AI-POWERED HEALTHCARE ASSISTANT: A RAG-BASED CHATBOT WITH HUGGING FACE TRANSFORMERS AND LANGCHAIN

Thesis Submitted
In Partial Fulfilment of the Requirements for the
Degree of

MASTER OF TECHNOLOGY

in
Software Engineering

Submitted by

Subhadip Das
(2K23/SWE/20)

Under the Supervision of
Mrs. Priya Singh
(Assistant Professor, SE, DTU)

To the
Department of Software Engineering

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India
JUNE, 2025

ACKNOWLEDGEMENTS

I would like to express my deep appreciation to **Mrs. Priya Singh**, Assistant Professor at the Department of Software Engineering, Delhi Technological University, for her invaluable guidance and unwavering encouragement throughout this research. Her vast knowledge, motivation, expertise, and insightful feedback have been instrumental in every aspect of preparing this research plan.

I am also grateful to **Prof. Ruchika Malhotra**, Head of the Department, for her valuable insights, suggestions, and meticulous evaluation of my research work. Her expertise and scholarly guidance have significantly enhanced the quality of this thesis.

My heartfelt thanks go out to the esteemed faculty members of the Department of Software Engineering at Delhi Technological University. I extend my gratitude to my colleagues and friends for their unwavering support and encouragement during this challenging journey. Their intellectual exchanges, constructive critiques, and camaraderie have enriched my research experience and made it truly fulfilling.

While it is impossible to name everyone individually, I want to acknowledge the collective efforts and contributions of all those who have been part of this journey. Their constant love, encouragement, and support have been indispensable in completing this M.Tech thesis.

Subhadip Das

(23/SWE/20)

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

CANDIDATE'S DECLARATION

I, Subhadip Das, Roll No's –2K23/SWE/20 students of M.Tech (Software Engineering), hereby certify that the work which is being presented in the thesis entitled "AI-Powered Healthcare Assistant: A RAG-Based Chatbot with Hugging Face Transformers and LangChain" in partial fulfilment of the requirements for the award of degree of

Master of Technology, submitted in the Department of Software Engineering, Delhi Technological University is an authentic record of my own work carried out during the period from Jan 2025 to June 2025 under the supervision of Mrs. Priya Singh .

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other institute.

Candidate's Signature

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor **Signature of External Examiner**

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road,Delhi-42

CERTIFICATE BY THE SUPERVISOR

I hereby certify that **Subhadip Das (Roll no 23/SWE/20)** has carried out their research work presented in this thesis entitled **"Ai-Powered Healthcare Assistant: A RagBased Chatbot With Hugging Face Transformers And Langchain"** for the award of **Master of Technology** from the Department of Software Engineering, Delhi Technological University, Delhi under my supervision. The thesis embodies results of original work, and studies are carried out by the student herself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution. 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
Date: 19 june,2025

Mrs. Priya Singh
Assistant Professor
Department of Software Engineering
DTU DELHI
India

LIST OF TABLES

LIST OF FIGURES

TABLE OF CONTENTS

AI-Powered Healthcare Assistant: A RAG-Based Chatbot with Hugging Face Transformers and Lang Chain

Subhadip Das

ABSTRACT

A significant issue in healthcare applications is the tendency of traditional AI language models to deliver convincing but factually wrong or hallucinated responses. This shortcoming has been brought to light by the growing demand for reliable medical information. False assertions about facts and an absence of citations for reliable sources are consequences of using just parametric data in conventional methods like GPT-3.5 and other fine-tuned LLMs. In complicated medical domains in particular, current implementations of retrieval-augmented generation (RAG) systems suffer from an imbalance between retrieval accuracy, semantic consistency, and response appropriateness, despite the fact that RAG systems were proposed as a solution to these problems. This research built a RAG-based medical AI assistant to help with these challenges. In order to generate responses quickly, it employs FAISS for vector retrieval and Mistral-7B-Instruct-v0.3. The AI has been programmed to only give responses that are backed by credible medical sources, as stated in The Gale Encyclopedia of Medicine. The quantitative measures that were utilized for the evaluation of our system were BERTScore (Precision: 0.8334, Recall: 0.8119, F1-Score: 0.8225), Answer Relevancy (0.9221), and Faithfulness (1.0). Its performance was much better than that of general-purpose models such as GPT-3.5-Turbo (Faithfulness: 0.89) and LLaMA-2-70B-Chat (Faithfulness: 0.95). While maintaining therapeutic relevance and high semantic consistency, our RAG framework successfully reduces hallucinations, according to the results. The fact that problems still arise when dealing with exceedingly complex or ambiguous inquiries further highlights the necessity for improved reasoning approaches. This paper highlights the possibilities of RAG systems that are customized for particular healthcare fields. Provides a trustworthy means of accessing accurate medical records and opens the door to developments in dynamic

knowledge integration and multi-hop retrieval. The results support the use of specialized AI helpers in fields like healthcare and academics where precision, openness, and dependability are critical.

Keywords: Retrieval-Augmented Generation (RAG), Medical AI, Mistral-7B, FAISS, HallucinationMitigation, Clinical Decision Support, BERTScore, Answer Relevancy,Faithfulness.

CHAPTER 1

INTRODUCTION

Global healthcare systems face mounting pressures due to aging populations, workforce shortages, and escalating costs. Traditional healthcare operations, especially manual diagnosis and paperwork, are becoming more and more seen as resource-intensive and inefficient. Research shows that physicians spend about half of their working hours on administrative tasks, which leaves less time for providing direct patient care (Bidemi, 2024). With its ability to automate administrative procedures, improve operational efficiency, and facilitate clinical decision-making, artificial intelligence (AI) has become a disruptive force in the healthcare industry (Alhashmi et al., 2020). The creation of AI-based health aides, who seek to increase the effectiveness, precision, and accessibility of healthcare services, is one of the most notable applications of AI (Manickam et al., 2022). AI-based tools provide a data-driven solution that is rapid, accurate and reliable as opposed to old fashioned healthcare solutions which are based almost entirely on manual documentation and subjective clinical judgement. AI makes it easier that way it can browse through patient history, scientific literature, and clinical regulations and then come up with solid evidence based recommendations.

With its ability to aid with illness diagnosis, medical data administration, treatment suggestions, and even direct patient engagement, AI-based health assistants have the potential to revolutionize the healthcare industry (J. Yang et al., 2022), (Secinaro et al., 2021). Healthcare is using big data, machine learning, and robotics to keep an eye on risks and benefits thanks to advances in AI (Hossen & Karmoker, 2020); (Dharani, 2021). When it comes to improving operations and speeding the delivery of care, medical data and analytics are the foundation. The variety and

volume of collected medical records have been greatly expanded in recent years. One such example is the enormous amount of data produced by patients, scientists and healthcare workers. Such data is a subset of this data and have been harvested from wide range of sources, like medical images, EHRs, health and lifestyle tracking apps and wearables. Beyond medicine, the knowledge is growingly applicable to the lay publicicasphalt Technology and other real-worldcases (Antoniou et al., 2017); (Xie et al., 2020).

In this context, AI technology has the potential to collect data, process it, perform dynamic analyses, and provide results that may be effectively applied to medical intervention (Comito et al., 2020). Constant examination of medical information, for example, can enable making reliable forecasts from patient patterns of behavior. Consequently, AI can offer hints in diagnosis, medical treatment, therapeutic knowledge, and prevention techniques to prevent deterioration in health. Patient outcomes across the range of sickness and diagnosis can be enhanced with its capability to assist with prevention to stop the deterioration of patients' problems. It may also be easier to prescribe and utilize medications. Significantly technologically advanced hospitals are now looking at using AI technology to improve medical procedure accuracy and lower operating costs (Sqalli & Al-Thani, 2019), (Zhou 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:** The recent developments in AI, machine learning, and deep learning have empowered AI systems with the ability to process and analyze possibly the largest volumes of healthcare data at extremely fast rates. The emergence of AI algorithms performing intricate medical tasks such as predictive analysis, image recognition, and natural language processing is the result of this.
- **Increasing Availability of Healthcare Data:** The proliferation of EHRs, genetic data, or even medical imaging and other kinds of healthcare data

has only fed massive repositories made available to AI systems. These massive datasets act as the training data that an AI system needs to improve its working with time.

- **Growing Demand for Healthcare Solutions:** The healthcare sector grapples with multiple issues, such as a growing number of patients, increasing costs, and the familiarity of chronic illnesses. Hence, new mechanisms to enhance healthcare delivery, manage resources, and improve patient outcomes are very much sought. The AI system may be useful in these areas by providing some recommendations and insights with a view to increasing operational efficiency as well as clinical decisions.
- **Demonstrated Success in Various Applications:** Drug discovery, personalized medicine, medical imaging, and diagnostics are just common examples of health- related areas in which AI has demonstrated usefulness. From prediction of patient outcome to cancer diagnosis on medical photographs, these AI systems perform better than human physicians. Human accomplishments in medicine have perhaps pushed interest and investments toward AI systems in health.
- **Supportive Regulatory Environment:** In the United States, institutions like the Food and Drug Administration (FDA) have even begun establishing protocols for testing and clearing medical devices and software incorporating artificial intelligence (AI) to market. These are policy steps paving the way toward right use of artificial intelligence, (AI) in healthcare, and may help to promote trust in AI solutions by health care professionals.

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: A.I. systems can be trained to read genetic data, as well as radiographic images and patient histories, to assist doctors in making more accurate diagnoses more quickly. Treatment may also be derived from

diagnoses from AI that serve as treatment recommendations, denting conditions at earlier stages, or may even be lifesaving. We can personalize treatment plans that cater to the needs of each patient with AI. Through analyzing unprecedented amounts of patient data, genetic markers, treatment results and medical history, AI systems can now offer the optimal plan for each individual patient. By minimizing side- effect risk and making treatment more effective, this personalized approach to health care is showing good preliminary data and is exciting.

Simplified Administrative Procedures: As administrative tasks like booking an appointment, billing, and record maintenance can get simpler with the help of AI, doctors get to spend more time on patient care. This is because AI can also have a roll in automating some routine work and making it easier for them, thus freeing the employers' time and decreasing paperwork and administration work done by a healthcare organization and ultimately their business performance and cost reduction. It could be that AI systems would mine patient data to find correlations or trends that would result in deducing.

Support for Remote Patient Monitoring: Changes or abnormalities in a patient's condition can be reported to healthcare professionals in real-time through artificial intelligence-based remote patient monitoring systems. Through the possibility of early intervention and proactive management of chronic conditions, this decreases the demand for hospitalizations and emergency room visits.

Research and medication Discovery: Artificial intelligence software has the capability to search through huge databases for new drugs, recognize possible interactions between drugs, and optimize existing treatment regimens. Streamlining the process of drug development and

reducing the length and expense of clinical trials are two means in which AI enablement can enhance patients' access to new treatments.

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 tools have shown great precision in evaluating numerous medical images, including X-ray, CT, and MRI scans. Since these algorithms are able to identify faint abnormalities that a human eye cannot detect, early and correct diagnoses of conditions like cancer, heart disease, and neurological disorders can be achieved. "AI ensures that patients receive timely and proper treatment by assisting radiologists and other medical professionals with medical image interpretation." Improved Disease Detection and Forecasting: Artificial intelligence algorithms can sort through mountains of patient data, such as genetic information, medical history, and biomarkers, to determine trends and patterns that could suggest a high risk of contracting certain diseases. By examining an individual's genetic composition, daily habits, and health history, AI could potentially forecast the probability of them contracting diabetes or heart disease. With the help of AI, doctors can detect possible health dangers sooner, which allows them to intervene and stop diseases from advancing. Personalized treatment programs for each patient's unique needs and characteristics are no longer a dream because of AI. To determine what is ideal treatment for each unique individual patient, artificial intelligence (AI) algorithms sift through mountains of information, like genetic information, response to treatment, and outcome. Improved therapy with fewer side effects are the outcome of this personalized method of treatment, in which patients are paired with drugs most likely to help in their specific disease.

Optimal Treatment Plans: Machine learning algorithms can today process a plethora of medical conditions by mining EHRs, clinical studies, and academic papers. By providing evidence-based recommendations, AI contributes to improve

decision making on treatment options, dosing regimens, and follow-up treatment plans for clinical physicians. This guarantees patients the best and most suitable treatment, based on the latest scientific findings and clinical guidelines. Assisting in the Development

Clinical Decisions: AI-driven disease diagnosis and treatment CDSS system integrate the information and, through artificial intelligence, offer medical doctors and healthcare professionals immediate recommendations to prescribe the most appropriate treatment to the patients. Physicians can leverage these systems to assist in diagnosis and treatment using patient data, clinical studies, and recommended best practice. When partnering clinical experience with AI-derived intelligence, CDSS improves the quality of patient care, decreases the amount of diagnostic errors, and improves treatment. Artificial intelligence has revolutionized healthcare by improving diagnosis, forecasting, detection of disease, therapy As time goes on and technology advances, AI will have an ever-increasing effect on healthcare delivery and patient outcomes, improving medicine and setting global standards for patient care.

1.4 The potential of virtual health assistants

With their emphasis on direct patient connection, virtual health assistants (VHAs) represent one of the most exciting new frontiers in artificial intelligence research. A number of essential tasks formerly performed by people in patient care have been taken over by virtual health assistants. These include appointment scheduling, answering questions, writing prescriptions, and helping with mental health issues (Chawda & Fatima, 2023). By providing patients with up-to-date, relevant health information over many channels (e.g., chatbots, voice recognition software, and mobile applications), many of these innovative AI-driven technologies may enhance doctor-patient interactions. In addition to offering patients unprecedented convenience, the VHAs relieve doctors and other healthcare professionals of a substantial administrative burden, allowing them to concentrate more on medical duties and decision- making (Sherani et al., 2024).

In this sense, VHAs may improve patient outcomes while saving time by doing basic duties. The majority of patient-provider interaction occurs during

consultations, like in traditional medical care systems, and questions and concerns are only brought up at that time. Nonetheless, a VHAs is available to help the patient at every stage of the procedure (Salunkhe et al., 2022). These systems have the ability to track patients' health information, alert patients when it's time to take a medicine, provide real-time answers to enquiries about medical issues, and provide recommendations for a specific patient's health issue (Husnain & Saeed, 2024). These ongoing exchanges strengthen the bond between the patient and the clinician and encourage patients to follow prescribed routines. Enhancing healthcare accessibility, particularly for underserved regions, is another significant contribution made by virtual health assistants (Khan et al., 2024).

Additionally, these applications help patients with self-management of chronic illnesses, prescription management, appointment scheduling, symptom assessment, and patient education (Javaid et al., 2023). Websites, voice assistants, and smartphone apps are just a few examples of the digital platforms that might make it easier to access care help. Virtual assistant use in healthcare is not without its difficulties, however, including problems with permission, privacy, security, data interpretation, ethics, and responsibility (Javaid et al., 2023).

1.4.1 Challenges and Barriers to VHA Adoption

Despite AI's rapid progress, there are still several substantial technological hurdles. The ability to understand subtleties of human emotion and language is essential for effective health coaching, but VHAs may struggle with this. It is also challenging to ensure that AI systems give reliable medical advice for a diverse array of specific health issues. Those who suffer from a number of chronic conditions, those whose medications, such as beta blockers, have unique effects on physical activity, and those whose diabetes is not well controlled are just a few instances. Furthermore, advanced AI models run the danger of overlooking important elements that were previously addressed since they primarily concentrate on more recent interactions, even if they are capable of remembering information throughout lengthy chats and across several sessions.

People are likely to provide a VHA sensitive health information, thus protecting the privacy and accuracy of their data is crucial. Data breaches will have serious repercussions for the data itself as well as for the public's confidence, which will eventually damage AI's capacity to be used in healthcare (Gillespie et al., 2023). VHAs must thus have strong cybersecurity features built in to guard against unwanted data access and preserve a usable experience. Regulatory problems connect with these concerns. There are several rules and regulations that virtual health assistants must abide by, and they differ from one nation to another. Specifically, several states prohibit the transmission of electronic health data abroad, which might make many AI systems in violation.

These difficulties might be addressed, for example, by creating regionalised AI platforms that allow data processing and storage inside the user's nation or area. Although certain jurisdictional limitations may be addressed by local solutions, issues may worsen on a larger scale as a result of some state actors' attempts to split the internet and the incompatibility of technological processes and governance frameworks (X. Xu et al., 2023).

Explainable AI, augmented reality, digital twins, closed systems, and synthetic data are some of the cutting-edge technological solutions that have the ability to overcome present constraints and boost clinician confidence in VHAs. A better comprehension of how AI choices are made is made possible by explainable AI, which improves interpretability and transparency. By acting as a reliable "second pair of eyes," augmented reality may assist medical professionals in more accurately interpreting complicated pictures (Harari et al., 2024). Virtual representations of real-world systems, or digital twins lower the likelihood of security or privacy breaches by enabling the representations to be modelled to forecast the possible behaviour of actual systems (Kenett & Bortman, 2022). The privacy and transparency concerns that open systems entail are addressed by system that is closed and relies on models and training data. One last option for training AI systems when real data is unavailable is to utilize synthetic data.

The possible loss of human interaction during caring and the substitution of human

occupations are further ethical concerns raised by VHAs. In behaviour change, where a patient's motivation is often tied to a feeling of responsibility to their physician (Eton et al., 2017) an aspect that may lessen with technology-based programs—the clinician-patient connection is crucial to the effectiveness of therapy. Healthcare organizations seeking cost savings may be enticed by the economic incentive of replacing human labor with automated systems, despite the fact that VHAs have substantial potential to enhance existing services or provide new ones, like digital coaches that use augmented reality to improve diagnosis.

Additionally, like other eHealth initiatives, VHAs run the danger of unintentionally expanding the health inequality gap even while they can make health information and services more accessible to all. People who are less tech-savvy or who have doubts about digital health may find this to be especially problematic.

Finally, VHAs need to maintain user engagement and gain users' confidence in order to be really effective. Even if the newest AI platforms show promise in this regard, A comprehensive grasp of human psychology, behavior, and needs is necessary for the development of AI systems that individuals feel comfortable and secure use as advisors for their health concerns.The long-term usage of these systems is threatened by growing evidence that users are especially forgiving of mistakes made by VHAs or when their enquiries are not addressed (Davis et al., 2020).

1.5 Chatbots in Health Care

Conversational agents, or chatbots, are leading the charge in the dynamic realm of information technology and digital communication, transforming the way that humans and technology communicate. Chatbots are computer programs designed to imitate how people talk to each other through text, images, audio, or video on websites, apps, or standalone software. (Kocaballi et al., 2019). Since its inception in 1994 (Mauldin, 1994), Chatbots have really come a long way. These days, they can help schedule appointments, answer patients' questions, and share information easily. Thanks to advancements in natural language processing and AI, chatbots

have moved on from just following set scripts with standard replies. Now, they can actually understand what users are asking and respond in a more fitting way (Nuruzzaman & Hussain, 2018), (Kumar et al., 2016). Journalism, education, online shopping, finance, healthcare, and entertainment are just a few areas that have figured out how to use them because they're so versatile. Take Amazon's Alexa, for example (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 well-known examples of these apps.

Chatbots could really help the healthcare industry by making treatment more effective, affordable, and better overall (Huisman et al., 2022), (Osipov & Skryl, 2021), (Jones et al., 2014),(Chandrashekar, 2018) as well as a wide range of applications and acceptance (Nadarzynski et al., 2019) ,(A. A. Abd-Alrazaq et al., 2021). More and more, chatbots are being used to get and offer health care services (A. Abd-Alrazaq et al., 2020),(Kretzschmar et al., 2019), (Cheng & Jiang, 2020), (Boucher et al., 2021), offering them different roles in diagnosis, prevention, and treatment support, which could impact the entire healthcare system.

Healthcare chatbots present particular difficulties despite their potential advantages (Luxton et al., 2016), (Denecke et al., 2021),(Parviainen & Rantala, 2022), (Palanica et al., 2019). Giving tailored advice is still a big challenge. Most chatbots offer broad answers that overlook local health issues, individual patient histories, and specific medical profiles, all of which are important for accurate diagnosis and personal care (Chang et al., 2024). Furthermore,

traditional chatbot systems usually do not possess the contextual understanding and flexibility required to access current medical information or function well across various medical areas. The blending of large language models (LLMs) such as ChatGPT with generative AI has been a notable breakthrough in the field of chatbots.

Their capacity to produce text that resembles that of a person allows for more organic and educational interactions (Ouyang et al., 2022), (OpenAI Achiam et al., 2023). With healthcare data being so complex and correct information being so important, LLMs represent a huge step forward in AI-assisted healthcare technology. Many consider the advancements in technology related to LLMs to be among the most significant technological milestones of the past several years. To generate an LLM, deep learning requires massive amounts of data and text for model training. This collection includes a variety of resources pertaining to languages, such as books, websites, articles, and video transcripts. As a result, chatbots and answer generators that employ LLMs to directly interpret linguistic material have made it possible to query online information more effectively and directly than in the past (based on document retrieval). However, by working with a variety of data sets, academics are always working to improve the calibre of LLMs and their applications. The main distinctions between existing LLMs such as GPT4o, LlaMA3/LlaMA3.1, Mistral, and Claude and pre-trained language models like BERT and GPT2 (Reimers & Gurevych, 2019) (T. Zhang et al., 2019)include step-by-step reasoning, instruction following, and in-context learning (ICL) (Topol, 2019). ICL is the capacity to produce an output without the need for more training or gradient adjustments, based on instructions or demonstrations. This implies that the model may pick up new skills by just following directions without any other help. One of the most important characteristics of LLMs is their capacity for sequential thinking. The Chain-of-Thought (CoT) prompting is a tool that LLMs use to solve complex problems. In CoT, users must provide part of the necessary information rather than asking the question directly. The methodology then divides the issue into more manageable stages. However, their use in healthcare is still developing. Errors and false information are a serious issue (Thirunavukarasu,

Hassan, et al., 2023), especially in the medical field where precision is essential. LLMs' one-size-fits-all methodology may not be a good match for the complex requirements of patient-centered treatment in the medical field (Thirunavukarasu, Ting, et al., 2023).

