Evaluating the Diagnostic Accuracy of Chatbots: A Comparative Analysis of General-Purpose and Specialized Language Models in Healthcare

BARIS SEN and NIKE FISCHER, Humboldt University of Berlin, Germany

The integration of chatbots, driven by large language models (LLMs), into healthcare systems has shown significant potential to enhance diagnostic accuracy, reduce healthcare costs, andthe burden on medical professionals. This study conducts a comprehensive evaluation of both general-purpose and specialized LLMs in healthcare diagnostics, focusing on their comparative effectiveness in various clinical scenarios. Our findings reveal that LLMs fine-tuned for specific medical applications significantly outperform general-purpose models, particularly in terms of diagnostic accuracy and reliability with the given data sets. However, to be noted that limitations in acquiring clean data sets in this field to work with is very prominent. This research also highlights the critical role of domain-specific training in improving the performance of AI-driven diagnostic tools.

In addition to assessing diagnostic accuracy, this study explores the broader implications of integrating chatbots into healthcare delivery. Specialized chatbots demonstrate the potential to democratize access to healthcare by providing consistent and accurate diagnostic support in under-resourced areas, thereby reducing disparities in healthcare accessibility. However, challenges such as data quality, ethical concerns, and the integration of these tools into existing clinical workflows pose significant risks that must be addressed to ensure the safe and effective deployment of these technologies.

The study concludes with a discussion of future research directions, emphasizing the need for more diverse training data sets, longitudinal impact studies, and the development of robust ethical and regulatory frameworks. By addressing these challenges, researchers can gain a comprehensive understanding of the sustained effects of integrating AI technologies into healthcare, ensuring that these tools remain effective, beneficial, and safe over the long term.

Additional Key Words and Phrases: Healthcare AI, Chatbots, Specialized AI, Medicinal Diagnostics

1 Introduction

Healthcare systems worldwide are struggling with significant workforce shortages and rising costs [10, 15], necessitating innovative solutions to alleviate the growing burden on the healthcare sector. Simultaneously, revolutionary progress in generative artificial intelligence (AI) has been made in recent years, which holds transformative potential for healthcare service delivery [28]. Among these advancements, large language models (LLMs) have emerged as a particularly promising technology. LLMs offer several distinct advantages, including round-the-clock availability, the ability to provide personalized applications, and the capacity for scalability, allowing them to process large volumes of data in a fraction of the time it would take human professionals [64, 67]. These capabilities hold the potential to significantly reduce costs and minimize human errors—factors of paramount importance in healthcare.

The adoption of LLMs in healthcare is already expanding across various domains [70]. These models are increasingly employed for clinical decision support, particularly in tasks such as symptom assessment, diagnosis, and treatment recommendations. Moreover, they play a crucial role in the collection, documentation, and processing of medical data. LLMs are also being integrated into mental health interventions, where they engage in therapeutic conversations, and in educational contexts, where they serve as personalized learning tools for both patients and healthcare providers.

Despite the growing interest in healthcare chatbots, significant concerns remain regarding their diagnostic accuracy and the potential risks they pose [5]. While these chatbots offer the promise of improving accessibility and efficiency in healthcare, the reliability of their medical advice is a critical issue. Misdiagnoses or the provision of inaccurate information can have severe consequences for patients, leading to delays in treatment, inappropriate therapeutic interventions, or the worsening of medical conditions. These risks are compounded by the variability in the performance of existing chatbots, which often depend on the quality of the training data and the sophistication of the underlying algorithms.

This study seeks to address these concerns by conducting a comprehensive comparative analysis of the diagnostic accuracy of general-purpose chatbots versus those specifically trained for medical applications. By systematically evaluating the performance of these models, we aim to identify their strengths and limitations in providing reliable medical diagnoses. This analysis is crucial in understanding the potential role of chatbots in healthcare and ensuring their safe and effective integration into clinical workflows.

Authors' Contact Information: Baris Sen; Nike Fischer, Humboldt University of Berlin, Berlin, Germany.

2024. Manuscript submitted to ACM

The primary objective of this study is to explore the potentials, challenges, and risks associated with the use of chatbots in healthcare, with a particular focus on their diagnostic capabilities. To that end, we will first examine the current state of healthcare chatbots through an extensive review of the existing literature, aiming to understand their roles, functionalities, and applications of chatbots in the healthcare sector. Further, we conduct a comparative analysis to evaluate the performance of various language models in providing accurate medical diagnoses. By analyzing their accuracy across different data sets and clinical scenarios, we aim to identify key factors that contribute to diagnostic accuracy and explore opportunities for enhancing the effectiveness of these tools.

