI/29** student of **M.Tech (Artificial Intelligence)**,hereby declare that the project Dissertation titled "**A Context-Aware Medical Support System Using Retrieval-Augmented Generation with LangChain, Hugging Face, Mistral LLM, and FAISS**" which is submitted by me to the Computer Science Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi Sumit Kumar Mishra

Date:

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled "**A Context-Aware Medical Support System Using Retrieval-Augmented Generation with LangChain, Hugging Face, Mistral LLM, and FAISS**" which is submitted by **Sumit Kumar Mishra**, Roll No's – **23/AFI/29**, Artificial Intelligence ,Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi Dr. Minni Jain

Date: **SUPERVISOR**

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

I wish to express my sincerest gratitude to **Dr.Minni Jain** for her continuous guidance and mentorship that she provided me during the project. She showed me the path to achieve my targets by explaining all the tasks to be done and explained to me the importance of this project as well as its industrial relevance. She was always ready to help me and clear my doubts regarding any hurdles in this project. Without her constant support and motivation, this project would not have been successful.

Place: Delhi Sumit Kumar Mishra

Date:

Abstract

An important problem in healthcare use cases is the inclination of conventional AI language models to provide realistic but factually incorrect or hallucinated answers. This limitation has been made apparent by increasing the need for credible medical information. Incorrect claims about facts and lack of citations to credible sources are effects of employing only parametric data in standard approaches such as GPT-3.5 and other fine-tuned LLMs. In complex medical domains specifically, existing instantiations of retrieval-augmented generation (RAG) systems are plagued by a trade-off between retrieval precision, semantic coherence, and response suitability, even though RAG systems were initially suggested as the solution to these issues. This work constructed a RAG-based medical AI assistant to assist with these obstacles.

To provide answers in a timely manner, it utilizes FAISS for the retrieval of vectors and Mistral-7B-Instruct-v0.3. The AI has also been trained to provide responses only that are supported by valid medical sources, according to The Gale Encyclopedia of Medicine. The quantitative metrics that were used for our system's evaluation were BERTScore (Precision: 0.8334, Recall: 0.8119, F1-Score: 0.8225), Answer Relevancy (0.9221), and Faithfulness (1.0). Its performance was significantly superior to that of general-purpose models like GPT-3.5-Turbo (Faithfulness: 0.89) and LLaMA-2-70B-Chat (Faithfulness: 0.95). While being still therapeutically relevant and with high semantic coherence, our RAG model effectively limits hallucinations, as per the outcome.

The observation that issues nonetheless occur when handling immensely complicated or unclear questions still supports the need for better reasoning methods. This paper sheds light on the potential of RAG systems that are tailored to specific areas of healthcare. Offers a reliable way of accessing correct medical records and gives way to innovation in dynamic knowledge fusion and multi-hop retrieval. The findings favor the application of specialist AI assistants in areas such as healthcare and academia where accuracy, transparency, and reliability are essential.

Contents

List of Tables

List of Figures

Chapter 1

INTRODUCTION

The world's healthcare systems are experiencing increasing pressures as a result of aging, workforce shortages, and rising costs. Core healthcare processes, particularly manual diagnosis and paperwork, are increasingly viewed as labor-intensive and not very effective. Studies indicate that doctors spend nearly half their working time on administrative activities, leaving them with less time to deliver direct patient care (Bidemi, 2024). With its capacity to automate administrative processes, enhance the operational effectiveness, and support clinical decision-making, artificial intelligence (AI) has been a disruptive force in the health sector (Alhashmi et al., 2020). Development of AI-powered health aides, who aim to enhance the efficacy, accuracy, and reach of healthcare services, is one of the most significant uses of AI (Manickam et al., 2022). AI-based tools offer a data-driven solution that is quick, precise and trustworthy compared to traditional healthcare options which are dependent nearly solely on paper documentation and individual clinical judgment. AI facilitates it simpler such that it can navigate through patient history, scientific evidence, and clinical guidelines and subsequently devise sound evidence based advice.

With its capability to support illness diagnosis, clinical data management, treatment recommendations, and even direct patient interaction, AI-powered health assistants can potentially transform the healthcare sector (J. Yang et al., 2022), (Secinaro et al., 2021). Healthcare is leveraging big data, machine learning, and robotics to monitor risks and benefits due to the developments in AI (Hossen Karmoker, 2020); (Dharani, 2021). With regards to enhancing operations and accelerating the delivery of care, medical data and analytics are the cornerstone. The diversity and volume of medical records gathered have been significantly enlarged over the past few years. An example is the vast amount of data generated by patients, scientists and healthcare professionals. This data is a subset of this data and have been reaped from broad variety of sources, such as medical images, EHRs, health and lifestyle monitoring apps and wearables. Outside of medicine, the knowledge is increasingly relevant to the lay publicicasphalt Technology and other real-worldcases (Antoniou et al., 2017); (Xie et al., 2020)

1.1 The Rise of AI in Healthcare

The rise of AI in healthcare can be attributed to several key factors:

- **Advancements in Technology:**Technological Developments: The most recent advances in artificial intelligence, machine learning, and

AI systems can now process and analyze potentially the greatest amounts of healthcare data at incredibly quick speeds thanks to deep learning. The appearance

The outcome of this is the use of AI algorithms to carry out complex medical tasks like image identification, natural language processing, and predictive analysis.

- **Increasing Availability of Healthcare Data:**Large archives made accessible to AI systems have only been fueled by the growth of EHRs, genetic data, medical imaging, and other types of healthcare data. These enormous datasets serve as the training material that an AI system requires in order to get better over time.
- **Growing Demand for Healthcare Solutions:**A expanding patient base, rising expenses, and the prevalence of chronic conditions are just a few of the problems facing the healthcare industry. Therefore, there is a great need for new ways to manage resources, improve patient outcomes, and improve healthcare delivery. The AI system could be helpful in these areas by offering suggestions and insights aimed at improving clinical judgments and operational efficiency.
- **Demonstrated Success in Various Applications:**AI has proven beneficial in a number of health-related fields, including drug development, customized medicine, medical imaging, and diagnostics. These AI systems outperform human doctors in everything from cancer detection on medical photos to patient outcome prediction. It's possible that advances in human medicine have increased interest in and funding for AI health systems.
- **Supportive Regulatory Environment:**Organizations such as the Food and Drug Administration (FDA) in the US have even started to set up procedures for testing and approving software and medical devices that use artificial intelligence (AI) before they are put on the market. These legislative actions are opening the door for the proper application of artificial intelligence (AI) in healthcare and could encourage medical practitioners to have more faith in AI solutions.

1.2 Importance of AI Enablement

Artificial intelligence's (AI) capacity to reshape different segments of healthcare towards improving patient outcomes, operational efficiency, and cost savings makes it a vital resource. Major points of significance for AI empowerment are as follows:

Improved Diagnostic Accuracy: To help physicians diagnose patients more rapidly and accurately, artificial intelligence (AI) systems can be trained to read genetic information, radiographs, and patient histories. AI-derived diagnoses can potentially be used to improve diseases at an earlier stage, provide treatment recommendations, or even save lives. With AI, we can create individualized therapy programs that meet the requirements of every patient. AI algorithms are now able to provide the best course of action for every patient by evaluating previously unheard-of volumes of patient data, genetic markers, treatment outcomes, and medical history. This individualized approach to healthcare is intriguing and displaying promising early results by reducing the risk of adverse effects and increasing the effectiveness of therapy.

Simplified Administrative Procedures: Doctors may devote more time to patient care as AI makes administrative duties like scheduling appointments, billing, and record keeping easier. This is due to the fact that AI can also be used to automate certain repetitive tasks and simplifying things for them, which will save the employers' time, reduce the amount of paperwork and administrative labor a healthcare organization does, and eventually improve business performance and cost savings. AI systems might analyze patient data to identify patterns or correlations that could lead to the deduction.

Support for Remote Patient Monitoring: Artificial intelligence-based remote patient monitoring systems provide real-time reporting of changes or anomalies in a patient's condition to medical personnel. This reduces the need for hospital stays and ER visits by offering the opportunity for early intervention and proactive management of chronic illnesses.

Research and medication Discovery: Software with artificial intelligence may look through massive databases for novel medications, identify potential drug interactions, and improve current treatment plans. AI enablement can improve patients' access to new medications in two ways: by expediting the drug development process and by lowering the duration and cost of clinical trials.

1.3 AI in Diagnosis and Treatment

Healthcare has been transformed by AI in diagnosis and treatment in a number of important ways:

Increased Diagnostic Accuracy:
Artificial intelligence algorithms have demonstrated remarkable accuracy in assessing a variety of medical pictures, such as CT, MRI, and X-ray scans. Because these algorithms can spot subtle anomalies that are invisible to the human eye, early and accurate diagnosis of illnesses like cancer, heart disease, and neurological conditions. "AI helps radiologists and other medical professionals interpret medical images, ensuring that patients receive timely and appropriate treatment." Better Disease Detection and Forecasting: Algorithms using artificial intelligence are able to sift through vast amounts of patient data, including genetic data, medical histories, and biomarkers, to identify trends and patterns that may indicate a high risk of developing specific diseases.AI may be able to predict a person's risk of developing diabetes or heart disease by looking at their genetic makeup, lifestyle, and medical history. Doctors can intervene and prevent diseases from progressing by using AI to identify potential health risks earlier. AI has made it possible to create customized treatment plans for every patient's particular requirements and traits. Artificial intelligence (AI) algorithms sort through mountains of data, including genetic information, reaction to treatment, and result, to decide the best course of action for each distinct individual patient. This individualized approach to treatment pairs patients with medications that are most likely to benefit with their particular condition, leading to better therapy with fewer side effects.

Optimal Treatment Plans: These days, machine learning algorithms can mine academic publications, clinical research, and electronic health records to process a wide range of medical disorders. AI helps clinical practitioners make better decisions about therapy alternatives, dosage schedules, and follow-up treatment plans by offering evidence-based suggestions. Based on the most recent scientific discoveries and clinical recommendations, this ensures patients receive the best and most appropriate care. Contributing to the Growth

Clinical Decisions:AI-powered illness detection and therapy By integrating data, the CDSS system uses artificial intelligence to provide doctors and other healthcare providers with prompt suggestions for the best course of action to treat their patients. Using clinical research, patient data, and suggested best practices, doctors can use these systems to help in diagnosis and therapy. CDSS enhances treatment, lowers the number of diagnostic errors, and improves patient care by combining clinical expertise with AI-derived information. Healthcare has been transformed by artificial intelligence, which has improved disease detection, predicting, diagnosis, and treatment. AI will have a growing impact on patient outcomes and healthcare delivery as technology develops, advancing medicine and establishing international standards for patient care.

1.4 The potential of virtual health assistants

With their focus on direct patient engagement, virtual health assistants (VHAs) are one of the most exciting new arenas in artificial intelligence research. Several critical tasks previously done by humans in patient care have been assumed by virtual health assistants. Some of these are scheduling appointments, answering questions, prescribing medicine, and assisting with mental health (Chawda Fatima, 2023). Through offering patients, contemporary, pertinent health information across various channels (i.e., chatbots, voice recognition software, and mobile apps), numerous such cutting-edge AI-based technologies might improve doctor-patient relationships. Beyond providing patients with unprecedented convenience, VHAs discharge physicians and other healthcare providers of a significant administrative workload, enabling them to focus more on medical tasks and decisionmaking (Sherani et al., 2024).

In this regard, VHAs can enhance patient outcomes as well as save time by performing basic tasks. Most patient-provider interaction takes place during consultations, such as in conventional medical care systems, and issues and concerns are raised only at that moment. Nevertheless, a VHAs is present to assist the patient at each step of the procedure (Salunkhe et al., 2022). These systems can monitor patients' medical data, notify patients when it is time to take a medication, offer real-time responses to questions regarding medical matters, and offer suggestions for a given patient's medical condition (Husnain Saeed, 2024). These constant interactions enhance the relationship between the patient and clinician and motivate patients to adhere to recommended schedules. Facilitating greater access to healthcare, especially for underprivileged regions, is another major contribution of virtual health assistants (Khan et al., 2024). Moreover, these applications assist patients in self-managing chronic diseases, managing prescriptions, scheduling appointments, assessing symptoms, and patient education (Javaid et al., 2023). Websites, voice assistants, and phone applications are just a few of the digital platforms that may facilitate more access to care assistance. Virtual assistant use in health is not without challenges, though, such as issues of permission, privacy, safety, interpretation of data, ethics, and responsibility (Javaid et al., 2023).

1.4.1 Challenges and Barriers to VHAAdoption

Even with AI's quick development, there are still a number of significant technological obstacles to overcome. Effective health coaching requires an understanding of the nuances of human emotion and language, which VHAs may find difficult to achieve. Making sure AI systems provide trustworthy medical advice for a wide range of particular health conditions is similarly difficult. A few examples include people with diabetes that is poorly managed, people with several chronic illnesses, and people whose drugs, like beta blockers, have particular impacts on physical activity. Furthermore, while very sophisticated AI models can retain information over extended talks and across platforms, they run the risk of missing crucial details that have already been covered because they mostly focus on more recent interactions.

Since people are likely to give a VHA sensitive health information, it is imperative to protect the accuracy and privacy of their data. Data breaches will have detrimental effects on both the data and public trust, ultimately undermining AI's potential for usage in healthcare (Gillespie et al., 2023). Strong cybersecurity measures must therefore be integrated into VHAs to prevent unauthorized data access and maintain a functional experience. These issues have a connection to regulatory issues. Virtual health aides are subject to a number of laws and guidelines, which vary from country to country. In particular, a number of governments forbid sending electronic health data overseas, which might put many AI systems in breach.

For instance, developing regionalized AI systems that enable data processing and storage within the user's country or region could help address these challenges. Some state actors' attempts to divide the internet and the incompatibility of technology processes and governance frameworks may exacerbate problems on a broader scale, even though local solutions may be able to address some jurisdictional limits (X. Xu et al., 2023).

Explainable AI, augmented reality, digital twins, closed systems, and synthetic data are some of the emerging technological solutions with potential for transcending current limitations and increasing clinician confidence in VHAs. An enhanced understanding of how AI decisions are arrived at is facilitated by explainable AI, enhancing interpretability and transparency. In playing the role of a trustworthy "second pair of eyes," augmented reality could help medical professionals in better interpreting complex images (Harari et al., 2024). Virtual representations of physical systems, or digital twins reduce the risk of security or privacy violation by allowing the representations to be modelled to predict the potential behaviour of real systems (Kenett Bortman, 2022). The privacy and transparency issues that open systems bring are handled by system that is closed and based on models and training data. Another final alternative to training AI systems where actual data are not available is the use of synthetic data. The potential loss of human interaction while caring and replacement of human occupations are additional ethical issues brought about by VHAs. In behaviour change, where a patient's motivation is frequently linked to a sense of responsibility to his physician (Eton et al., 2017) an element that can reduce with technology programs—the clinician-patient relationship is central to the success of therapy. Healthcare organizations in pursuit of cost containment may be attracted by the economic motive of substituting human labor with automated processes, even though VHAs hold great promise to augment current services or offer new ones, such as digital coaches based on augmented reality for enhanced diagnosis. Moreover, as with other eHealth projects, VHAs have the risk of inadvertently widening the gap in health inequality even as they provide health information and services to all. Those who are less familiar with technology or who are skeptical of digital health might find this particularly troublesome. Lastly, VHAs must keep users interested and win users' trust in order to be truly effective. Although even the latest AI systems appear to be promising in this sense, A full understanding of human psychology, behavior, and requirements is required to create AI systems that people are willing to be confident and safe using as health advisors.The long-term use of these systems is jeopardized by accumulating evidence showing that users are particularly tolerant of error on the part of VHAs or when their questions are not answered (Davis et al., 2020).

1.5 Chatbots in Health Care

Chatbots or conversational agents are at the forefront of the vibrant world of information technology and digital communication, revolutionizing how human beings and technology interact. Chatbots are software programs that simulate the manner in which humans converse with one another using words, images, sounds, or video on websites, applications, or independent software. (Kocaballi et al., 2019). From its start in 1994 (Mauldin, 1994), Chatbots have truly arrived a great distance. Nowadays, they can assist in making appointments, respond to queries of patients, and share data effortlessly. Due to improvements in natural language processing and AI, chatbots have progressed far from simply going through pre-programmed lines with generic responses.

Now,they are able to fully comprehend what users ask and reply appropriately (Nuruzzaman Hussain, 2018), (Kumar et al., 2016). Journalism, education, e-shopping, finance, healthcare, and entertainment are but some of the fields that have managed to utilize them because they're so versatile. The case of Amazon's Alexa (Laymouna et al., 2024), Apple's Siri (Laymouna et al., 2024), Google Assistant (Sprengholz Betsch, 2021), Microsoft's Cortana (Shaikh Cruz, 2023), and Samsung's Bixby (Laymouna et al., 2024) are just some popular examples of these apps. Chatbots would actually be able to assist the medical field in making the treatment more effective, affordable, and overall better (Huisman et al., 2022), (Osipov Skryl, 2021), (Jones et al., 2014),(Chandrashekar, 2018) along with numerous applications and approval (Nadarzynski et al., 2019) ,(A. A. Abd-Alrazaq et al., 2021). Increasingly, chatbots are utilized to receive and provide healthcare services (A. Abd-Alrazaq et al., 2020),(Kretzschmar et al., 2019), (Cheng Jiang, 2020), (Boucher et al., 2021), providing them with varying functions in diagnosis, prevention, and treatment assistance, which may affect the whole healthcare system.