1.6 Advancing Chatbots with Retrieval-Augmented Generation (RAG)

Considering the limitation of LLMs, the application of large language models (LLMs) based on Retrieval-Augmented Generation (RAG) is a significant milestone in AI-driven healthcare technology. Compared to traditional LLMs that solely rely on available training data, RAG models pull external information in real-time and incorporate it into the content creation process, unlike traditional LLMs that simply use available training data to create content**(Toukmaji & Tee, 2024).**

RAG is a state-of-the-art strategy for raising response quality. This paradigm allows the chatbot to access external data from a large corpus in real-time by integrating information retrieval (IR) techniques with the natural language production characteristic of transformer models such as GPT. To improve the relevance and accuracy of its responses, RAG relies on the system actively searching through specific databases or papers for material relevant to the discussion, rather than relying exclusively on the model's pre-trained expertise.

By enhancing chatbot performance for application situations, such as providing timely responses with well-documented explanations, RAG integration may pave the way for new customer service applications. It is clear from 1 that RAG is a straightforward approach. A query is sent to the system by the user. With this query, the system may get the most important and pertinent documents from the database. These papers will be chosen according to how closely they match the question. After that, the system takes the user's inquiry and the collected documents and runs them through an LLM to generate an answer that sounds human. Consequently, the user gets this data in the form of an accurate and natural answer to his question.

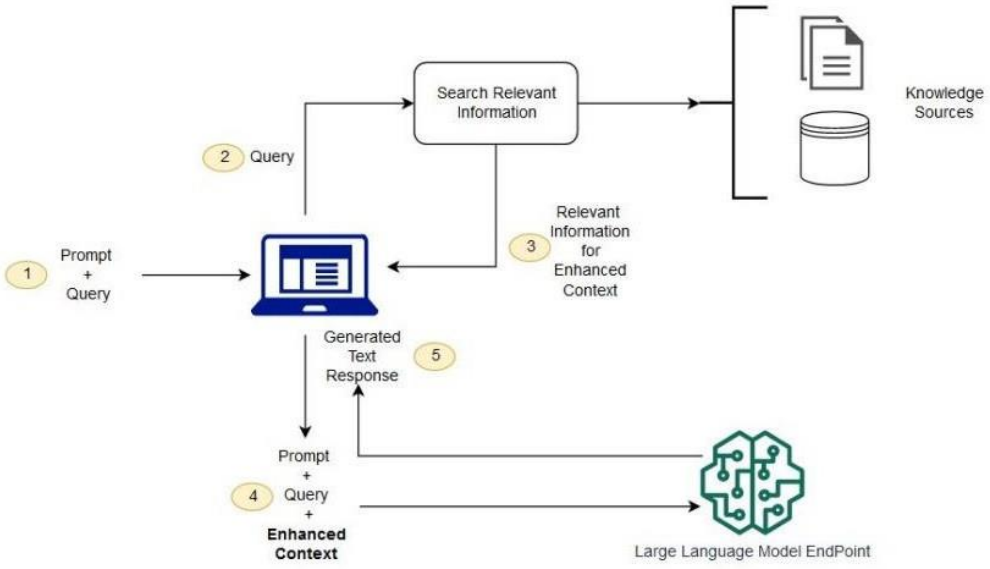

Figure 1 Steps of Retrieval Augmented Generation (RAG)

The ability of RAG-based LLMs to extract massive amounts of medical data from organized databases enables them to provide more accurate and contextually relevant responses. An important challenge for artificial intelligence in healthcare is the generation of reliable, accurate, and contextually relevant information; RAG systems offer a possible solution to this issue. Nevertheless, strong and flexible NLP frameworks are necessary for the successful deployment of such systems in healthcare. Here, Hugging Face and LangChain stand out as cornerstone technologies that pave the way for the creation of smart, RAG-based medical aides. Hugging Face has quickly become an industry leader in natural language processing (NLP). Modern pretrained models like as BERT, RoBERTa, GPT, and T5 are readily available in its Transformers library. Existing models are pre-trained for most natural language processing tasks, so they can be implemented with little programming effort. Filtering, translation, and question answering are all facilitated with ease using the Transformers library. For information extraction tasks, hugging face models perform efficiently in identifying names, dates, question answers, etc (Pol et al., 2024). These models are critical to the RAG architecture because they translate and process inputs in natural language, identify relevant medical information, and derive sensible responses. At the same time, LangChain acts as

the orchestration layer, facilitating seamless integration of these models into a conversational pipeline. It oversees the flow of conversation through entities such as ChatModel and Message, making it possible to conduct natural, multi-turn conversations (Singh et al., 2024). Hugging Face Transformers bring language skills, while LangChain helps manage conversations. Together, they create a useful AI healthcare assistant that's responsive, understands context, and fits well with medical standards.

This work lays out a fresh way to build an intelligent health chatbot using Hugging Face Transformers and LangChain. Unlike traditional language models that just stick to their training data, this approach pulls in health info from outside sources while it’s running. While past research either focused on Hugging Face models or chatbots separately, this combines both to provide answers that are relevant, accurate, and specific to health. An inventive architecture designed to satisfy the complex requirements of healthcare communication is shown by the smooth synchronization of real-time retrieval, sophisticated natural language processing, and organized conversation management.

Although RAG research and development is still under progress, academics are now categorising RAG into three categories, which are shown in Figure 2: Modular RAG, Advanced RAG, and Naive RAG. Advanced RAG employs embedding models like BERT and its generalisations, while Naive RAG utilises simple retrieval models like BM25. Depending on the job, many retrieval procedures are possible with modular RAG. The Advanced RAG type is the most often used in chatbot creation research out of all of them. Because context awareness is very important in a medical chatbot situation. When discussing health issues, it is common to need precise, up-to-the-minute access to detailed information, such as drug interactions, recent study findings, or revised clinical recommendations. Unlike Modular RAG, which separates retrieval and production but doesn't necessarily integrate them deeply, Naïve RAG gathers information without considering the importance of context. However, the Advanced RAG type provides

a more nuanced connection between production and retrieval. By firmly integrating retrieval into the generating process, Advanced RAG is able to manage such complexity and continually refine the relevance of the retrieved documents depending on the current interaction. This aids in error reduction and ensures that responses are up-to-date and medically sound, depending on the quality of the external or recovered data.

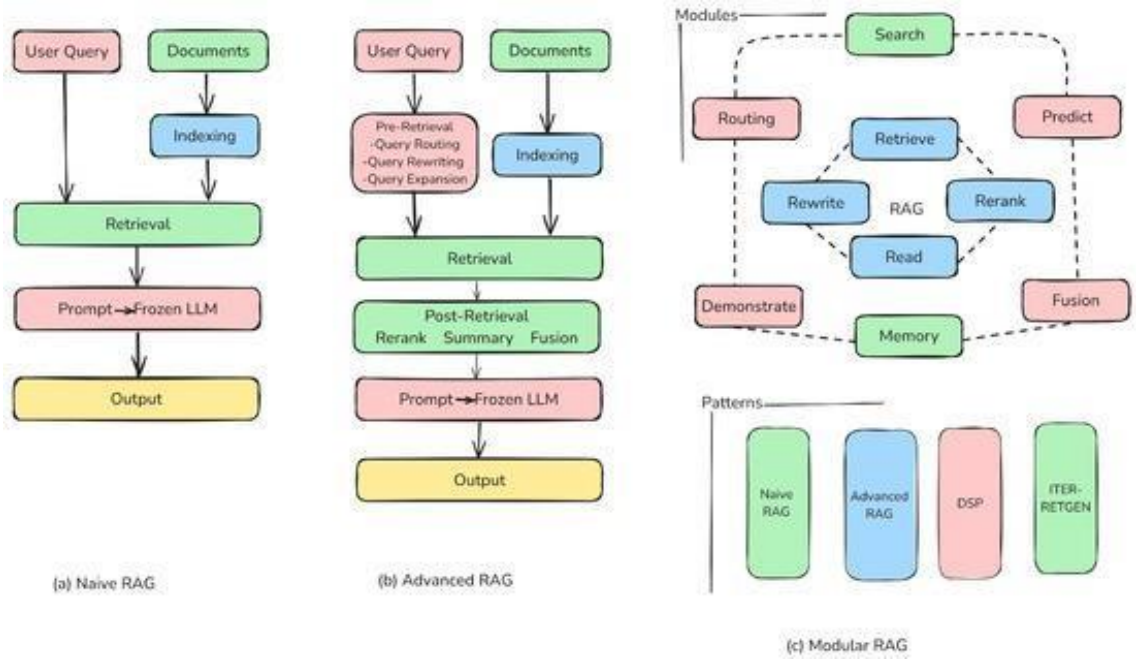

Figure 2 Illustrative comparison of the three RAG paradigms.

Indexing, retrieval, and generation are the three main phases that make up Naive RAG, as shown on the left (a). Midway through (b), Advanced RAG incorporates a number of pre- and post-retrieval optimization algorithms into its linear, chain-like process, which is otherwise comparable to Naive RAG. Modular RAG, as seen on the right (c), expands upon earlier methods by allowing for the addition or removal of functional modules as needed, thereby increasing the system's adaptability. Methods like iterative and adaptive retrieval are now a part of its process, expanding beyond sequential retrieval and production (Qu et al., 2025).

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

Large domain of computer science called artificial intelligence (AI) focuses on the study, development, and testing of software and methodology that allows machine

perception. Hugging Face tools and architecture are Python-based. To help medical professionals understand this programming language and AI processes, there is a need for a professional such as a computational technician. For clinical doctors, this will enable smoother creation of customized and task-specific models. Simply put, this platform has opened up AI to everyone by flipping the development process and accelerating it, making it more collaborative and accessible. It's taken a complicated and messy process and made it simpler for everyone. This change helps pros build advanced AI models easily, while also letting beginners and AI fans jump in without much hassle. So, Hugging Face isn't just a tool; it's a big step forward for AI.

1.8 Lang Chain

LangChain brings a fresh way to build apps that make the most of advances in natural language processing. It allows developers to create custom applications that work with language models and various data sources by providing easy-to-use chains and modular parts **(Ioannidis et al., 2023).**

- **Data knowledge:** One of the great things about LangChain is how it can link language models with all sorts of data. This ability to understand and handle information from different places really helps apps boost the accuracy and usefulness of what the model produces.
- **Agentic interaction:** To really use a language model, it needs to interact with its surroundings. This means it can answer user questions, talk to other software or systems, and perform tasks based on what the user asks.
- **Modular components:** Modularity makes it easier to code apps in a flexible way. LangChain gives developers easy-to-use building blocks that they can adjust and expand depending on what their projects need.
- **Standard chains:** These pre-built chains help developers get started and make it easier to finish common tasks quickly. LangChain really shines when it comes to growing and adjusting these networks to fit the needs of a project.

1.9 Motivation of the Research

Because of the growing complexity of healthcare data and the critical need for accurate, up-to- the-minute, context-specific medical information, smart digital assistants are becoming more important in modern healthcare systems. Though effective, classical LLMs have limitations because to their reliance on static training data, which can lead to false positives and out-of- date information. Retrieval-Augmented Generation (RAG) is a game-changing method that improves the quality and usefulness of produced material by continuously querying external medical information sources.

There is a unique opportunity to create a smart, conversational healthcare chatbot due to the growing presence of advanced NLP models through platforms such as Hugging Face and the modular, developer-centric design of libraries such as LangChain. Such a system can enhance health outcomes and decision-making by closing the knowledge gap between clinical information and patients or providers.

1.10 Problem Statement

In recent years, there has been a tremendous growth in the need for accessible, reliable, and real-time medical information, fuelled by an increasing dependency on online platforms for health-related information. Nonetheless, most current AI chatbots within the healthcare industry are not equipped to sustain contextual knowledge over multi-turn interactions and tend to output generic or unreliable information. These weaknesses result in lower user confidence and possible harm via misleading answers. Furthermore, typical NLP models are weak when handling domain-specific vocabulary and are unable to fetch up-to-date, evidence-based facts. This research fulfills the urgent necessity for an intelligent, contextual, and more reliable AI medical assistant with Retrieval-Augmented Generation (RAG), Hugging Face Transformers, and LangChain to develop a chatbot that can provide correct, personalized, and contextually appropriate medical answers in conversational format.

1.11 Novelty of the Study

This project is unique because it makes use of LangChain architecture and Retrieval- Augmented Generation (RAG) and Hugging Face Transformers to build a health chatbot. The chatbot is able to hold context-dialogue, through more than one turn, and to offer correct medical evidence-based information. To more effectively control the relevance and authority of responses, we study in this work the retrieval and generation as a whole rather than using generative models or static retrieval that are widely adopted by healthcare assistances nowadays. Applying LangChain to the conversational.

1.12 Objectives of the Research

- The aim is to see how Retrieval-Augmented Generation (RAG) can improve how AI medical assistants work.
- We want to achieve natural language understanding and generation in healthcare using Hugging Face Transformers.
- To allow the chatbot to handle smart, ongoing conversations, we'll use LangChain, a framework for chatting.
- We'll build a healthcare chatbot that uses RAG and gives accurate, current, and relevant answers to medical queries.
- Finally, we'll evaluate how well this system works in real-life healthcare situations, focusing on accuracy, relevance, and user satisfaction.

1.13 Scope of the Study

This project is about making an AI medical assistant. We'll build and test it using LangChain, Hugging Face Transformers, and the Retrieval-Augmented Generation model. Except for emergency measures and clinical tests, it is confined to the healthcare industry and handles minor medical questions, symptom information, and health advice. The goal of the project is to develop a functional prototype with the ability to comprehend medical context and respond accordingly via multi-turn dialogue. Safe and ethical testing environments also exist with evaluating the system performance based on answer correctness, contextual relevance, and user satisfaction using publically available or simulated medical data.

1.14 Significance of the Study

This research is critical in the rapidly evolving field of AI-enabled healthcare IT. This work will complete the gap between the need for appropriate, context-aware medical assistance and the availability of technology that can mimic human discourse through the union of cutting- edge tools such as Retrieval-Augmented Generation (RAG), Hugging Face Transformers, and LangChain.

- **Improve Healthcare Communication:** The study lends credence to the building of smart virtual medical care assistants that can understand and answer medicine questions with contextual adequacy. This will have the ability to promote the accessibility and effectiveness of preliminary medical counsel, especially in rural areas with limited access to medical professionals.
- **Technological Progress in AI and NLP Research:** Using RAG and the most recent Hugging Face NLP models, this research demonstrates a tangible use of cutting-edge AI technologies in the sensitive and high-risk field of medicine. It assists in maintaining momentum towards improving the interpretability and trustworthiness of AI-generated content.
- **Enhanced Contextual Understanding:** With LangChain integrated, the chatbot can properly manage multi-turn conversations and keep context across conversations. This is especially crucial in healthcare since patient inquiries are usually multifaceted and evolve as a conversation progresses.
- **Reliable and Evidence-Based Support**: With the use of the RAG framework combining retrieval from authoritative medical sources and generative capabilities, the system aims to provide answers based on authentic information, thereby ensuring that there is no chance of misinformation.
- **Foundation for Future Developments:** The project lays a platform for the development of future AI-based systems that can operate within the confines of real- world care settings. The prototype constructed can serve as a template for further enhancements, including multilingualism, electronic health record integration, and real- time symptom analysis.
- **User-Centric Assessment:** Through the assessment of the system according to precision, relevance, and user satisfaction, the study

emphasizes the importance of placing end-user experience at the forefront in the design of AI healthcare tools. This ensures technological innovation to meet the practical needs and demands of patients and healthcare workers.

Essentially, this research is a technical piece of work on AI and healthcare integration as well as a move towards providing trusted, intelligent health care assistance to a larger community.

1.15 Organization of the Thesis

A clearly defined framework with six divisions introduces a logical path of study in this thesis. Creating a RAG-based AI healthcare assistant is the focus of Section 1, which also defines the study's context, issue description, research aims, scope, and importance. In Section 2, we review the relevant literature on healthcare chatbots powered by artificial intelligence (AI), outlining its present strengths and weaknesses as well as the gaps that our study aims to solve. In Section 3, we lay out the theoretical and technical groundwork for the core technologies used in the study. These include the RAG architecture, Hugging Face Transformers for NLP tasks, and the LangChain environment for context-aware, multi-turn dialogue handling. Part 4 details the study approach, including the system design, data sources, preprocessing steps, and components integrated into a functional chatbot pipeline. Using metrics including answer accuracy, contextual relevance, and user satisfaction as well as comparisons to baseline models, Section 5 details the experimental setup and evaluates the system's performance. To wrap up the thesis, Section 6 summarizes the results and suggests directions for future research to enhance the use and implementation of AI-based healthcare assistants.

CHAPTER 2

LITERATURE REVIEW

Intelligent and readily available healthcare has entered a new era with the introduction of AI into healthcare systems. The potential for AI-driven healthcare assistants to speed up the healing process, increase patient engagement, and alleviate doctors' workloads has piqued the imagination of many. These systems have matured as chatbots, moving from rule-based models to ones based on deep learning and machine learning for natural language processing (NLP). More effective and context-aware health care interactions are now possible because to transformer models like Hugging Face Transformers, which have greatly improved robots' text- reading and writing abilities. One of the most promising developments in the field is the use of Retrieval-Augmented Generation (RAG), which combines the strength of information retrieval with the ability of generative models to provide relevant and accurate results. When it comes to healthcare, where precise and evidence-based answers are crucial, a hybrid architecture like this one is invaluable. Lang Chain, among the leading frameworks for chaining language model operations with other sources of information and tools, has a crucial part to play in increasing the modularity, control, and real-time responsiveness of such chatbots. This survey of literature discusses the history and the state of affairs with AI-based healthcare assistants, particularly transformer models, RAG architecture, and Lang Chain framework. It amalgamates key contributions, research areas with gaps, and calling for robust, privacy-preserving, and explainable AI solutions to the high-stakes and sensitive use of healthcare.

2.1 Introduction to AI in Healthcare

(Santosh et al., 2021) AI's potential to transform healthcare, particularly in resource-poor areas, is intriguing. Human transformation might be transformed by it. AI is transforming medication research and public health. Brain signals translation helps restore speech after stroke or other neurological problems. AI subfields like

computer vision outperform practitioners in several areas. Disease diagnosis and epidemic prediction are pioneering uses of deep learning in medicine. The core idea, "Transfer learning—a process that uses general datasets to create learning to specific problems," spurred more creative thinking. Talks on artificial intelligence in the workplace revolve around public health, economic performance, and social transformation. Since AI is getting more proficient at automating learning, how can society mould it for public good? How will patients, healthcare systems, and the community be affected? Famous application examples, benefits, risks, and AI issues in human society transformation are covered in this chapter.

(Väänänen et al., 2021) analysed essential components to develop effective AI-based healthcare services and provide a narrative assessment of healthcare services that incorporate AI-based services as part of their operations. Some signs of AI's benefits in healthcare include its ability to improve outcomes, aid caregivers in their work, and reduce healthcare costs. The healthcare industry's artificial intelligence market is expanding at a CAGR of 28%, so there's plenty of space for expansion. This article will explore several aspects of the healthcare industry, including financial, health, and care outcomes, and will provide suggestions and key considerations for the effective use of AI methods in healthcare.It proves that artificial intelligence (AI) has the potential to improve healthcare while reducing expenses.

(Kannelønning, 2024) AI is supposed to tackle healthcare problems, however ethics, legislation, data access, human trust, and poor clinical proof are obstacles. To traverse this complicated ecosystem, players from diverse backgrounds and affiliations must collaborate. an informal professional network enables Artificial intelligence in publicly funded healthcare in Norway. Digital meetings and interviews are seen by others who did not participate in the qualitative longitudinal case study. It concludes that further healthcare AI deployments may alleviate certain uncertainties, but others may develop. Mobilising spokespersons from non-discussing parties may strengthen hybrid knowledge generation, detect, mitigate, and monitor uncertainties, and ensure sustainable AI deployments.

According to (Arora, 2020) the innovative breakthrough, artificial intelligence (AI) is surpassing human healthcare providers in the diagnosis of some medical disorders, especially in image analysis in radiology and dermatology. Machine learning algorithms may learn from a patient's medical history, genetic data, and data collected in real-time. Materials for medical education or the integration of robots may be made using it. The potential implementation hurdles and real-world impacts on healthcare services are other sources of concern. Regulators are taking an interest in AI because of the "Software as a Medical Device" perception of it. Risks include algorithmic biases, data privacy, long-term personnel issues, automation bias, over-reliance, and corrigibility. Clinicians need to keep control of the diagnostic process and comprehend the algorithmic procedures that provide diagnoses when AI crosses-examines datasets.

(A. Kaur & Goyal, 2025) In healthcare, explainable artificial intelligence (XAI) improves AI decision-making trust and transparency. It helps patients and doctors comprehend AI models' diagnosis, treatment recommendations, and projections. Provide interpretable outputs to promote collaboration between human specialists and AI systems, encouraging accountability and ethics. This openness boosts AI adoption in healthcare settings, particularly for critical choices like diagnosis and therapy. XAI improves interpretability, cooperation, responsibility, and trust in healthcare decision-making, making it more trustworthy and informed. XAI in healthcare is new and needs further research.