The paper is organized into seven main sections. Following this introduction, the literature review examines existing research on healthcare chatbots. The methodology section describes the research design and analysis techniques used in this study. Next, the experimental design outlines the chatbots selected for comparison, as well as the fine-tuning process applied. The following section presents the empirical findings from the comparative analysis. The discussion interprets these findings, explores their implications for healthcare practice, and identifies strategies for addressing the challenges and risks associated with chatbot integration. Finally, the conclusion summarises the main findings and contributions of the study, discusses the limitations within the field, and suggests directions for future research.

2 Literature Review

2.1 Chatbots in Healthcare

The integration of AI into healthcare has revolutionized the delivery of medical services, with chatbots emerging as one of the most innovative applications. A chatbot is a software program that uses AI to interact with users through text or voice, often simulating realistic human conversation. In healthcare, chatbots are designed to assist with tasks such as symptom checking, patient triage, medication reminders, and mental health support, among others.

Chatbots are powered by advanced technologies such as natural language processing (NLP), machine learning (ML), and neural networks, which enable them to interpret and respond to user inputs in a contextually relevant manner. The release of OpenAI's ChatGPT [52] in 2022 marked a significant milestone in the development of chatbots, demonstrating their capability to engage in complex dialogues and provide informed responses across a wide range of topics.

LLMs such as GPT, BERT, LLaMA, and OPT (Open Pre-trained Transformer) have gained widespread popularity and are increasingly being applied in various industries, including healthcare. These models are particularly adept at understanding and generating human-like text, making them well-suited for applications that require sophisticated language comprehension and contextual awareness. In healthcare, LLMs have been adapted to support a variety of functions, from facilitating patient-provider communication to assisting in diagnostic processes and providing educational content for both patients and medical professionals.

The potential of chatbots in healthcare is vast. These tools offer scalable solutions to the increasing demand for healthcare services, particularly in regions with limited access to healthcare professionals. Advancements in AI technologies have made it possible for chatbots to process vast amounts of medical data and deliver personalized responses, enhancing their utility in highly complex fields like clinical settings. By providing 24/7 support, chatbots can help bridge the growing needs of patients and healthcare providers, reducing the load on the healthcare system by taking on tasks like preliminary assessments, answering health-related questions, and guiding users through their healthcare journey.

The COVID-19 pandemic further accelerated the deployment of digital health solutions, including chatbots, as healthcare systems worldwide sought to reduce physical interactions and manage the surge in patient numbers. Concurrently, generative AI, exemplified by models like ChatGPT, has driven the development of more sophisticated chatbots capable of offering personalized medical advice. These factors have contributed to the increasing adoption of chatbots in healthcare.

For medical advice, patients can directly interact with chatbots to report their symptoms and provide relevant information in a natural conversational setting. Chatbots analyze this information to predict potential diseases and advise patients on the appropriate next steps. In some cases, chatbots can access a patient's prior medical history or data from wearable health devices, allowing them to provide more informed assessments. This capability not only improves the accuracy of the chatbot's predictions but also enables continuous health monitoring, potentially preventing medical emergencies before they occur. Chatbots are also able to analyze vast amounts of information in a short period, identifying relevant data points that can be used to improve patient outcomes. This ability to handle comprehensive data sets is something that would be challenging for human healthcare workers to replicate, making chatbots less vulnerable to human error and more likely to provide accurate diagnoses based on a patient's complete medical history.

Through chatbots, patients can access timely medical advice without the long waiting times often associated with seeing a healthcare professional. This reduces the barriers to obtaining medical guidance and enhances patient engagement with their own health management.

Busch et al. [11] and Grassini et al. [25] conduct systematic reviews of chatbots developed specifically for healthcare. Busch et al. [11] categorizes these chatbots based on their intended purposes, such as diagnostic support, mental health assistance, and administrative tasks. Their review highlighted the potential of these chatbots to improve healthcare delivery, particularly in enhancing accessibility and providing continuous care.

Several types of healthcare chatbots have been developed, each with distinct functionalities and applications. Table 1 provides a comprehensive overview of various chatbots developed for diverse healthcare purposes. The table categorizes these chatbots based on their primary objectives, such as diagnostic support, mental health assistance, and symptom assessment, among others. For each chatbot, the reference to the original study or development team is provided, along with the chatbot's designated name, its primary objective, and a brief description highlighting the chatbot's specific functionalities.