Healthcare chatbots offer specific challenges in spite of their possible benefits (Luxton et al., 2016), (Denecke et al., 2021),(Parviainen Rantala, 2022), (Palanica et al., 2019). Providing individualized advice remains a significant challenge. Most chatbots provide general responses that ignore local health problems, patient medical histories, and individual medical profiles, which are all significant to proper diagnosis and personal treatment (Chang et al., 2024). Additionally, conventional chatbots traditionally lack contextual understanding and adaptability needed to retrieve up-to-date medical information or perform well across multiple areas of medicine. The merger of large language models (LLMs) like ChatGPT with generative AI has been a significant advancement in chatbot technology. Their ability to generate text that is similar to that of a human enables more natural and educational conversations (Ouyang et al., 2022), (OpenAI Achiam et al., 2023). With medical data so complicated and accurate data so critical, LLMs are a significant leap forward in medical AI technology. Most view the technological advancements associated with LLMs to be among the greatest technological achievements of the last few years. Deep learning uses enormous amounts of text and data for training models to create an LLM. The resources belong to a mix of materials related to language, including books, websites, articles, and transcripts of videos.

Consequently, chatbots and answer generators that utilize LLMs to translate linguistic content directly have enabled us to ask online information more efficiently and directly than before (on the basis of document retrieval). Yet, by experimenting with a range of data sets, scholars are constantly striving to enhance the calibre of LLMs and their applications. The key differences between current LLMs such as GPT4o, LlaMA3/LlaMA3.1, Mistral, and Claude and pre-trained language models such as BERT and GPT2 (Reimers Gurevych, 2019) (T. Zhang et al., 2019)such as step-by-step reasoning, instruction following, and in-context learning (ICL) (Topol, 2019). ICL is the ability to generate an output with or without additional training or gradient tuning from instructions or demonstrations. This would mean that the model can acquire new skills by merely complying with instructions without any additional assistance. A key feature of LLMs is their ability to sequence. Chain-of-Thought (CoT) prompting is a mechanism that LLMs utilize to resolve intricate problems. In CoT, users have to supply a portion of the information required as opposed to asking the query directly. The approach then splits the problem into smaller practicable phases. But their application in healthcare is not yet mature. Errors and misinformation are a critical problem (Thirunavukarasu, Hassan, et al., 2023), particularly in the medical profession where accuracy is of utmost importance. The one-size-fits-all approach of LLMs might not be an appropriate fit for the intricacies of patient-centered therapy in the medical profession (Thirunavukarasu, Ting, et al., 2023).

1.6 Advancing Chatbots with Retrieval-Augmented Generation (RAG)

The use of large language models (LLMs) based on Retrieval-Augmented Generation (RAG) is a noteworthy advancement in AI-driven healthcare technology, given the limitations of LLMs. RAG models pull external information in real-time and integrate it into the content generation process, in contrast to typical LLMs that only use available training data to generate content (Toukmaji Tee, 2024).

RAG is a cutting-edge method for improving response quality. By combining information retrieval (IR) methods with the natural language production feature of transformer models like GPT, this paradigm enables the chatbot to access external data from a sizable corpus in real-time. Instead of on solely on the model's pre-trained knowledge, RAG uses

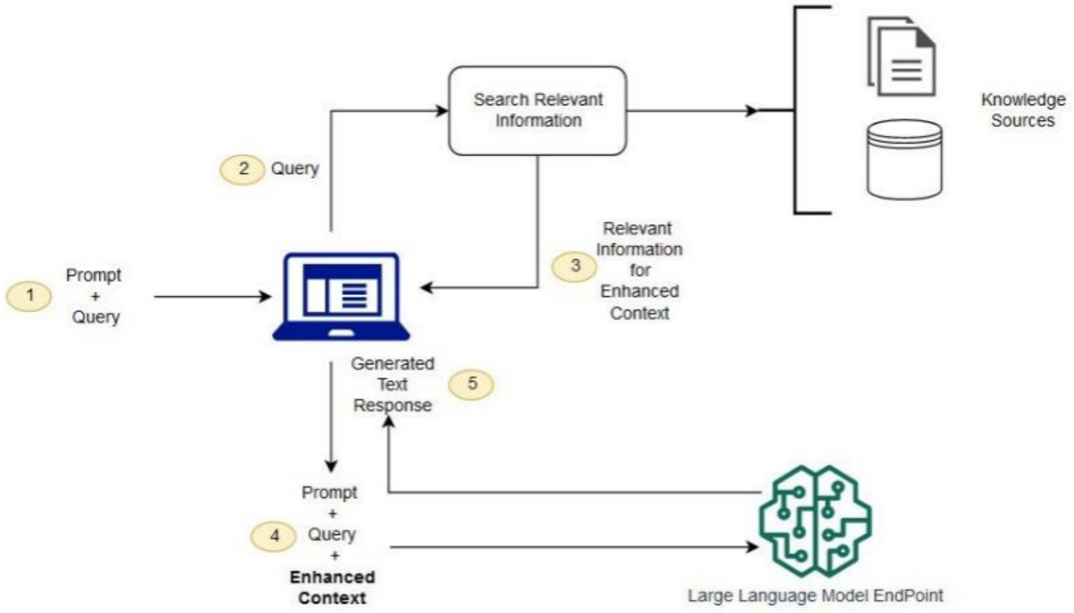

Search Relevant Information
Knowledge Sources
2 Query
Relevant Information for Enhanced Context
3
1 Prompt + Query
Generated Text Response
5
Prompt + Query + Enhanced Context
4
Large Language Model EndPoint

Figure 1.1: Steps of Retrieval Augmented Generation (RAG)

the system to actively search through particular databases or papers for content pertinent to the topic in order to increase the relevance and accuracy of its responses.

RAG integration could open the door for new customer care applications by improving chatbot performance for application scenarios, such as giving prompt answers with thorough justifications. 1 makes it obvious that RAG is a simple method. The user submits a query to the system. The system may retrieve the most significant and relevant documents from the database using this query. The degree to which these papers address the question will determine their selection. The machine then uses an LLM to produce a response that sounds human by taking the user's query and the gathered documents. As a result, the user receives this information in the form of a precise and organic response to his query.

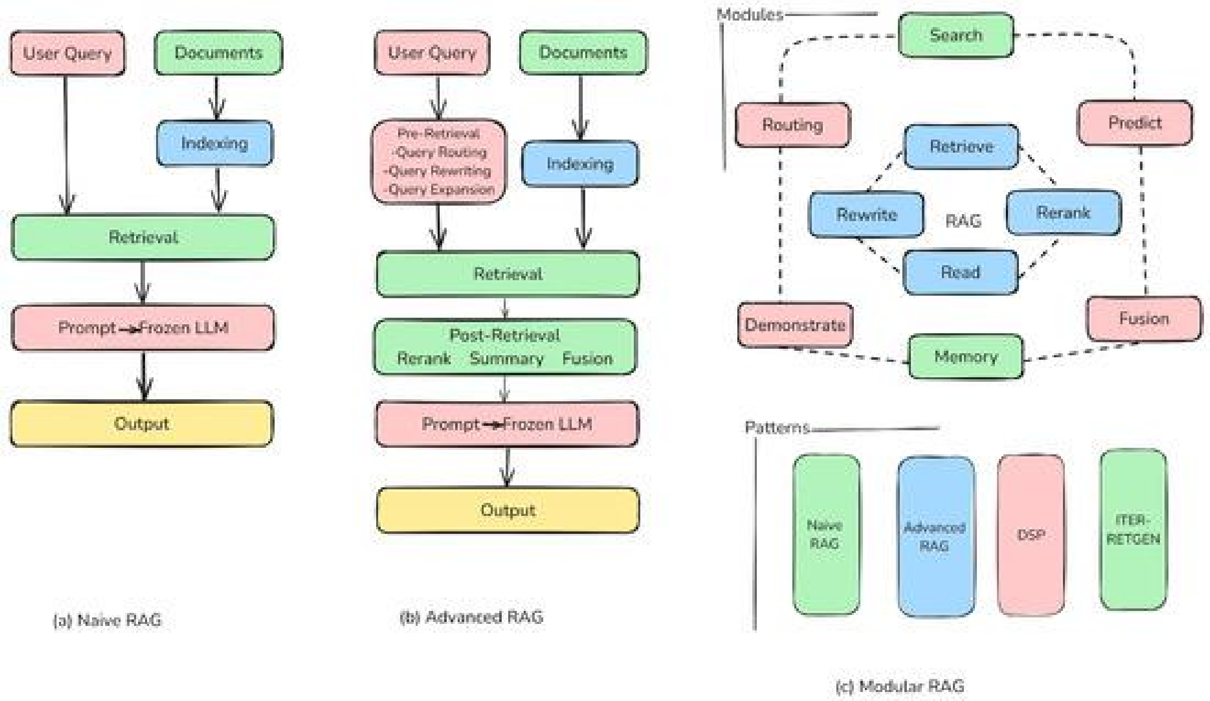
The capability of RAG-based LLMs to retrieve huge volumes of medical information from structured databases makes them more accurate and contextually correct response providers. A significant challenge for healthcare artificial intelligence is generating safe, accurate, and contextually correct information; RAG systems present a potential solution for this challenge. However, robust and adaptable NLP architectures are required for the successful deployment of such systems within healthcare. Hugging Face and LangChain are herein cornerstone technologies that lay the groundwork for building smart, RAG-based medical assistants. Hugging Face has emerged in a short while as an industry standard for natural language processing (NLP). Latest pretrained models such as as BERT, RoBERTa, GPT, and T5 can be easily accessed within its Transformers library. Pre-trained models are available for most natural language processing tasks, and thus can be utilized with minimal programming. Filtering, translation, and question answer are all easy to do with the help of the Transformers library. For info extraction tasks, hugging face models work well in extracting names, dates, question answers, etc (Pol et al., 2024).

The models are instrumental to the RAG architecture since they translate and process input in natural language, detect relevant medical information, and obtain reasonable responses. Concurrently, LangChain serves as the orchestration layer, enabling easy integration of these models into a conversational pipeline. It coordinates the flow of conversation through entities like ChatModel and Message, enabling one to have natural, multi-turn conversations (Singh et al., 2024). Hugging Face Transformers provide language abilities, while LangChain assists in handling conversations. They combine to form helpful AI medical assistant that's reactive, context-aware, and integrates well with medical standards.

This paper details a new approach to crafting a smart health chatbot with Hugging Face Transformers and LangChain. In contrast to other language models that merely adhere to their training data, this method draws in health information from external sources as it runs. Although previous studies either concentrated on Hugging Face models or chatbots individually, this does both to offer responses that are accurate, precise, and relevant to health. A creative architecture tailored to meet the intricate needs of health communication is demonstrated by the seamless harmonization of real-time retrieval, advanced natural language processing, and structured conversation management. Though RAG research and development continues to be ongoing, scholars now categorizing RAG into three categories, namely, shown in Figure 2: Modular RAG, Advanced RAG, and Naive RAG.

Advanced RAG uses embedding models such as BERT and its generalisations, whereas Naive RAG uses basic retrieval models such as BM25. With modular RAG, numerous retrieval processes are feasible depending on the job. The Advanced RAG type is the most frequently utilized in chatbot generation research among all of them. Due to the fact that context awareness is highly crucial within a medical chatbot context. When speaking about medical conditions, it is usual to require exact, up-to-the-second access to comprehensive information, for example, drug interactions, latest research results, or updated clinical guidelines. Contrary to Modular RAG, which keeps retrieval and production distinct but does not necessarily stitch them deeply, Naïve RAG collects data without taking into account the significance of context. Yet the Advanced RAG variety allows for a richer association between generating and retrieval. By embedding retrieval strongly into the process of generating, Advanced RAG can cope with such complexity and constantly adapt the retrieved documents' relevance based on current interaction. This contributes to reducing errors and making sure that responses are current and medically sound, depending on the type of external or recovered data.

As seen on the left (a), Naive RAG is composed of three primary stages: indexing, retrieval, and creation. The linear, chain-like process of Advanced RAG, which is otherwise equivalent to Naive RAG, includes several pre- and post-retrieval optimization methods halfway through (b). As noted on the right (c), modular RAG builds on previous approaches by enabling the addition or deletion of functional modules as required, en-



(a) Naive RAG

(b) Advanced RAG

(c) Modular RAG

Figure 1.2: Illustrative comparison of the three RAG paradigms

hancing the system's versatility. Its method has evolved beyond sequential retrieval and production to include techniques like iterative and adaptive retrieval (Qu et al., 2025).

1.7 Hugging Face's impact on medical applications of artificial intelligence

Artificial intelligence (AI) is a broad field of computer science that focuses on the investigation, creation and evaluation of methods and software that enable machine perception. The Hugging Face architecture and tools are based on Python. To support medical practitioners

An expert like a computational specialist is required to comprehend this programming language and AI procedures. This will make it easier for clinical physicians to create task-specific and personalized models. In other words, by inverting and speeding up the development process, this platform has made AI more accessible and collaborative, hence opening it up to everyone. It has simplified a convoluted and disorganized procedure for all parties involved.This modification makes it easier for experts to create sophisticated AI models while also enabling novices and AI enthusiasts to get started with little difficulty. Hugging Face is therefore more than simply a tool; it represents a significant advancement in AI.

1.8 Lang Chain

Using the latest developments in natural language processing, LangChain offers a novel approach to developing applications. By offering user-friendly chains and modular components, it enables developers to produce unique applications that integrate with language models and multiple data sources (Ioannidis et al., 2023).

Data knowledge: The ability of LangChain to connect language models with various types of data is one of its many wonderful features. This capacity to comprehend and manage data from many sources greatly aids apps in improving the precision and utility of the model's output.

Agentic interaction:A language model must interact with its environment in order to be truly used. This implies that it can respond to user inquiries, communicate with other programs or systems, and carry out actions in response to user commands.

Modular components: Flexible app coding is made easier by modularity. Developers can modify and expand LangChain's user-friendly building pieces to suit the needs of their projects.

Standard chains: These pre-built chains facilitate developers' first steps and speed up the completion of routine activities. When it comes to expanding and modifying these networks to meet project requirements, LangChain truly excels.

1.9 Motivation of the Research

Smart digital assistants are becoming more and more crucial in contemporary healthcare systems because to the increasing complexity of healthcare data and the urgent need for precise, current, context-specific medical information. Despite their effectiveness, traditional LLMs have drawbacks due to their dependency on static training data, which may result in outdated knowledge and false positives. Retrieval-Augmented Generation (RAG) is a revolutionary technique that continuously queries external medical information sources to enhance the caliber and use of generated content.

The increasing availability of sophisticated NLP models via platforms like Hugging Face and the modular, developer-centric architecture of libraries like LangChain present a special chance to develop an intelligent, conversational healthcare chatbot. By bridging the knowledge gap between patients or providers and clinical data, such a system can improve health outcomes and decision-making.

1.10 Problem Statement

The demand for easily available, trustworthy, and up-to-date medical information has skyrocketed in recent years. However, the majority of AI chatbots used in the healthcare sector today are ill-equipped to maintain contextual knowledge across multi-turn exchanges and frequently produce inaccurate or generic data. These flaws lead to decreased user confidence and potential harm from providing false information. Additionally, standard NLP models struggle to retrieve current, fact-based information and struggle to handle domain-specific terminology. Using Retrieval-Augmented Generation (RAG), Hugging Face Transformers, and LangChain, this research addresses the pressing need for a more intelligent, contextual, and dependable AI medical assistant by creating a chatbot that can deliver accurate, customized, and contextually relevant medical responses in a conversational format.

1.11 Novelty of the Study

This project is special because it builds a health chatbot using Hugging Face Transformers, Retrieval Augmented Generation (RAG), and LangChain architecture. The chatbot can carry on a context-based conversation, make multiple turns, and provide accurate information based on medical evidence. In this work, we examine retrieval and generation as a whole rather than utilizing generative models or static retrieval, which are now commonly used by healthcare assistances, in order to more effectively manage the relevance and authority of responses. utilizing LangChain in conversational settings.

1.12 Objectives of the Research

- The goal is to determine how AI medical aides can function better with Retrieval-Augmented Generation (RAG).
- Our goal is to use Hugging Face Transformers to generate and interpret natural language in the healthcare industry.
- We will use LangChain, a chat framework, to enable the chatbot to manage intelligent, continuous discussions.
- Using RAG, we will develop a healthcare chatbot that provides precise, up-to-date, and pertinent responses to medical questions.
- Lastly, we will assess the system's performance in actual healthcare scenarios, paying particular attention to accuracy, relevance, and user satisfaction.

1.13 Scope of the Study

The goal of this project is to create an AI medical assistant. We'll use the Retrieval-Augmented Generation model, Hugging Face Transformers, and LangChain to construct and test it. It is limited to the healthcare sector and deals with basic medical inquiries, symptom information, and health advice, with the exception of emergency procedures and clinical testing. The project's objective is to create a working prototype that can understand medical context and react appropriately through multi-turn discourse. Using publicly accessible or generated medical data, safe and ethical testing environments are also available for assessing system performance based on answer correctness, contextual relevance, and user satisfaction.

1.14 Significance of the Study

In the quickly developing field of AI-enabled healthcare IT, this research is essential. This project will bridge the gap between the demand for context-aware, appropriate medical care and the availability of technology that can simulate human speech by combining state-of-the-art tools like Hugging Face Transformers, Retrieval-Augmented Generation (RAG), and LangChain.