(Tekkeşin, 2019) Examined the AI aspires to match human cognitive processes. Increasing healthcare data availability and advanced analytics approaches are driving a paradigm change within the medical field. what artificial intelligence (AI) can do for healthcare now and in the future. Healthcare data, whether organized or unstructured, may benefit from AI. Popular artificial intelligence (AI) methods for structured and unstructured data, respectively, include methods for processing natural language, deep learning, neural networks, and support vector machines. Neurology, cardiology, and cancer are three major fields of AI-driven disease research. We take a close look at the many ways AI is helping with stroke care, from diagnosis and early detection to treatment and prognosis evaluation. Lastly, we

discuss cutting-edge AI systems such as IBM Watson and the practical difficulties associated with implementing them.

(Shaheen, 2021b) AI technologies are transforming healthcare by predicting, grasping, learning, and acting, from identifying genetic code correlations to controlling surgery-assisting robots. Machine learning for medical treatment. three sectors of healthcare that are using AI to revolutionize the industry: medication discovery, clinical trials, and patient care. that artificial intelligence has empowered pharmaceutical firms to expedite drug discovery and automate target detection. AI has the potential to enhance data monitoring methods that are labor- intensive. Clinical trials aided by AI can process massive volumes of data with confidence, and medical AI companies provide tools to help patients in every way. Through the analysis of medical data, clinical intelligence provides insights that can improve patients' quality of life. (Quaranta et al., 2024) used of AI tools in many industries have been brought up by the revised language of the Artificial Intelligence Act (AI Act). With the establishment of new legal and procedural restrictions, A difficult issue has arisen about the usage of AI technology, namely in medical applications. There are currently rules governing AI medical devices, such as the Medical Device Regulation, and it is vital to determine how much overlap there is between these laws and the AI Act. The primary goal, with many levels of applicable rules for AI medical devices, is to provide realistic criteria for the integration of AI into healthcare systems while guaranteeing their adherence to the law.In order to have a complete picture of the problem of using AI in healthcare, we also want to demonstrate legal short circuits associated with AI medical technologies.

(Menaga & Paruvathavardhini, 2022) Explored the “AI in Healthcare" examines medical analytic algorithms and technology in several disciplines. It shows how AI can extract sophisticated medical information from books to help doctors make better judgements. To make disease detection easier, AI is being trained using medical tools and technical algorithms. Artificial intelligence is crucial in the fields of radiology, pathology, immuno-oncology, neurology, neurodegenerative diseases, chronic illnesses, and the development of pharmaceutical and chemical drugs. Both traditional support vector machines and neural networks are discussed

in relation to structured data processing. Deep learning and NLP are discussed in relation to handling unstructured data. To illustrate the use of AI in telehealth, we look at chatbots like Sky, WebMD, Ada, and Skin Vision. This chapter also covers artificial intelligence (AI) pandemic quick testing kits.

(Briganti & Le Moine, 2020) Quickly emerging clinical solutions are medical devices powered by artificial intelligence. As the amount of health data collected by smartphones, wearables, and other mobile monitoring devices continues to rise, deep learning algorithms will be able to keep up. Artificial intelligence is useful in just a subset of therapeutic scenarios; they include the detection of atrial fibrillation, epileptic episodes, hypoglycemia, and medical imaging and histopathology-based ailment diagnoses. Doctors who weren't ready for it hate it, but patients are looking forward to improved medicine since technology provides them greater control and individualized treatment. This issue highlights the need to validate new technologies with traditional clinical trials, update medical education to reflect digital medicine, and think about the ethics of connected monitoring. Covering recent studies, it delves into the pros, cons, opportunities, and threats posed by clinical AI applications to healthcare providers, hospitals, universities, and the field of bioethics. (Knapič et al., 2021) looked at medical image analysis decision support using Explainable AI approaches. The results of the Convolutional Neural Network (CNN) were made more comprehensible by using three distinct explainable methods to the identical set of medical imaging data. To make black-box predictions more reliable, the study looked at stomach images taken by video capsule endoscopy. The researchers used two methods for making the machine learning results easier to understand: SHAP and LIME. They also checked out another method called Contextual Importance and Utility (CIU) for explanations. The results found that the CIU method was clearer than LIME and SHAP, helping people make better choices.

(Eskandar, 2023) AI can now spot illnesses in medical images with more than 90% accuracy, and it's changing the game in healthcare, diagnostics, drug research, and treatment. One of the many ways AI can help is by making hospital management and surgeries run more smoothly, as shown in this interesting research

review.While opening up fascinating possibilities, issues like biases and ethics need careful consideration. Inspiring case studies and captivating data highlight AI's present effects and encourage further cooperation in this astounding technological transformation, advancing us towards a day when AI works in unison with medical professionals to provide better patient outcomes.

(Khalifa & Albadawy, 2024) Clinical prediction is essential in healthcare for predicting patient outcomes. In this field, By improving the precision of diagnosis, treatment plans, illness prevention, and individualized care, AI improves healthcare efficiency and patient outcomes. In eight different areas of clinical prediction—diagnosis, prognosis, risk assessment, therapy response, disease progression, readmission risks, complications, and mortality prediction— artificial intelligence (AI) improves accuracy when compared to human researchers.AI helps predict clinical outcomes in oncology and radiology. The paper highlights AI's disruptive influence on diagnosis, prognosis, personalised therapy, and patient safety. Improving data quality, multidisciplinary cooperation, ethical AI practices, AI education, clinical trials, regulatory control, patient involvement, and AI system monitoring and improvement are recommendations.

(Longoni et al., 2019) AI is revolutionising healthcare, yet consumer receptivity to AI in medicine is unknown. AI-provided healthcare is unpopular in actual and hypothetical options, individual and collaborative assessments. Study 1 shows that consumers are less likely to use healthcare, study 2 shows reduced reservation fees, studies 3A–3C show less sensitivity to provider performance, and research 4 shows negative utility from automated providers. Consumer opposition to medical AI stems from uniqueness neglect, the belief that AI providers cannot account for individuals' particular qualities and situations. Consumers who feel distinctive oppose medical AI more (research 5). Uniqueness neglect mediates medical AI resistance (study 6), but it disappears when AI provides personalised (study 7), consumer-based (study 8), or supporting (study 9) healthcare provider decisions. These results advance the psychology of automation and medical decision-making and provide ways to improve consumer acceptability of AI in medicine.

(N. Sharma & Kaushik, 2025) Artificial intelligence has substantially improved the accuracy, validity, and trustworthiness of medical diagnoses. algorithms, review components, criteria, urgent follow-up approaches, causal inference methods, synthetic adversarial systems, and advancements to the performance and robustness of AI infection localization models.Combining GANs that synthesise real-world tests to decrease bias and increase program generality reduces symptomatic mistakes dramatically. Consideration components improve interpretability and reliability by highlighting impacted clinical performance or measures. Causal inference approaches demonstrate promiscuous, personalised healing tools. Unified learning improves collaborative tutoring by addressing security and administrative compliance. For full AI control in healthcare transportation, future thinking must include cross- models, ethical suggestions, flexibility, and integration with rising advances.

(Shaheen, 2021a) discussed the advantages and difficulties of AI in healthcare, with a particular emphasis on the financial and medical aspects. Sentiment analysis can understand and react to human emotions, surgical robots increase surgical accuracy, intraoperative assistance via video photos and communication systems is helpful, and intelligent data inclusion enhances the quality of decision-making. Additionally, AI may help professionals manage their workload so they have more time to contact with patients. Data biases, the need for huge datasets, possible confidentiality issues, and the possibility of inaccurate AI systems harming patients all obstacles, nevertheless. highlights the potential benefits of new technology and the need for future study to overcome these issues in order to guarantee the greatest results.

(Lainjo, 2024) The text talks about how artificial intelligence (AI) can be used in healthcare. It points out that AI can improve patient care, make processes smoother, and help with accurate diagnoses. It also mentions the impact of AI on clinical decision support and personalized care. On the flip side, there are discussions about legal and ethical concerns, like patient privacy and the need for solid tech systems. The importance of having good quality data for accurate decision-making is highlighted, along with the idea that there should be a sense of responsibility in

keeping patient data safe. Overall, AI has the potential to really change the healthcare field.

2.2 Conversational AI and Chatbots in Healthcare

(Nadarzynski et al., 2024) The goal was to find ways to make AI designs fairer and cut down on bias in the tech. They put together a framework based on 17 principles of AI use, using a research method that looked at people's experiences. They talked with 33 individuals from different backgrounds, including people working in AI, business leaders, and doctors. They came up with a ten-step plan that includes activities aimed at promoting fair AI. The focus was on working together and involving patient groups, leading to practical suggestions to improve fairness in conversational AI in healthcare.

According to (Lal & Neduncheliyan, 2024) A promising tech in healthcare is conversational AI. It helps create personalized chats between virtual assistants, experts, and patients. This technology uses patterns and gives thoughtful responses based on what's happening in the conversation. To improve the performance of sentiment analysis models, a new method known as Generative Pretrained based Recurrent Neural Network (GPbRNN) was created. By streamlining decision-making, improving treatment delivery, and delivering customized data, this software has the potential to radically alter the healthcare system. Additional investment in healthcare R&D should lead to better service delivery and better patient outcomes.

(Milne-Ives et al., 2020) An increase in the need for these services, along with developments in AI, has led to the creation of conversational bots to assist with healthcare-related activities. Agents might automate mundane operations, expand people's access to healthcare, and allow physicians to concentrate on more complicated situations. Thirteen publications on the topic of healthcare conversational bots with unrestricted natural language processing have been evaluated since 2008. A third of the studies had positive or mixed effectiveness, whereas 27 out of 30 trials had good usability and 26 out of 31 trials had good satisfaction. User opinions vary in terms of quality, nevertheless. The agents' health

care efficacy should be carefully assessed and opportunities for improvement should be identified through better research design and reporting, as several trials were poor. Additional research on agent privacy, security, and cost-effectiveness is required.

(J. Gupta et al., 2022) The use of cutting-edge conversational AI systems by healthcare companies has been on the rise in recent years. Automated AI systems mainly serve as interfaces to enhance the quality of human-computer interaction. Both patients and physicians are feeling the effects of the dramatic shift in the healthcare business brought about by conversational AI. Rasa, Google Dialogflow, and IBM Watson are just a few examples of the NLU-using natural language processing (NLP) systems used by conversational AI. Ainume, a powerful conversational AI agent, was developed on the Google Cloud Platform (GCP) utilizing the Google Dialogflow architecture. Ainume determines the symptoms of common and chronic diseases before suggesting nutraceutical remedies to relieve them. heart conditions, an area where Ainume has great success.

(NV et al., 2023) Healthy people are physically, mentally, and socially healthy. Chatbots have been widely used in this area, yet there is still potential for innovative applications. Healthcare conversational AI applications are adaptable and industry-specific. Patients may use them to learn more about their ailment, potential treatments, and insurance coverage. that healthcare chatbots may enhance patient happiness and cut wait times, prompting several businesses to investigate using them. Chatbots in healthcare provide several benefits, including monitoring, anonymity, personalisation, and in-person engagement. This case study uses user input on patient symptoms to identify the probable disease type. A specialist doctor will be referred to the patient based on the kind of illness and recommended measures. Symptoms were extracted using a sequential model, and the KNN approach was used to predict the patient's illness type.

According to (Meshram et al., 2021) the development of technologies such as artificial intelligence (AI), big data, and the internet of things (IoT) has led to several technical breakthroughs during the last ten years. The applications of these

technologies are many. "Chatbot," often known as "Chatterbot," is one such program. Conversational AIs that imitate human speech are called chatbots. Synthesizing AI with NLP is the heart of the method. Chatbots have contributed to technological progress by automating mundane operations and reducing the need for human participation. Numerous sectors make use of chatbots, such as the commercial, medical, and academic spheres.a number of publications and spoke about the different kinds of chatbots and their benefits and drawbacks as part of the research. the assessment, chatbots are universally applicable due to their accuracy, lack of reliance on human resources, and round-the-clock availability.

According to (Bhirud et al., 2019) research, 80% of basic, small-scale illnesses that account for 60% of medical visits may be treated at home with easy fixes. Healthcare services are provided by chatbots; however, they only answer generic healthcare FAQs and don't have the ability to communicate naturally with people. Work is underway to make chatbots able to converse like people, so enhancing communication and creating a virtual companion. This may be achieved by incorporating Natural Language Processing (NLP), Natural Language Understanding (NLU), and Machine Learning (ML) techniques into common programmed chatbots. The healthcare chatbot system is covered in this study, along with a comparison of the different NLU, NLG, and ML algorithms that should be used.

(D. Sharma et al., 2022) AI has been around for almost 50 years. AI has advanced due to processor power, data accessibility, and algorithm enhancements. AI-powered chatbots replicate user interactions. Chatbots reduce hospital wait times, appointments, and consultation meets, helping people find the proper clinician quickly. In reducing hospital visits and unneeded treatments and offering ideas and warnings, chatbots lessen healthcare personnel' burden. However, chatbots in healthcare present several obstacles. Research on healthcare chatbots is systematically reviewed in this study. provide general information about the application type, technologies, and evaluation methods used to evaluate healthcare chatbots and serve as a research guideline for chatbot development in other fields.

(A. Abd-Alrazaq et al., 2020) the technological parameters that other research has used to assess health care chatbots is the goal of this study. Seven bibliographic databases and reference list verification were used in the study. 27 technical parameters, such as usability, classifier performance, speed, response production, response comprehension, and aesthetics, were assessed across 65 included research. Global usability measures and survey designs dominated the technical metrics, which were varied. It is challenging to assess the effectiveness of health chatbots because to the absence of standardisation and objective metrics, which might impede progress. In addition to creating a framework of technical measures with suggestions for certain situations for their inclusion in chatbot studies, the paper advises researchers to use metrics calculated from conversation logs more regularly.

(L. Xu et al., 2021) assessed the developments in medical chatbot technology, with a particular emphasis on cancer treatments and its uses in patient assistance, diagnosis, treatment, monitoring, workflow efficiency, and health promotion. Limitations and areas of concern are examined throughout the essay, including technological, ethical, moral, security, and regulatory requirements. Diagnoses, treatments, monitoring, support, workflow, and health promotion were the main categories used in our literature search, which we ran across a number of databases. Chatbots might save expenses, simplify process, and enhance patient outcomes in clinical practice. There has been a lack of comprehensive research into other potential applications in global health, education, and pandemic relief. The quality of patient treatment may be greatly enhanced, clinician burden could be balanced, and medical practice could be revolutionised with more research and multidisciplinary cooperation.

(Abbas et al., 2025) investigated the use of chatbots in medicine during the last 20 years, mostly in cancer, psychiatry, and chronic illness management. They found 38 clinically relevant papers via a thorough literature search. Studies indicate that chatbots provide effective early symptom identification, personalised treatment planning, and emotional support for mental health needs. how chatbots may enhance communication between patients and healthcare practitioners, leading to

better access to medical information and treatment adherence. the possibilities for chatbots to enhance the provision of healthcare, but additional research is needed in radiology and sophisticated diagnostics. As tech keeps moving forward, chatbots are set to play a big role in making health care easier and helping patients understand their health better. It's important to keep putting money into chatbot tech in healthcare to really make the most of what they can do for patient care.

(Fan et al., 2021) The study looked into how AI chatbots can be used in healthcare, especially for self-diagnosis. It analyzed over 47,000 consultations that about 16,500 people had with a self-diagnosis chatbot in China over six months. Many users were middle-aged and older, seeking advice on various health issues. The study highlighted some challenges, like users reporting health concerns and dropping out of sessions. It suggests that for these chatbots to be more helpful and patient-friendly, they need to be more reliable, provide better information, and be easier to use. Improvements in the onboarding process could also help engage users more effectively.

(Basharat & Shahid, 2024) looked into the ethics of using AI chatbots in healthcare, stressing how important trust and reliability are. As part of the methodology for the qualitative study, A diverse range of participants, including patients, academic researchers, medical professionals, ethicists, and legal experts, were interviewed in thirteen separate semi-structured interviews. Trust, reliability, ethical issues, and possible ethical repercussions are the four main topics that emerge from the research. Data security, patient privacy, prejudice, and responsibility are some of the many problems that might be highlighted, it contributes to our understanding of AI- enabled healthcare chatbots.Trust and dependability are crucial in reducing ethical concerns. (Li, 2023) Bots like ChatGPT and Google Bard employ AI and natural language processing to interpret client enquiries and respond in a conversational manner. ChatGPT, an OpenAI chatbot, has proven popular online. Patient and public health might benefit from AI chatbots. The huge volumes of data they are educated on may contain sensitive patient and corporate data. Increased chatbot usage raises data security concerns that should be addressed but are understudied. This study identifies AI chatbots' biggest security issues and

proposes methods for securing sensitive health data. The influence of ChatGPT on healthcare is examined. It also lists ChatGPT's main security threats and mitigating strategies. It closes with health care AI chatbot policy implications.

(Sepahpour, 2020) Examined the AI-powered mental health chatbots might transform healthcare. Conversational entities replicate human discussions and interact with people like humans. This technology may overcome regional, economical, and social stigma-related treatment hurdles. The favourable results are encouraging despite the risks. As a therapeutic, the technique is experimental and unapproved. It addresses ethical issues of chatbots for mental health care, emphasising the need to recognise damages and avoid them. Chatbots' therapeutic potential as a barrier-reduction tool is promising, and ethical concerns are essential for their ethical growth in mental health.

(Garcia Valencia et al., 2023) discussed the moral ramifications of using chatbots in nephrology, highlighting the need of strong rules for data gathering, sharing, and preservation in order to preserve privacy and guarantee data security. It recommends establishing suitable degrees of data access, investigating anonymisation strategies, and putting encryption techniques into practice. Informed permission and open data use policies are essential ethical factors. To tackle any biases, it's a good idea to regularly review algorithms, use diverse strategies, and keep an eye on things constantly. It’s also important to make chatbots work better, ensure they're easy to use, and have clear permission rules. The study points out the need for cultural awareness, support in different languages, and a good mix of automation with human touch. Ongoing research and new ideas are really key to making the most of chatbot technology and improving patient care.

2.3 Transformer Models in Natural Language Processing

(Kalyan et al., 2021) Transformer-based pretrained language models, like GPT and BERT, have shown great results in different language tasks. They learn from a ton of text data using a self-supervised approach, which helps them understand language better and apply that knowledge to new problems. This survey takes a closer look at T-PTLMs, breaking down key concepts like how they're trained, their

methods, the tasks they handle, and how they can be adapted for specific uses. It wraps up with some benchmarks for testing these models and points out helpful libraries for working with them, along with suggestions for future research.

(Wolf et al., 2020) Looked into how improvements in model pretraining and design have played a big role in recent progress in natural language processing. Transformer topologies have made it easier to create models with more capacity, and pretraining has enabled the efficient use of this capacity for a range of applications. To make these developments available to the larger 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 main objective of this study. This work demonstrates that by fixing segmentation and labelling job limits on four datasets, the performance of XLNet and vanilla RoBERTa models may be greatly improved. in order to guide the development of future deep NLP architectures, we will first suggest ways to make the Transformer architecture more expressive.

(Kim & Awadalla, 2020) A state-of-the-art method for NLU applications is the transformer model. Over time, models become better at a wide variety of activities. Computationally challenging are transformer models due to their slower inference time compared to typical procedures. Optimal inference-time performance of Transformer-based models in NLU applications is achieved with the help of Fast

Formers, a set of recipes introduced in this study. Using numerical optimization, structured pruning, and knowledge distillation, we show that inference efficiency may be significantly improved. Optimal settings for natural language understanding challenges and pretrained models may be easily achieved with the aid of our superb recipes. compared to baseline CPU models, we get a 9.8x to 233.9x speedup using the Superglue test's recommended recipes. The strategies provided can increase GPU speed by as much as 12.4 times. By using Fast Formers on an Azure F16s v21 instance, the cost of handling 100 million requests may be reduced from 4,223 USD to just 18 USD. A sustainable runtime was achieved, according to the SustaiNLP 2020 shared task metrics, with an energy reduction of 6.9x - 125.8x.

According to (Canchila et al., 2024) the use of human language in computer systems, known as NLP, is gaining importance in different fields such as research, everyday activities, trade, and entrepreneurship. Numerous IT businesses invest in NLP approach, model, and product development. Additionally, open-source contributions to the field are increasing. Despite advancements, it might be tough to comprehend the present state of NLP and the most effective models. To assist people navigating the ever-changing NLP environment, they have compiled a detailed overview of the current research and achievements.

(Turner, 2023) Elements of neural networks, such as the transformer, have the potential to acquire meaningful representations of data sequences. The transformer has recently been the engine that has propelled advances in computer vision, spatio-temporal modeling, and natural language processing. The mathematical depictions of architectural and design intuitions are often lacking in introductions to transformers. 1 In addition, the winding course of study may provide novel explanations for transformer components. The purpose of this essay is to explain transformer building in a way that is easy to understand while still being mathematically precise. Since training is normal, they will not address it. They take it as read that you know your way around linear transformations, multi-layer Perceptrons, softmax functions, and the fundamentals of probability and machine learning.