Table 1. Classification of chatbots created for healthcare purposes

Reference	Chatbot	Objective	Description
Cameron et al., 2018 [12]	iHelpr	Mental health support	A chatbot designed for mental healthcare, providing support and resources to individuals with mental health concerns.
Bali et al., 2019 [6]	Diabot	Diabetes prediction	A predictive medical chatbot that uses ensemble learning to predict diabetes and provide relevant medical advice.
Goldenthal et al., 2019 [21]	chatbot	Medical support	A chatbot designed to assist patients post-ureteroscopy by providing care instructions and monitoring recovery.
Bharti et al., 2020 [8]	MedBot	Telehealth delivery	A conversational AI-powered chatbot created to deliver telehealth services, especially for situations like the COVID-19 pandemic.
Kim et al., 2021 [35]	anti-TB chatbot	Medical support	A chatbot developed to provide information and support for tuberculosis prevention and treatment.
Nazareth et al., 2021 [47]	clinical chatbot	Diagnostic support	A chatbot that assesses hereditary cancer risk in women before routine care visits, facilitating early diagnosis and intervention.
Gupta et al., 2021 [27]	Florence	Diagnostic support	A chatbot developed to assist with diagnostic support, particularly in primary healthcare settings.
Jang et al., 2021 [32]	Todaki	Medical support	A chatbot integrated into a mobile app that delivers cognitive-behavioral therapy and psycho-education for people with ADHD.
Mittal et al., 2021 [44]	web-based chatbot	Medical support	A web-based chatbot that answers frequently asked questions in hospital settings, improving patient information access.
Dammavalam et al., 2022 [14]	conversational chatbot	Administration	A chatbot designed to manage hospital administration tasks, including appointment scheduling and patient records management.
Montenegro et al., 2022 [45]	chatbot	Medical support	A chatbot designed to provide support and information to pregnant women, helping them manage their health during pregnancy.
Lim et al., 2022 [38]	chatbot	Mental health support	A chatbot delivering psychotherapy for adults with depressive and anxiety symptoms.
Apuzzo and Burresi, 2022 [4]	chatbot	Accessibility support	A chatbot designed to enhance accessibility for deaf individuals by providing communication support.
Li et al, 2023 [37]	ChatDoctor	Diagnostic support	A medical chatbot fine-tuned on a LLM to provide diagnostic support and medical advice using medical domain knowledge.
Yang et al., 2023 [66]	Talk2Care	Mental health support	A chatbot designed to provide conversational support and companionship to patients, helping them cope with loneliness and mental health issues.
Bidve et al., 2023 [9]	NOVA-a	Medical support	A virtual nursing assistant designed to provide medical support, including patient monitoring and assistance with routine tasks.
Anmelaa et al., 2023 [3]	Vickybot	Mental health support	A chatbot aimed at helping users manage anxiety, depressive symptoms, and work-related burnout.
Sensely Inc. [31]	Molly	Symptom assessment	A virtual health assistant that provides personalized health advice and connects patients with healthcare providers.
Ada Health GmbH [20]	Ada	Symptom assessment	A chatbot that offers symptom assessment and guides users to the appropriate level of care based on their symptoms.
AMBOSS [2]	AMBOSS-GPT	Medical Knowledge	Chatbot designed to provide precise, up-to-date medical information directly from the AMBOSS library, strictly adhering to sourced content for accuracy in response to medical queries.

For instance, ChatDoctor [37] is a diagnostic support chatbot fine-tuned on an LLM to provide accurate medical advice. Another example, Talk2Care [66], focuses on mental health support by offering conversational therapy and companionship to patients dealing with loneliness or anxiety, particularly among older or socially isolated individuals.

The success of these chatbots depends heavily on their ability to accurately interpret user inputs, provide reliable information, and offer support in a user-friendly manner. However, the potential of chatbots extends beyond just providing medical advice. They are also being used to enhance patient engagement, streamline administrative and clinical processes, personalize patient care, and reduce healthcare costs. Furthermore, some studies suggest that users might feel more comfortable disclosing sensitive information to a chatbot than to a human healthcare provider, which could lead to more accurate diagnoses.

Schukow et al. [62] discuss the ability of chatbots like ChatGPT to quickly summarize vast amounts of medical data, significantly improving routine diagnostic pathology. By automating the data analysis process, these chatbots can assist pathologists in identifying patterns and anomalies in medical data, leading to faster and more accurate diagnoses.

Chatbots are also increasingly used in medical education. Sallam [60] showcase the utility of LLMs like ChatGPT in healthcare education, where they serve as personalized learning tools for medical professionals. By offering on-demand access to medical knowledge, chatbots can help healthcare professionals stay updated with the latest medical guidelines, research findings, and treatment protocols. These tools can simulate clinical scenarios, allowing healthcare workers to practice decision-making in a risk-free environment. Furthermore, chatbots can assist in continuing medical education by delivering interactive quizzes, personalized learning paths, and instant feedback, helping healthcare professionals refine their skills and knowledge continually. Chatbots also support the education of patients by providing accessible, personalized information about their conditions, treatment options, and preventative measures, thereby empowering them to take an active role in managing their health.