- **Improve Healthcare Communication:**The study supports the development of intelligent virtual medical care assistants that are capable of comprehending and providing contextually appropriate answers to medical queries. This could improve the effectiveness and accessibility of initial medical advice, particularly in remote locations with little access to healthcare providers.
- **Technological Progress in AI and NLP Research:**This study shows a practical use of state-of-the-art AI technology in the delicate and dangerous field of medicine using RAG and the most recent Hugging Face NLP models. It helps keep the momentum going in the direction of making AI-generated material more comprehensible and reliable.
- **Enhanced Contextual Understanding:**The chatbot can effectively handle multi-turn discussions and maintain context throughout them thanks to LangChain integration. This is particularly important in the healthcare industry because patient questions are typically complex and change as the discussion goes on.
- **Reliable and Evidence-Based Support:**By utilizing the RAG framework, which combines generative capabilities with retrieval from reliable medical sources, the system seeks to deliver responses based on accurate knowledge, eliminating the possibility of false information.
- **Foundation for Future Developments:**The research creates a foundation for the creation of AI-based solutions in the future that can function inside actual healthcare environments. The developed prototype can be used as a model for future improvements, such as real-time symptom analysis, electronic health record integration, and multilingualism.

In essence, this research represents a technical effort on the integration of AI and healthcare, as well as a step toward offering a wider community reliable, intelligent health care support.

1.15 Organization of the Thesis

A well-established structure of six divisions provides an orderly process of study in this thesis. Section 1 is dedicated to developing a RAG-based AI healthcare assistant and also defining the context of study, issue description, research objectives, scope, and significance. We discuss the background literature on artificial intelligence (AI)-powered healthcare chatbots in Section 2, summarizing its current strengths and weaknesses and the shortfalls that our research seeks to address. In Section 3, we provide the theoretical and technical foundations for the foundational technologies employed in the research. They are the RAG architecture, Hugging Face Transformers for NLP tasks, and the LangChain environment for context-dependent, multi-turn conversation management. Part 4 describes the research approach, the system design, data sources, preprocessing methodologies, and modules integrated into an operational chatbot pipeline. Employing metrics like answer accuracy, context relevance, and user satisfaction along with baseline model comparisons, Section 5 outlines the experimental setup and analyzes the performance of the system. To conclude the thesis, Section 6 summarizes the findings and proposes avenues for future work to increase the utilization and application of AI-assisted healthcare assistants.

Chapter 2

LITERATURE REVIEW

With the integration of AI into healthcare systems, intelligent and accessible healthcare has entered a new era. Many people are excited about the potential of AI-driven healthcare aides to reduce physician workloads, improve patient involvement, and expedite the healing process. From rule-based models to ones based on deep learning and machine learning for natural language processing (NLP), these systems have developed into chatbots. Hugging Face Transformers and other transformer models have significantly enhanced robots' text reading and writing skills, enabling more efficient and context-aware health care interactions.The application of retrieval-augmented generation (RAG), which blends the power of information retrieval with the capacity of generative models to provide accurate and pertinent outputs, is one of the most exciting advancements in the industry. A hybrid design like this one is extremely helpful in the healthcare industry, where accurate and evidence-based solutions are essential. Increasing the modularity, control, and real-time responsiveness of such chatbots is made possible by Lang Chain, one of the top frameworks for connecting language model activities with other information and tool sources.The history and current status of AI-based healthcare aides, in particular transformer models, RAG architecture, and Lang Chain framework, are covered in this literature review. It combines important contributions, research needs, and calls for strong, private, and transparent AI solutions for the delicate and high-stakes application of healthcare.

2.1 Introduction to AI in Healthcare :

It's exciting to think about how AI might change healthcare, especially in places with little resources. It could transform human transformation. AI is revolutionizing public health and drug research. Translation of brain signals aids in speech recovery following a stroke or other neurological issues. In a number of domains, AI subfields like as computer vision do better than practitioners. Two of the first applications of deep learning in medicine are the diagnosis of diseases and the forecasting of epidemics.More original thought was sparked by the central concept, "Transfer learning—a process that uses general datasets to create learning to specific problems." Discussions about AI in the workplace center on social transformation, economic performance, and public health. Given AI's increasing ability to automate learning, how can society shape it for the benefit of all? What effects will it have on the community, healthcare systems, and patients? This chapter discusses well-known applications, advantages, dangers, and AI concerns in the evolution of human society.

In order to create successful AI-based healthcare services, (Väänänen et al., 2021) examined key elements and offered a narrative evaluation of healthcare services that use AI-based services. AI's potential to lower healthcare costs, assist caregivers in their work, and enhance results are some indications of its advantages in the field. There is a lot of room for growth as the artificial intelligence market in the healthcare sector is growing at a CAGR of 28%.This article will examine a number of healthcare-related topics, such as care, health, and financial outcomes, and offer recommendations and important factors for the successful application of AI techniques in healthcare. It demonstrates how artificial intelligence (AI) has the ability to lower costs while improving healthcare.

(Kannelønning, 2024) AI is meant to solve healthcare issues, yet ethics, regulations, data availability, human confidence, and weak clinical evidence are challenges. To navigate this complex landscape, actors from various backgrounds and constituencies need to cooperate. An informal professional network facilitates Artificial intelligence in publicly funded healthcare in Norway. Virtual sessions and interviews are viewed by others who did not join the qualitative longitudinal case study. It opines that subsequent deployments of healthcare AI will reduce some uncertainties but potentially create others. Mobilizing spokesmen from the nondiscussing parties can fortify hybrid knowledge generation, identify, elimitate, and track uncertainties and facilitate sustainable AI deployments.

Artificial intelligence (AI) is exceeding human healthcare professionals in diagnosing some medical conditions, particularly in image analysis in radiology and dermatology, according to the groundbreaking discovery (Arora, 2020). Real-time data collection, genetic information, and a patient's medical history can all be used to train machine learning algorithms. It can be used to create materials for medical education or the incorporation of robotics. Concerns also exist regarding practical effects on healthcare services and possible implementation challenges.The "Software as a Medical Device" view of AI is attracting the attention of regulators. Algorithmic bias, data privacy, long-term staffing problems, automation bias, over-reliance, and corrigibility are among the risks. When AI crosses-examines datasets, clinicians must maintain control over the diagnostic process and understand the algorithmic processes that yield diagnoses.

(A. Kaur & Goyal, 2025) In medicine, explainable artificial intelligence (XAI) enhances AI decision-making trust and transparency. It assists patients and physicians to understand AI models' diagnosis, treatment suggestions, and forecasting. Give interpretable results to enhance collaboration among human experts and AI systems, promoting responsibility and ethics. This transparency enhances AI uptake in healthcare environments, especially for critical deciding factors such as diagnosis and treatment. XAI enhances interpretability, collaboration, accountability, and trust in healthcare decision-making, becoming more reliable and informed. Healthcare XAI is emerging and warrants further studies.

(2019, Tekkeşin) Analyzed AI aims to mimic human thought processes. Within the medical industry, a paradigm shift is being fueled by growing access to healthcare data and sophisticated analytics techniques. the potential benefits of artificial intelligence (AI) for healthcare both today and in the future. AI may help with both structured and unstructured healthcare data. Deep learning, neural networks, support vector machines, and natural language processing are popular artificial intelligence (AI) techniques for structured and unstructured data, respectively. Three main areas of AI-driven disease research are neurology, cardiology, and cancer.We closely examine the various ways AI is assisting in the treatment of stroke, ranging from early detection and diagnosis to prognosis assessment and treatment. Finally, we go into state-of-the-art AI systems like IBM Watson and the challenges of putting them into practice.

Shaheen (2021b) By anticipating, understanding, learning, and acting—from recognizing genetic code correlations to managing surgery-assistance robots—AI technologies are revolutionizing the healthcare industry. Medical applications of machine learning. Medication discovery, clinical trials, and patient care are the three areas of healthcare that are utilizing AI to transform the sector. that pharmaceutical companies may now automate target detection and speed up drug discovery thanks to artificial intelligence. AI has the potential to improve labor-intensive data monitoring techniques. AI-assisted clinical trials can confidently handle enormous amounts of data, and medical AI firms offer resources to assist patients in every manner. Clinical intelligence offers insights that can enhance patients' quality of life by analyzing medical data.

(Quaranta et al., 2024) application of AI instruments across various sectors has been raised by the new language of the Artificial Intelligence Act (AI Act). By implementing new legal and procedural limitations, A challenging problem has emerged regarding the use of AI technology, i.e. in medical uses. Existing regulations on AI medical devices, including the Medical Device Regulation, and it is important to assess how much overlap exists between these legislations and the AI Act. The overarching aim, with numerous levels of relevant rules for AI medical devices, is to make realistic requirements for the integration of AI into healthcare systems while ensuring that they comply with the law.To get an overall view of problem of applying AI in medicine, we also wish to illustrate legal short circuits involved with AI medical technologies.

(Paruvathavardhini Menaga, 2022). The study "AI in Healthcare" looks at medical analytics algorithms and technology across a number of fields. It demonstrates how artificial intelligence (AI) can extract complex medical data from books to assist physicians in making better decisions. AI is being trained utilizing technical algorithms and medical instruments to facilitate disease identification. In the domains of radiography, pathology, immuno-oncology, neurology, neurodegenerative diseases, chronic illnesses, and the creation of chemical and pharmaceutical medications, artificial intelligence is essential.Neural networks and conventional support vector machines are both examined in connection with structured data processing. The handling of unstructured data is examined in connection to deep learning and natural language processing. Sky, WebMD, Ada, and Skin Vision are a few examples of chatbots that demonstrate the application of AI in telehealth. Quick testing kits for artificial intelligence (AI) pandemics are also covered in this chapter.

(Le Moine Briganti, 2020) Artificial intelligence-powered medical gadgets are rapidly developing clinical solutions. Deep learning algorithms will be able to keep up with the growing volume of health data being gathered by wearables, smartphones, and other mobile monitoring devices. A small number of therapeutic scenarios—such as the detection of atrial fibrillation, epileptic episodes, hypoglycemia, and medical imaging and histopathology-based disease diagnoses—benefit from artificial intelligence.Patients are looking forward to better care since technology gives them more control and tailored therapy, but doctors who weren't prepared for it despise it. This problem emphasizes the necessity of considering the ethics of linked monitoring, updating medical education to incorporate digital medicine, and validating novel technologies through conventional clinical trials. Examining current research, it explores the benefits, drawbacks, opportunities, and risks that clinical AI applications present to hospitals, healthcare practitioners, academic institutions, and the bioethics community.

(Knapič et al., 2021) examined Explainable AI techniques for medical image analysis decision support. Three different explainable approaches were applied to the same set of medical imaging data in order to make the Convolutional Neural Network's (CNN) output easier to understand. The study examined stomach images captured by video capsule endoscopy in order to increase the accuracy of black-box predictions. To make the machine learning results more comprehensible, the researchers employed two techniques: SHAP and LIME. For clarifications, they also looked at Contextual Importance and Utility (CIU), another approach. According to the findings, the CIU approach helped people make better decisions since it was more understandable than LIME and SHAP.

Eskandar (2023) AI is revolutionizing healthcare, diagnosis, medication research, and therapy by being able to identify diseases in medical imaging with over 90% accuracy. As demonstrated in this intriguing scientific review, AI can be useful in a variety of ways, including improving hospital administration and surgical procedures. Although they present exciting opportunities, problems like biases and ethics must be carefully considered. Motivating case stories and compelling data illustrate AI's current impacts and promote additional collaboration in this remarkable technological revolution, moving us closer to a time when AI collaborates with healthcare providers to improve patient outcomes.

(Khalifa Albadawy, 2024) Clinical prediction is critical to healthcare for patient outcome prediction. In healthcare, By enhancing the accuracy of diagnosis, treatment planning, disease prevention, and personalized care, AI enhances healthcare efficacy and patient outcomes. In eight clinical prediction domains—diagnosis, prognosis, risk assessment, response to therapy, disease progression, readmission risk, complications, and mortality prediction—artificial intelligence (AI) enhances accuracy compared to human researchers.AI assists forecast clinical outcomes in radiology and oncology. The article underscores the disruptive in impact of AI on diagnosis, prognosis, personalized therapy, and patient safety. Enhancing data quality, multidisciplinary collaboration, ethical AI practices, AI training, clinical trials, regulatory oversight, patient engagement, and AI system monitoring and refinement are recommendations.

In 2019, Longoni et al. Healthcare is being revolutionized by AI, but it is unknown how receptive consumers will be to this technology. In both real-world and hypothetical scenarios, as well as in both individual and group evaluations, AI-provided healthcare is unpopular. Consumers are less likely to utilize healthcare, according to study 1, reservation costs are lower, provider performance is less sensitive, and automated providers are less useful, according to studies 3A–3C.Uniqueness neglect, the idea that AI providers cannot take into consideration each person's unique characteristics and circumstances, is the root cause of consumer resistance to medical AI. Medical AI is more opposed by consumers who feel unique (study 5). Medical AI resistance is mediated by uniqueness neglect (research 6), but it vanishes when AI supports (study 9), personalizes (study 7), or is consumer-based (study 8). These findings contribute to the psychology of automation and medical judgment and offer strategies for enhancing patient acceptance of AI in healthcare.

(Kaushik N. Sharma, 2025) Medical diagnosis are now far more accurate, legitimate, and reliable thanks to artificial intelligence. algorithms, review elements, criteria, quick follow-up strategies, methods for causal inference, artificial adversarial systems, and improvements in the robustness and performance of AI infection localization models. Symptomatic errors are significantly reduced when GANs that synthesize real-world tests are combined to reduce bias and improve program generality. By emphasizing how clinical performance or measurements are affected, consideration components increase interpretability and reliability. Promiscuous, individualized healing techniques are demonstrated using causal inference approaches. By addressing administrative compliance and security, unified learning enhances collaborative tutoring. Future planning must incorporate cross-models, ethical recommendations, adaptability, and interaction with emerging technologies in order to achieve complete AI control over healthcare transportation.

(Shaheen, 2021a) talked on the benefits and challenges of AI in healthcare, focusing especially on the medical and financial aspects. Intelligent data inclusion improves the quality of decision-making, surgical robots improve surgical precision, intraoperative aid through video, photographs, and communication systems is useful, and sentiment analysis can recognize and respond to human emotions. AI may also assist professionals in managing their workload so they may spend more time interacting with patients. However, there are a number of challenges, including data biases, the requirement for large datasets, potential confidentiality concerns, and the potential for patients to suffer from erroneous AI systems. draws attention to the potential advantages of new technology and the necessity of overcoming these challenges in future research to ensure the best outcomes.

(Lainjo, 2024) The application of artificial intelligence (AI) in healthcare is discussed in the text. It highlights how AI may enhance patient care, streamline procedures, and support precise diagnosis. It also discusses how AI is affecting individualized treatment and clinical decision support. Conversely, there are debates over moral and legal issues, such as patient privacy and the requirement for reliable technological systems. The significance of having high-quality data for precise decision-making is emphasized, as is the notion that patient data security requires accountability. All things considered, AI has the potential to significantly alter the healthcare industry.

2.2 Conversational AI and Chatbots in Healthcare :

In 2024, Nadarzynski et al. Finding strategies to reduce bias in the technology and make AI designs more equitable was the aim. They used a research approach that examined people's experiences to create a framework based on 17 AI usage principles. They spoke with 33 people from a variety of backgrounds, including doctors, corporate executives, and AI professionals. They developed a ten-step strategy that includes actions meant to advance equitable AI. The emphasis on collaboration and patient groups resulted in useful recommendations to enhance equity in conversational AI in healthcare.

Conversational AI is a promising technology in healthcare, claim Lal and Neduncheliyan

(2024). It facilitates the creation of tailored conversations between patients, specialists, and virtual assistants. Based on the chat, this technology looks for patterns and provides intelligent answers. A novel technique called the Generative Pretrained based Recurrent Neural Network (GPbRNN) was developed to enhance the functionality of sentiment analysis models. This software has the ability to drastically change the healthcare system by delivering personalized data, enhancing therapy delivery, and expediting decision-making. Better patient outcomes and service delivery should result from increased investment in healthcare R&D.

(Milne-Ives et al., 2020) Growing demand for such services, coupled with advances in AI, has prompted the development of conversational bots to help carry out healthcare-related tasks. Agents could automate routine operations, increase individuals' exposure to healthcare, and enable doctors to focus on more complex cases. Thirteen articles on the subject of healthcare conversational bots with free natural language processing have been assessed since 2008. Thirty percent of the studies were positive or had mixed effectiveness, while 27 out of 30 trials were good in terms of usability and 26 out of 31 trials were good for satisfaction. User opinions differ regarding quality, however. The efficacy of the agents' health care must be evaluated well and areas for improvement must be determined based on improved research design and reporting, since some trials were substandard. Further research into agent privacy, security, and cost-effectiveness is needed.

(J. Gupta and others, 2022). In recent years, healthcare organizations have been using more advanced conversational AI systems. The primary function of automated AI systems is to improve the caliber of human-computer interaction through interfaces. The profound change in the healthcare industry brought about by conversational AI is having an impact on both patients and doctors. Conversational AI uses natural language processing (NLP) systems that leverage NLUs, such as IBM Watson, Google Dialogflow, and Rasa. The Google Dialogflow architecture was used to create Ainume, a potent conversational AI agent, on the Google Cloud Platform (GCP). Before recommending nutraceutical treatments to alleviate the symptoms of common and chronic ailments, Ainume assesses them. cardiac problems, a field in which Ainume excels.