(Xiao & Zhu, 2023) Empirical machine learning models of natural language processing have been dominated by transformers. In this work, they outline the fundamental ideas of transformers and highlight important methods that have contributed to their recent development. a number of model improvements, typical applications, and an explanation 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 understanding of the advantages and disadvantages of these models, they also provide a summary of the major concepts influencing this topic.(Min et al., 2022) One popular AI technique that has shown promise in modelling graph- structured data is the Transformer model. A thorough analysis of the literature and a methodical assessment of Transformer variations for graphs are, however, absent. Graph Transformer models are thoroughly reviewed in this examination from the standpoint of architectural design. There are three common methods for adding graph information to the Transformer: Enhanced Attention Matrix from Graphs, Improved Positional Embedding from Graphs, and GNNs as Auxiliary Modules. Additionally, the research evaluates these elements on well-known graph data benchmarks, demonstrating the advantages of Transformer's present graph-specific modules.

(Pol et al., 2024) Hugging Face is a key player in the world of AI and natural language processing. They're well-known for their open-source tools that help people build and use advanced NLP models. This analysis takes a closer look at Hugging Face, focusing on their main technologies, like the Transformers library, and their mission to make AI accessible to everyone. It covers various applications of Hugging Face models and shows how they can boost productivity and creativity in areas like healthcare, finance, customer service, and education. The piece also discusses what’s next for Hugging Face in the AI and NLP space, their strong community, and how they connect with other tech. In the end, it wraps up by touching on Hugging Face's future and its impact on AI and NLP.

(Pourkeyvan et al., 2024) Finding and treating mental health issues early can really help with recovery and prevent serious problems later on. This research looks into how social media and language models can help predict signs of mental illness based on what users post. We compare and contrast four separate Hugging Face BERT models with popular machine learning techniques utilized in automated depression detection research today. that, with an accuracy rate of up to 97%, the new models perform better than the earlier strategy. Upon analysing the data, we discover that even little pieces of information—such as user biographical descriptions—have the capacity to forecast mental illnesses, supporting previous findings. that social media data is a great way to screen for mental health issues, and that this important work can be efficiently automated using pre-trained models.

(Chhabra et al., 2023) adjustments to pricing forecasting, product maintenance, and further managing the company's sales and marketing department, sentiment analysis (SA) is required based on customer opinions. Multiple ML models have been trained using dataset variability as their basis. When two companies accomplish comparable work, they usually see it as a plus.

However, the time limitation of training data creates a gap region. The amount of SA's practicability is largely determined by social media. A brief case study is shown here, using the Python transformer library and a pre-trained model to collect data in a more efficient manner. The same is obtained with older machine learning techniques, albeit the accuracy varies depending on the dataset and the goal structure. SA now has another, more accurate platform thanks to Hugging Face.

(G. Kaur et al., 2024) investigated how well four well-known Python sentiment analysis libraries—Text Blob, Vader, Flair, and Hugging Face Transformer—perform in identifying the polarity and strength of feelings in romantic letters. We looked at 500 sentences out of 300 love letters. The accuracy and quality of the sentiment annotations were assessed by human specialists. Low to moderate agreement was indicated by Cohen's Kappa values, and each tool demonstrated distinct advantages in managing emotional complexity. Additionally, the research identifies shortcomings and suggests unique methods for assessing sentiment analysis techniques in romantic correspondence. The results add to the burgeoning area of sentiment analysis and provide guidance for creating natural language models that are more appropriate for sensitive and intimate domains.

(Chow et al., 2024) discussed how conversational AI using Natural Language Processing (NLP) may revolutionise the way artificial intelligence (AI) and healthcare are integrated. This study looks at Large Language Models (LLMs) and covers several areas. It starts with an overview of conversational AI and its role in healthcare. Next, it goes over basic natural language processing techniques, focusing on how they help create better conversations in healthcare settings. We'll also explore how LLMs have developed within NLP frameworks, discussing the benefits and challenges of using these models in healthcare. From systems that assist healthcare professionals to tools that focus on patients, such as diagnostic and treatment suggestions, applications that are relevant to healthcare debates are detailed. Patient confidentiality, ethical considerations, and conformity with regulations are some of the legal and ethical concerns addressed. Conclusions drawn from the paper highlight the revolutionary potential of LLMs and NLP in changing healthcare interactions, while also recognizing current challenges and

forecasting future advances.

(Nerella et al., 2023) The ubiquitous nature of artificial intelligence (AI) in society, especially in healthcare, is causing a revolution in a wide range of applications in the Transformers neural network architecture. Although it was developed to address problems in natural language processing (NLP), a deep learning architecture known as Transformer has discovered uses in other domains, including healthcare. Among the many kinds of data surveyed here are those pertaining to physiological signals, biomolecular sequences, social media, electronic health records, and medical imaging. Potential applications of these models include drug and protein production, data reconstruction, clinical diagnosis, and report generation. In order to find studies that were relevant, we employed PRISMA guidelines. Computation cost, model interpretability, fairness, ethical considerations, and environmental impact are some of the benefits and drawbacks of using transformers in healthcare.

(Nerella et al., 2024) A number of industries, including healthcare, are embracing the fast evolving Transformers neural network architecture, which was originally designed for NLP tasks. Studies including clinical natural language processing, electronic health records, social media, bio-physiological signals, and biomolecular sequences have all made use of this design. Critical care adverse outcome prediction and the generation of surgical instructions are two further areas where it has found utility. Clinical diagnostics, report writing, data reconstruction, and protein and drug manufacturing are some of the many applications of transformers. But problems like computing cost, interpretability of models, equity, ethical considerations, and environmental effect must be resolved.

(Bird & Lotfi, 2023) investigated how chatbots may be used to help those who are depressed or anxious. The study finds an efficient hyperparameter set by topology optimisation that can predict tokens with an accuracy of 88.65% and 96.49% and 97.88% for the proper tokens. Despite the stigma attached to asking for assistance, the research shows how chatbots may provide anonymous, easily available support. It is necessary to recognise the drawbacks and difficulties of using chatbots to assist with mental health, and more research is recommended to fully grasp both their

potential and constraints, guaranteeing its growth and implementation in an ethical and responsible manner.

(Y. Zhang et al., 2023) ChatGPT shows a lot of potential for healthcare. It's part of the rapid growth in AI. While it could help make healthcare more efficient and improve things like education and diagnosis, there are still some problems to deal with, such as accuracy, privacy, and ethical concerns. Moving forward, research should focus on boosting the model's performance, tackling data gaps, and sorting out copyright and ethical issues. This way, AI in healthcare can be more effective.

2.4 Retrieval-Augmented Generation (RAG) Framework

(Pradeep et al., 2025) The TREC 2024 RAG Track focuses on boosting creativity in how we evaluate retrieval augmented generation (RAG) systems. We'll also work on setting standard input and output definitions, building a reusable framework, selecting new MS MARCO V2.1 collections, and sharing development topics with the public. Along with a web-based user interface for an interactive arena, the framework will uncover important industry standards such as OpenAI's GPT-4o and Cohere's Command R+. In order to lay the groundwork for future RAG systems to adhere to, the Ragnarök architecture and baselines are intended to be made publicly available. Search engines that integrate real-time data into extensive language models would benefit from this.

(Sudhi et al., 2024) The size and complexity of large language models (LLMs) make it such that they don't explain anything in their answer. As a result, less people will trust LLM-based tools, particularly RAG for QA tasks. With RAG-Ex, users may get rough explanations on why LLMs generated a written response from user input, regardless of the model or language used. allows both commercial and open-source LLMs to function. We evaluate the significance scores of our generic explainer's approximated explanations in English and German QA tests, along with their link with LLM performance. Our explanation achieves an F1-score of 76.9% according to end user annotations, comparable to model-intrinsic techniques in comprehensive user surveys.

(Jiang et al., 2024) the classic RAG structure, the retriever must explore a wide

corpus for the "needle" unit, whereas readers simply need to construct replies from short retrieval units. This mismatch might cause suboptimal performance because short units lose contextual information, introducing hard negatives during retrieval. To address this, Long RAG employs what it calls a "long reader" and a "long retriever." Long RAG takes the complete corpus and uses it to create 4K-token units that are 30 times longer. It then applies this to two datasets based on Wikipedia, NQ and HotpotQA. The number of units decreases from 22 million to 600,000, leading to robust retrieval with just a few of elite units remaining. In the absence of training, Long RAG possesses 64.3% HotpotQA EM and 62.7% NQ. On non-Wikipedia datasets, it scores 25.9% on Qasper and 57.5% on Multifield-en.

(S. Gupta et al., 2024) examined the RAG architecture, emphasising retrieval and generation integration for knowledge-intensive jobs. A recent study on RAG technology looks at new advances in retrieval-augmented language models and their use in tasks like information lookup, answering questions, and writing summaries. It covers the latest methods for boosting retrieval efficiency. The research also talks about ongoing challenges like scalability, bias, and ethical issues. Future research aims to make RAG models more dependable, broaden their use, and tackle social concerns. This study can help researchers and industry folks check out RAG's potential in NLP and where it might head next.

(Saad-Falcon et al., 2023) Hand annotations are often used to improve input searches, find passages, and generate responses when testing retrieval-augmented generation (RAG) systems. The Automated RAG Evaluation System (ARES) rates these systems based on three main things: how relevant the answers are, how true the responses are, and how relevant the context is. ARES creates training data to check the quality of RAG parts and helps train lightweight language model judges. For prediction-powered inference (PPI), ARES uses a small number of human-annotated data points to reduce errors in predictions. ARES accurately evaluates RAG systems across eight tasks that need a lot of knowledge, like KILT, Super GLUE, and AIS, with minimal human input. The assessments from ARES stay reliable even when the domains change, regardless of the altered queries or

documents used in the RAG systems tested.

(Fleischer et al., 2024) Setting up a Retrieval-Augmented Generation (RAG) system can be tough because it involves looking at a lot of data, understanding different use cases, and figuring out design details. It's not an easy task, and you need to take a careful approach to check how well the system retrieves information and the quality of what it produces. That's why we've created RAG FOUNDRY—an open-source framework that helps improve RAG big language models, and it's available for free. RAG FOUNDRY aims to make it easier to create RAG environments using data-augmented datasets. It streamlines things like data generation, training, inference, and evaluation. This means users can quickly create datasets and train models with their own expertise. We've tested this framework with the Llama-3 and Phi-3 models using different RAG setups, and we've seen solid improvements across three data- heavy datasets..

(Gao et al., 2023) Large language models (LLMs) are powerful, but they have some issues like making up information, not having the latest updates, and giving unclear answers. Retrieval- Augmented Generation (RAG) helps fix these problems by improving accuracy, reducing errors, and keeping the info fresh. It works by blending data from LLMs with outside knowledge sources. This article explains RAG's three main approaches—Naive, Advanced, and Modular—and its three important parts—retriever, generator, and augmentation tools. It also looks at how to evaluate these systems and the metrics used in the current automated evaluation setup.

(Han et al., 2024) We looked into how we could automate tasks like pulling data and spotting trends in systematic literature reviews (SLRs) using something called Retrieval-Augmented Generation (RAG) in large language models (LLMs). RAG helps cut down on mistakes by blending what LLMs generate with real-time data. Our study focuses on three key parts of the RAG system: creating, improving, and retrieving information. We also suggest some future directions to explore, like using LLMs that fit specific fields, processing different types of data, and trying out other retrieval methods. It's important to find a good balance between quick engineering,

RAG techniques, training methods, and picking the right LLM.

(Ru et al., 2024) Even though Retrieval-Augmented Generation (RAG) has shown a potential capacity to use external information, its modular architecture, evaluation of long-form replies, and measurement reliability make a thorough assessment of RAG systems difficult. provide RAGCHECKER, a fine-grained assessment system that includes a number of diagnostic measures for the production and retrieval modules. Compared to other assessment measures, RAGCHECKER has noticeably superior associations with human judgements, according to meta evaluation. assess eight RAG systems and do a thorough performance study using RAGCHECKER, exposing valuable trends and trade-offs in the RAG architectural design decisions. Researchers and practitioners may use RAGCHECKER's measurements to help them create more efficient RAG systems.

(X. Zhang et al., 2024) Large Language Models (LLMs) reflect human-level thinking, knowledge retention, and conversation skills. Even the most skilled LLMs encounter obstacles including hallucinations and constantly upgrading their knowledge. Research aims to solve this barrier by arming LLMs with external information using Retrieval Augmented Generation (RAG). However, two major obstacles hampered RAG development. Lack of detailed and unbiased comparisons across innovative RAG algorithms is a developing issue. Second, open- source technologies like Llama Index and Lang Chain use high-level abstractions, limiting openness and innovation in algorithms and assessment measures. present RAGLAB, to fill this need with an open-source library that is both modular and geared for research. RAGLAB is a comprehensive environment for RAG algorithm research and can reproduce six methods. They make good use of RAGLAB to compare six RAG algorithms over ten benchmarks. Scientists may evaluate existing algorithms and develop whole new ones with the help of RAGLAB.

(P. Zhao et al., 2024) Advancements in model algorithms, core models, and access to high- quality datasets are directly responsible for the growth of AI-generated content (AIGC). Ongoing challenges for AIGC include managing high training and inference costs, avoiding data leaks, dealing with long-tail data, and maintaining

up-to-date information.To tackle these issues, a new paradigm called Retrieval-Augmented Generation (RAG) has arisen. The information retrieval technique is introduced by RAG, which improves accuracy and robustness. By categorizing RAG foundations according to the manner in which the retriever augments the generator, this research surveys previous attempts to incorporate RAG approaches into AIGC situations. Practical applications, benchmarks, constraints, and future research goals are also covered, in addition to various RAG augmentation strategies.

(Yu et al., 2024) The use of Retrieval-Augmented Generation (RAG) to improve generative models is becoming more common in NLP. The hybrid nature and dependence on dynamic information sources of these systems, however, make evaluation of them a unique difficulty. In order to give a thorough review of RAG system assessment and benchmarks, this study used a Unified assessment Process of RAG (Auepora). Present standards are compared in this study, along with their shortcomings and possible future approaches for RAG benchmarks. Quantitative measures like as relevance, correctness, and fidelity are considered.

(Singh et al., 2025) A revolution in artificial intelligence has been ignited by Large Language Models (LLMs), which replicate human speech patterns and grasp spoken language. However, because their training data is static, they are unable to reply to dynamic requests and will thus provide inaccurate or out-of-date replies. Retrieval-Augmented Generation (RAG) use real- time data retrieval to deliver contextually relevant replies in real-time. Traditional RAG systems lack multi-step reasoning and complicated job management due to static processes. Embedding autonomous AI agents within the RAG pipeline allows dynamic retrieval strategy management, contextual understanding refining, and adaptive workflows. This overview covers Agentic RAG designs, applications in healthcare, banking, and education, and implementation methodologies.

(Zheng et al., 2025) By utilizing retrieval-augmented generation (RAG), a significant AI tool, LLMs are able to tap into other information sources and enhance their abilities. Because it adds meaningful information to model outputs, it has

proven very helpful in AI-Generated Content (AIGC). To overcome the drawbacks of depending entirely on internal model information, RAG has recently been incorporated into CV. Visual comprehension and visual production are the primary foci of this survey, which examines the present status of retrieval-augmented approaches in CV. Subjects covered include multimodal question responding, medical reporting, and simple image recognition. Recent developments in RAG for embodied AI have centered on domain-specific improvements, interaction, planning, task-execution, and multimodal perception.

(Gan et al., 2024) Even though knowledge-heavy jobs are pretty successful, they haven't really made the most of large language models (LLMs) because they're too expensive and slow. Retrieval-augmented generation (RAG) helps some, but most models still just match queries with documents based on similarity. This study introduces a new system called METRAG, which stands for Multi-layered Thoughts. METRAG enhances RAG by adding a focus on compactness, a simpler utility model, and a smarter approach that combines similarity with utility. Tests in knowledge-intensive jobs show that METRAG is the best choice.

(R. Yang, 2024) Case GPT is a system that helps improve case-based reasoning in the legal and medical fields. It uses advanced tech to help with searching and generating language. By allowing searches based on context, it makes data more useful and easier to access. The system can identify complex patterns in existing case data, which helps it find relevant examples and provide useful analysis. It's already made impressive strides in medical diagnostics and retrieving legal cases, showing its potential to change how we find information and support decisions.

(Alkhalaf et al., 2024) used zero-shot quick engineering on generative artificial intelligence (AI) models to summarise organised and unstructured EHR data and extract nutritional data. The Llama 2 13B model uses 40 Australian RACFs' unstructured and organised EHRs. The model was tested for its capacity to create structured summaries of a client's nutritional status and extract malnutrition risk indicators. Zero-shot learning applied to generative AI models accurately summarised and extracted nutritional information from RACF customers with

93.25% accuracy. Adding retrieval augmented generation (RAG) boosted summarisation accuracy to 99.25%. The model also extracted risk variables with 90% accuracy. When specifics were not supplied, the model may have hallucinatory limits.

(Amugongo et al., 2024) Although LLMs have demonstrated promise in healthcare, they're not without their problems, including as using out-of-date training data and an absence of transparency. External knowledge sources can be used using retrieval augmented generation (RAG) to bolster LLM replies. On the other hand, healthcare evaluation frameworks, readily available statistics, and RAG procedures are all understudied.This study examines RAG-based methodologies in healthcare, focusing on retrieval, augmentation, and generation processes. It highlights the need for more research and development to address ethical implications and ensure responsible RAG use in the medical field.

(Gargari & Habibi, 2025) To enhance LLM performance by integrating data from other sources, an AI approach called retrieval-augmented generation (RAG) is employed. Clinical decision support, patient care, and accurate diagnosis are just a few of the medical fields that could gain from its use. Medical data extraction, clinical decision-making, and patient diagnosis are just a few domains where RAG-based systems have demonstrated to outperform more traditional techniques. Model assessment, cost-efficiency, and lowering the threshold for AI hallucinations are still areas of concern, nevertheless. This paper highlights the potential of RAG to advance medical AI applications and calls for further cooperation between AI researchers and healthcare practitioners, as well as further improvement of retrieval processes and embedding models.

(Puhakka, 2025) investigated how retrieval-augmented generation (RAG) in conjunction with locally deployed large-language models (LLMs) might be used to solve healthcare issues such as data privacy concerns and resource limitations. The prototype personal health assistant that integrates LLMs utilizing RAG and dynamic information retrieval from synthetic patient data and clinical guidelines is the main focus of the work. Installing the system locally on-premises allows users

to comply with GDPR and lessens their reliance on cloud-based services. In this study, three open-source LLMs—Deepseek-R1-Distill-Llama, Phi-4-miniinstruct, and LlamaMedicine—are tested using curated clinical recommendations and FHIR-organized simulated patient data. Although it required more time to process the data, the Deepseek-R1- Distill-Llama model achieved the best accuracy. Additional domain-specific adjustments, automated evaluation methods, and standardized formats are needed to improve scalability and practical applicability, according to the study.

(Anandavally, 2024) examined how Retrieval-Augmented Generation (RAG) and GPT-4- powered virtual health assistants might enhance patient outcomes, operational effectiveness, and healthcare accessibility. By automating repetitive processes and providing precise health information, In particular, the assistant helps alleviate healthcare disparities in underserved areas by reducing the burden of medical experts. The results show that virtual health assistants powered by AI have the ability to revolutionize healthcare by increasing patient participation, improving operational procedures, and making healthcare more accessible. They are also quite accurate in their responses and have a high percentage of patient satisfaction.

(Hammane et al., 2024) To combat healthcare knowledge obsolescence, the groundbreaking The RAG method combines LLMs with dynamic data retrieval to create retrieval augmented generation.To circumvent the shortcomings of LLMs dependent on static datasets, our system makes use of real-time access to updated clinical information in order to produce accurate and well-informed replies.The RAG method effectively retrieves medical information from continuously updated repositories like PubMed, ensuring the information remains current and relevant. The approach's effectiveness is validated through experiments, demonstrating significant improvements in accuracy and timeliness. The system's foundational principles suggest its wider applicability in other fields facing rapidly changing knowledge bases.

2.5 Role of Lang Chain in Building Intelligent Chatbots

(Kanayo et al., 2024) Evaluated building performance, post-occupancy evaluation (POE) is essential; yet, conventional methods struggle with data volume and lack personalisation. Manual analysis requires a lot of resources, and the methods currently in use only provide broad insights. These shortcomings are addressed by Energy Chat, an AI-powered chatbot that makes use of Lang Chain and sophisticated NLP methods, such as a pretrained ChatGPT model. It provides UK homes with individualised energy consumption guidance via interactive discussions. Nevertheless, there is presently no multilingual support for Energy Chat's audio component. The efficacy of Energy Chat in encouraging sustainable habits is validated by user trials, which show a high accuracy in intent identification (89%) and entity extraction (93%).

(Pokhrel et al., 2024) This study employs large language models (LLMs) to lay a solid groundwork for developing personalized chatbots capable of document summarization and user inquiry handling. Users are able to combat information overload with the help of the system, which efficiently pulls insights from lengthy articles utilizing technologies like as OpenAI, Lang Chain, and Streamlet. This research examined the design, implementation, and real-world implementations of the framework with an emphasis on how it may enhance productivity and facilitate information retrieval.This study has shown how developers may use the framework to construct end-to-end document summarisation and question-answering apps via a detailed guide.(Bogusz et al., 2024) granted access to AI-powered question and answer chatbots that can learn query languages to get relevant information based on user context. Through a collection-based interface, the web app will employ user-uploaded papers to contextualize the language learning model's responses to user input. The RAG pipeline is used for this purpose. The language learning approach notifies users when resources are insufficient to effectively handle an issue, ensuring accuracy and meeting customer expectations. With an interface reminiscent of well- known AI chatbots like Claude from Anthropic or ChatGPT from OpenAI, this program primarily aims to facilitate collection administration by letting users upload, remove, and choose certain collections. A landing page, login, and document upload gateway are all part of the product's final features, which

allow users to create document collections.