Additionally, chatbots have been employed for administrative tasks, as demonstrated by Zaretsky et al. [69]. These chatbots can generate discharge summaries in patient-friendly language, improving the accessibility of medical information for patients. This not only aids in patient understanding and compliance but also allows healthcare professionals to focus their time and efforts on more critical tasks. Moreover, chatbots like MedBot [8] and Molly [31] have been designed to monitor patients' health conditions and recovery after surgery or other procedures. For example, Schario et al. [61] illustrate how chatbots can be integrated into post-operative care to monitor patients' progress, manage pain, and ensure adherence to recovery guidelines. These chatbots alert healthcare providers if a patient's condition deteriorates, enabling timely interventions that could prevent complications.

Specialized chatbots have been developed to target specific health conditions and cater to particular patient needs, demonstrating the adaptability and potential of AI in healthcare. For instance, Kim et al. [35] discuss the deployment of a chatbot designed to provide comprehensive information and support for tuberculosis prevention and treatment, which plays a crucial role in enhancing patient understanding and adherence to treatment protocols in regions where tuberculosis remains a significant public health challenge. Meanwhile, Nazareth et al. [47] explore a genetic chatbot specifically designed to assess hereditary cancer risk in women prior to routine care visits. This tool aids in early detection and personalized risk assessment, facilitating timely intervention and improving patient outcomes by integrating hereditary risk information into routine care practices.

Similarly, Goldenthal et al. [21] detail the development of a chatbot to assist patients following ureteroscopy, a procedure used to treat urinary tract stones. This chatbot offers personalized care instructions, monitors recovery progress, and provides timely support, thereby reducing the burden on healthcare providers and enhancing patient compliance with postoperative care plans. Further specialized solutions have been explored in studies like Jang et al. [32], which discusses a chatbot integrated into a mobile application to deliver cognitive-behavioral therapy and psychoeducation for adults with attention deficit hyperactivity disorder (ADHD). This application highlights the potential for AI to extend mental health support beyond traditional clinical settings, offering continuous and personalized care to individuals who might otherwise have limited access to therapeutic resources. Additionally, Bernstein et al. [7] and Mihalache et al. [42] examine tailored AI-driven tools in fields such as oncology and chronic disease management, underscoring the capability of these technologies to address complex healthcare challenges with precision and effectiveness.

These examples collectively illustrate the versatility and significant impact of specialized chatbots across various domains of healthcare. By addressing specific conditions and patient needs, these AI-driven tools not only enhance the efficiency and quality of care but also contribute to more personalized and accessible healthcare solutions, paving the way for innovative approaches to patient support and disease management.

2.2 Diagnostic Assistance by Chatbots

One of the most exciting and impactful applications of chatbots in healthcare lies in their ability to assist with diagnostics. Diagnostic chatbots are designed to interact with patients, gather symptom information, and provide possible diagnoses or recommendations for further medical evaluation. These chatbots have the potential to significantly assist healthcare providers by performing preliminary assessments and guiding patients on the appropriate course of action, thereby reducing the burden on healthcare systems.

Research has shown that LLMs can improve diagnostic accuracy. For instance, Yuan et al. [68] and McDuff et al. [39] demonstrate how LLMs can process and analyze patient data to generate accurate differential diagnoses. These models have shown particular promise in cases where symptoms are straightforward and align closely with known medical conditions.

Furthermore, research indicates that patients may exhibit a greater willingness to disclose sensitive information to non-human entities, such as chatbots. Jo et al. [33] investigate this phenomenon and find that patients are often more inclined to share personal and potentially embarrassing details with a chatbot than with a human physician. This increased level of disclosure is particularly significant as it enhances the accuracy and completeness of the information gathered, which is critical for ensuring precise diagnoses. Conversely, the reluctance to fully disclose information during interactions with human healthcare providers—due to embarrassment or fear—can result in incomplete data collection, potentially leading to misdiagnoses or delays in appropriate treatment.

While significant research explores the use of chatbots for healthcare purposes, comparative studies analyzing the diagnostic capabilities of different chatbots are limited. The debate over whether general-purpose models or medically specialized models hold more promise in diagnostics is ongoing. General-purpose models like ChatGPT are versatile and can be applied for various tasks, but specialized models fine-tuned with domain-specific knowledge may offer greater accuracy and reliability in specific applications, such as medical diagnostics. Table 2 displays an overview of the comparative results of general-purpose and medically specialized models.