(NV et al., 2023) Healthy individuals are healthy physically, mentally, and socially. Chatbots have found numerous applications in this field, but there is still scope for novel implementations. Healthcare conversational AI applications are versatile and sector-specific. They can be used by patients to learn more about their condition, possible treatments, and coverage. which such healthcare chatbots can bring patient joy and reduce wait times, leading various companies to investigate with them. Healthcare chatbots offer various advantages, such as monitoring, anonymity, personalisation, and face-to-face interaction. This case study uses the input provided by users of patient symptoms to determine the likely type of disease. A specialist physician will be referred to the patient according to the type of disease and suggested actions. Symptoms were retrieved based on a sequential model, and KNN was employed in predicting the patient's disease type.

Meshram et al. (2021) assert that during the past 10 years, a number of technological advancements have been made possible by the growth of technologies like artificial intelligence (AI), big data, and the internet of things (IoT). These technologies have a wide range of applications. A program called "Chatbot," sometimes referred to as "Chatterbot," is one example. Chatbots are conversational AIs that mimic human speech. The core of the approach is combining AI with NLP. By automating repetitive tasks and lowering the need for human involvement, chatbots have advanced technology.Chatbots are used in a wide range of industries, including academia, medicine, and business. discussed the many types of chatbots and their advantages and disadvantages as part of the study in a number of articles. According to the assessment, chatbots are widely applicable because of their precision, independence from human resources, and 24/7 accessibility.

80% of common, minor ailments that cause 60% of doctor visits can be easily managed at home, per research by Bhirud et al. (2019). Chatbots are used to deliver healthcare services, however they are unable to engage in genuine conversation with humans and can only respond to general healthcare FAQs. In order to improve communication and create a virtual companion, efforts are being made to enable chatbots to speak like humans. Commonly developed chatbots can accomplish this by integrating Natural Language Processing (NLP), Natural Language Understanding (NLU), and Machine Learning (ML) approaches. This study discusses the healthcare chatbot system and compares the various NLU, NLG, and ML algorithms that ought to be applied.

D. Sharma and associates, 2022 AI has existed for about half a century. Processor power, data accessibility, and algorithm improvements have all contributed to the advancement of AI. Chatbots driven by AI mimic user interactions. Chatbots cut down on consultations, appointments, and hospital wait times.

meetings, assisting individuals in rapidly locating the right physician. Chatbots relieve the strain on medical staff by lowering hospital stays and unnecessary procedures while also providing suggestions and alerts. Chatbots in healthcare, however, come with a number of challenges. This article does a systematic review of healthcare chatbot research. serve as a research guideline for chatbot creation in other domains and include broad details regarding the application type, technologies, and evaluation techniques utilized to assess healthcare chatbots.

The objective of this study is to evaluate health care chatbots using the same technological characteristics that have been utilized in prior studies (A. Abd-Alrazaq et al., 2020). The study made use of seven bibliographic databases and reference list checking. Across 65 included studies, 27 technical factors were evaluated, including usability, classifier performance, speed, response creation, response comprehension, and aesthetics. The technical measurements, which were diverse, were dominated by survey designs and global usability assessments. The lack of objective criteria and standardization makes it difficult to evaluate the efficacy of health chatbots, which could obstruct advancement. The study recommends that researchers employ metrics derived from conversation logs more frequently and develops a framework of technical measures with recommendations for specific scenarios for their inclusion in chatbot studies.

2.3 Transformer Models in Natural Language Processing

(Kalyan and others, 2021). GPT and BERT, two transformer-based pretrained language models, have demonstrated excellent performance on a variety of language tasks. They use a self-supervised approach to learn from vast amounts of text data, which improves their comprehension of language and allows them to apply that

information to novel issues. This survey examines T-PTLMs in further detail, dissecting important ideas like as their training, techniques, tasks they do, and adaptability for particular applications. It concludes with recommendations for more study as well as benchmarks for evaluating these models and identifying useful libraries for using them.

2.4 Retrieval-Augmented Generation (RAG) Framework

(Kalyan et al., 2021) Transfromer-based pre-trained language models, such as GPT and BERT, have achieved excellent performance in a variety of language tasks. They learn through an enormous amount of text data in a self-supervised manner, thereby gaining a better understanding of language and using that insight to solve new problems. This survey brings T-PTLMs into sharper focus, dissecting important ideas such as how they're trained, their approach, the work they do, and how they can be adapted for particular purposes. It concludes with some testing benchmarks for these models and indicates useful libraries for using them, as well as ideas for future research.

(Wolf et al., 2020) Examined how advancements in model pretraining and design have played an important role in recent advances in natural language processing. Transformer topologies have facilitated the development of models with greater capacity, and pretraining has made it possible to efficiently use this capacity in a variety of applications. To make these technologies available to the wider machine learning community, Transformers is an open-source library. A typical API unifies the libraries meticulously designed cutting-edge Transformer structures. A carefully chosen set of pretrained models created by the community and made accessible to everyone supports this library. Transformers are designed to be easily expanded by researchers, quick and reliable in industrial settings, and easy for practitioners to use.

(Chernyavskiy et al., 2021) Modern neural architectures such as BERT and large-scale pre trained models like Transformer have greatly enhanced Natural Language Processing (NLP). New models such as XLNet, RoBERTa, and ALBERT may have been introduced, but they still can't handle all types of data or represent all types of data to the same extent as older models. Focusing on the theoretical constraints of pre-trained BERT-style models based on Transformers is the primary goal of this research. This paper proves that by setting segmentation and labelling job constraints on four datasets, the performance of XLNet and vanilla RoBERTa models can potentially be significantly enhanced. to inform the construction of future deep NLP architectures, we will first propose mechanisms for making the Transformer architecture more expressive.

(Kim Awadalla, 2020) One of the cutting-edge approaches to NLU tasks is the transformer model. Models improve over time at a very large diversity of tasks. Transformer models are computationally difficult because they have slower inference time than standard processes. Best inference-time performance of Transformer-based models for NLU tasks is obtained with the assistance of Fast Formers, a collection of recipes proposed in this work. With numerical optimization, structured pruning, and knowledge distillation, we demonstrate that inference efficiency can be greatly enhanced. Pretrained models and optimal configurations for natural language understanding tasks can be easily obtained with the help of our superb recipes. compared to baseline CPU models, we have a 9.8x to 233.9x speedup when using the Superglue test's recommended recipes. The approaches presented can boost GPU speed up to as much as 12.4 times. With the utilization of Fast Formers on an Azure F16s v21 instance, 100 million requests' handling cost can be lowered from 4,223 USD to 18 USD. An eco-friendly runtime was experienced, as per SustaiNLP 2020 shared task metrics, with an energy saving of 6.9x - 125.8x.

As per (Canchila et al., 2024) the application of human language in computer systems, termed as NLP, becomes increasingly significant in various fields like research, daily life, trade, and entrepreneurship. Various IT companies invest in NLP model, approach, and product development. Open-source contributions to the technology are also on the rise. With all developments, it may be difficult to understand the current status of NLP and the best models. To help individuals go through the constantly changing world of NLP, they have gathered a comprehensive description of recent work and accomplishments.

(Turner, 2023) Components of neural networks, i.e., the transformer, can potentially learn useful representations of sequences of data. The transformer has been the driving engine that has led developments in computer vision, spatio-temporal modeling, and natural language processing in recent times. The mathematical descriptions of architectural and design insights are too often short of introductions to transformers. 1 Furthermore, the winding course of study can offer new accounts for transformer parts. The aim of this essay is to describe transformer construction in terms that anyone can comprehend without sacrificing mathematical accuracy. As training is to be expected, they will not mention it. They assume you know your path through linear transformations, multi-layer Perceptrons, softmax functions, and the basics of probability and machine learning.

(Xiao Zhu, 2023) Empirical models of natural language processing have been dominated by transformers. Here, they present the basic concepts of transformers and point out significant methods that have helped in their recent development. a series of model enhancements, common applications, and a description of the basic Transformer design. They are unable to go into every component of the model or address every technical aspect since Transformers and similar deep learning approaches may be developing in previously unimaginable ways. Rather, we only concentrate on the ideas that are essential to comprehending Transformers and their variations. In order to provide some knowledge of the pros and cons of such models, but they also present a summary of the key concepts shaping this subject.

(Min et al., 2022) One of the most widely used AI methods to date with some success in graph modeling structured data is the Transformer model. A comprehensive review of the literature and a systematic evaluation of Transformer variants for graphs are, however, lacking. Graph Transformer models are comprehensively examined herein from the perspective of architectural design. There exist three typical ways of incorporating graph knowledge into the Transformer: Graph-enhanced Attention Matrix, Graph-enhanced Positional Embedding, and GNNs as Auxiliary Modules. Moreover, the study compares these components on popular graph data benchmarks, showing the merits of Transformer's current graph-specialized modules.

(Pol et al., 2024) Hugging Face is a major contributor to the global AI and natural language processing community. They're famous for their open-source solutions that enable users to develop and utilize sophisticated NLP models. This study delves deeper into Hugging Face, zooming in on their core technologies, such as the Transformers library, and their vision of democratizing AI for everyone. It discusses different use cases of Hugging Face models and how they are capable of enhancing productivity and creativity in fields such as healthcare, finance, customer service, and education. The article also talks about what's in the future for Hugging Face in the AI and NLP world, their great community, and how they are integrated with other tech. At last, it concludes by discussing Hugging Face's future and how it will affect AI and NLP.

(Pourkeyvan et al., 2024) It can actually aid recovery and avoid serious complications later if mental health conditions are found early on. This study examines how language models and social media can aid in predicting indicators of mental illness from what people post on them. We contrast and compare four distinct Hugging Face BERT models to well-validated machine learning techniques used in automated depression detection research currently being conducted. that, with an accuracy rate of up to 97%, the new models outperform the previous strategy. After studying the data, we find that even small pieces of information, such as user biographical descriptions, can be used to predict mental illnesses, confirming previous findings. that social media information is an excellent means of screening mental health conditions, and that such valuable work can be effectively automated with pre-trained models.

(Chhabra et al., 2023) price forecasting adjustments, product upkeep, and additional managing the sales and marketing division of the company, sentiment analysis (SA) is necessary based on the views of the customers. Various ML models have been trained with dataset variability as their foundation. When two firms achieve similar work, they typically view this as an advantage. But the time constraint of training data forms a gap area. Social media plays an important role in determining the degree of SA's practicability. A concise case study is provided below, utilizing the Python transformer library and a pre-trained model to gather data in a more economical way. Similarly, the same is achieved with traditional machine learning methods, but the accuracy differs with the dataset and structure of the goal. SA now also has an additional, more precise platform due to Hugging Face.

(G. Kaur et al., 2024) examined the performance of four popular Python sentiment analysis libraries—Text Blob, Vader, Flair, and Hugging Face Transformer—when it comes to detecting the polarity and intensity of emotions in love letters. We examined 500 sentences out of 300 love letters. Human specialists evaluated the accuracy and quality of sentiment annotations. Low to moderate agreement was represented by Cohen's

Kappa scores, and each tool revealed different strengths in handling emotional intricacy. The research also identifies gaps and proposes novel approaches for evaluating sentiment analysis techniques in romantic letters. The findings contribute to the growing field of sentiment analysis and offer insights for developing more suitable natural language models for sensitive and personal areas.

(Chow et al., 2024) explained how conversational AI with Natural Language Processing (NLP) can transform the future of artificial intelligence (AI) and medicine. This study examines Large Language Models (LLMs) and touches on a number of different topics. It begins with an overview of conversational AI and healthcare. It then reviews some of the fundamental natural language processing methods and how they assist in building improved conversations within healthcare environments. We will also discuss the progression of LLMs within NLP models, including advantages and limitations of utilizing these models within healthcare. Starting with systems that support healthcare practitioners to tools concentrating on patients, like diagnostic and treatment recommendations, applications appropriate to healthcare issues are outlined. Patient confidentiality, moral considerations, and compliance with legislation are among the legal and ethical issues discussed. Implications based on the paper identify the revolutionary possibilities of LLMs and NLP to transform healthcare interactions, while acknowledging existing challenges and anticipating future developments.

(Nerella et al., 2023) The ubiquity of artificial intelligence (AI) across society, particularly in healthcare, is bringing a revolution in numerous applications in the Transformers neural network model. While it was invented to solve issues related to natural language processing (NLP), a deep learning model called Transformer has found applications in other areas, such as healthcare. Some of the numerous types of data covered in this survey are those relating to physiological signals, biomolecular sequences, social media, electronic health records, and medical imaging. Some of the applications for these models are drug and protein manufacturing, reconstruction of data, clinical diagnosis, and generation of reports. We used PRISMA guidelines to locate studies that were applicable. Computation cost, model interpretability, fairness, ethical issues, and environmental sustainability are some of the advantages and limitations of transformer application in healthcare.

(Nerella et al., 2024) Various sectors, such as healthcare, are adopting the rapid transforming Transformers neural network model, which was initially developed for NLP applications. Clinical natural language processing, electronic health records, social media, biophysiological signals, and biomolecular sequences are some studies that have utilized this design. Critical care adverse outcome prediction and generation of surgery instructions are two more domains where it has been of use. Clinical diagnostics, report generation, data reconstruction, and protein and drug production are some of the numerous uses of transformers. But issues such as calculation of cost, model interpretability, fairness, ethical issues, and environmental impact need to be addressed.

(Bird Lotfi, 2023) examined how chatbots can be employed to assist depressed or anxious individuals. The research identifies an effective hyperparameter set by topology optimisation that is capable of predicting tokens at the rate of 88.65% and 96.49% and 97.88% for the correct tokens. As much as there is stigma about seeking help, the study establishes how chatbots can offer easy, anonymous assistance. There has to be an acknowledgement of the limitations and challenges of employing chatbots to help with mental health, and further studies are suggested in order to best understand both its possibilities and limitations, ensuring its development and application in a responsible and ethical way. (Y. Zhang et al., 2023) ChatGPT has immense potential in healthcare. It's one of the quick increase in AI. Although it might make healthcare more efficient and enhance such aspects as education and diagnosis, there are certain issues that still have to be addressed, including accuracy, privacy, and ethics. In the future, research would have to concentrate on increasing the performance of the model, addressing holes in data, and ironing out copyright and ethical problems. In this manner, AI in health care can be more productive.

2.5 Role of Lang Chain in Building Intelligent Chatbots

(Kanayo et al., 2024) Assessed building performance, post-occupancy evaluation (POE) is necessary; however, traditional approaches are challenged by data quantity and missing personalization. Manual assessment involves significant resources, and the approaches utilized today only offer general insights. These limitations are solved by Energy Chat, which is an AI chatbot utilizing Lang Chain and advanced NLP techniques, including a pretrained ChatGPT model. It offers UK households personalized energy usage advice through interactive conversation. However, there is currently no multilingual assistance available for Energy Chat's audio feature. The effectiveness of Energy Chat in promoting sustainable behavior is supported by user trials, which report a high degree of accuracy in intent detection (89%) and entity recognition (93%).

(Pokhrel et al., 2024) This research utilizes large language models (LLMs) to provide a sound foundation for the creation of customized chatbots to perform document summarization and user query management. The users are able to fight against information overload through the assistance of the system, which effectively extracts knowledge from long papers using technologies such as as OpenAI, Lang Chain, and Streamlet. This work investigated the design, implementation, and real-world deployments of the framework focusing on how it can improve productivity and ease information lookup.This research has demonstrated how the framework can be utilized by developers to build end-to-end document summarization and question-answering applications through a step-by-step guide.

(Bogusz et al., 2024) provided access to AI-driven question and answer chatbots able to learn query languages in order to retrieve pertinent information based on user context. Through a collection-based interface, the web app will utilize user-uploaded papers to contextualize language learning model's reaction to input from the user. The RAG pipeline is applied for this use. The language learning method reminds users when resources are lacking to deal with an issue effectively, promising precision and customer satisfaction. With an interface similar to popular AI chatbots such as Claude by Anthropic or ChatGPT by OpenAI, this software mainly seeks to make collection administration possible by allowing users to upload, delete, and select some collections. A landing page, login, and document uploading gateway are all included in the product's last features, which enable users to create document collections.

(Mavroudis, 2024) says according to the framework named Lang Chain, it is simpler to build, generate, and deploy applications that utilize large language models (LLMs). It offers resources for managing conversation models, including RAG, and making secure API communications. Its modularity and integration, though, complicate it and introduce potential security concerns. Its architecture and core components, including Lang Graph, Lang Serve, and Lang Smith, are discussed in this research along with its applications in numerous domains. It evaluates its usability, security, and scalability limitations. For developers and scholars seeking to leverage Lang Chain for artistic and secure LLM-based applications, this piece is a invaluable resource.

(Mahadevan Raman, 2023) Automated Essay Score (AES) is innovative technology. Methods of scoring are used to accomplish many functions. There are dependable scores that have been derived from important considerations. These considerations can be calculated using domain-based methods. Our research is interested in what the user has learned concerning a subject. Utilizes the score index of Large Language Models. Enables users to compare and contrast how much they know a new subject. used in learning analytics and boosting learning capacities. is interested in summarizing PDF files and quantifying user comprehension. The process entails the use of a Lang chain instrument to summarize and extract key data from the PDF. The study uses the approach for measuring user comprehension of summarized information.