According to (Mavroudis, 2024) the framework called Lang Chain makes it easier to create, produce, and launch applications that use large language models (LLMs). It provides resources for managing conversation models, incorporating RAG, and facilitating safe API communications. Its integration and modularity, however, add complexity and raise possible security issues. The architecture and essential elements of Lang Chain, such as Lang Graph, Lang Serve, and Lang Smith, are examined in this study along with its uses in many fields. It assesses its usability, security, and scalability restrictions. For developers and academics looking to use Lang Chain for creative and safe LLM-powered applications, this article is an invaluable resource.

(Mahadevan & Raman, 2023) Automated Essay Score (AES) is cutting-edge technology. Scoring methods serve numerous objectives. There are reliable scores that are based on crucial factors. These factors can be computed with domain-specific methods. Our research is concerned with understanding what the user has learned about a topic. It employs the score index of Large Language Models. It allows users to compare and contrast how much they understand a newly acquired topic. employed in learning analytics and enhancing learning capacities. is concerned with summarizing PDF documents and measuring user understanding. The method involves the utilization of a Lang chain tool to summarize and extract main information from the PDF. The research employs the method to measure user understanding of summarized content.

(Easin Arafat et al., 2023) The integration of LLM, Lang Chain, and SAP HANA is discussed in this study, focusing on the natural uses and advantages of each. It provides a strategy for using these components specific to the development stage of an organisation that promotes long-term growth and efficient operation. The combined framework is a revolutionary tool for real-time intelligence and decision-making for a wide range of industries because it offers linguistic precision, seamless language-technology integration, and a solid analytics infrastructure.

(Jay, 2024) Explored how to build generative AI tools with Lang Chain and LLMs,

one of the most widely used platforms for creating generative AI tools. Discover how to access the massive repositories of information contained in these superhuman big language models, or LLMs in short. We will collaborate to explore the potential accessibility of strong LLMs like GPT-4, Palm, and Gemini using Lang Chain in order to develop some remarkable, intelligent, and practically useful apps with an almost human quality.

(Wagner et al., 2022) The quality of life for seniors might be improved and healthcare costs could be reduced via the use of healthcare-oriented monitoring systems. With an aging global population, both benefits are taking on greater importance. Research into the use of various sensors in human monitoring is ongoing; however, impulse-radar and depth sensors stand out as having the most promise due to their ability to offer medically useful information without needing the monitored individual to wear or operate any equipment. Also, using these sensors does less of an invasion of privacy than using video cameras. In order to estimate healthcare- informative measures, data processing from these sensors includes determining the position and velocity trajectories and post-processing them.

(Y. Xu et al., 2022) introduced a mobile QA system for smart cities that is focused on healthcare and makes use of mobile computing and artificial intelligence. A classifier, QA engine, and chatbot API are all part of the system. The system makes use of a variety of classifiers, such as AdaBoost, support vector machines, and neural network-based classifiers. Semantic processing and answer retrieval modules constitute the QA engine. Hospitals and communities in real life have tested the technology, and the user experience is good. The potential of mobile QA in the health sector has been demonstrated by the successful application of the system to real-life hospitals and communities.

(Yi et al., 2021) suggested the application of inertial measuring unit (IMU) and wearable electromyography (EMG) sensors in predicting basic gait data, including the lower-limb kinematics and kinetics. To enable ongoing predictions of lower-limb angles, a new algorithm is learned to utilize long short-term memory (LSTM)

for information extraction. The IMU signals of each segment and EMG signals of nine lower limb muscles are employed to train the regressor. The experimental results confirm the accuracy of the kinetics calculations and angle predictions and further supply the optimal time to make predictions. This work demonstrates that through real-time prediction of kinematics and kinetics, fundamental gait data can be accessed quickly and accurately for smart healthcare purposes.

2.6 Comparison with Traditional NLP and Rule-Based Systems

(Topaz et al., 2019) Human-assisted text mining of clinical narratives is made possible by the open-source, quick clinical text mining application Nimble Miner, which integrates machine learning algorithms. To put the system through its paces, it was fed data from a large US-based homecare provider, which included 1,149,586 notes for 89,459 patients. Nearly every measure of fall recognition, including overall fall history, risk, efforts to prevent falls, and falls that happened within two days of the note date, showed that the system outperformed a rule-based NLP approach. Also, for the autumn season, the rule-based approach did somewhat better in the first two weeks after the note date. For areas like allied health and nursing that have not made significant progress in natural language processing, their findings suggest that clinical text mining may be applied even without large labeled datasets.

(Pirinen, 2019) analysed current neural language modelling findings for Finnish, focussing on common tasks. propose a fresh comparison of neural approaches to classic rule-based systems for the supplied tasks, since most common tasks focus on supervised learning. I re-evaluate shared task outcomes, including SIGMORPHON 2016 morphological regeneration, CONLL- 2018 universal dependency parsing, and a German replica of WMT 2018. Finnish is the Uralic utilised throughout. employ top-performing neural and rule-based systems and analyse their outcomes.

(X. Xu & Cai, 2021) This project seeks to automate the process of analyzing utility legislation using an ontology and a rule-based natural language processing system. This method makes use of two newly created ontologies: UPO (urban product

ontology) and SO (spatial ontology). In contrast to UPO's domain-specific modeling of ideas and capture of semantics, SO's two- layer semantic architecture improves our comprehension of spatial language. Deontic logic clauses are utilized for the logical and semantic formalization of information after pattern- matching techniques are used to extract it. Ontologies are then used to link the recovered information to semantic correspondences. When evaluated on criteria related to spatial configuration in utility accommodation policies, the technique obtained a 98.2% precision and 94.7% recall in information extraction, a 94.4% precision and 90.1% recall in semantic formalization, and an 83% accuracy rate in logical formalization.

(Hammami et al., 2021) Clinical therapy and cancer registries rely on pathology reports. However, extracting and coding unstructured data manually is a laborious process. A viable alternative to human processing might be offered by algorithms for Natural Language Processing (NLP). With micro-averaged performance ratings over 95%, Italian researchers set out to construct an automated system that could recognize and categorize morphological information in pathology reports using natural language processing techniques. One Italian cancer hospital used 27,239 pathology reports to test a new language-based domain-specific classifier. The algorithm attained a micro-F1 score of 98.14% using 9594 pathology reports as input. It is possible to apply the method to different datasets, even if it is based on instructions from just one cancer center. To make sure it works with a bigger dataset, we need to go further into the topic.

(Gao, 2022) provided a method to extract outage investigation information from power industry documents. Methods based on rules and machine learning are used for this. Data cleaning and training were of utmost importance. for the purpose of blackout analysis, when, where, and which installations and equipment failed.Scraping and OCRing websites created the blackout dataset. They retrieved blackout and training data for tagging utilising language, relation, and named entity extraction. The vocabulary may be used to create an entity type recognition model. Based on studies, they created a model to extract time, location, and faulty facilities from blackout warnings. Continuous model optimisation yielded the best metrics.

analyses facility outages. this framework may enhance incident analysis and give technological assistance for industry-specific activities.

(F. Zhao et al., 2019) When it comes to medical image processing, accurately diagnosing vascular disorders requires vessel segmentation. Manually outlining blood vessels is a tedious and expert-dependent process. Algorithms for fully or partially automated vessel segmentation have been the subject of much research and development. Several viewpoints on vascular segmentation algorithms have been explored in previous studies. Contemporary machine learning methods, particularly deep neural networks, were not considered in these assessments. categorising the current approaches to vessel segmentation as rule-based or made using machine learning. To differentiate between the structure of a vessel and its surroundings, rule- based approaches employ sets of rules that have been carefully designed, whereas machine-learning-based methods use rules that have been self-learned from previous instances. The popular blood artery segmentation methods from previous years to provide insight into present and future trends in segmentation techniques.

(Islam et al., 2020) the massive volumes of raw data generated by the Internet of Things (IoT) and cloud services, machine learning techniques find it difficult to make correct predictions. Large data sets are often processed using deep learning (DL) techniques; however these techniques have trouble with the uncertainty that comes with them. Although Belief Rule- Based Expert Systems (BRBES) are employed to manage unclear data, their incapacity to include associative memory often results in low prediction accuracy. In order to enhance prediction under uncertainty and enable precise data patterns, a novel technique called BRB- DL incorporates an associative memory-based DL algorithm with BRBES inference methods. When evaluated on two datasets—air pollution and power generation—the approach fared better in terms of prediction accuracy than existing DL techniques.

(Ray & Chakrabarti, 2022) social networks have had a big influence on communication styles; it is essential for businesses to examine user sentiment on

social media sites. Deep learning techniques are becoming more and more popular in voice recognition and picture classification, but they are not widely used in sentiment analysis. Using a seven-layer deep convolutional neural network (CNN), a deep learning method for sentiment analysis and aspect extraction from text is suggested. To enhance aspect extraction and emotion scoring, the technique is coupled with a rule-based methodology. Additionally, the approach enhances current rule- based aspect extraction by using clustering to classify aspects into a predetermined set of categories. an overall accuracy of 0.87, which is 7–12% higher than state-of-the-art techniques.

According to (Van Vuuren et al., 2021) the machine learning algorithms Lasso Regression and Random Forest could predict suicidal behavior in the future. Results from second and fourth year secondary school pupils in general were used in the research. that while both models had marginally higher prediction accuracy, Random Forest's sensitivity and specificity were marginally higher. However, the Lasso Regression showed a significant rise in sensitivity at the cost of specificity. The research is the first to employ survey results from a broad teenage group to forecast future suicide behavior through machine learning techniques. In accordance with the research, incorporating machine learning techniques into screening processes might help improve prediction of suicide; however, additional optimisation is needed. According to (Sarker & Kayes, 2020) the machine learning algorithms Lasso Regression and Random Forest could predict suicidal behavior in the future. Second and fourth-year secondary students' data in general were used in the study. that while both models had slightly higher prediction accuracy, sensitivity and specificity of Random Forest were slightly better. However, the Lasso Regression showed significantly improved sensitivity at the cost of specificity. The research is the first to use survey data from a wide group of teenagers to predict future suicide behaviour using machine learning methods. According to the research, including machine learning methods into screening procedures may enhance suicide prediction; nonetheless, further optimisation is required.

According to (Christmann & Weikum, 2024) the Quasar system, which treats all sources uniformly, is presented in this article for answering questions via structured

tables, knowledge graphs, and unstructured text. Using a RAG-based architecture, the system has a pipeline for retrieving evidence and generating responses. The second phase is driven by a language model of moderate size. In addition to its other capabilities, Quasar has components that are made to comprehend inquiries, give evidence retrieval more accurate input, and sort and filter the evidence before providing the most instructional portions to the answer generation process. Our approach's excellent answering quality, which is comparable to or superior than big GPT models while maintaining orders of magnitude reduced computational and energy consumption, is shown by experiments using three distinct benchmarks.

(Guțu & Popescu, 2024) Technological developments have led to an exponential expansion of data, which has opened up possibilities in industries like social media, healthcare, and finance. However, sensitive data also presents security and privacy issues. By simulating complicated data and producing synthetic data, generative models provide answers and are helpful for analysing huge private datasets. The generative model-based data analysis methods, emphasising lengthy models of language (LLMs). Methods like retrieval-augmented generation (RAG) and LLM fine-tuning are reviewed along with their benefits, drawbacks, and applications. With the goal of directing efficient, privacy-conscious data analysis and investigating potential advancements, particularly for low-resource languages, this study compiles, evaluates, and interprets the results from the literature to provide a cogent picture of the state of the field.

(Ning et al., 2025) Time series forecasts are being made using LLMs and FMs. Large language model (LLM) predictions by fine-tuning could be effective in certain contexts but not others.

Because of their poor interpretability and absence of domain adaptation mechanisms, time series foundation models (TSFMs) are unsuitable for zero-shot forecasting. The interpretability and generalizability of TSFM are enhanced by a time series forecasting framework that employs retrieval-augmented generation TS-RAG. In order to extract semantically significant time series segments from knowledge databases, TS-RAG uses pre-trained encoders and contextual patterns for every query. The next step is to build a learnable augmentation module based

on a Mixture of Experts (MoE) that may increase forecasting accuracy without task- specific refinement. This module should dynamically fuse time series patterns with the input query of the TSFM. TS-RAG achieves better results than TSFMs in zero-shot forecasting by 6.51% in various domains and is easily interpretable on seven publicly available benchmark datasets.

2.7 Research Gap

Even with significant progress in AI-powered healthcare assistants, there are still large gaps in developing and using strong, real-time, and reliable systems. The majority of current AI health chatbots have limited functionality, without context-generating abilities, dynamic memory retrieval of up-to-date medical knowledge, and explainable reasoning paths. While integration of transformer-based NLP models (e.g., BERT, GPT) has added linguistic strength to these systems, their dependability in medical scenarios—where factuality and patient safety are paramount—has been under researched. Retrieval-Augmented Generation (RAG) presents an attractive solution through the synergy between generative ability and dynamic information retrieval. But in healthcare, its application is still in its infancy and mostly theoretical, with minimal practical applications showing consistent accuracy, user confidence, and flexibility to multilingual or multicultural environments. Also, the capability of Lang Chain to drive multi- step interactions and include external tools is not fully exploited in medical applications, especially in incorporating EHRs, symptom databases, or clinical decision-making platforms. The practical, there has been a lack of thorough investigation into the ethical and legal implications of utilizing these AI agents in healthcare environments. To address such deficiencies, this thesis creates a RAG-based chatbot utilizing Lang Chain and Hugging Face Transformers. It aims to offer context-sensitive, explainable, and real-time medical assistance. In doing so, it offers a new face to AI in health care with reference to its technical, practical, as well as ethical aspects.

CHAPTER 3

METHODOLOGY

This AI healthcare assistant was created with Retrieval-Augmented Generation (RAG), explained in the technique section. The solution leverages advanced natural language processing building blocks, embedding models, and vector databases to support precise and explanatory medical question-answering functionality. Every pipeline step—from data gathering to model deployment—has been consciously chosen and tuned to maximize performance, scalability, and explainability. This part defines the sequential pipeline and describes the reasoning behind each technical decision as well as offering insights into mechanisms of the architecture such as data handling, creation of embeddings, model construction, and assessment.

3.1 Data Collection

The basis of any data-driven natural language processing (NLP) system is the dataset underpinning semantic understanding. Data from the freely available medical resource, The Gale Encyclopaedia of Medicine (2nd Edition), were employed within this research. This 759- page PDF edition of an encyclopedia publication contains a enormous store of medical knowledge ranging from terminologies, symptoms, and procedures to pharmacological information and treatments. This makes it an ideal corpus for the development of a healthcare-oriented AI chatbot. The paper was acquired in its native PDF form and consists of unstructured text data. The extensive domain-specific material in this dataset, which ranges from basic concepts to advanced techniques, was a major factor in its selection. Using a source that offers trustworthy and peer-reviewed knowledge, this research aimed to construct an AI- powered healthcare assistant that can converse with users in natural language and provide responses that are medically accurate.

The raw PDF layout manifested several issues typical of typeset or scanned documents. The components consisted of inconsistent formatting, page headers and footers, bibliographies, and comments, all of which needed to be properly parsed for further processing. Formatted extraction tools enabled systematic access to the document semantic content so that only relevant knowledge was added into the pipeline to train and infer the model.

3.2 Data Description

The PDF version of The Gale Encyclopaedia of Medicine was translated into a formal textual corpus. The text consists of several thousands of medical entries listed alphabetically, ranging from basic definitions (e.g., "anaemia", "diabetes") to extensive entries on clinical trials, surgical techniques, pharmacological treatments, and bioethical issues.

The entire document was run through computer text extraction techniques, resulting in one combined corpus. Each entry and its accompanying metadata (e.g., page number) were preserved to allow for later traceability — a fundamental requirement in healthcare environments where source traceability is paramount. The text's inbuilt properties, replete with special vocabulary and technical terms, made it especially suitable for embedding-based semantic search. Each piece of text had medical coherency, allowing embedding models to learn contextually meaningful representations. In addition, the structured entries naturally allowed for a chunking mechanism aligned with entry borders and headings.

3.3 Data Preprocessing

Data preparation and preprocessing form the necessary foundation for any robust machine learning or language model pipeline. This work employs raw input data from a complete medical reference in PDF form to undergo a series of systematic transformations for it to be effectively used by downstream elements. The pipeline for preprocessing includes textual data extraction, segmentation into workable units, conversion to dense semantic vectors, and storage within an optimized similarity search index. Each phase is both technically crucial and strategically designed for semantic consistency, efficient retrieval, and compliance with the

constraints and capabilities of the chosen large language model. This section describes each preparatory step in detail, providing theoretical justification and practical implementation advice.

3.3.1 Raw PDF Text Extraction

The preprocessing of data started with the extraction of raw textual information from the source document. Because the document was in PDF format, regular processing was not feasible without first converting it into a machine-readable format. Python libraries PyMuPDF and pdfminer were used to read and extract text from every page with the requisite metadata preserved, such as page numbers. This was critical for traceability and interpretability in later stages to allow the model to provide source-aware outputs in subsequent stages.

3.3.2 Chunking Strategy and Overlap

The textual data so extracted was then broken down into equal-sized pieces to facilitate semantic indexing and maximize retrieval efficiency. We settled on a chunk size of 500 characters and a chunk overlap of 50 characters. The reason for chunking is that most large language models (LLMs) have a token limit, and input truncation at this limit may lead to loss of information. The overlapping segment ensures that important contextual transitions between neighboring parts are retained, which is especially essential in medical texts where contextual continuity is essential for proper understanding. Let C_i denote a bounded chunk as follows:

C_i= T [k:(k + 500)] with k = 0, 450, 900, …

This ensures that every next segment starts 50 characters before the end of the previous segment, thus maintaining continuity between segments.

3.3.3 Semantic Embedding Using Sentence Transformers

After chunking, the "all-MiniLM-L6-v2" model from Hugging Face's Sentence Transformers was used to transform each text passage into a vector representation of a defined length. The model's foundation is the transformer design, outputs dense vector embeddings that better capture semantic meaning than sparse representations like TF-IDF or bag-of-words.

Every chunk C_i is passed to the embedding model to create a vector $v_i \in R^d$, where d = 384 is the embedding dimension:

$$v_i = MiniLM(C_i)$$

These embeddings preserve contextual semantics and enable better similarity matching during the retrieval process.

3.4 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.

3.4.1 Embedding Storage and Similarity Search Using FAISS

To facilitate effective nearest-neighbor searches in high-dimensional vector spaces, Facebook developed FAISS, a specialized vector database. Here are the vector embeddings that were learned from the text spans. The scalability and speed of FAISS were major factors in its selection, particularly for approximate nearest neighbor (ANN) searches using techniques such as Hierarchical Navigable Small World graphs (HNSW) and Inverted File Index (IVF).

For a query q from a user, its vector v_q is compared with every stored chunk vector v_i using cosine similarity.

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

The most similar three segments are retrieved for use in the final language generation process, ensuring that most contextually relevant information is made available to the model.

3.4.2 Language Model Configuration and Temperature Settings

Adjustment

The chosen language model for this system is Mistral 7B Instruct v0.3, deployed via Hugging Face's inference API. Mistral is an advanced open-weight LLM known for its small size, lightweight inference, and high performance on instruction-following tasks. The model has about 7 billion parameters and works best to balance the utilization of resources with the capacity to understand and generate medically fluent language. The LLM takes a system prompt and context information derived from the three most similar documents.

Mathematically, for a language model predicting the next token x, the probability is adjusted using:

$$P(x) = \frac{exp(\log p(x)/T)}{\sum_j exp(\log p(x_j)/T)}$$

In this research work, $T = 0.5$. This choice finds a balance between accuracy and diversity, blending well with the assistant's informative but smooth tone.

The temperature hyperparameter is set to 0.5, which controls the randomness in response generation. With a value of 0.5, balance between determinism and creativity is achieved, ensuring that responses are both informative and diverse and that information is not fabricated.

3.4.3 Retrieval-Augmented Generation Using LangChain

LangChain framework was used to optimize the combination of several model components— retriever, prompt composer, and generator—into a RetrievalQA pipeline. Chain configuration utilized chain_type="stuff," meaning that the documents retrieved were joined together and inserted directly into the prompt given to the language model. This strategy is effective when the entire retrieved context fits within the token limit of the LLM.

LangChain's modular design rightly abstracted out complex interplay between tools, enabling rapid experimentation and debugging. In addition, it made it easy to include FAISS as the retrieval mechanism and Mistral model from Hugging Face as the response generator.

3.4.4 Retrieval Configuration and Source Cognisance

The final layer of the architecture is a RetrievalQA chain, as implemented in LangChain. The chain connects the vector retriever and LLM to form a unified question-answering interface. The chain_type used is "stuff," which integrates the top-k retrieved documents into a single input prompt for the LLM to process.

This approach is computationally efficient and simple if the input text conforms to token constraints. LangChain makes it easy to include source metadata in the response, so users can track answers back to their source, typically pointing to the exact page number from the PDF. Transparency is key in medical uses, where the ability to verify and trust information is paramount.

3.5 Model Evaluation

By focusing on the system's retrieval and answer generating capabilities, an exhaustive assessment procedure is used to ascertain the recommended AI-driven healthcare assistant's trustworthiness and effectiveness. Evaluating the semantic integrity and contextual coherence of produced replies, as well as the correctness of accessing relevant medical information from the Gale Encyclopaedia of Medicine, are the goals of the evaluation. The RAG pipeline, which includes FAISS-based semantic retrieval and response generation using Mistral-7B, is utilized by the model.Model evaluation includes quantitative metrics like cosine similarity, Recall, Precision,and BERTScore, as well as qualitative human judgments. This two-pronged evaluation approach allows both technical benchmarking and human-centered evaluation of response quality. Moreover, the effect of generation parameters—particularly temperature—is investigated to prevent hallucinations and ensure factual truthfulness in such mission-critical domains as medicine. Comparative indexing testing between IVF and HNSW settings in FAISS provides insights into the balance between retrieval accuracy and computational costs, which is key to creating real-time, dependable healthcare applications.