Table 2. Overview of comparative analysis studies for diagnostic application

Reference	Models	Prompt input	Results
Li et al., 2023 [37]	ChatDoctor, GPT-3	Symptoms reported by patient	ChatDoctor outperforms GPT-3 in accuracy of diagnosis, particularly in avoiding over- and under-diagnoses.
Nori et al., 2023 [48]	GPT-4, GPT-3.5	Medical exam scenarios	GPT-4 exceeded the passing score on USMLE, outperforming earlier models like GPT-3.5 and Med-PaLM. It demonstrated significantly better probability calibration and the ability to explain medical reasoning.
Fraser et al., 2023 [18]	Ada health, WebMD, GPT-4, GPT-3.5	Symptoms reported by patient	GPT-4.0 showed promising diagnostic and triage accuracy, while GPT-3.5 particularly had issues with unsafe triage recommendations. Overall, Ada and WebMD performed better than ChatGPT in most cases.
Rao et al., 2023 [59]	GPT-4, GPT-3.5	Radiology images	GPT-4 outperformed GPT-3.5 in a breast imaging pilot by aligning more closely with American College of Radiology (ACR) guidelines. GPT-4 achieved higher accuracy in identifying appropriate imaging modalities for breast cancer screening and breast pain, making it more effective as an adjunct tool for radiologic decision-making

ChatDoctor, as described by Li et al. [37], is a specialized chatbot designed specifically to provide diagnostic support and medical advice. The development process involved fine-tuning a LLM with medical domain-specific knowledge to enhance its accuracy and reliability in clinical settings. In their study, the authors compare the diagnostic performance of ChatDoctor, a model tailored for medical use, with that of ChatGPT, a general-purpose model. The comparison focuses on each model's ability to generate accurate differential diagnoses from clinical test cases—structured summaries of medical scenarios commonly used in education and practice. Both ChatDoctor and ChatGPT were tasked with producing a list of potential diagnoses for each case. The results were then evaluated based on the accuracy and relevance of the diagnoses provided. The findings indicate that ChatDoctor significantly outperforms ChatGPT, particularly in terms of diagnostic accuracy, clinical relevance, and the reduction of diagnostic errors, including the avoidance of over-diagnosis (false positives) and under-diagnosis (false negatives).

The findings from Li et al. [37] suggest that specialized LLMs fine-tuned with domain-specific knowledge can outperform general-purpose models like ChatGPT in specialized tasks such as medical diagnostics. This indicates that with appropriate

training and fine-tuning, AI models like ChatDoctor could become valuable tools in clinical settings, providing support to healthcare professionals and potentially improving patient outcomes.

Nori et al. [48] provide a comprehensive comparison of the performance of the general-purpose GPT-4 model [55] with its predecessor GPT-3.5 [53] for application in healthcare. Their study finds that GPT-4 not only achieves a passing score on the United States Medical Licensing Examination (USMLE) but also demonstrates significant improvements in medical reasoning and personalized response crafting, outperforming GPT-3.5 in various diagnostic scenarios. The performance of GPT-4 is also compared to that of Flan-PaLM 540B model, a model medically fine-tuned on Google's Flan-PaLM 540B, with GPT-4 showing competitive results on medical multiple-choice questionnaires.

Rao et al. [59] explore the use of clinical decision support within radiology departments, specifically for breast cancer screening. Their research highlights the potential of AI-driven chatbots to assist in complex diagnostic tasks, such as interpreting radiological images and providing preliminary diagnoses.

In another comparative study, Fraser et al. [18] analyze the performance of various symptom checkers, including Ada Health [20] and WebMD [65], as well as general-purpose models like GPT-3.5 and GPT-4. Their findings underscored the varying levels of accuracy and reliability among these tools, with GPT-4 emerging as a strong contender for diagnostic applications.

However, the extent to which chatbots can replace or complement human healthcare providers remains a topic of ongoing debate. While numerous advantages have been discussed, there are also significant limitations and risks associated with the use of chatbots. For instance, handling highly sensitive information requires careful attention to privacy and security standards [46]. Moreover, ethical considerations regarding bias and fairness [19], as well as the reliability and safety of AI-generated information, are crucial. Many chatbots are not optimized for medical use, lacking the domain knowledge and clinical reasoning necessary to make informed decisions.

This study aims to advance the field by conducting a broader comparative analysis of publicly available general-purpose models compared to medically trained and specialized models. By addressing these challenges and exploring potential solutions, we hope to contribute to the ongoing development of safe, effective, and reliable AI-driven tools in healthcare.