(Easin Arafat et al., 2023) The combination of LLM, Lang Chain, and SAP HANA is explained in this research, highlighting the natural applications and benefits of each. It presents a methodology for utilizing these elements relevant to the developmental phase of an organization that fosters long-term growth and optimal operation. The integrated framework is a game-changing tool for real-time intelligence and decision-making across diverse industries since it provides linguistic accuracy, frictionless language-technology integration, and robust analytics infrastructure.

(Jay, 2024) Discussed how to develop generative AI tools using Lang Chain and LLMs, one of the most popular platforms used to develop LLMs-based generative AI tools. Learn how to get access to vast stores of information in these superhuman giant language models, or LLMs for short. Together we will test the feasibility of accessing strong LLMs such as GPT-4, Palm, and Gemini with Lang Chain for creating some fantastic, smart, and practically useful apps with near-human character.

(Wagner et al., 2022) The lives of elderly people may be enhanced and the costs of healthcare may be minimized through the utilization of healthcare-targeted monitoring systems. With the aging of the global population, both advantages are assuming greater significance. Studies of the application of different sensors in monitoring humans continue; however, impulse-radar and depth sensors emerge as the most promising because they can provide medically helpful information without requiring the monitored person to wear or use any device. In addition, the use of these sensors intrudes less into one's privacy compared to the use of video cameras. To approximate healthcare informative measures, processing data from these sensors involves establishing the position and velocity trajectories and post-processing them.

(Y. Xu et al., 2022) presented a mobile QA system for smart cities that is aimed at healthcare and utilizes mobile computing and artificial intelligence. Classifier, QA engine, and chatbot API are all included in the system. The system utilizes a range of classifiers, including AdaBoost, support vector machines, and neural network-based classifiers. Semantic processing and answer retrieval modules make up the QA engine. Hospitals and communities in real life has tested the technology, and the user interface is good. The promise of mobile QA in the health field has been shown with successful testing of the system with actual-life hospitals and communities.

(Yi et al., 2021) posited the use of inertial measuring unit (IMU) and wearable electromyography (EMG) sensors in the prediction of basic gait information, including the lower-limb kinematics and kinetics. For the purpose of ongoing prediction of lower-limb angles, a new algorithm is trained to use long short-term memory (LSTM) for extracting information. Each segment's IMU signals and nine muscles' EMG signals of the lower limb are used in training the regressor. The results of the experiment validate the calculation accuracy of the kinetics and angle predictions and also provide the best time to make predictions. This paper proves that by real-time prediction of kinematics and kinetics, core gait data can be obtained soon and accurately for intelligent healthcare.

2.6 Comparison with Traditional NLP and Rule-Based Systems

(Topaz et al., 2019) Human-aided text mining of clinical narratives becomes possible due to the open-source, rapid clinical text mining tool Nimble Miner, which incorporates machine learning algorithms. In order to test the system, it was trained on data from a large US-based homecare organization, which contained 1,149,586 notes for 89,459 patients. Almost all of these measures of fall identification, such as total fall history, risk, fall prevention efforts, and falls occurring within two days of the note date, demonstrated that the system performed better than a rule-based NLP method. Furthermore, for the fall season, the rule-based method did somewhat better during the initial two weeks following the note date. For departments such as allied health and nursing that have not had made great strides in natural language processing, yet their results indicate that clinical text mining can even be used without big labeled datasets.

(Pirinen, 2019) examined recent neural language modelling results for Finnish with emphasis on common tasks. cite a new comparison of neural methods with traditional rule-based systems for the provided tasks, as the majority of common tasks are based on supervised learning. I re-examine shared task results, such as SIGMORPHON 2016 morphological regeneration, CONLL 2018 universal dependency parsing, and a German copy of WMT 2018. Finnish is the Uralic language used throughout. use best-performing neural and rule-based models and examine their results.

(X. Xu Cai, 2021) The task of this project is to automate analyzing utility legislation using an ontology and a rule-based natural language processing system. This approach makes application of two newly developed ontologies: UPO (urban product ontology) and SO (spatial ontology). Unlike UPO's domain-specific conceptualization of concepts and semantics capture, SO's two layer semantic structure enhances our understanding of spatial language. Deontic logic clauses are employed for logical and semantic formalization of data following pattern matching methods being applied to extract it. Ontologies are then employed to connect the restored information to semantic correspondences. On its evaluation on the basis of spatial configuration criteria in utility accommodation policies, the method achieved a 98.2% accuracy and 94.7% recall in information extraction, a 94.4% accuracy and 90.1% recall in semantic formalization, and an accuracy rate of 83% in logical formalization.

(Hammami et al., 2021) Clinical therapy and cancer registries are dependent upon pathology reports. Yet, manually extracting and coding unstructured information is a tedious process. One feasible alternative to human processing could be provided by Natural Language Processing (NLP) algorithms. With micro-averaged performance scores above 95%, Italian researchers aimed to build an automated system that would be able to identify and classify morphological details within pathology reports through the use of natural language processing methods. An Italian cancer hospital employed 27,239 pathology reports to evaluate a new language-based domain-specific classifier. The algorithm achieved a micro-F1 score of 98.14% with 9594 pathology reports as input. The method can be applied to other datasets, although it results from instructions by only one cancer center. To ensure it operates with a larger dataset, we must delve deeper into the subject.

(Gao, 2022) gave a technique to mine outage investigation data from power sector documents. Rule- and machine-learning-based methods are applied for this. Training and data cleaning were most critical. for blackout analysis reasons, when and where and which installations and equipment broke down.Scraping and OCRing websites generated the blackout dataset. They extracted blackout and training data for tagging using language, relation, and named entity extraction. The lexicon can be utilized to train an entity type recognition model. From research, they developed a model to extract time, place, and defective facilities from blackout notifications. Upgrading the model constantly provided the best results. analyses facility outages. this system can improve incident analysis and provide technological support for activity-specific tasks.

(F. Zhao et al., 2019) In medical image processing, the correct diagnosis of vascular disorders is dependent on vessel segmentation. Blood vessel segmentation manually is a time-consuming and expert-based process. Fully or partially automated vessel segmentation algorithms have been the focus of intensive research and development. Various perspectives of vascular segmentation algorithms were tested in earlier research. Modern machine learning techniques, especially deep neural networks, were not taken into consideration in such evaluations. classifying the prevailing vessel segmentation methods as rule-based or constructed through machine learning. To distinguish the vessel structure from that of its environment, rule based methods utilize sets of rules that are well crafted, while machine learning-based methods utilize rules that have been learned themselves from past experiences. The common blood artery segmentation techniques of the past years to present an insight into current and future trends in segmentation processes.

(Islam et al., 2020) the enormous amounts of raw data produced by the Internet of Things (IoT) and cloud services, machine learning methods struggle to make accurate predictions. Big data sets are processed by techniques of deep learning (DL); however these techniques struggle with uncertainty that accompanies them. While Belief Rule Based Expert Systems (BRBES) are used to handle ambiguous data, their inability to accommodate associative memory tends to lead to poor prediction accuracy. To improve prediction under uncertainty and allow for accurate data patterns, a new technique BRB DL integrates an associative memory-based DL algorithm with BRBES inference techniques. When tested on two datasets—power generation and air pollution—the method performed better in terms of prediction compared to current DL methodologies.

(Ray Chakrabarti, 2022) social networks have influenced communication patterns significantly; it is crucial for companies to analyze user sentiment on social media platforms. Voice recognition and image classification are becoming increasingly popular using deep learning techniques but are not generally utilized in sentiment analysis. Based on a seven-layer deep convolutional neural network (CNN), a deep learning approach to sentiment analysis and aspect extraction from text is proposed. To further improve aspect extraction and emotion scoring, the method is combined with a rule-based approach. The method also improves existing rule based aspect extraction by applying clustering to sort aspects into a predefined set of categories. a global accuracy of 0.87, a 7–12% increase over state-of-the-art methods. Based on (Van Vuuren et al., 2021) the machine learning models Lasso Regression and Random Forest were able to predict suicidal behavior in the future. Outcomes of second and fourth year secondary school students in general were utilized in the study. that although both models exhibited marginally improved prediction accuracy, Random Forest's sensitivity and specificity were marginally better. The Lasso Regression, however, had a large increase in sensitivity at the expense of specificity. The study is the first to utilize survey data from a large adolescent group to predict future suicide behavior using machine learning methods. In alignment with the study, integrating machine learning methods into screening protocols may assist in suicide prediction improvement; however, further optimisation is required.

As per (Sarker Kayes, 2020) the machine learning models Lasso Regression and Random Forest would be able to forecast suicidal behavior in the future. Data of second and fourth-year secondary students in general were utilized in the study. which though both models performed a bit better in prediction accuracy, sensitivity and specificity of Random Forest was a bit better. Nonetheless, the Lasso Regression experienced greatly better sensitivity at the expense specificity. This study is the first to employ survey data from a large sample of teenagers to forecast subsequent suicide behaviour with the aid of machine learning techniques. Based on the study, the inclusion of machine learning techniques in screening methods can improve suicide forecasting; however, further optimization is needed.

Based on (Christmann Weikum, 2024) the Quasar system, which treats all sources consistently, is introduced here in this article for question answering through structured tables, knowledge graphs, and unstructured text. With a RAG-based architecture, the system possesses a pipeline for evidence retrieval and response generation. The second component is powered by a language model of moderate size. Apart from its other functionalities, Quasar possesses components that are designed to understand questions, provide more precise evidence retrieval with, and pre-sort and pre-filter the evidence prior to delivering the most pedagogical sections to the answer generation mechanism. Our method's superior answering quality comparable to or better than large GPT models with orders of magnitude lower computational and energy usage is demonstrated with three different benchmarks.

(Guțu Popescu, 2024) Advances in technology have created an exponential growth of data, which has created opportunities in social media, healthcare, and financial industries. Sensitive information also poses security concerns and privacy issues. Through the simulation of complex data and creation of synthetic data, generative models deliver solutions and are useful in analysing enormous private datasets. The data analysis techniques based on generative models, highlighting long models of language (LLMs). Approaches such as retrieval-augmented generation (RAG) and fine-tuning of LLMs are discussed along with their advantages, limitations, and uses. Aimed at informing effective, privacy-oriented data analysis and probing imminent advances, especially for low-resource languages, the research synthesizes, assesses, and interprets findings from the literature to present an informed snapshot of the field.

(Ning et al., 2025) LLMs and FMs are being utilized for time series predictions. Fine-tuned large language model (LLM) predictions may work in some but not all contexts. Due to their lack of interpretability and failure to have domain adaptation mechanisms, time series foundation models (TSFMs) are not fit for zero-shot forecasting. TSFM's interpretability and generalizability are improved upon by a time series forecasting framework that uses retrieval-augmented generation TS-RAG. In an effort to obtain semantically meaningful time series segments from knowledge bases, TS-RAG employs pre-trained encoders and contextual patterns for each query. The following task is constructing a learnable augmentation module based on a Mixture of Experts (MoE) that can possibly enhance forecasting accuracy without task specific tuning. This module ought to dynamically combine time series patterns with the input question of the TSFM. TS-RAG outperforms TSFMs in zero-shot forecasting by 6.51% across different domains and is interpretable with ease on seven publicly available benchmark datasets.

2.7 Research Gap

Despite considerable advancements in AI-based health assistants, considerable gaps remain in developing and employing robust, real-time, and dependable systems. Most existing AI health chatbots lack functionality with no context-generating capabilities, dynamic memory-based retrieval of current medical knowledge, and understandable reason paths. Although integration of transformer-based NLP models (e.g., BERT, GPT) has imparted linguistic power to these systems, their reliability in clinical situations—where truthfulness and patient safety are of prime concern—remains understudied. Retrieval-Augmented Generation (RAG) is a promising solution based on the synergy between generative capacity and dynamic information retrieval. Yet in medicine, its use is in its early stages and largely hypothetical, with few practical applications having consistent accuracy, user trust, and adaptability to multilingual or multicultural settings. Moreover, the ability of Lang Chain to enable multi step interactions and incorporate external tools is not yet utilized in medical purposes, particularly in integrating EHRs, symptom databases, or clinical decision-making systems. The practical, there has been a shortage of extensive examination of the ethical and legal consequences of using these AI agents in healthcare settings. To overcome such deficits, this thesis develops a RAG-based chatbot using Lang Chain and Hugging Face Transformers. It is designed to provide context-aware, explainable, and real-time medical help. Doing so, it provides a new face for AI in health care with reference to its technical, practical, as well as ethical dimensions.

Chapter 3

METHODOLOGY

Retrieval-Augmented Generation (RAG), which is described in the technology section, was used to build this AI healthcare assistant. The system supports accurate and illuminating medical question-answering functionality by utilizing vector databases, embedding models, and sophisticated natural language processing building pieces. From data collection to model deployment, every stage of the pipeline has been carefully selected and adjusted to optimize explainability, scalability, and performance. This section outlines the sequential pipeline, explains the rationale behind each technical choice, and provides details on the architecture's workings, including data processing, embedding creation, model building, and evaluation.

3.1 Data Collection

The dataset is the foundation of every data-driven natural language processing (NLP) system supporting semantic comprehension. This study used information from The Gale Encyclopaedia of Medicine (2nd Edition), a freely accessible medical resource. This 759-page An encyclopedia publication's PDF version includes a vast amount of medical information, covering everything from pharmacological details and treatments to terms, symptoms, and procedures. This makes it the perfect corpus for creating an AI chatbot with a healthcare focus. The document, which contains unstructured text data, was obtained in its original PDF format. This dataset's vast domain-specific content, which covers everything from fundamental ideas to sophisticated methods, had a significant role in its selection.

The goal of this project was to develop an AI-powered healthcare assistant that can communicate with users in normal language and provide medically accurate responses by using a source that provides reliable, peer-reviewed knowledge. A number of problems common to typeset or scanned documents were present in the raw PDF layout. The elements included bibliographies, comments, page headers and footers, and irregular formatting, all of which required appropriate parsing before being processed further. Only pertinent information was contributed to the pipeline for model training and inference thanks to formatted extraction techniques, which allowed for systematic access to the document semantic content.

3.2 Data Description

The Gale Encyclopaedia of Medicine's PDF edition was converted into a formal text corpus. From simple definitions (such as ”anemia” and ”diabetes”) to in-depth entries on clinical trials, surgical methods, pharmaceutical therapies, and bioethical concerns, the text's hundreds of medical entries are arranged alphabetically.

A single unified corpus was produced when the complete document was subjected to computer text extraction algorithms. To enable later tracing, which is essential in healthcare settings where source traceability is crucial, each input and the metadata that goes with it (such as the page number) were saved. Because of its inherent characteristics, which include a large vocabulary and technical terminology, the text was particularly well-suited for embedding-based semantic search. Because of the medical coherency of each text, embedding models were able to acquire representations that were contextually relevant. Furthermore, a chunking technique in line with entry boundaries and titles was naturally made possible by the structured entries.

3.3 Data Preprocessing

Any strong pipeline for machine learning or language models must start with data preparation and preprocessing. In order for downstream components to use the raw input data from a comprehensive medical reference in PDF format, it must first go through a number of methodical transformations. Textual data extraction, segmentation into manageable chunks, conversion to dense semantic vectors, and storage within an optimized similarity search index are all part of the preprocessing pipeline. Every stage is strategically planned for semantic consistency, effective retrieval, and adherence to the limitations and capabilities of the selected large language model, in addition to being technically essential. This section goes into great detail about each phase in the preparation process, offering both theoretical support and helpful implementation tips.

3.4 Raw PDF Text Extraction

The first step in the data preparation process was to extract the source document's raw textual information. Regular processing was not possible without first transforming the document into a machine-readable format because it was in PDF format. Text from each page was read and extracted using the Python tools PyMuPDF and pdfminer, ensuring that all necessary metadata, including page numbers, was retained. This was essential for later phases of traceability and interpretability so that the model could produce source-aware outputs.

3.5 Chunking Strategy and Overlap

To improve retrieval performance and enable semantic indexing, the retrieved textual material was then divided into equal-sized parts. We decided on a 500 character chunk size and a 50 character chunk overlap. Chunking is used because input truncation at the token limit of the majority of large language models (LLMs) may result in information loss. Important contextual transitions between adjacent sections are preserved because

to the overlapping segment, which is particularly crucial in medical texts where good comprehension depends on contextual continuity. Let C_i represent a bounded piece in this way: $C_i = T[k : (k + 500)]$ considering k = 0, 450, 900,
This maintains continuity between segments by guaranteeing that each subsequent section begins 50 characters before the conclusion of the preceding segment.

3.6 Semantic Embedding Using Sentence Transformers

After chunking, each text passage was converted into a vector representation of a predetermined length using Hugging Face's Sentence Transformers' "all-MiniLM-L6-v2" model. The transformer architecture, which forms the basis of the model, produces dense vector embeddings that are superior than sparse representations like TF-IDF or bag-of-words in terms of capturing semantic content. The embedding model receives each chunk C_i to produce a vector $v_i \epsilon \mathrm{R}^d$, where d = 384 is the embedding dimension:

$$v_i = MiniLM(C_i)$$

Better similarity matching is made possible during the retrieval process by these embeddings, which also maintain contextual semantics.