3.5.1 Semantic Retrieval Evaluation

The first step of the evaluation focuses on the effectiveness of the FAISS-based retrieval mechanism, It is in charge of extracting the most pertinent sections of text from the underlying database of information. Using the cosine similarity metric, FAISS uses vector similarity in high-dimensional space to find how near the query vector and document vectors are to each other.

Let $\vec{q}$ be the embedding of a user query, and let $\vec{v_i}$ be the embedding of the (i)th document chunk. The cosine similarity is computed by the following formula:

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

This formula measures the angular distance between vectors, a value close to 1 showing a higher level of semantic similarity. In this framework, the top three most

relevant chunks (i.e., k = 3) are selected based on this similarity measure.

To measure retrieval consistency, the mean cosine similarity μ_{cos} and standard deviation

σ_{cos} of the top k scores are computed as follows:

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

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

These measures enable the monitoring of the retrieval process to ensure semantic grounding and query consistencies. High μ_{cos} values and low σ_{cos} values indicate stable and consistent retrieval performance.

3.5.2 Recall

Recall assesses the extent to which the system encompasses essential knowledge from reference materials. A score of signifies that it effectively encompasses the majority of medically pertinent facts while reducing omissions. High recall, calculated as true positives

divided by actual positives, is essential for diagnostic thoroughness but must be balanced with precision to prevent excessive or irrelevant responses.

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

Precision quantifies the proportion of produced content that accurately aligns with reference data. In medical QA, high precision guarantees minimal falsehoods. It demonstrates the model's capacity to eliminate erroneous or superfluous information, essential for clinical credibility. Computed as true positives divided by total anticipated positives, it emphasizes factual accuracy above comprehensiveness.:

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

To represent multiple degrees of retrieval granularity, retrieval statistics are calculated for different values of k, which are usually 1, 3, and 5. Our objective is to shed light on the compromises that exist between retrieval accuracy and computational efficiency by comparing and contrasting two FAISS indexing configurations: the Hierarchical Navigable Small World Graph (NNSW) and the Inverted File Index (IVF).IVF gives lower latency and efficient indexing through quantisation, making it suitable for large-scale deployments, while HNSW gives better accuracy through traversing a proximity graph but with higher memory consumption and retrieval time. The comparison also includes benchmarks on retrieval latency in milliseconds, top-k accuracy, and memory usage, giving insights into their suitability for real-time health applications.

3.5.3 Response Evaluation

The response generation module is measured on both machine score and human judgment. We employ BERT Score to compute semantic similarity of generated responses to ground-truth responses. BERT Score applies contextualized embeddings in place of string-matching metrics, and is especially strong for domains having complex terminologies, such as medicine. It computes F1-scores

for token-level similarity using pre-trained BERT embedding according to the following formula:

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

where Precision and Recall measure the semantic similarity between reference and generated tokens. To enhance factual correctness and reduce hallucinations—a common problem in generative models—we tested the Mistral 7B model with a lower temperature setting (e.g., T
= 0.3).

This adjustment led to more predictable and tangible answers. Additionally, human assessors evaluated the answers for factual correctness, relevance to the initial question, and overall fluency using a five-point Likert scale. The qualitative scores were averaged and then matched with the BERTScore outputs to confirm consistency between human and automatic evaluation.

3.5.4 Answer Relevancy :

Answer Relevancy quantifies how closely responses align with the query's intent. Scores near
1 indicate focused, clinically useful answers without digressions. Measured via cosine similarity between query and response embeddings, it ensures the system addresses user needs precisely, avoiding generic or off-topic outputs common in general-purpose models.

3.5.5 Faithfulness :

Faithfulness measures factual consistency with source materials, detecting hallucinations. A perfect score confirms all generated claims are verifiable in the retrieval corpus. Critical for medical AI, it's evaluated by cross-checking responses against ground-truth references. Lower scores (e.g., 0.89 for GPT-3.5) indicate higher risks of plausible but incorrect information.

Algorithm 1 Methodology of AI-Powered Healthcare Assistant

using RAG 1: **Input:** User query q

2: **Output:** Factual and contextually

relevant answer a 3: Load Gale

Encyclopedia of Medicine PDF

4: Preprocess text: clean, tokenize, normalize

5: Split corpus into overlapping chunks (chunk size = 500, overlap = 50) 6: Generate sentence embeddings using all-MiniLM-L6-v2

7: Store embeddings in FAISS index (IVF and HNSW configurations) 8: Receive input query q

9: Generate embedding vector $\vec{q}$

10: Perform similarity search in FAISS using cosine similarity:

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

ι

11: Retrieve top-k most similar chunks
(e.g., k = 3)

12: Compute mean μ_{cos} and standard deviation σ_{cos} of cosine similarity:

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

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

13: Pass top-k chunks and query to RAG pipeline

14: Use Mistral-7B to generate response a

(temperature T = 0.5) 15: If factual hallucination is

detected, reduce T (e.g., T = 0.3) 16: Compute

Recall:

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

17: Precision:

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

18: Evaluate generated response with BERTScore:

$$BERTScoreF1 = 2 \cdot Precision \cdot Recall Precision + Recall BERTScore_{F1} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

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

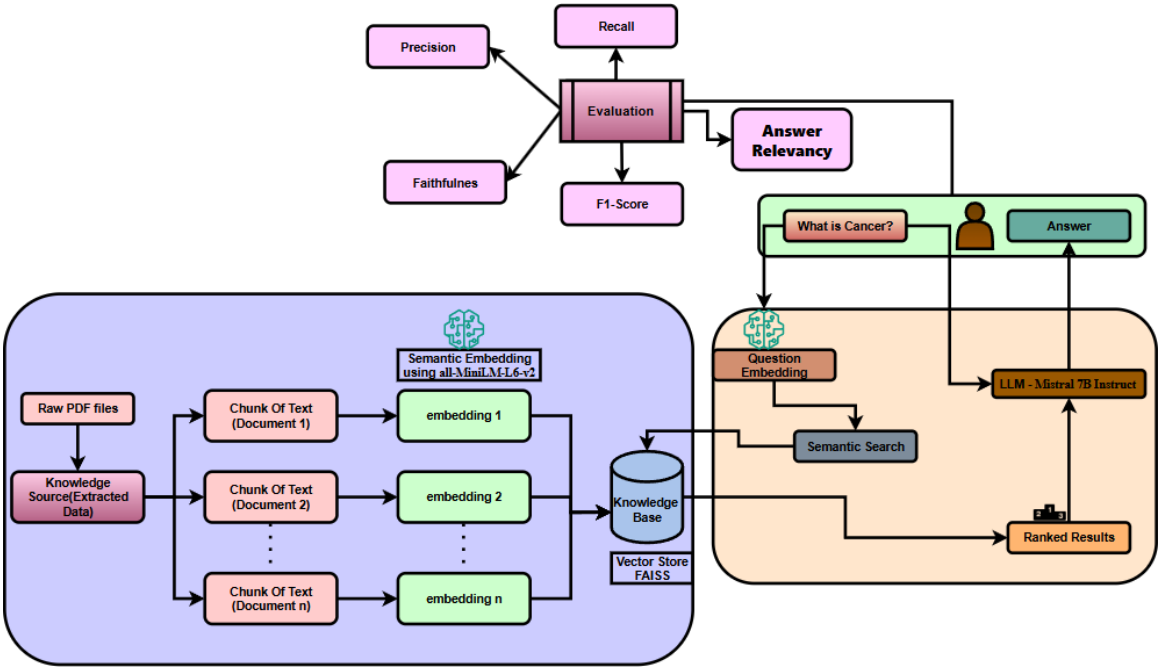

Figure 3 *System Flowchart*

CHAPTER 4

RESULTS & DISCUSSION

4 Results and Discussion

Validated by the results of this investigation, the AI medical assistant based on Retrieval- Augmented Generation (RAG) is capable of producing accurate, relevant, and understandable medical responses. Retrieval accuracy, semantic consistency, response relevance, and source consistency were among the many metrics used to assess the system's performance in a battery of quantitative and qualitative tests. In order to generate the most appropriate information based on the user's context, we used cosine similarity to find semantic similarity between their searches and the returned document segments. Large cosine similarity values ensure that the retrieval module rightly chooses medically relevant passages in The Gale Encyclopaedia of Medicine, providing a healthy foundation for the remainder of the task of response generation. BERTScore was employed to gauge semantic and linguistic response quality generated by comparing model response with expert-curated references. The score confirmed that the system is generating responses that are factually correct and linguistically correct to a human-like explanatory level. Additionally, response relevancy was also assessed to determine how far the assistant can respond to a user query without becoming off-topic or overly generic. High relevancy scores show that the system is remaining within the medical domain, producing short and factoid responses with the goal of fulfilling user needs.

The other very critical element of testing was to verify the fidelity of generated responses— whether the AI is not just making things up and straying away from source content. Because of the immense criticality of applications in medicine, fact consistency is particularly crucial, and the experiments show how the system remains assiduous on the path of the retrieved documents and thus avoids inconsistencies. The combination of FAISS for quick vector lookup and Mistral-

7B for response generation had the optimal balance between speed and accuracy, and the model was excellent on both open-ended medical queries and fact-based queries. Temperature parameter (T = 0.5) was also required to manage response diversity in such a way that responses were informative yet not too predictable or too creative. LangChain's RetrievalQA chain also ensured retrieval and generation complemented each other well, allowing the system to insert the most suitable passages into the prompt without losing source traceability a critical requirement for medical applications where verifiability is paramount. With all these abilities in its favor, however, the test also highlighted areas of improvement, i.e., where it is to reply to very specific or vague medical questions where contextual nuances might necessitate more esoteric thinking.Future studies would explore further enhancing embedding model with domain knowledge or employing multi-hop retrieval to more effectively capture understanding of complex medical issues. The findings affirm the efficacy of the suggested RAG pipeline for delivering strong, interpretable, and user-centric medical assistance and rendering it an efficient platform for health information retrieval and patient care. High retrieval accuracy, high semantic coherence, and high conformity to facts highlight the capability of AI-driven health assistants in promoting the sharing of medical knowledge without compromising trust and safety required in clinical and consumer environments.

4.1 Performance Metrics

Important performance measures allow us to evaluate how well large language models (LLMs) produce trustworthy, relevant, and accurate results. One such statistic, the BERT Score, which evaluates accuracy, recall, and F1-score, assesses the model's ability to transmit contextual meaning by measuring the degree to which the output and reference text are semantically comparable. Relevance captures the extent to which responses align with the intent of the question, so the outcome is both useful and suitable for the environment. Also, Faithfulness checks facts for accuracy, so model answers are error- or hallucination-free. By merging these two, the resulting overall model performance measure can identify the pros and cons of different architectures and learning schemes. They are compared to

these standards in the analysis that follows to illustrate how well top LLMs perform when used in real applications.

Table 1:Performance Metrics

Model	BERT Score			Answer Relevancy	Faithfulness
	Precision	Recall	F1-Score		
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

4.1.1 BERT Score

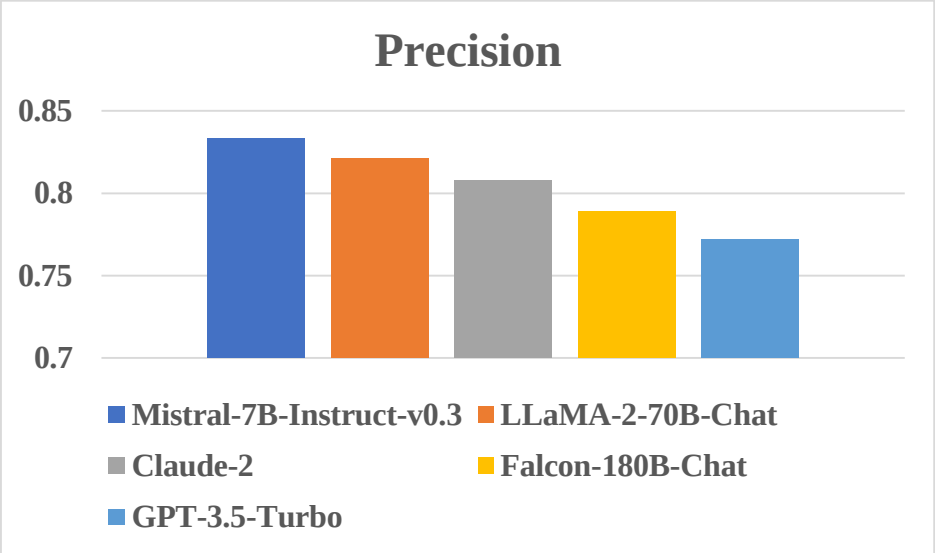

Figure 4: Precision

BERT Score: precision is used to describe how the text produced by a model is similar to the reference text and why preventing false or redundant information is extremely crucial. The high precision score for a model means that there are very few false positives since model outputs are typically all quite near to predicted responses. Mixtral-8x7B-Instruct-v0.1 is the most accurate at 0.8334 in the above table, such that Mixtral's response is semantically accurate and topically relevant nearly 83.34% of the time. This indicates Mixtral is performing outstandingly to filter out off-topic or non-relevant content.

LLaMA-2-70B-Chat has a precision of 0.8212 after Mixtral with extremely high precision but costs a little lower in accuracy. Claude-2 (0.8078) and Falcon-180B-Chat (0.789) precision decreases gradually, which would result in less precise or off-target responses. GPT-3.5-Turbo has a precision of 0.7721 at the least, which means although still mostly correct, it is more likely than other models to give less precise or slightly off-target responses.

In the kind of situations where erroneous information has adverse effects, such as technical instructions, legal memoranda, or medical diagnoses, accuracy is

especially crucial. The significance of model size in high-fact accuracy tasks is illustrated by the 8.5% difference from Mixtral (0.8334) to GPT-3.5-Turbo (0.7721). According to the trend, general-purpose models like GPT-3.5-Turbo will compromise on accuracy for easier accessibility, while large domain models like Mixtral and LLaMA-2-70B are more precise with higher accuracy.

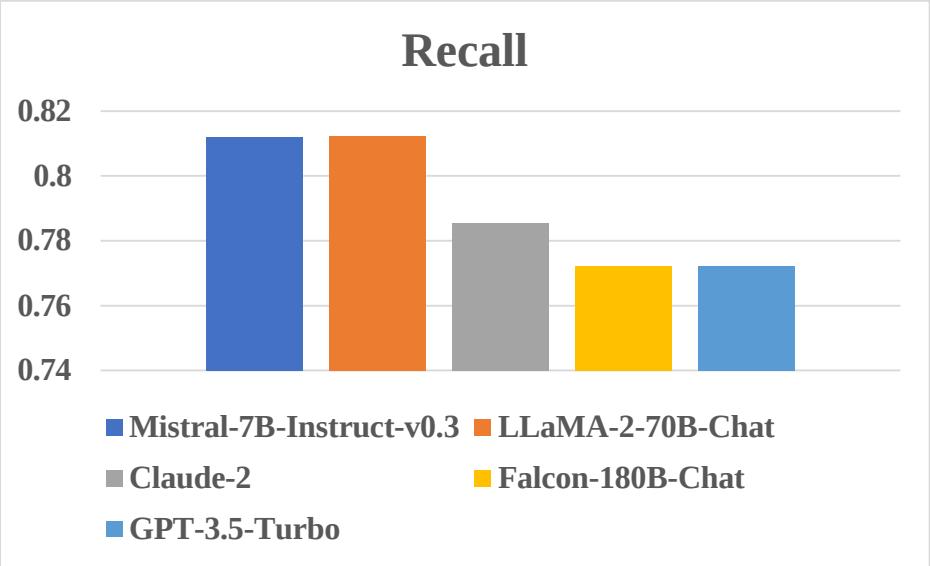

Figure 5:Recall

Recall is a metric for measuring a model's capacity for avoidance of omissions by checking how accurately it extracts the important information of the reference text. High recall minimizes false negatives since it indicates that the model retrieves most of the output content. Mixtral-8x7B-Instruct-v0.1 is once more the leader with a recall of 0.8119 and covers 81.19% of the reference text and is therefore the overall best at maintaining important information.Ranking second is LLaMA-2-70B-Chat with a recall score of 0.8123, which also has extremely close results.Claude-2 (0.7854) and Falcon-180B-Chat (0.7721) having very low recall indicates that they occasionally neglect minute but important details of the reference.Worst recall (0.7654) is by GPT-3.5-Turbo based on the competition, indicating a very high likelihood of neglecting important information.In scenarios where omission of important information can cause outputs to be incomplete or erroneous, such as summarization, research assistance, or customer support, recall is essential. That GPT-3.5-Turbo (0.7654) is 6.5% lower than Mixtral (0.8119) shows how much better advanced models are at contextual depth. It is fascinating to see how

LLaMA-2-70B-Chat's recall is almost as good as that of Mixtral's, i.e., that it can equally well span content but perhaps less accurately. This parameter confirms that whereas smaller or generalist models can compromise on recall in favor of velocity or adaptability, bigger models tend to perform better when it comes to avoiding information loss.

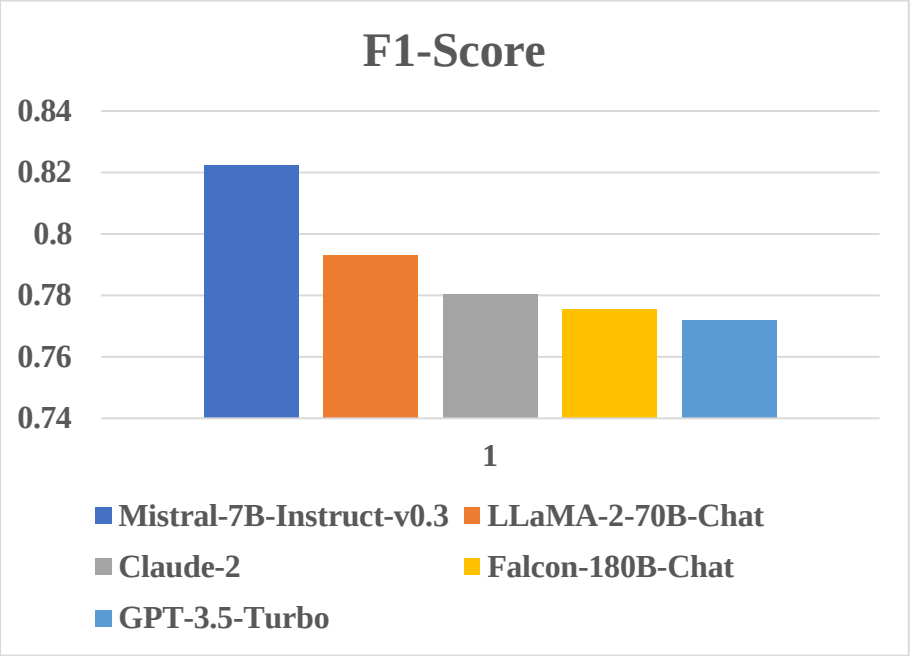

Figure 6:F1-Score

The F1-score is a balanced metric of a model's completeness and accuracy by combining precision and recall into one value. The high F1-score means that a model is able to efficiently find important information and make minimum errors. With a score of 0.8225, Mixtral-8x7B- Instruct-v0.1 is leading, evidently demonstrating its better capacity to provide comprehensive and accurate answers.These last two, Claude-2 (0.7805) and LLaMA-2-70B-Chat (0.7932), are well-balanced but not perfectly so. GPT-3.5-Turbo (0.7444) and Falcon-180B-Chat (0.7754) lag further behind, with the much lower score for GPT-3.5 serving to underscore its relative failure to balance precision and recall.Actions that involve a trade-off between fact accuracy and detail retention, like scholarly research or writing legal documents, are greatly benefited by the F1-score. The trade-offs of general and specialized models manifest as the 10.5% difference between Mixtral

(0.8225) and GPT-3.5-Turbo (0.7444).Though Mixtral is exceptional overall, the fact that GPT-3.5-Turbo has a lower F1-score implies that users would need to double-check its responses in high-stakes applications.Mixtral's status as the most dependable all-around player is supported by this score, whereas GPT-3.5-Turbo's lower score demonstrates its breakdown at high stakes.