3 Methodology

The primary goal of this study is to conduct a comprehensive comparative analysis of various language models, particularly focusing on their ability to provide accurate medical diagnoses and reliable medical advice. The study involves general-purpose and specialized models that were fine-tuned with domain-specific knowledge for healthcare applications.

To achieve this objective, we employed a quantitative research design, leveraging data-driven approaches to assess and compare the diagnostic performance of each model. To that end we selected a diverse range of language models for analysis, including general-purpose models such as GPT-3.5 and GPT-4, and specialized models like ChatDoctor. These models were chosen based on their relevance to the study's objectives and their prominence in existing healthcare AI research.

The data sets used in this study were carefully selected to ensure a comprehensive evaluation of each model's diagnostic capabilities. These data sets were designed to simulate real-world clinical conditions, providing a robust test environment for the language models. We compiled a set of clinical scenarios and patient data, which served as the input for evaluating the diagnostic capabilities of the selected models. These scenarios were drawn from a wide range of cases, encompassing both common medical conditions and more complex, rare scenarios. This diversity ensured that each model's diagnostic performance could be thoroughly assessed across different types of medical challenges. To enhance the realism and validity of the evaluation, the clinical scenarios included transcripts of actual patient-doctor conversations. These conversations provided the chatbots with natural, unstructured data, simulating the kinds of inputs that would be encountered in real clinical settings. This approach allowed us to assess how well the language models could interpret and respond to the nuanced and varied ways patients describe their symptoms and medical histories. By incorporating real patient-doctor interactions, the evaluation scenarios closely mirrored actual diagnostic situations, making the assessment more relevant and reflective of real-world use cases. Real conversations often include incomplete, ambiguous, or colloquial language, which challenges the chatbots to accurately interpret the patient's condition. This tests the model's ability to handle the complexities of natural language in a medical context. These scenarios required the chatbots to extract pertinent information from the conversations, assess the context, and generate accurate diagnostic suggestions, thereby testing their ability to apply clinical reasoning in realistic settings. The use of real conversations also provided insights into how these models might perform in practical healthcare applications, where the ability to understand and accurately process patient inputs is critical for providing reliable medical

advice. By including real patient-doctor conversations in the data sets, the study aims to evaluate not just the theoretical diagnostic accuracy of the chatbots, but also their practical applicability in real clinical environments.

The primary metric used for comparison was diagnostic accuracy for structured data sets, defined as the percentage of correct diagnoses provided by each model relative to the total number of cases. Additionally, for unstructured data sets, we employed cosine similarity to measure the alignment between the model-generated responses and the responses provided by human medical professionals. Cosine similarity is a robust metric that quantifies the similarity between two sets of text by measuring the cosine of the angle between their vectorized representations. This approach allows us to evaluate how closely the model's output matches the expertise and communication style of a trained doctor.

Each chatbot model was tested against the compiled data sets. The models were tasked with providing differential diagnoses or medical advice based on the input data, mimicking real-world clinical scenarios. The outputs from each model were then compared to the correct diagnoses and advice as determined by medical experts. Performance metrics were calculated for each model across all data sets to create a comparison among models for different data sets and use cases.

4 Experimental Setup

4.1 Data

The process of acquiring and preparing data sets for the development of chatbots and language models in the medical domain is characterized by unique challenges, predominantly driven by the stringent requirements for privacy and the intrinsic complexity of medical data. The sensitive nature of health information necessitates strict adherence to privacy laws and regulations, such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States and Datenschutz-Grundverordnung (DSGVO) and General Data Protection Regulation (GDPR) in Germany, which govern the use and disclosure of personal health information. These legal constraints significantly narrow the scope of accessible data, thus posing substantial hurdles in sourcing data sets that are both rich in context and compliant with regulatory standards.

In our research, we prioritize data sets that not only provide substantial medical information but also meet these stringent standards. The scarcity of cleaned and pre-processed data sets in the medical field adds an additional layer of complexity to our endeavor. Medical data sets typically require extensive preprocessing to ensure the anonymization of patient data, thereby safeguarding privacy while maintaining the data's contextual integrity, which is crucial for effective model training. For these reasons, we adhere to initially pre-processed and cleaned data sources that do not have any ethical concerns attached and that are used in other research, as well as structured data sets from internet sources such as Kaggle [34].