3.7 Model Building

The RAG pipeline, which is at the heart of an AI-powered healthcare assistant, integrates semantic vector search and reasoning from massive language models to provide precise and contextually relevant medical replies. First, there is the embedding generation mechanism; second, there is the FAISS-based semantic retrieval engine; third, there is the large language model (LLM) for creation; and last, there is the RetrievalQA chain, which integrates all of these components into a unified end-to-end system.

3.8 Embedding Storage and Similarity Search Using FAISS

Facebook created FAISS, a specialized vector database, to enable efficient nearest-neighbor searches in high-dimensional vector spaces. The vector embeddings that were discovered from the text spans are shown here. FAISS was chosen in large part because to its scalability and performance, especially for approximate nearest neighbor (ANN) searches employing methods like Inverted File Index (IVF) and Hierarchical Navigable Small World graphs (HNSW). Cosine similarity is used to compare the vector v_q of a user query q with each stored chunk vector v_i.

$$cosinesim(v_q, v_i) = \frac{v_q \cdot v_i}{\|v_q\| \|v_i\|}$$

To ensure that the model has access to the most contextually relevant information, the three most similar segments are obtained for use in the final language creation process.

3.9 Language Model Configuration and Temperature Settings Adjustment

Mistral 7B Instruct v0.3 is the selected language model for this system, and it is implemented using Hugging Face's inference API. Mistral is a sophisticated open-weight LLM that excels at obeying instructions and is renowned for its compact size and light inference. With almost 7 billion parameters, the model performs best when resource usage and the ability to comprehend and produce medically fluent language are balanced. The three most comparable documents provide the system prompt and context information that the LLM uses. In mathematics, the likelihood of a language model predicting the next token x is modified using:

$$P(x) = \frac{\exp\left(\frac{\log p(x)}{T}\right)}{\sum \exp\left(\frac{\log p(x_j)}{T}\right)}$$

In the present study, T = 0.5. This selection strikes a good balance between diversity and accuracy, complementing the assistant's pleasant yet educational tone. The temperature hyperparameter, which regulates response generation's randomness, is set to 0.5. A number of 0.5 strikes a compromise between creativity and determinism, guaranteeing that replies are varied and instructive and that no information is falsified.

3.10 Retrieval-Augmented Generation Using LangChain

The merging of many model components—retriever, prompt composer, and generator—into a RetrievalQA pipeline was optimized using the LangChain framework. Chain configuration used chain_type="stuff" which meant that the retrieved documents were combined and put straight into the language model's prompt. When the whole recovered context falls inside the LLM's token limit, this technique works well. Rapid testing and debugging were made possible by LangChain's modular design, which appropriately abstracted out intricate tool interactions. It also makes it simple to incorporate Hugging Face's Mistral model as the response generator and FAISS as the retrieval method.

3.11 Retrieval Configuration and Source Cognisance

A RetrievalQA chain, as used in LangChain, is the last layer of the architecture. The chain creates a single question-answering interface by joining the vector retriever and LLM. Using the chain_type "stuff," the top k retrieved documents are combined into a single input prompt that the LLM can process. If the input text complies with token requirements, this method is straightforward and computationally efficient. In order for users to trace replies back to their original source, which is usually the precise page number from the PDF, LangChain makes it simple to include source metadata in the response. In medical applications, where the capacity to validate and trust information is crucial, transparency is essential.

3.12 Model Evaluation

By concentrating on the system's ability to retrieve and provide answers, an extensive evaluation process is employed to determine the suggested AI-based healthcare assistant's trustworthiness and efficiency. Measuring semantic integrity and contextual coherence of generated responses, and correctness in retrieving applicable medical data from the Gale Encyclopaedia of Medicine, are the objectives of the evaluation. The RAG pipeline, which involves FAISS-based semantic retrieval and response generation with Mistral-7B and is used by the model.Model assessment involves quantitative measures such as cosine similarity, Recall, Precision,and BERTScore, along with qualitative human assessment. This dual-pronged assessment permits both technical benchmarking and human-centric assessment of response quality. Additionally, generation parameters' influence—specifically temperature—is explored to avoid hallucinations and factual truthfulness in such mission-critical domains as medicine. Comparative indexing testing between IVF and HNSW settings in FAISS reveals the trade-off between retrieval accuracy and computational expenses, which is important in developing real-time, reliable healthcare applications.

3.13 Semantic Retrieval Evaluation

The evaluation's initial phase centers on how well the FAISS-based retrieval is effective. Essentially, it is responsible for extracting the most relevant text passages from the information database. FAISS employs vector similarity with the cosine similarity measure in high-dimensional space to determine the proximity of the document and query vectors. Let $\vec{q}$ represent a user query embedding, and let $\vec{v_t}$ represent the embedding of the (i)th chunk of documents. The following formula is used to calculate the cosine similarity:

$$CosineSimilarity(\vec{q}, \vec{v_i}) = \frac{\vec{q} \cdot \vec{v_i}}{\|\vec{q}\| \cdot \|\vec{v_i}\|}$$

The mean cosine similarity μ_{cos} and standard deviation σ_{cos} of the top k scores are calculated as follows to assess retrieval consistency:

$$\mu_{cos} = \frac{1}{k} \sum_{i=1}^{k} CosineSimilarity(\vec{q}, \vec{v_i})$$

$$\sigma_{cos} = \sqrt{\frac{1}{k} \sum_{i=1}^{k} \left(CosineSimilarity(\vec{q}, \vec{v_i}) - \mu_{cos}\right)^2}$$

These steps make it possible to keep an eye on the retrieval process in order to guarantee semantic grounding and query consistency. Good retrieval performance is indicated by high μ_{cos} values and low σ_{cos} values.

3.14 Recall

Recall evaluates how well the system incorporates key information from reference materials. A score of indicates that it successfully reduces omissions while incorporating the majority of medically relevant material. For diagnostic thoroughness, high recall—which is determined by dividing true positives by actual positives—is necessary, but it must be balanced with precision to avoid giving too many or irrelevant answers.

$$Recall = \frac{|Relevant_k \cap Retrieved_k|}{|Relevant_{total}|}$$

The percentage of generated material that precisely matches reference data is measured by precision. High precision ensures fewer fallacies in medical quality assurance. It illustrates the model's ability to remove inaccurate or unnecessary data, which is crucial for clinical credibility. It prioritizes factual correctness above comprehensiveness and is calculated as true positives divided by total projected positives.

$$Precision = \frac{|Relevant_k \cap Retrieved_k|}{|k|}$$

Retrieval statistics are computed for several values of k, often 1, 3, and 5, in order to represent various levels of retrieval granularity. By contrasting and comparing two FAISS indexing configurations—the Hierarchical Navigable Small World Graph (NNSW) and the Inverted File Index (IVF)—we want to clarify the trade-offs between retrieval accuracy and computing efficiency. Large-scale deployments can benefit from IVF's reduced latency and effective indexing through quantization, whereas HNSW offers superior precision by navigating a proximity graph at the expense of increased memory usage and retrieval time. In order to provide information about their applicability for real-time health applications, the comparison also includes benchmarks on memory utilization, top-k accuracy, and retrieval latency in milliseconds.

3.15 Response Evaluation

Both human assessment and machine score are used to evaluate the response generation module. To determine the semantic closeness of generated responses to ground-truth responses, we use the BERT Score. BERT Score is particularly effective in fields with complicated terminology, like medicine, where contextualized embeddings are used in place of string-matching measurements. Using pre-trained BERT embedding, it calculates F1-scores for token-level similarity using the formula below:

$$BERTScore_{F1} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

where the semantic similarity between reference and generated tokens is measured by Precision and Recall. We run the Mistral 7B model with a lower temperature (e.g., T = 0.3) to improve factual correctness and lessen hallucinations, a typical issue in generative models. As a result of this modification, the responses became more tangible and predictable. Additionally, using a five-point Likert scale, human assessors assessed the responses for factual accuracy, relevancy to the original query, and general fluency. To ensure consistency between human and automated review, the qualitative scores were averaged and compared with the BERTScore results.

3.16 Answer Relevancy

Answer Relevancy measures how closely answers match the purpose of the query. Answers with a score close to 1 are targeted, clinically helpful, and free of distractions. It guarantees that the system accurately meets user demands by eliminating generic or off-topic outputs, which are typical of general-purpose models, and is measured by the cosine similarity between query and response embeddings.

3.17 Faithfulness

Faithfulness detects hallucinations by measuring factual congruence with source materials. A perfect score indicates that every generated claim in the retrieval corpus can be verified. It is crucial for medical AI and is assessed by comparing answers to references to ground truth. Higher chances of believable but inaccurate information are indicated by lower scores (e.g., 0.89 for GPT-3.5).

Algorithm 1 Methodology of AI-Powered Healthcare Assistant using RAG

```
Input: User query q
Output: Factual and contextually relevant answer a
Load Gale Encyclopedia of Medicine PDF
Preprocess text: clean, tokenize, normalize
Split corpus into overlapping chunks (chunk size = 500, overlap = 50)
Generate sentence embeddings using all-MiniLM-L6-v2
Store embeddings in FAISS index (IVF and HNSW configurations)
Receive input query q
Generate embedding vector $\vec{q}$
Perform similarity search in FAISS using cosine similarity:
    $CosineSimilarity(\vec{q}, \vec{v_i}) = \frac{\vec{q} \cdot \vec{v_i}}{\|\vec{q}\| \cdot \|\vec{v_i}\|}$
Retrieve top-k most similar chunks (e.g., k = 3)
Compute mean $\mu_{cos}$ and standard deviation $\sigma_{cos}$ of cosine similarity
    $\mu_{cos} = \frac{1}{k} \sum_{i=1}^{k} CosineSimilarity(\vec{q}, \vec{v_i})$
    $\sigma_{cos} = \sqrt{\frac{1}{k} \sum_{i=1}^{k} \left(CosineSimilarity(\vec{q}, \vec{v_i}) - \mu_{cos}\right)^2}$
Pass top-k chunks and query to RAG pipeline
Use Mistral-7B to generate response a (temperature T = 0.5)
If factual hallucination is detected, reduce T (e.g., T = 0.3)
Compute Recall:
    $Recall = \frac{|Relevant_k \cap Retrieved_k|}{|Relevant_{total}|}$
Compute Precision:
    $Precision = \frac{|Relevant_k \cap Retrieved_k|}{|k|}$
```

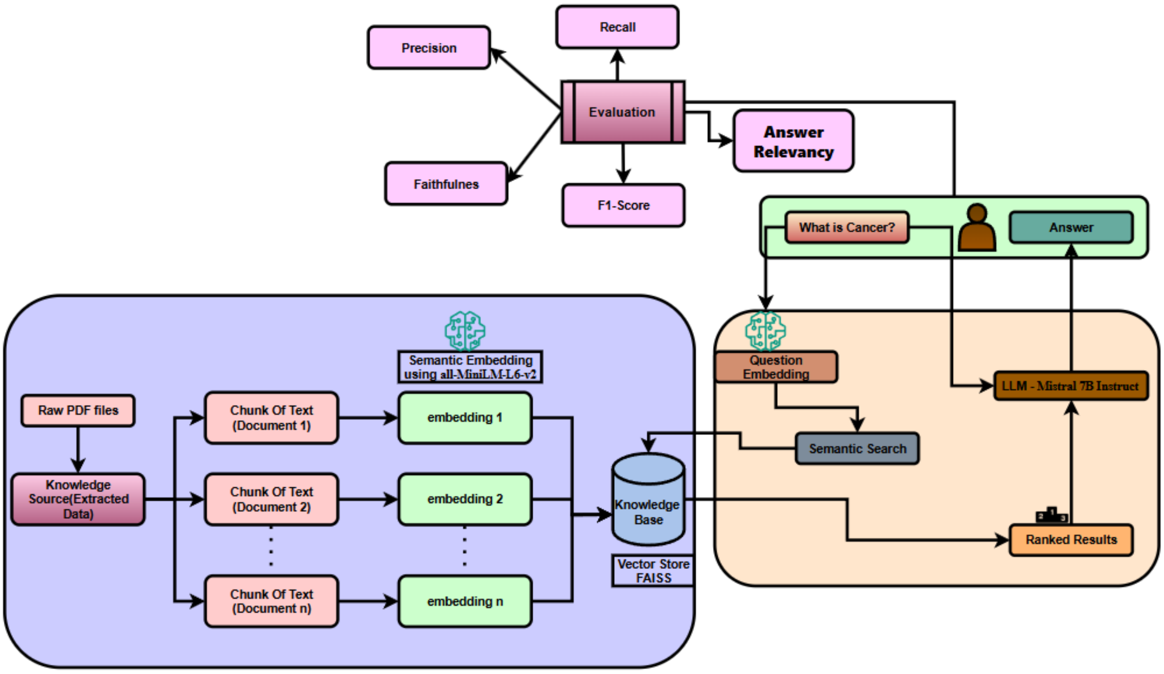


Figure 3.1: System FlowChart

```
18: Evaluate generated response with BERTScore:
        $BERTScore_{F1} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$
19: Human assessors rate answer for:
20: a) Factual Accuracy
21: b) Contextual Relevance
22: c) Coherence
23: Calculate composite Response Quality Score (RQS):
Answer Relevancy , Faithfulness
24: Compare FAISS index types (IVF vs. HNSW) on latency, memory, top-k accuracy
25: Return final response
```

Chapter 4

RESULTS AND DISCUSSION

Supported by the findings of this study, the RAG-based AI medical assistant is able to generate accurate, relevant, and comprehensible medical answers. Retrieval accuracy, semantic coherence, response relevance, and source consistency were some of the numerous measures employed to test the performance of the system in a set of quantitative and qualitative tests. To produce the most suitable information in response to a given medical question based on the context of the user, we applied cosine similarity to identify semantic similarity in their search and the retrieved document passages. High values of cosine similarity guarantee the retrieval module appropriates medically applicable passages in The Gale Encyclopaedia of Medicine, creating a healthy basis for the rest of the task of response generation. BERTScore was used to measure semantic and linguistic response quality produced by comparing model response with expert-curated references. The score reaffirmed that the system is producing responses that are factually correct and linguistically correct to a human-like explanatory level. The relevancy of response was also measured to see how far the assistant can answer a user query without going off-topic or too generic. High relevancy scores indicate that the system is staying in the medical field, creating short and factoid answers with the aim of answering user requirements

The other equally important aspect of testing was to check the fidelity of generated responses—whether the AI is not merely spouting things at random and getting off course from source material. The fact consistency is especially important due to the vast criticality of uses in medicine, and the experiments demonstrate how it stays diligent on the track of the retrieved documents and thereby steers clear of contradictions. The combination of FAISS for rapid vector search and Mistral-7B for response generation had the perfect blend of speed and accuracy, and the model was top-notch on open-ended medical questions and fact-based questions. Temperature parameter (T = 0.5) was also needed to control response diversity in such a manner that responses were informative but not overly predictable or overly creative. LangChain's RetrievalQA chain also made retrieval and generation harmonize well with one another, so the system could insert the most appropriate passages into the prompt without sacrificing source traceability an essential requirement for medical use cases where verifiability is key. With all these capabilities in its arsenal, though, the test also pointed to areas of improvement, i.e., where it is to respond to extremely specific or ambiguous medical queries where contextual subtleties could require more mystical reasoning.Future research would investigate further improving embedding model with domain knowledge or using multi-hop retrieval to better capture comprehension of intricate medical problems. The results confirm the effectiveness of the proposed RAG pipeline for providing strong, interpretable, and user-oriented medical help and making it an effective platform for health information retrieval and patient care.

High accuracy in retrieval, high semantic coherence, and high fact conformity emphasize the potential of AI-powered health assistants in facilitating the exchange of medical knowledge without undermining trust and safety needed in clinical and consumer settings.

4.1 Performance metrics

We can assess the effectiveness of large language models (LLMs) by using key performance metrics generate accurate, pertinent, and reliable results. One such metric, the BERT Score, measures the model's capacity to convey contextual information by analyzing accuracy, recall, and F1-score meaning by calculating the degree of semantic similarity between the reference text and the output. Relevance measures how closely answers match the purpose of the inquiry, ensuring that the result is both practical and environmentally friendly. Additionally, Faithfulness verifies the veracity of the facts, ensuring that model responses are devoid of errors or hallucinations. These two can be combined to create an overall model performance measure that can show the advantages and disadvantages of various learning techniques and architectures.The analysis that follows compares them to these criteria to show how well the best LLMs function in practical settings.

Model	**Prec.**	**Rec.**	**F1**	**Rel. Ans.**	**Faith.**
Mixtral-8x7B-Instruct-v0.1	0.8334	0.8119	0.8225	0.9221	1
LLaMA-2-70B-Chat	0.8212	0.8123	0.7932	0.914	0.95
Claude-2	0.8078	0.7854	0.7805	0.8641	0.93
Falcon-180B-Chat	0.789	0.7721	0.7754	0.8465	0.9
GPT-3.5-Turbo	0.7721	0.7654	0.7444	0.8341	0.89

Table 4.1: Comparison of models based on precision, recall, F1 score, relevance, and faithfulness.

4.2 BERT Score

Precision, as measured by the BERT Score, explains how closely a model's output resembles the reference text and why it is so important to avoid providing superfluous or inaccurate information. Since model outputs are usually all fairly close to projected responses, a model with a high precision score has very few false positives. In the above table, Mixtral-8x7B-Instruct-v0.1 has the highest accuracy of 0.8334, meaning that Mixtral's response is topically relevant and semantically accurate almost 83.34% of the time. This shows that Mixtral is doing a fantastic job at weeding out irrelevant or off-topic content.