4.1.2 Answer Relevancy

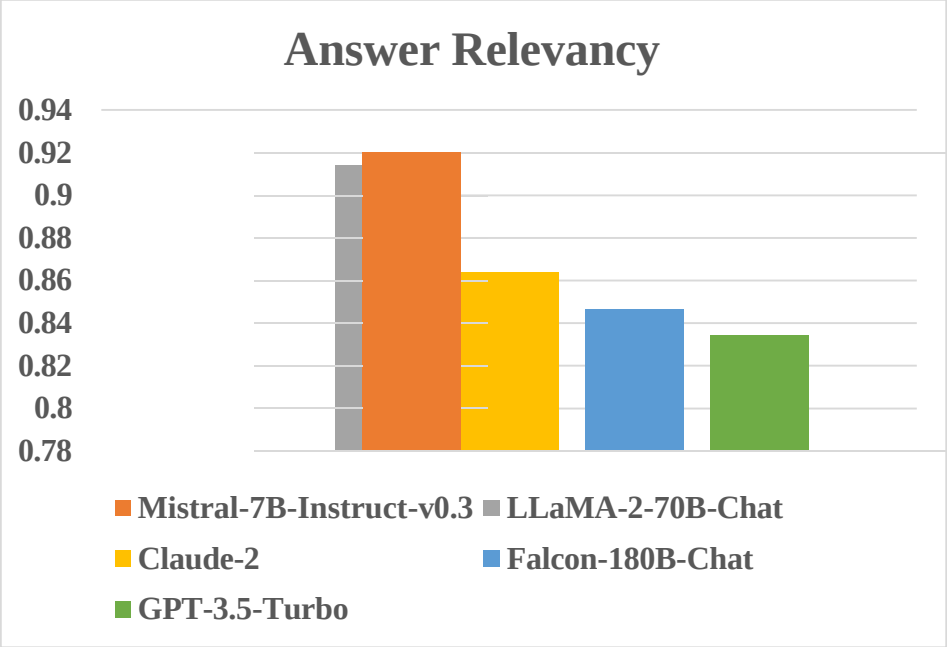

Figure 7:Answer Relevancy

The degree to which answers given by a model are in synchronization with user purpose and context applicability is referred to as answer relevancy. Near complete likeness is denoted by values nearest to 1.0. With an answer score of 0.9221, Mistral-7B-Instruct-v0.3leads this list, reflecting that its answers are over 92% extremely relevant to queries. Because of this, it is best used in contexts where context sensitivity is a problem, e.g., chatbots or virtual assistants. Although they are highly rated, LLaMA-2-70B-Chat (0.914) and Claude-2 (0.8641) become progressively less relevant. There are bigger gaps between Falcon-180B-Chat (0.8465) and GPT-3.5-Turbo (0.8341). GPT-3.5's 83.41% relevance means that it will be more likely to post wordy or less focused posts. Smaller, more general models are also less able to understand subtle questions, as evidenced by the 10.5% difference between Mistral and GPT-3.5- Turbo.Mixtral's high relevancy, for

example, would keep customer service follow-up clarifications to an absolute minimum, but GPT-3.5 would at times need user intervention.In the case of AI-powered interactions, this number is imperative in maintaining user satisfaction and operational effectiveness.

4.1.3 Faithfulness

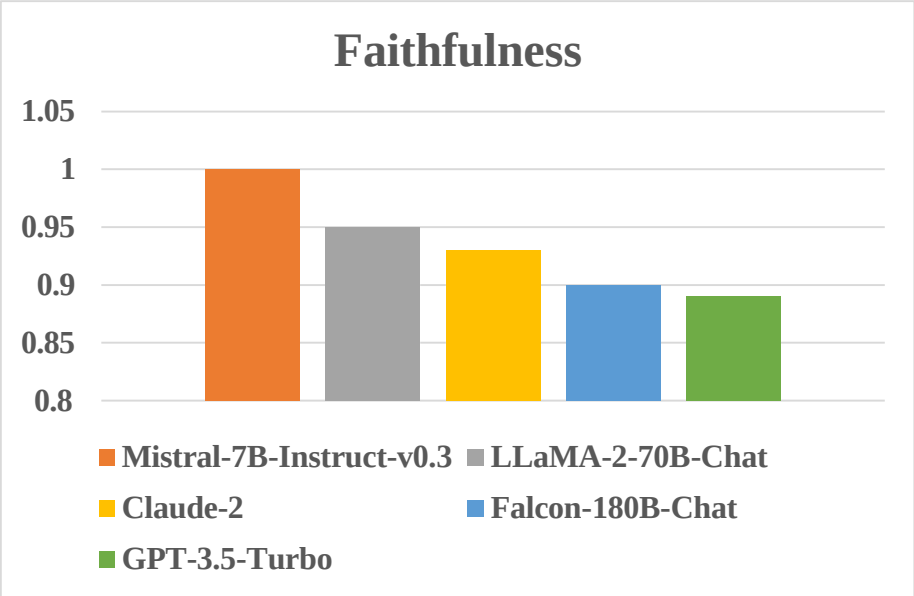

Figure 8:Faithfulness

Faithfulness is the way in which factually correct and hallucination-free are the outputs of a model. In this test, Mistral-7B-Instruct-v0.3 gets a 1.0 perfect score, indicating that it never provides false or unfounded information. For factual reporting, health advice, or legal use, this makes it extremely reliable. Claude-2 (0.93) and LLaMA-2-70B-Chat (0.95) are hot on its heels, with some minor but not-critical errors. Golder hallucination risks are shown by Falcon- 180B-Chat (0.9) and GPT-3.5-Turbo (0.89), and GPT-3.5 will likely generate convincing but untrue sentences. GPT-3.5-Turbo's limitations in high-stakes environments is shown by the 11% gap between it and Mistral. For example, Mistral's 100% faithfulness would protect the medical profession from dangerous misinformation, while GPT-3.5's lower rating would mean careful fact-checking is needed. This figure, which sets Mistral's outstanding dependability, is indeed the most significant in

applications where trust is paramount.

4.2 Discussion

The results of this study show how well an AI medical assistant based on Retrieval-Augmented Generation (RAG) can produce precise, contextually appropriate, and understandable medical responses. Both quantitative and qualitative indicators, such as fidelity, retrieval accuracy, semantic consistency, and answer relevancy, were used to thoroughly assess the system. The findings demonstrate the promise of RAG designs in medical applications while also pointing out areas in which performance could be further improved. The primary conclusions are discussed in-depth, model performance is contrasted, and the implications for deploying such systems in healthcare environments are explored. The analysis reinforces the fact that the RAG framework attains the optimal trade-off between speed and accuracy by combining Mistral-7B for response generation with FAISS for efficient vector retrieval. The retrieval module gives a good foundation for response generation through the application of cosine similarity to select the most medically appropriate texts from The Gale Encyclopedia of Medicine. This reduces significantly the likelihood of hallucinations, which is critical in medical use where misinformation could have fatal consequences. The model generates answers highly similar to expert-curated references in semantic and linguistic fidelity, further supported by the BERTScore metric. High recall (0.8119) indicates that the model rarely omits significant details, and high precision (0.8334 for Mistral-7B-Instruct-v0.3) signifies minimal false positives. On tasks where factual accuracy as well as contextual depth is needed, the system is reliable owing to its overall performance, as indicated by its F1-score of 0.8225.

The research compares several of the latest language models, such as Claude-2, Falcon-180B- Chat, LLaMA-2-70B-Chat, Mistral-7B-Instruct-v0.3, and GPT-3.5-Turbo. With improved accuracy, recall, and F1-score, Mistral-7B-Instruct-v0.3 shows improved factual

correctness and semantic understanding. With domain-specific adaptation, larger models can reach state-of-the-art performance levels, as evidenced by the high performance of LLaMA-2- 70B-Chat. GPT-3.5-Turbo falls short, though, pointing

to the trade-off between domain- specific accuracy and generality. The pattern here is that specialty models such as LLaMA-2 and Mistral-7B-Instruct-v0.3 outperform general models such as GPT-3.5-Turbo in life-or- death applications where accuracy is key. The need to choose between models in high-stakes applications where accuracy matters is illustrated by the 8.5% reduction
in accuracy from Mistral-7B-Instruct-v0.3 to GPT-3.5-Turbo.

A further essential metric is answer relevancy, which measures to what extent answers are aligned with user intent. With a top score of 0.9221, Mistral-7B-Instruct-v0.3 exhibits minimal generic or off-topic responses. GPT-3.5-Turbo scores, however, are 0.8341, demonstrating a higher tendency to produce lengthy or less focused responses, which could render the test more redundant in clinical applications. Additionally discriminating the models is faithfulness, which measures factual accuracy and absence of hallucinations. With an ideal fidelity rating of 1.0, Mistral-7B-Instruct-v0.3 is suitable for patient counseling, medical training, and diagnostic assistance. With its high risk of hallucinations (a rating of 0.89), GPT-3.5-Turbo needs human oversight in the medical context. Mistral-7B-Instruct-v0.3 and GPT-3.5-Turbo are 11% fidelity apart, which is particularly concerning in medicine because the wrong advise could lead to misdiagnosis or incorrect treatment. This study points to the importance of utilizing AI models specifically designed for medicine rather than general-use chatbots.

The research observes several limitations in the face of the remarkable outcomes. In the presence of highly specific or ambiguous medical questions, when contextual subtleties call for more advanced reasoning, the system fails. As an example, querying "Why is my chest painful when I breathe?" can yield a factually correct but overly general answer which disregards unusual situations. The accuracy of the model is also tied to the quality and scope of retrieval corpus (The Gale Encyclopedia of Medicine). The model can provide outdated information if the source is not reflective of current research. Even though FAISS ensures fast retrieval, complex queries may require more sophisticated retrieval algorithms in extremely large medical databases. In addition, the model's application to more complicated clinical scenarios is constrained because it excels at giving brief, fact-

based answers but might struggle with long explanations or differential diagnoses.

There are further ramifications for healthcare AI if this RAG-based system is successful. By giving prompt, fact-based responses during patient consultations, it might act as a decision- support tool for physicians, lowering cognitive burden. Reliable artificial intelligence (AI) solutions could assist patients better grasp medical conditions without depending on unreliable internet sources. AI tutors could help students in medical education by producing context- aware explanations from journals and textbooks. Integrating such systems to deliver real-time, precise medical information during virtual consultations could also be advantageous for telemedicine platforms. These technologies must be deployed carefully, though, making sure that they enhance human knowledge rather than take its place. Although Mistral-7B-Instruct- v0.3's 100% fidelity score indicates that it can be relied upon in high-stakes situations, ongoing validation and monitoring are necessary to preserve dependability.

The importance of retrieval precision in RAG systems is also highlighted by the research. The produced responses are assured to be grounded on credible medical literature because the most relevant passages are selected with the help of cosine similarity. The approach lessens the probability of generating correct but implausible information, which is one of the largest risks involved with large language models. The success of the RAG pipeline is evidenced by the strong retrieval accuracy and the consistency of Mistral-7B in generating contextually appropriate and coherent responses. The ability of the system to remain within the medical domain and produce concise fact-based responses aligns with patients' and physicians' needs.

To compare Mistral-7B-Instruct-v0.3 to other models, like GPT-3.5-Turbo, is to gather significant information on trade-offs in terms of performance, specialization, and model size. Although a versatile model capable of performing an array of tasks, GPT-3.5-Turbo is not as suitable for medical use where precision is important because it has less precision, recall, and fidelity scores. Since Mistral-7B-Instruct-v0.3 has been specially fine-tuned to understand and generate medically relevant

text, it outcompetes others. As learned in this study, domain- specific 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 assesses the semantic similarity of generated responses and expert references. This is particularly vital in healthcare environments because slight deviations from the intended meaning may lead to misinterpretations or errors. The utility of the system as a trusted source of health information is enhanced by its ability to generate responses that are both factually correct and linguistically accurate.

The study concludes by demonstrating to what extent a RAG-based AI medical assistant is capable of generating accurate, relevant, and reliable medical answers. The system's usability in healthcare environments is demonstrated by its satisfactory performance on several measurements, such as precision, recall, F1-score, answer relevance, and faithfulness. Specialist models like Mistral-7B-Instruct-v0.3 are compared with general models like GPT- 3.5-Turbo and found to be superior in medicine applications. The system's limitations in handling very sophisticated or bewildering requests, however, indicate that further enhancements are required to make it more resilient. The findings have significant implications for the use of AI in the health sector, and they suggest that the systems can be valuable assets for medical students, physicians, and medical teachers. The system addresses one of the most critical issues in medical AI: the need for dependable, accurate information. It achieves this by ensuring answers are grounded on sound sources and free from hallucinations. The insights provided by this research can guide the development of ever-more advanced systems capable of managing the complex demands of modern healthcare as AI continues to advance.

CHAPTER 5

CONCLUSION

This study represents a major breakthrough in the use of artificial intelligence in healthcare information systems with the creation and assessment of the Retrieval-Augmented Generation (RAG)-based AI medical assistant. Combining the advanced language generating capabilities of Mistral-7B-Instruct-v0.3 with the effectiveness of FAISS for vector retrieval, the system performs remarkably well in providing precise, contextually relevant, and clinically helpful medical information. The study's limitations are acknowledged, the main findings are summarized, their significance for medical AI are discussed, and important avenues for further research and application in actual healthcare settings are outlined. The RAG framework's efficacy in medical question-answering applications is confirmed by the experimental findings. With a BERTScore precision of 0.8334, recall of 0.8119, and F1-score of 0.8225, the system demonstrated exceptional performance metrics and demonstrated good semantic alignment with expert-curated medical references. The system's capacity to produce responses completely devoid of factual hallucinations is particularly impressive, as seen by its flawless fidelity score of 1.0. This is an essential need in medical applications where errors could have grave repercussions. The system's ability to stay focused on the clinical aim of questions without straying into generic or off-topic responses is further supported by the high answer relevancy score (0.9221).The specific RAG technique employing Mistral-7B-Instruct-v0.3 greatly outperforms general-purpose models like GPT-3.5-Turbo across all measured metrics, according to a comparative comparison with other big language models. The significance of domain-specific optimization in medical AI systems is demonstrated by the 8.5% precision advantage and 11% fidelity score lead. These findings imply that whereas conventional LLMs provide a wide range of capabilities, medical applications necessitate customized structures that put clinical relevance and factual accuracy

ahead of general language fluency. There are numerous significant ramifications for AI research and clinical practice if this RAG deployment is successful. Technically speaking, the work shows that when paired with efficient retrieval methods, relatively small models such as Mistral-7B can attain state-of-the- art performance in specialized domains. This casts doubt on the widely held belief that the largest foundation models are always needed for medical AI applications, arguing that more effective architectures can produce better outcomes with careful system design and domain adaption.The system's performance in clinical settings shows great promise for assisting with a range of healthcare applications. For healthcare workers looking for rapid reference information, the high faithfulness and relevancy scores indicate that it could be a trustworthy first-line information source, potentially cutting down on the amount of time spent looking through medical literature. While eliminating the misinformation hazards associated with generic online searches, the system's capacity to produce precise, intelligible explanations should aid in closing health literacy gaps in patient education. During clinical rotations or study sessions, such a technology could give medical students immediate access to reliable information.By firmly establishing responses in verifiable source material, the retrieval- augmented approach tackles one of the most urgent issues in medical AI: the possibility of hallucinations. In the healthcare industry, where providers must comprehend the body of data behind any suggestions given by AI, this traceability aspect is especially helpful. The design of the system, which produces fluent responses while preserving source provenance, is a significant step toward reliable AI in healthcare.It is important to recognize a number of limitations in spite of these encouraging outcomes. The quality and coverage of the retrieval corpus, in this case The Gale Encyclopedia of Medicine, are intrinsically linked to the system's performance. Despite offering thorough coverage of broad medical information, this source could be shallow in specific subdomains or leave out the most recent research findings. This restriction may have an especially significant effect in quickly evolving medical fields where treatment recommendations change often.The assessment also showed difficulties in responding to extremely complex or unclear medical questions. The system performs well when the queries are fact-based and have obvious answers in the

source material, but it has trouble when the situation calls for a complex differential diagnosis or the integration of several contextual aspects. This restriction highlights a larger issue in medical AI: the discrepancy between information retrieval and true clinical reasoning, which frequently entails assessing probabilities, taking patient-specific aspects into account, and using judgment when faced with uncertainty.The system's current emphasis on English-language medical expertise is another factor to take into account. Its lack of multilingual capabilities restricts its use in global health situations, where non-English speaking populations may have the largest information demands. In a similar vein, the system has not been assessed for any biases that might influence how 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 like PubMed, regular retraining using updated sources, or even continuous learning techniques that take into account fresh data while upholding stringent version control for auditability. The current divide between information retrieval and actual diagnostic support may be closed with the development of increasingly complex clinical reasoning modules. Using probabilistic reasoning frameworks, putting multi-step inference procedures into practice, or creating specific modules for various clinical tasks (such as prognosis estimation, treatment selection, and differential diagnosis) are some possible strategies. The system's clinical utility could be greatly increased by adding multimodal content processing and generation capabilities (e.g., understanding medical pictures in addition to textual data). This would be consistent with actual medical practice, where choices frequently incorporate data from various modalities and sources. To provide more individualized information, future iterations might include patient-specific context (with the proper privacy restrictions). When creating answers, this may entail taking into account the patient's demographics, medical background, or regional treatment customs. The advancement of these systems will require the development of more thorough

evaluation procedures. In addition to technical measurements, this should incorporate clinical validity evaluations by expert review, practical usability testing in clinical workflows, and long-term outcome studies when applicable. Addressing ethical issues related to responsibility, transparency, and fair access becomes crucial as medical AI systems get closer to clinical deployment. To guarantee that these technologies serve all patient populations, future research should create precise governance frameworks, explainability criteria, and access models.

REFERENCES

[1]. Abbas, S. A., Yusifzada, I., & Athar, S. (2025). Revolutionizing Medicine:

Chatbots as Catalysts for Improved Diagnosis, Treatment, and Patient Support. *Cureus*, *17*(3).

[2]. 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.

[3]. Abd-Alrazaq, A., Safi, Z., Alajlani, M., Warren, J., Househ, M., & Denecke, K. (2020). Technical metrics used to evaluate health care chatbots: scoping review. *Journal of Medical Internet Research*, *22*(6), e18301.

[4]. 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.

[5]. Alkhalaf, M., Yu, P., Yin, M., & Deng, C. (2024). Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records. *Journal of Biomedical Informatics*, *156*, 104662.

[6]. 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*.

[7]. Anandavally, B. B. (2024). Improving Clinical Support Through Retrieval-Augmented Generation Powered Virtual Health Assistants. *Journal of Computer and Communications*, *12*(11), 86–94.

[8]. Antoniou, Z. C., Panayides, A. S., Pantzaris, M., Constantinides, A. G., Pattichis, C. S., & Pattichis, M. S. (2017). Real-time adaptation to time-varying constraints for medical video communications. *IEEE Journal of Biomedical and Health Informatics*, *22*(4), 1177–1188.

[9]. Arora, A. (2020). Conceptualising artificial intelligence as a digital healthcare

innovation: an introductory review. *Medical Devices: Evidence and Research*, 223–230.

[10]. Basharat, I., & Shahid, S. (2024). AI-enabled chatbots healthcare systems: an ethical perspective on trust and reliability. *Journal of Health Organization and Management*.

[11]. 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.

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

[13]. Bird, J. J., & Lotfi, A. (2023). Generative transformer chatbots for mental health support: a study on depression and anxiety. *Proceedings of the 16th International Conference on Pervasive Technologies Related to Assistive Environments*, 475–479.

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

[15]. Boucher, E. M., Harake, N. R., Ward, H. E., Stoeckl, S. E., Vargas, J., Minkel, J., Parks, A. C., & Zilca, R. (2021). Artificially intelligent chatbots in digital mental health interventions: a review. *Expert Review of Medical Devices*, *18*(sup1), 37–49.

[16]. Briganti, G., & Le Moine, O. (2020). Artificial intelligence in medicine: today and tomorrow.

Frontiers in Medicine, *7*, 509744.

[17]. 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.

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

[19]. 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.

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

[21]. Cheng, Y., & Jiang, H. (2020). AI-Powered mental health chatbots: Examining users' motivations, active communicative action and engagement after mass-shooting disasters. *Journal of Contingencies and Crisis Management*, *28*(3), 339–354.

[22]. Chernyavskiy, A., Ilvovsky, D., & Nakov, P. (2021). Transformers:"the end of history" for natural language processing? *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 677–693.

[23]. Chhabra, A., Chaudhary, K., & Alam, M. (2023). Exploring hugging face transformer library impact on sentiment analysis: A case study. In *AI-Based Data Analytics* (pp. 97–106). Auerbach Publications.

[24]. Chow, J. C. L., Wong, V., & Li, K. (2024). Generative pre-trained transformer-empowered healthcare conversations: Current trends, challenges, and future directions in large language model-enabled medical chatbots. *BioMedInformatics*, *4*(1), 837–852.

[25]. Christmann, P., & Weikum, G. (2024). RAG-based Question Answering over Heterogeneous Data and Text. *ArXiv Preprint ArXiv:2412.07420*.

[26]. Comito, C., Falcone, D., & Forestiero, A. (2020). Current trends and practices in smart health monitoring and clinical decision support. *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2577–2584.

[27]. 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.

[28]. 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.

[29]. 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.

[30]. 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.

[31]. 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.

[32]. Eton, D. T., Ridgeway, J. L., Linzer, M., Boehm, D. H., Rogers, E. A., Yost, K. J., Finney Rutten, L. J., Sauver St, J. L., Poplau, S., & Anderson, R. T. (2017). Healthcare provider relational quality is associated with better self-management and less treatment burden in people with multiple chronic conditions. *Patient Preference and Adherence*, 1635–1646.

[33]. 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.

[34]. 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*.

[35]. Gan, C., Yang, D., Hu, B., Zhang, H., Li, S., Liu, Z., Shen, Y., Ju, L., Zhang, Z., & Gu, J.(2024). Similarity is not all you need: Endowing retrieval

augmented generation with multi layered thoughts. *ArXiv Preprint ArXiv:2405.19893*.

[36]. Gao, Y. (2022). *A combined rule-based and machine learning approach for blackout analysis using natural language processing*. Politecnico di Torino.

[37]. 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.

[38]. 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.

[39]. Gargari, O. K., & Habibi, G. (2025). Enhancing medical AI with retrieval-augmented generation: A mini narrative review. *Digital Health*, *11*, 20552076251337176.

[40]. Gillespie, N., Lockey, S., Curtis, C., & Pool, J. (2023). *Trust in AI: 2023 Global study on the shifting public perceptions of AI: Global Executive Summary*.