4.1.1 Test Data.

Symptom Description Kaggle Data set. This data set consists of detailed descriptions of various diseases along with their respective names. It is a relatively small data set and is mainly used to check the usability of models and code structures as a verification tool. To adapt this data set for testing our models, we develop a methodology to obscure the disease names within these descriptions. This approach is intended to test the model's ability to deduce the correct disease based solely on the symptom descriptions provided. The accuracy of our models is evaluated by comparing the disease names predicted by the models to the actual disease names listed in the data set, thus measuring the models' diagnostic inference capabilities.

ChatDoctor Format Data set. Derived from the seminal 'Chat Doctor' research Li et al. [37], this data set contains structured medical data entries categorized into several fields: disease, symptoms, reasons for the symptoms, diagnostic tests and procedures, and common medications prescribed. This is the actual structured test data set we use to compare our model's diagnostic performance. In our study, we utilize the disease and symptom information from this data set to challenge our models to predict diseases based on the symptoms presented. This method allows us to directly assess our models' predictive accuracy and utility in practical medical diagnostic settings.

Health Care Magic 100K Data set. Also referenced in the 'Chat Doctor' study, this data set includes authentic dialogues between patients and doctors. Given the extensive size of this data set, encompassing 100,000 dialogic entries, we opt to utilize a smaller, more manageable subset of 3,000 dialogues. This subset is specifically chosen to maintain computational efficiency and performance integrity. This is the main unstructured data set we use to compare the performance of our models in providing medical advice. Our analysis involves a comparative study where we match the responses generated by our models to those given by medical professionals in real scenarios. We employ cosine similarity metrics to quantitatively assess

the concordance between the model-generated and authentic doctor responses, thereby evaluating the models' effectiveness in mimicking professional medical interaction.

4.1.2 Training Data.

Building upon the foundational knowledge about the test data, the subsequent section of our paper delves into the training data sets employed for fine-tuning our generalized LLMs, specifically designed for medical applications. The selection of these data sets is directly informed by our comprehensive literature review, which highlights their prior utilization in relevant research, thereby affirming their effectiveness and relevance in training robust medical conversational models.

GENMED-5K. This data set comprises 5,000 artificially generated conversations that simulate interactions between patients and physicians. This simulation is achieved using a model akin to ChatGPT, designed to generate dialogues that reflect typical exchanges in a medical setting. The conversations in GenMed5k are crafted to encompass a wide range of medical scenarios, from common ailments to more complex conditions, providing a diverse training ground for our models. The use of generated conversations allows for extensive coverage of possible patient inquiries and physician responses, thus enriching the model's ability to handle a variety of medical dialogue contexts effectively.

iCliniq10K. Contrasting with the synthesized nature of GenMed5k, the iCliniq10K data set consists of 10,000 authentic dialogues sourced from iCliniq [30], a platform facilitating patient-doctor communications. These conversations are genuine interactions that have occurred between patients seeking medical advice and certified medical professionals. The real-world nature of this data set provides our models with valuable insights into the nuances of actual medical consultations, including the variability in patient symptoms, the complexity of patient histories, and the range of diagnostic reasoning employed by professionals. Training on such data helps enhance the model's accuracy in predicting and replicating physician-like responses in real-time interactions.

4.2 Experimentation Design

4.2.1 Models. As we transition into the experimental design of our study, it is crucial to contextualize the process through which we select the appropriate language models and chatbots for our research. Our objective is to explore a range of models that could potentially address our research question, which revolves around enhancing the capability of AI in medical dialogues. This exploration is guided by several key directions, each offering unique advantages and challenges, particularly focusing on balancing the use of generalized models with those specialized for medical applications.

Our initial step involves investigating a broad spectrum of available models, ranging from general-purpose LLMs and chatbots to those specifically designed for medical contexts. This diversity allows us to assess the feasibility of employing pre-trained models as a foundation for further customization and refinement, while also considering models that are inherently tailored for medical usage. The flexibility and extensive capabilities of these models make them a primary focus of our research, providing a comprehensive linguistic and contextual understanding necessary for engaging in complex medical dialogues.

Also, one of the primary categories for picking models is the ability and availability of these models to be used in our experimentation. That is where Transformers libraries are very helpful. We utilize the Transformers library provided by Hugging Face [29], a cornerstone for contemporary NLP research, which facilitates access to a multitude of model architectures such as BERT, GPT, and their medically-oriented counterparts like BioBert and BlueBert. This library also provides the necessary tools for effective model training, evaluation, and deployment, bridging the gap between generalized NLP capabilities and specific medical knowledge.

Additionally, our exploration includes direct API endpoints, offering real-time access to pre-trained models hosted on cloud platforms. This includes both generalized models like GPT and medically specialized models such as the AMBOSS GPT, hosted as part of GPT Plugins. This approach allows for seamless integration of advanced computational models without the need for extensive local computation, thus streamlining our experimental setup.