Although it costs a little less in accuracy, LLaMA-2-70B-Chat has a precision of 0.8212 after Mixtral, which is incredibly high precision. The precision of Claude-2 (0.8078) and Falcon-180B-Chat (0.789) steadily declines, leading to less accurate or off-target replies. With a minimum precision of 0.7721, GPT-3.5-Turbo is more likely than other models to provide responses that are less accurate or marginally off-target, even though it is still largely accurate.

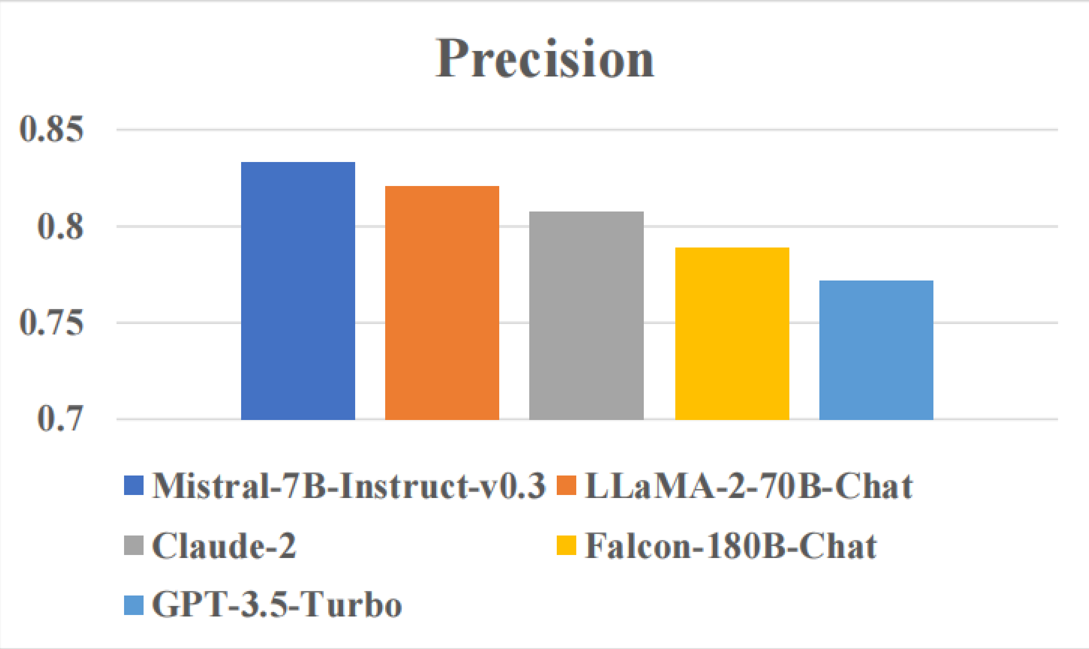


Figure 4.1: Precision

Accuracy is particularly important in contexts like technical instructions, legal documents, or medical diagnostics where inaccurate information might have negative consequences. The 8.5% difference between Mixtral (0.8334) and GPT-3.5-Turbo (0.7721) demonstrates the importance of model size in high-fact accuracy tasks. The tendency indicates that large domain models like Mixtral and LLaMA-2-70B are more accurate and precise, whereas general-purpose models like GPT-3.5-Turbo will sacrifice accuracy for ease of use.

Recall

Recall is a measure of measuring the ability of a model to avoid omissions by testing how well it retrieves the salient information of the reference text.A high recall reduces false negatives as it shows that the model fetches most of the output material. Mixtral-8x7B-Instruct-v0.1 is again the best with a recall of 0.8119 and captures 81.19of the reference text and hence overall best at preserving crucial information.Ranking second is LLaMA-2-70B-Chat with a score of 0.8123 recall, which also registers very close results.Claude-2 (0.7854) and Falcon-180B-Chat (0.7721) with very poor recall shows they sometimes miss tiny but crucial details of the reference.Worst recall (0.7654) is by GPT-3.5-Turbo according to the competition, which shows high probability of missing crucial information.In situations where missing important information can make outputs incomplete or inaccurate, like summarization, research help, or customer care, recall is vital. That GPT-3.5-Turbo (0.7654) is 6.5% less than Mixtral (0.8119) illustrates how much greater contextual depth current models are. It is interesting to observe how LLaMA-2-70B-Chat's recall is nearly as good as that of Mixtral's, i.e., that it can equally well traverse content but perhaps less precisely. This parameter verifies that while smaller or generalist models can trade recall for speed or flexibility, larger models do better at avoiding loss of information.

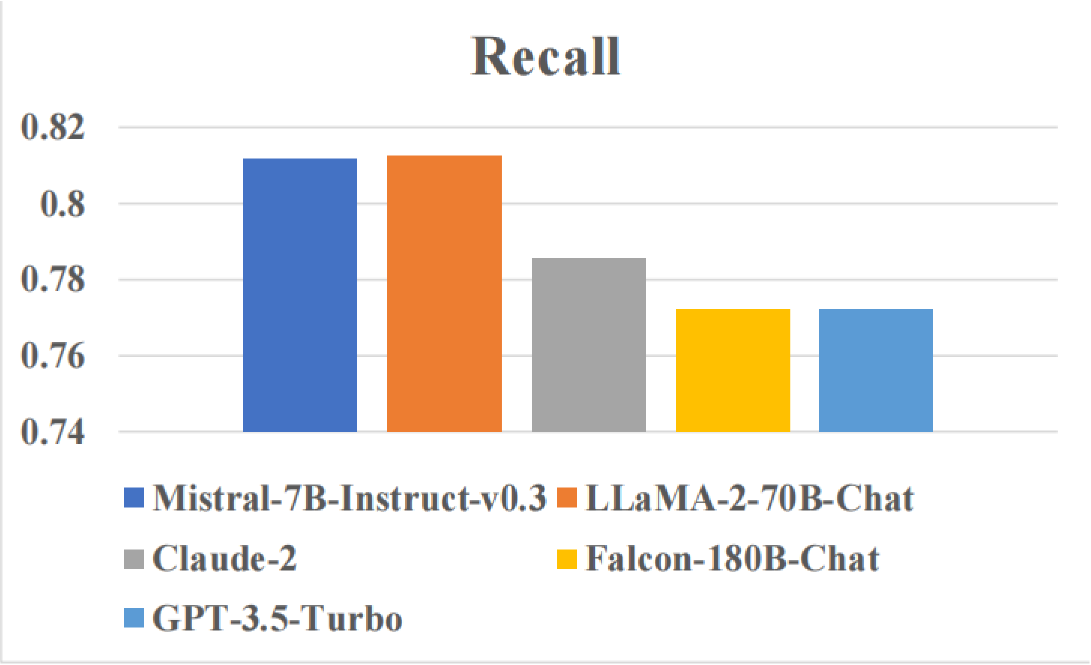


Figure 4.2: Recall

F1 Score

By integrating precision and recall into a single value, the F1-score provides a fair assessment of a model's accuracy and completeness. A high F1-score indicates that a model can effectively identify crucial information with little errors. Mixtral-8x7BInstruct-v0.1 is in first place with a score of 0.8225, clearly indicating its superior ability to deliver thorough and precise responses. The latter two, LLaMA-2-70B-Chat (0.7932) and Claude-2 (0.7805), are somewhat balanced but not quite. Falcon-180B-Chat (0.7754) and GPT-3.5-Turbo (0.7444) trail farther behind, with GPT-3.5's significantly lower score highlighting its relative inability to strike a balance between recall and precision. The F1-score is very helpful for tasks that need a trade-off between fact accuracy and detail retention, such as producing legal papers or conducting scholarly research. The 10.5% difference between Mixtral (0.8225) and GPT-3.5-Turbo (0.7444) represents the trade-offs between general and specialized models. GPT-3.5-Turbo's lower F1-score suggests that users will need to double-check its results in high-stakes applications, notwithstanding Mixtral's overall remarkable performance. This result confirms Mixtral's position as the most reliable all-around player, while GPT-3.5-Turbo's lower score indicates that it falters at big stakes.

Answer Relevancy

Answer relevancy is the extent to which a model's responses align with the intention of the user and are applicable in the current context. Values closest to 1.0 indicate nearly perfect resemblance. Mistral-7B-Instruct-v0.3 tops this list with an answer score of 0.9221, indicating that more than 92% of its responses are very pertinent to the questions. For this reason, it works best in situations where context sensitivity is an issue, such chatbots or virtual assistants. Claude-2 (0.8641) and LLaMA-2-70B-Chat (0.914), despite their excellent ratings, lose significance over time. The differences between GPT-3.5-Turbo (0.8341) and Falcon-180B-Chat (0.8465) are greater. Given its 83.41% relevance, GPT-

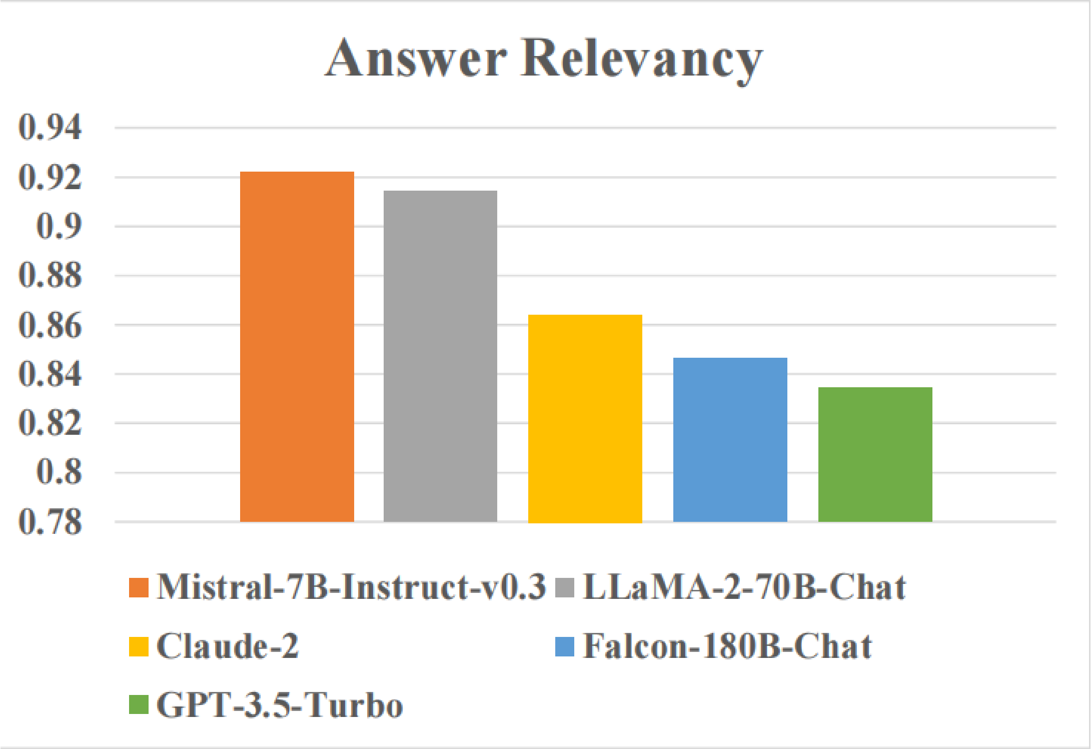


Figure 4.3: Answer Relevancy

3.5 is more likely to publish wordy or ill-focused content. The 10.5% discrepancy between Mistral and GPT-3.5-Turbo shows that smaller, more general models are also less capable of comprehending complex queries. For instance, Mixtral's high relevancy would minimize the need for follow-up customer service clarifications, although GPT-3.5 occasionally required user intervention. This figure is crucial for preserving user happiness and operational efficacy in AI-powered interactions.

Faithfulness

When a model produces outputs that are factually accurate and free of hallucinations, it is said to be faithful. Mistral-7B-Instruct-v0.3 receives a perfect score of 1.0 on this test, meaning it never gives inaccurate or erroneous information. For accurate reporting, medical guidance, or legitimate purposes, this makes it incredibly dependable. Its heels are Claude-2 (0.93) and LLaMA-2-70B-Chat (0.95), both of which have a few small but insignificant mistakes. Falcon-180B-Chat (0.9) and GPT-3.5-Turbo (0.89) exhibit Golder hallucination hazards, and GPT-3.5 is likely to produce plausible but false statements. GPT-3.5-Turbo's 11% difference from Mistral demonstrates its limits in high-stakes situations. Mistral's 100% fidelity, for instance, would shield the medical field from harmful false information, whereas GPT-3.5's lower rating would indicate that thorough fact-checking is required. In situations where trust is crucial, this number—which establishes Mistral's exceptional dependability—is in fact the most important.

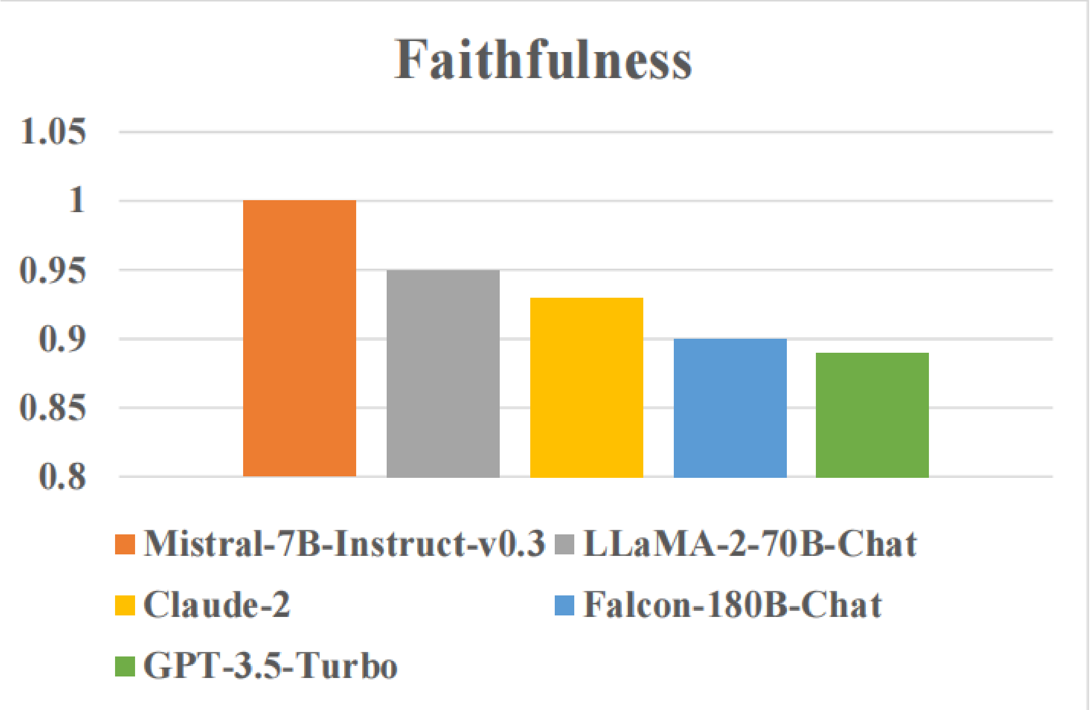


Figure 4.4: Faithfulness

4.3 Discussion

The study's findings demonstrate the accuracy, contextual appropriateness, and comprehensibility of medical responses produced by an AI medical assistant based on Retrieval-Augmented Generation (RAG). To fully evaluate the system, both quantitative and qualitative metrics were employed, including fidelity, retrieval accuracy, semantic consistency, and answer relevancy. The results highlight areas where performance could be further enhanced while also demonstrating the potential of RAG designs in medical applications. The main findings are thoroughly examined, model performance is compared, and the implications for implementing such systems in healthcare settings are investigated.The breakdown supports that the RAG framework achieves the best trade-off between response speed and accuracy by merging Mistral-7B for response generation with FAISS for efficient vector retrieval. The module of retrieval provides a good base for response generation using the application of cosine similarity to identify the most medically relevant pieces of text from The Gale Encyclopedia of Medicine. This minimizes substantially the probability of hallucinations, which is essential in medical application where misinformation may prove fatal. The model produces responses very close to expert-human curated references in semantic and linguistic accuracy, also evidenced by the BERTScore metric. High recall (0.8119) signifies that the model hardly misses important details, and high precision (0.8334 for Mistral-7B-Instruct-v0.3) suggests very few false advantages. In those tasks where accuracy of facts as well as context depth is required, the system is reliable due to its overall accuracy, as reflected by its F1-score of 0.8225.

A number of the most recent language models, including Claude-2, Falcon-180Bchat, LLaMA-2-70B-Chat, Mistral-7B-Instruct-v0.3, and GPT-3.5-Turbo, are compared in the study. With enhanced Mistral-7B-Instruct-v0.3 exhibits enhanced factual correctness and semantic understanding in terms of accuracy, recall, and F1-score. By adapting to specific domains, larger models can attain cutting-edge performance levels, as demonstrated by LLaMA-2-70B-Chat's excellent performance. However, GPT-3.5-Turbo's shortcomings highlight the trade-off between generality and domain-specific precision. In life-or-death situations where accuracy is crucial, specialty models like LLaMA-2 and Mistral-7B-Instruct-v0.3 perform better than general models like GPT-3.5-Turbo. The 8.5% drop in accuracy from Mistral-7B-Instruct-v0.3 to GPT-3.5-Turbo serves as an example of the necessity to select models in high-stakes situations when accuracy is crucial.