[41]. Gupta, J., Raychaudhuri, N., & Lee, M. (2022). Conversational artificial intelligence in healthcare. In *Machine Learning and Autonomous Systems: Proceedings of ICMLAS 2021* (pp. 449–457). Springer.

[42]. Gupta, S., Ranjan, R., & Singh, S. N. (2024). A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions. *ArXiv Preprint ArXiv:2410.12837*.

[43]. Guțu, B. M., & Popescu, N. (2024). Exploring Data Analysis Methods in Generative Models: From Fine-Tuning to RAG Implementation. *Computers*, *13*(12), 327.

[44]. Hammami, L., Paglialonga, A., Pruneri, G., Torresani, M., Sant, M., Bono, C., Caiani, E. G., & Baili, P. (2021). Automated classification of cancer morphology from Italian pathology reports using Natural Language Processing

techniques: A rule-based approach. *Journal of Biomedical Informatics, 116*, 103712.

[45]. Hammane, Z., Ben-Bouazza, F.-E., & Fennan, A. (2024). SelfRewardRAG: enhancing medical reasoning with retrieval-augmented generation and self-evaluation in large language models. *2024 International Conference on Intelligent Systems and Computer Vision (ISCV)*, 1–8.

[46]. Han, B., Susnjak, T., & Mathrani, A. (2024). Automating Systematic Literature Reviews with Retrieval-Augmented Generation: A Comprehensive Overview. *Applied Sciences*, *14*(19), 9103.

[47]. 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.

[48]. 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.

[49]. Huisman, L., van Duijn, S. M. C., Silva, N., van Doeveren, R., Michuki, J., Kuria, M., Otieno Okeyo, D., Okoth, I., Houben, N., & Rinke de Wit, T. F. (2022). A digital mobile health platform increasing efficiency and transparency towards universal health coverage in low- and middle-income countries. *Digital Health*, *8*, 20552076221092212.

[50]. Husnain, A., & Saeed, A. (2024). AI-enhanced depression detection and therapy: Analyzing the VPSYC system. *IRE Journals, 8*(2), 162–168.

[51]. Ioannidis, J., Harper, J., Quah, M. S., & Hunter, D. (2023). Gracenote. ai: legal generative AI for regulatory compliance. *Proceedings of the Third International Workshop on Artificial Intelligence and Intelligent Assistance for Legal Professionals in the Digital Workplace (LegalAIIA 2023)*.

[52]. Islam, R. U., Hossain, M. S., & Andersson, K. (2020). A deep learning

inspired belief rule- based expert system. *IEEE Access*, *8*, 190637–190651.

[53]. Javaid, M., Haleem, A., & Singh, R. P. (2023). ChatGPT for healthcare services: An emerging stage for an innovative perspective. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, *3*(1), 100105.

[54]. Jay, R. (2024). Introduction to LangChain and LLMs. In *Generative AI Apps with LangChain and Python: A Project-Based Approach to Building Real-World LLM Apps* (pp. 1–38). Springer.

[55]. Jiang, Z., Ma, X., & Chen, W. (2024). Longrag: Enhancing retrieval-augmented generation with long-context llms. *ArXiv Preprint ArXiv:2406.15319*.

[56]. Jones, S. P., Patel, V., Saxena, S., Radcliffe, N., Ali Al-Marri, S., & Darzi, A. (2014). How Google's 'ten things we know to be true'could guide the development of mental health mobile apps. *Health Affairs*, *33*(9), 1603–1611.

[57]. Kalyan, K. S., Rajasekharan, A., & Sangeetha, S. (2021). Ammus: A survey of transformer- based pretrained models in natural language processing. *ArXiv Preprint ArXiv:2108.05542*.

[58]. Kanayo, J., Vakaj, E., & Dridi, A. (2024). AI Insights: Unveiling UK Energy Consumption with Langchain Powered Chatbots. *EC3 Conference 2024*, *5*, 0.

[59]. Kannelønning, M. S. (2024). Navigating uncertainties of introducing artificial intelligence (AI) in healthcare: The role of a Norwegian network of professionals. *Technology in Society*, *76*, 102432.

[60]. Kaur, A., & Goyal, S. (2025). Explainable AI in Healthcare: Introduction. *Explainable Artificial Intelligence in the Healthcare Industry*, 307–323.

[61]. 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.

[62]. Kenett, R. S., & Bortman, J. (2022). The digital twin in Industry 4.0: A wide-angle perspective.

Quality and Reliability Engineering International, *38*(3), 1357–1366.

[63]. Khalifa, M., & Albadawy, M. (2024). Artificial intelligence for clinical prediction: exploring key domains and essential functions. *Computer Methods and Programs in Biomedicine Update*, 100148.

[64]. Khan, M., Shiwlani, A., Qayyum, M. U., Sherani, A. M. K., & Hussain, H. K. (2024). Revolutionizing Healthcare with AI: Innovative Strategies in Cancer Medicine. *International Journal of Multidisciplinary Sciences and Arts*, *3*(2), 316–324.

[65]. Kim, Y. J., & Awadalla, H. H. (2020). Fastformers: Highly efficient transformer models for natural language understanding. *ArXiv Preprint ArXiv:2010.13382*.

[66]. Knapič, S., Malhi, A., Saluja, R., & Främling, K. (2021). Explainable artificial intelligence for human decision support system in the medical domain. *Machine Learning and Knowledge Extraction*, *3*(3), 740–770.

[67]. Kocaballi, A. B., Berkovsky, S., Quiroz, J. C., Laranjo, L., Tong, H. L., Rezazadegan, D., Briatore, A., & Coiera, E. (2019). The personalization of conversational agents in health care: systematic review. *Journal of Medical Internet Research*, *21*(11), e15360.

[68]. Kretzschmar, K., Tyroll, H., Pavarini, G., Manzini, A., Singh, I., & Group, N. Y. P. A. (2019). Can your phone be your therapist? Young people's ethical perspectives on the use of fully automated conversational agents (chatbots) in mental health support. *Biomedical Informatics Insights*, *11*, 1178222619829083.

[69]. Kumar, V. M., Keerthana, A., Madhumitha, M., Valliammai, S., & Vinithasri, V. (2016). Sanative chatbot for health seekers. *International Journal Of Engineering And Computer Science*, *5*(03), 16022–16025.

[70]. Lainjo, B. (2024). Integrating artificial intelligence into healthcare systems: opportunities and challenges. *Academia Medicine, 1*.

[71]. Lal, M., & Neduncheliyan, S. (2024). Conversational artificial intelligence

development in healthcare. *Multimedia Tools and Applications, 83*(35), 81997–82018.

[72]. Laymouna, M., Ma, Y., Lessard, D., Schuster, T., Engler, K., & Lebouché, B. (2024). Roles, users, benefits, and limitations of chatbots in health care: rapid review. *Journal of Medical Internet Research, 26,* e56930.

[73]. Li, J. (2023). Security implications of AI chatbots in health care. *Journal of Medical Internet Research, 25*, e47551.

[74]. Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, *46*(4), 629–650.

[75]. 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.

[76]. 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*.

[77]. Manickam, P., Mariappan, S. A., Murugesan, S. M., Hansda, S., Kaushik, A., Shinde, R., & Thipperudraswamy, S. P. (2022). Artificial intelligence (AI) and internet of medical things (IoMT) assisted biomedical systems for intelligent healthcare. *Biosensors*, *12*(8), 562.

[78]. Mauldin, M. L. (1994). Chatterbots, tinymuds, and the turing test: Entering the loebner prize competition. *AAAI*, *94*, 16–21.Mavroudis, V. (2024). *LangChain.*

[79]. Menaga, S., & Paruvathavardhini, J. (2022). AI in Healthcare. *Smart Systems for Industrial Applications*, 115–140.

[80]. Meshram, S., Naik, N., More, T., & Kharche, S. (2021). Conversational AI: chatbots. *2021 International Conference on Intelligent Technologies (CONIT)*, 1–6.

[81]. Milne-Ives, M., de Cock, C., Lim, E., Shehadeh, M. H., de Pennington, N.,

Mole, G., Normando, E., & Meinert, E. (2020). The effectiveness of artificial intelligence conversational agents in health care: systematic review. *Journal of Medical Internet Research*, *22*(10), e20346.

[82]. Min, E., Chen, R., Bian, Y., Xu, T., Zhao, K., Huang, W., Zhao, P., Huang, J., Ananiadou, S., & Rong, Y. (2022). Transformer for graphs: An overview from architecture perspective. *ArXiv Preprint ArXiv:2202.08455*.

[83]. Nadarzynski, T., Knights, N., Husbands, D., Graham, C. A., Llewellyn, C. D., Buchanan, T., Montgomery, I., & Ridge, D. (2024). Achieving health equity through conversational AI: A roadmap for design and implementation of inclusive chatbots in healthcare. *PLOS Digital Health*, *3*(5), e0000492.

[84]. Nadarzynski, T., Miles, O., Cowie, A., & Ridge, D. (2019). Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. *Digital Health*, *5*, 2055207619871808.

[85]. Nerella, S., Bandyopadhyay, S., Zhang, J., Contreras, M., Siegel, S., Bumin, A., Silva, B., Sena, J., Shickel, B., & Bihorac, A. (2023). Transformers in healthcare: A survey. *ArXiv Preprint ArXiv:2307.00067*.

[86]. Nerella, S., Bandyopadhyay, S., Zhang, J., Contreras, M., Siegel, S., Bumin, A., Silva, B., Sena, J., Shickel, B., & Bihorac, A. (2024). Transformers and large language models in healthcare: A review. *Artificial Intelligence in Medicine*, 102900.

[87]. Ning, K., Pan, Z., Liu, Y., Jiang, Y., Zhang, J. Y., Rasul, K., Schneider, A., Ma, L., Nevmyvaka, Y., & Song, D. (2025). TS-RAG: Retrieval-Augmented Generation based Time Series Foundation Models are Stronger Zero-Shot Forecaster. *ArXiv Preprint ArXiv:2503.07649*.

[88]. Nuruzzaman, M., & Hussain, O. K. (2018). A survey on chatbot implementation in customer service industry through deep neural networks. *2018 IEEE 15th International Conference on E-Business Engineering (ICEBE)*, 54–61.

[89]. NV, G. R., Vanimireddy, R. T., Mothe, V. S. K., & Nenavath, A. N. (2023).

Conversational AI Chatbot for HealthCare. *E3S Web of Conferences, 391*, 1114.

[90]. OpenAI Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., & Anadkat, S. (2023). GPT-4 technical report. arXiv. *ArXiv Preprint ArXiv:2303.08774*.

[91]. Osipov, V. S., & Skryl, T. V. (2021). Impact of digital technologies on the efficiency of healthcare delivery. In *IoT in healthcare and ambient assisted living* (pp. 243–261). Springer.

[92]. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., & Ray, A. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

[93]. Palanica, A., Flaschner, P., Thommandram, A., Li, M., & Fossat, Y. (2019). Physicians' perceptions of chatbots in health care: cross-sectional web-based survey. *Journal of Medical Internet Research*, *21*(4), e12887.

[94]. Parviainen, J., & Rantala, J. (2022). Chatbot breakthrough in the 2020s? An ethical reflection on the trend of automated consultations in health care. *Medicine, Health Care and Philosophy*, *25*(1), 61–71.

[95]. Pirinen, T. A. (2019). Neural and rule-based Finnish NLP models—expectations, experiments and experiences. *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, 104–114.

[96]. 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.

[97]. Pol, U. R., Vadar, P. S., & Moharekar, T. T. (2024). *Hugging Face:*

Revolutionizing AI and NLP.

[98]. Pourkeyvan, A., Safa, R., & Sorourkhah, A. (2024). Harnessing the power of hugging face transformers for predicting mental health disorders in social networks. *IEEE Access*, *12*, 28025–28035.

[99]. Pradeep, R., Thakur, N., Sharifymoghaddam, S., Zhang, E., Nguyen, R., Campos, D., Craswell, N., & Lin, J. (2025). Ragnarök: A reusable RAG framework and baselines for TREC 2024 retrieval-augmented generation track. *European Conference on Information Retrieval*, 132–148.

[100]. Puhakka, O. (2025). *Usability of local large language models and retrieval augmented generation in health care*. O. Puhakka.

[101]. Qu, C., Dai, S., Wei, X., Cai, H., Wang, S., Yin, D., Xu, J., & Wen, J.-R. (2025). Tool learning with large language models: A survey. *Frontiers of Computer Science*, *19*(8), 198343.

[102]. Quaranta, M., Amantea, I. A., & Molinari, M. (2024). The Introduction of AI in Healthcare: A Multi-layered Issue. *EPIA Conference on Artificial Intelligence*, 233–244.

[103]. Ray, P., & Chakrabarti, A. (2022). A mixed approach of deep learning method and rule-based method to improve aspect level sentiment analysis. *Applied Computing and Informatics*, *18*(1/2), 163–178.

[104]. Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert- networks. *ArXiv Preprint ArXiv:1908.10084*.

[105]. Ru, D., Qiu, L., Hu, X., Zhang, T., Shi, P., Chang, S., Jiayang, C., Wang, C., Sun, S., & Li, H. (2024). Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. *Advances in Neural Information Processing Systems*, *37*, 21999–22027.

[106]. Saad-Falcon, J., Khattab, O., Potts, C., & Zaharia, M. (2023). Ares: An automated evaluation framework for retrieval-augmented generation systems. *ArXiv Preprint ArXiv:2311.09476*.

[107]. Salunkhe, V., Chintha, V. R., Pamadi, V. N., Jain, A., & Goel, O. (2022). AI-powered solutions for reducing hospital readmissions: A case study on ai-driven patient engagement. *International Journal of Creative Research Thoughts*, *10*(12), 757–764.

[108]. Santosh, K. C., Gaur, L., Santosh, K. C., & Gaur, L. (2021). Introduction to ai in public health. *Artificial Intelligence and Machine Learning in Public Healthcare: Opportunities and Societal Impact*, 1–10.

[109]. Sarker, I. H., & Kayes, A. S. M. (2020). ABC-RuleMiner: User behavioral rule-based machine learning method for context-aware intelligent services. *Journal of Network and Computer Applications*, *168*, 102762.

[110]. 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.

[111]. Sepahpour, T. (2020). *Ethical considerations of chatbot use for mental health support*. Johns Hopkins University.

[112].Shaheen, M. Y. (2021a). AI in Healthcare: medical and socio-economic benefits and challenges. *ScienceOpen Preprints*.

[113]. Shaheen, M. Y. (2021b). Applications of Artificial Intelligence (AI) in healthcare: A review.

ScienceOpen Preprints.

[114]. Shaikh, S. J., & Cruz, I. F. (2023). AI in human teams: Effects on technology use, members' interactions, and creative performance under time scarcity. *AI & SOCIETY*, *38*(4), 1587– 1600.

[115]. Sharma, D., Kaushal, S., Kumar, H., & Gainder, S. (2022). Chatbots in healthcare: challenges, technologies and applications. *2022 4th International Conference on Artificial Intelligence and Speech Technology (AIST)*, 1–6.

[116]. 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.

[117]. Sherani, A. M. K., Qayyum, M. U., Khan, M., Shiwlani, A., & Hussain, H. K. (2024). Transforming Healthcare: The Dual Impact of Artificial Intelligence on Vaccines and Patient Care. *BULLET: Jurnal Multidisiplin Ilmu*, *3*(2), 270–280.

[118]. Singh, A., Ehtesham, A., Kumar, S., & Khoei, T. T. (2025). Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG. *ArXiv Preprint ArXiv:2501.09136*.

[119]. Singh, A., Ehtesham, A., Mahmud, S., & Kim, J.-H. (2024). Revolutionizing mental health care through langchain: A journey with a large language model. *2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC)*, 73–78.

[120]. Sprengholz, P., & Betsch, C. (2021). Ok Google: Using virtual assistants for data collection in psychological and behavioral research. *Behavior Research Methods*, 1–13.

[121]. Sqalli, M. T., & Al-Thani, D. (2019). AI-supported health coaching model for patients with chronic diseases. *2019 16th International Symposium on Wireless Communication Systems (ISWCS)*, 452–456.

[122]. Sudhi, V., Bhat, S. R., Rudat, M., & Teucher, R. (2024). Rag-ex: A generic framework for explaining retrieval augmented generation. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2776–2780.

[123]. Tekkeşin, A. İ. (2019). Artificial intelligence in healthcare: past, present and future. *Anatol J Cardiol*, *22*(Suppl 2), 8–9.

[124]. Thirunavukarasu, A. J., Hassan, R., Mahmood, S., Sanghera, R., Barzangi, K., El Mukashfi, M., & Shah, S. (2023). Trialling a large language model (ChatGPT) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. *JMIR*

Medical Education, *9*(1), e46599.

[125]. Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S.

W. (2023). Large language models in medicine. *Nature Medicine*, *29*(8), 1930–1940.

[126]. Topaz, M., Murga, L., Gaddis, K. M., McDonald, M. V, Bar-Bachar, O., Goldberg, Y., & Bowles, K. H. (2019). Mining fall-related information in clinical notes: Comparison of rule-based and novel word embedding-based machine learning approaches. *Journal of Biomedical Informatics*, *90*, 103103.

[127]. Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44–56.

[128]. 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.

[129]. Turner, R. E. (2023). An introduction to transformers. *ArXiv Preprint ArXiv:2304.10557*.

[130]. Väänänen, A., Haataja, K., Vehviläinen-Julkunen, K., & Toivanen, P. (2021). AI in healthcare: A narrative review. *F1000Research*, *10*, 6.

[131]. Van Vuuren, C. L., Van Mens, K., de Beurs, D., Lokkerbol, J., Van der Wal, M. F., Cuijpers, P., & Chinapaw, M. J. M. (2021). Comparing machine learning to a rule-based approach for predicting suicidal behavior among adolescents: Results from a longitudinal population-based survey. *Journal of Affective Disorders*, *295*, 1415–1420.

[132]. Wagner, J., Mazurek, P., & Morawski, R. Z. (2022). Introduction to Healthcare-Oriented Monitoring of Persons. In *Non-invasive Monitoring of Elderly Persons: Systems Based on Impulse-Radar Sensors and Depth Sensors* (pp. 1–39). Springer.

[133]. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., & Funtowicz, M. (2020). Transformers: State-of-the-

art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.

[134]. Xiao, T., & Zhu, J. (2023). Introduction to transformers: an nlp perspective. *ArXiv Preprint ArXiv:2311.17633*.

[135]. Xie, X., Zang, Z., & Ponzoa, J. M. (2020). The information impact of network media, the psychological reaction to the COVID-19 pandemic, and online knowledge acquisition: Evidence from Chinese college students. *Journal of Innovation & Knowledge*, *5*(4), 297– 305.

[136]. 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.

[137]. Xu, X., & Cai, H. (2021). Ontology and rule-based natural language processing approach for interpreting textual regulations on underground utility infrastructure. *Advanced Engineering Informatics*, *48*, 101288.

[138]. Xu, X., Yu, A., Jonker, T. R., Todi, K., Lu, F., Qian, X., Evangelista Belo, J. M., Wang, T., Li, M., & Mun, A. (2023). Xair: A framework of explainable ai in augmented reality. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1– 30.

[139]. Xu, Y., Jiang, Y., Li, R., Gao, H., Guo, J., Liu, Y., Hei, L., & Wang, Y. (2022). A healthcare- oriented mobile question-and-answering system for smart cities. *Transactions on Emerging Telecommunications Technologies*, *33*(10), e4012.

[140]. Yang, J., Luo, B., Zhao, C., & Zhang, H. (2022). Artificial intelligence healthcare service resources adoption by medical institutions based on TOE framework. *Digital Health*, *8*, 20552076221126030.

[141]. Yang, R. (2024). CaseGPT: a case reasoning framework based on language models and retrieval-augmented generation. *ArXiv Preprint ArXiv:2407.07913*.

[142]. 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 data- driven neural signal decoding. *Future Generation Computer Systems*, *114*, 96–105.

[143]. Yu, H., Gan, A., Zhang, K., Tong, S., Liu, Q., & Liu, Z. (2024). Evaluation of retrieval- augmented generation: A survey. *CCF Conference on Big Data*, 102–120.

[144]. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *ArXiv Preprint ArXiv:1904.09675*.

[145]. Zhang, X., Song, Y., Wang, Y., Tang, S., Li, X., Zeng, Z., Wu, Z., Ye, W., Xu, W., & Zhang,

Y. (2024). Raglab: A modular and research-oriented unified framework for retrieval- augmented generation. *ArXiv Preprint ArXiv:2408.11381*.

[146]. Zhang, Y., Pei, H., Zhen, S., Li, Q., & Liang, F. (2023). Chat generative pre-trained transformer (ChatGPT) usage in healthcare. *Gastroenterology & Endoscopy*, *1*(3), 139–143.

[147]. Zhao, F., Chen, Y., Hou, Y., & He, X. (2019). Segmentation of blood vessels using rule-based and machine-learning-based methods: a review. *Multimedia Systems*, *25*, 109–118.

[148]. Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., Yang, L., Zhang, W., Jiang, J., & Cui,

B. (2024). Retrieval-augmented generation for ai-generated content: A survey. *ArXiv Preprint ArXiv:2402.19473*.

[149]. 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*.

[150]. Zhou, L., Li, Z., Zhou, J., Li, H., Chen, Y., Huang, Y., Xie, D., Zhao, L., Fan, M., & Hashmi, S. (2020). A rapid, accurate and machine-agnostic

segmentation and quantification method for CT-based COVID-19 diagnosis. *IEEE Transactions on Medical Imaging*, *39*(8), 2638–2652.