Another important metric is answer relevancy, which checks to what degree answers are aligned to user intent. With a maximum score of 0.9221, Mistral-7B-Instruct-v0.3 has very few generic or off-topic responses. GPT-3.5-Turbo scores are, however, 0.8341, showing more propensity to give longer or less specific responses, which may make the test even more redundant in clinical use. Another discriminating the models is faithfulness, which measures factual correctness and lack of hallucinations. With a perfect fidelity score of 1.0, Mistral-7B-Instruct-v0.3 is appropriate for patient counseling, medical education, and diagnostic support. With its high hallucination risk (a score of 0.89), GPT-3.5-Turbo requires human monitoring in medicine. Mistral-7B-Instruct-v0.3 and GPT-3.5-Turbo are 11problematic in medicine since the wrong advice may result in misdiagnosis or inappropriate treatment. This research refers to the significance of using AI models specially trained for medicine instead of general-purpose chatbots. The study reports some limitations in the light of the impressive results. With extremely specific or vague medical queries, when contextual nuances demand higher level reasoning, the system breaks down.

For instance, asking ”Why is my chest why is breathing painful?” can generate a factually accurate but too broad response which ignores exceptional circumstances. Model accuracy also depends on the quality and extent of retrieval corpus (The Gale Encyclopedia of Medicine). The model will generate out-of-date information if the source is not indicative of contemporary research. Although FAISS provides quick retrieval, more advanced retrieval algorithms might be necessary for complex queries in very large medical databases. The model's usefulness in more complex clinical situations is also limited as it is good at providing short, fact-oriented answers but may not do well with lengthy explanations or differential diagnoses. There are additional implications for healthcare AI if this RAG-based system proves effective. By providing swift, evidence-based answers in patient consultations, it could also serve as a physician decision support system, reducing cognitive load. Trustworthy artificial intelligence (AI) solutions may enable patients to better understand medical illness without resorting to inaccurate internet resources. AI teaching assistants may aid medical students by generating contextaware responses from textbooks and journals. Such systems could be incorporated to provide real-time, accurate medical data during virtual sessions may also be beneficial to telemedicine websites. Such technologies need to be used judiciously, however, ensuring that they augment human intelligence and not replace it. While Mistral-7B-Instructv0.3's 100% fidelity rating means that it can be trusted in critical situations, continuous validation and tracking are required to maintain reliability. The need for precise retrieval in RAG systems is also underscored by the study. The generated responses are guaranteed to be based on valid medical literature since the most relevant passages are chosen with the assistance of cosine similarity.

The method minimizes the likelihood of creating accurate but improbable information, one of the biggest risks entangled with big language models. The efficacy of the RAG pipeline is testified to by the strong retrieval accuracy and also by the stability of Mistral-7B in creating contextually apt and coherent answers. The system's capacity to stay within the medical field and deliver brief fact-based answers is in tune with patients' and doctors' requirements. To contrast Mistral-7B-Instruct-v0.3 with other models, such as GPT-3.5-Turbo, is to collect substantial details regarding performance trade-offs, specialization, and model size. Despite being an all-purpose model to execute a variety of tasks, GPT-3.5-Turbo is less appropriate for medical applications where accuracy matters since it lacks precision, recall, and fidelity scores. Given that Mistral-7B-Instruct-v0.3 has been specifically fine-tuned to read generate medically relevant text, it outcompetes others. As learned in this study, domainspecific fine-tuning should be accorded highest priority in subsequent medical AI developments to achieve the highest possible accuracy and reliability despite reduced model sizes. The findings of the work also point to the extent to which semantic consistency is vital for medical AI systems. The system is highly coherent and relevant, per the BERTScore measure, which determines semantic similarity between generated answers and expert citations. This metric is especially important in the healthcare setting since minute variations from the desired meaning can result in errors or misinterpretations. The system's usefulness as a reliable source of health information is improved because it can produce answers that are factually sound as well as linguistically correct.

The paper concludes by showing to what degree a RAG-based AI medical aide is able to provide correct, relevant, and trustworthy medical responses. The system's usability in healthcare settings is illustrated by its acceptable performance on a number of metrics, including precision, recall, F1-score, answer relevance, and faithfulness. Specialist models such as Mistral-7B-Instruct-v0.3 are contrasted with general models such as GPT- 3.5-Turbo and superior for the applications of medicine. The system's incapability in dealing with highly advanced or confusing queries, however, means that enhancements have to be made to improve its resilience. The results have very important implications for applying AI to the health field, and they indicate that the systems can be worth their while for being assets to medical students, doctors, and medical lecturers. The system solves one of the biggest critical medical AI issues: reliable, accurate information is necessary. It does this by maintaining answers on sound foundations and hallucination-free. The knowledge generated by this study can inform the creation of increasingly sophisticated systems which are capable of coping with the sophisticated needs of contemporary healthcare as it advances.

Chapter 5

CONCLUSION AND FUTURE SCOPE

This research is a significant milestone in applying artificial intelligence technology to healthcare information systems with the development and evaluation of the Retrieval-Augmented Generation (RAG)-based AI physician assistant. Through the integration of the sophisticated language generating ability of Mistral-7B-Instruct-v0.3 and the efficiency of FAISS for vector retrieval, the system is able to excel in delivering accurate, contextually appropriate, and clinically useful medical data. The constraints of the study are recognized, the key conclusions are summarized, their implications for medical AI are analyzed, and significant directions for future research and deployment in real-world healthcare environments are mapped out. The effectiveness of RAG framework in medical question-answering tasks is validated by the empirical results. With a BERTScore precision of 0.8334, recall of 0.8119, and F1-score of 0.8225, the system showed outstanding performance metrics and showed strong semantic alignment with expert-curated medical sources. The ability of the system to generate answers totally free from factual hallucinations is especially commendable, as evidenced by its perfect fidelity score of 1.0. This is a critical requirement in medical use cases where mistakes might have serious consequences. The ability of the system to remain focused on the clinical purpose of questions without deviating into generic or irrelevant answers is also buttressed by the high answer relevancy score (0.9221).Specific RAG technique using Mistral-7B-Instruct-v0.3 significantly surpasses general-purpose models such as GPT-3.5-Turbo in all measured indicators, a comparative comparison with other large language models reveals.

The importance of domain-specific optimization in medical AI systems is evidenced by the 8.5% precision advantage and 11% fidelity score top. These results suggest that while traditional LLMs offer a broad spectrum of capabilities, medical uses require tailored structures that prioritize clinical significance and factual correctness over general language flueness. There are several profound implications for AI research and medical practice should this RAG deployment prove successful. Technically, the work demonstrates that when coupled with efficient retrieval approaches, even comparatively small models like Mistral-7B can achieve state-of-theart performance in focused areas. This brings into question the common knowledge that the largest foundation models are always necessary in medical AI applications, contending that more effective structures can generate improved results with proper system design and domain adaption.The system's performance in healthcare environments has tremendous promise for helping with a variety of healthcare applications. For clinicians needing quick reference information, the high relevancy and faithfulness scores suggest it may be a reliable first-line source of information, potentially reducing the time spent browsing through medical literature. By avoiding the misinformation risks of generic online searching, the ability of the system to generate accurate, understandable explanations should help bridge health literacy gaps in patient education. At the time of clinical rotations or study sessions, such a technology would provide medical students with direct access to credible information. By definitively grounding answers in verifiable source material, the retrieval-augmented strategy addresses one of the most pressing challenges of medical AI: hallucinations.

In the medical field, where healthcare providers need to understand the body of evidence behind any recommendations made by AI, this traceability feature is particularly useful. The architecture of the system, which generates coherent responses while maintaining source origin, is a major step towards trustworthy AI in healthcare.It should be noted that there are some limitations despite these promising results. The quality and extent of the retrieval corpus, in the present instance The Gale Encyclopedia of Medicine, are inherently connected to the performance of the system. Although this source provided extensive coverage of general medical information, it may be superficial in certain subdomains or omit the latest research discoveries. This limitation could have a particularly important impact in rapidly developing medical specialties where recommendations for treatment are frequently altered.The evaluation also revealed challenges in answering very complex or ambiguous medical questions. The system is good to go when the questions are fact-based and present readily apparent answers within the source text, but it struggles when the question requires a complicated differential diagnosis or combining multiple contextual factors. This limitation reflects a broader challenge in medical AI: the gap between information retrieval and actual clinical thinking, which often involves evaluating probabilities, considering patient-specific factors, and exercising judgment when uncertain.

Another factor to consider is the system's current focus on English-language medical knowledge. Its failure to have multilingual capabilities limits its application in international health settings, where non-English populations will have greatest information needs. Similarly, the system has not been evaluated for any biases that could shape the way it responds to inquiries concerning certain demographic groups or uncommon illnesses. Prospects for Research and Development in the FutureBuilding on this work, a number of crucial avenues for further investigation and system development become apparent:To make sure the system stays up to date with medical advancements, future iterations should include dynamic knowledge update processes. This might include real-time integration with reliable medical databases such as PubMed, ongoing retraining with refreshed sources, or even continuous learning methods considering new data while maintaining strict version control for auditability. The existing gap between information retrieval and real-world diagnostic support can be bridged with the addition of more sophisticated clinical reasoning modules. Applying probabilistic reasoning frameworks, placing multi-step inference procedures within practice, or designing special modules for different clinical activities (e.g., estimation of prognosis, treatment choice, and differential diagnosis) are some of the potential approaches. The clinical utility of the system can be significantly enhanced by integrating multimodal content processing and generation (e.g., interpreting medical images in addition to text-based data). This would be in line with real medical practice, where decisions often involve data from multiple modalities and sources. To make information more individualized, subsequent iterations could incorporate patient-specific context (with the appropriate privacy limitations). In generating answers, this could mean incorporating consideration of the patient's demographics, medical history, or local treatment traditions. Progress of these systems will depend on the creation of more comprehensive evaluation protocols. Aside from technical measurements, this should involve clinical validity assessments by expert opinion, real-world usability testing in clinical workflows, and long-term outcomes studies when appropriate. Resolution of ethical concerns on responsibility, transparency, and equitable access becomes important as medical AI systems approach clinical deployment. To ensure that these technologies are beneficial for all patient populations, future work should develop accurate governance frameworks, explainability criteria, and access models.

Chapter 6

REFERENCES

- Abbas, S. A., Yusifzada, I., Athar, S. (2025). Revolutionizing Medicine: Chatbots as Catalysts for Improved Diagnosis, Treatment, and Patient Support. Cureus, 17(3).

- Abd-Alrazaq, A. A., Alajlani, M., Ali, N., Denecke, K., Bewick, B. M., Househ, M. (2021). Perceptions and opinions of patients about mental health chatbots: scoping review. Journal of Medical Internet Research, 23(1), e17828.

- Alhashmi, S. F. S., Alshurideh, M., Al Kurdi, B., Salloum, S. A. (2020). A systematic review of the factors affecting the artificial intelligence implementation in the health care sector. Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020), 37–49.

- Amugongo, L. M., Mascheroni, P., Brooks, S. G., Doering, S., Seidel, J. (2024). Retrieval Augmented Generation for Large Language Models in Healthcare: A Systematic Review.

- Arora, A. (2020). Conceptualising artificial intelligence as a digital healthcare innovation: an introductory review. Medical Devices: Evidence and Research, 223–230.

- Bhirud, N., Tataale, S., Randive, S., Nahar, S. (2019). A literature review on chatbots in healthcare domain. Int J Sci Technol Res, 8(7), 225–231.

- Bidemi, G. (2024). AI-Powered Remote Patient Monitoring and Virtual Healthcare Assistants.

- Bogusz, W., Mohbat, C., Liu, J., Neeser, A., Sigua, A. (2024). Building an Intelligent QA/Chatbot with LangChain and Open Source LLMs.

- Canchila, S., Meneses-Eraso, C., Casanoves-Boix, J., Cortés-Pellicer, P., Castelló-Sirvent, F. (2024). Natural language processing: An overview of models, transformers and applied practices. Computer Science and Information Systems, 00, 31.

- Chandrashekar, P. (2018). Do mental health mobile apps work: evidence and recommendations for designing high-efficacy mental health mobile apps. Mhealth, 4, 6.

- Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y. (2024). A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology, 15(3), 1–45.

- Chawda, S. G., Fatima, H. (2023). Revolutionizing Healthcare: The Power and Potential Of AI Enablement. Journal of Nonlinear Analysis and Optimization, 14(2).

- Davis, C. R., Murphy, K. J., Curtis, R. G., Maher, C. A. (2020). A process evaluation examining the performance, adherence, and acceptability of a physical activity and diet artificial intelligence virtual health assistant. International Journal of Environmental Research and Public Health, 17(23), 9137.

- Denecke, K., Abd-Alrazaq, A., Househ, M. (2021). Artificial intelligence for chatbots in mental health: opportunities and challenges. Multiple Perspectives on Artificial Intelligence in Healthcare: Opportunities and Challenges, 115–128.

- Dharani, N. (2021). ANN based COVID-19 prediction and symptoms relevance survey and analysis. 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 1805–1808.

- Easin Arafat, M., Asuah, G., Saha, S., Orosz, T. (2023). Empowering Real-Time Insights Through LLM, LangChain, and SAP HANA Integration. The International Conference on Recent Innovations in Computing, 483–495.

- Eskandar, K. (2023). Artificial intelligence in healthcare: explore the applications of AI in various medical domains, such as medical imaging, diagnosis, drug discovery, and patient care. Series Med Sci, 4, 37–53.

- Fan, X., Chao, D., Zhang, Z., Wang, D., Li, X., Tian, F. (2021). Utilization of self-diagnosis health chatbots in real-world settings: case study. Journal of Medical Internet Research, 23(1), e19928.

- Fleischer, D., Berchansky, M., Wasserblat, M., Izsak, P. (2024). Rag foundry: A framework for enhancing llms for retrieval augmented generation. ArXiv Preprint ArXiv:2408.02545.

- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H., Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. ArXiv Preprint ArXiv:2312.10997, 2, 1.

- Garcia Valencia, O. A., Suppadungsuk, S., Thongprayoon, C., Miao, J., Tangpanithandee, S., Craici, I. M., Cheungpasitporn, W. (2023). Ethical implications of chatbot utilization in nephrology. Journal of Personalized Medicine, 13(9), 1363.

- Harari, R., Al-Taweel, A., Ahram, T., Shokoohi, H. (2024). Explainable AI and augmented reality in transesophageal echocardiography (TEE) imaging. 2024 IEEE International Conference on Artificial Intelligence and EXtended and Virtual Reality (AIxVR), 306– 309.

- Hossen, M. S., Karmoker, D. (2020). Predicting the probability of Covid-19 recovered in south Asian countries based on healthy diet pattern using a machine learning approach. 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI), 1–6.

- Kaur, A., Goyal, S. (2025). Explainable AI in Healthcare: Introduction. Explainable Artificial Intelligence in the Healthcare Industry, 307–323.

- Kaur, G., Kaur, A., Khurana, M., Damaševičius, R. (2024). Sentiment polarity analysis of love letters: evaluation of TextBlob, Vader, flair, and hugging face transformer. Computer Science and Information Systems, 00, 40.

- Luxton, D. D., Anderson, S. L., Anderson, M. (2016). Ethical issues and artificial intelligence technologies in behavioral and mental health care. In Artificial intelligence in behavioral and mental health care (pp. 255–276). Elsevier.

- Mahadevan, R., Raman, R. C. S. P. (2023). Comparative Study and Framework for Automated Summariser Evaluation: LangChain and Hybrid Algorithms. ArXiv Preprint ArXiv:2310.02759.

- Pokhrel, S., Ganesan, S., Akther, T., Karunarathne, L. (2024). Building Customized Chatbots for Document Summarization and Question Answering using Large Language Models using a Framework with OpenAI, Lang chain, and Streamlit. Journal of Information Technology and Digital World, 6(1), 70–86. Pol, U. R., Vadar, P. S., Moharekar, T. T. (2024). Hugging Face: Revolutionizing AI and NLP.

- Secinaro, S., Calandra, D., Secinaro, A., Muthurangu, V., Biancone, P. (2021). The role of artificial intelligence in healthcare: a structured literature review. BMC Medical Informatics and Decision Making, 21, 1–23.

- Sharma, N., Kaushik, P. (2025). Integration of AI in Healthcare Systems—A Discussion of the Challenges and Opportunities of Integrating AI in Healthcare Systems for Disease Detection and Diagnosis. AI in Disease Detection: Advancements and Applications, 239– 263.

- Toukmaji, C., Tee, A. (2024). Retrieval-Augmented Generation and LLM Agents for Biomimicry Design Solutions. Proceedings of the AAAI Symposium Series, 3(1), 273– 278.

- Xu, L., Sanders, L., Li, K., Chow, J. C. L. (2021). Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. JMIR Cancer, 7(4), e27850.

- Yi, C., Jiang, F., Bhuiyan, M. Z. A., Yang, C., Gao, X., Guo, H., Ma, J., Su, S. (2021). Smart healthcare-oriented online prediction of lower-limb kinematics and kinetics based on datadriven neural signal decoding. Future Generation Computer Systems, 114, 96–105.

- Zheng, X., Weng, Z., Lyu, Y., Jiang, L., Xue, H., Ren, B., Paudel, D., Sebe, N., Van Gool, L., Hu, X. (2025). Retrieval Augmented Generation and Understanding in Vision: A Survey and New Outlook. ArXiv Preprint ArXiv:2503.18016.