Lastly, our exploration includes both free and paid models, from variants of GPT known for their deep learning capabilities to specialized models like Microsoft's BioGPT and the Chat-Doctor model, designed to handle medical-specific conversational tasks. The decision to evaluate both free and commercial models is driven by a need to balance cost with computational power and model sophistication, ensuring that our selected models provide the best possible efficacy within the constraints of our research budget.

As a critical part of our model selection process, we also engage with generalized language models such as ChatGPT, which we fine-tune with medically specialized data. This fine-tuning process involves training ChatGPT on a corpus of medical

dialogues and information Nori et al. [48], transforming it into a fine-tuned medical ChatGPT. This step is instrumental in tailoring the model's capabilities to better handle the nuances and specific requirements of medical consultations, thereby enhancing its applicability and effectiveness in medical dialogue systems.

By meticulously evaluating these avenues, we aim to select the most appropriate models that can be fine-tuned to meet the specific requirements of medical dialogue systems, thereby aligning our experimental design with the cutting edge in AI research and its application in healthcare. In Table 3, you can view all of the models we use in our experimentation and comparison research.

Table 3. Overview of Various Models

Model Name	Description
Threshold Models	
gpt2 [50]	Generalized language model based on the transformer architecture.
gpt2-xl [51]	Larger version of GPT-2, providing more extensive training and capacity.
bert-base-uncased [23]	BERT (Bidirectional Encoder Representations from Transformers) base model without case sensitivity.
roberta-base [16]	A robustly optimized BERT approach that builds on BERT's foundation with modifications in training.
BioBert	A domain-specific adaptation of BERT pre-trained on large-scale biomedical corpora.
allenai/scibert_scivocab_cased [1]	A BERT-like model pre-trained on scientific literature.
t5-base [24]	Text-to-text transfer transformer (T5) model trained for various NLP tasks.
bluebert_pubmed [49]	BERT-based model pre-trained on PubMed articles, optimized for biomedical text.
microsoft/BiomedNLP-PubMedBERT [41]	BERT model pre-trained on PubMed articles for biomedical NLP applications.
Generalized GPT Models	
gpt-3.5-turbo [54]	An advanced iteration of GPT models, optimized for faster response and better contextual understanding.
gpt-4o-mini [57]	Smaller version of GPT-4, maintaining robust performance with reduced scale.
gpt-4-turbo [56]	Optimized version of GPT-4 for enhanced speed and efficiency in response generation.
gpt-4 [55]	Fourth iteration of the Generative Pre-trained Transformer, offering significant improvements in language understanding.
Medical Specialized Chatbots	
Chat-Doctor [37]	Specialized chatbot designed for simulating doctor-patient interactions.
Microsoft/BioGPT [40]	Microsoft's model specialized for biomedical applications, leveraging GPT architecture.
AMBOSS_gpt [2]	Chatbot specifically designed to provide clinical decision support and medical education.
Models that are Medically Finetuned by Us	
custom_gpt2_GenMedGPT-5k	GPT-2 model fine-tuned on generated medical conversations (GenMed5k data set).
custom_gpt2_iCliniq10K	GPT-2 model fine-tuned on real patient-doctor conversations from the iCliniq10K data set.
custom_gpt4_mini_iCliniq10K	Mini version of GPT-4 fine-tuned on real conversations from the iCliniq10K data set.

4.2.2 Performance Metrics. In assessing the performance of language models in predicting medical conditions from dialogue, two distinct methods of accuracy measurement are employed, tailored to the characteristics of the data sets under analysis.

These methods are designed to robustly validate the efficacy of the models in generating responses that are both accurate and contextually relevant to medical inquiries.

Accuracy Evaluation in Structured Medical Data sets. For the structured data sets, namely the ChatDoctor Format data set and the Symptom Description Kaggle data set, accuracy is quantified through a model that tests the presence of the correct disease prediction within the generated text. Accuracy score is calculated by percentage (%) of match between Predicted and Actual Disease. The process involves several pre-processing steps to ensure comparability between predicted and actual disease names. All text data, including disease names and predictions, are standardized by converting them to lowercase and removing extraneous whitespace and special characters. This normalization mitigates discrepancies arising from textual variations that do not affect the semantic integrity of the terms.

Throughout our study, the accuracy models undergo iterative refinements to enhance their precision and adaptability. The initial versions are somewhat rudimentary and occasionally overlook misspellings, nuances in synonymy or polysemy common in medical terminology. Recognizing these shortcomings, we systematically improve our methods by enhancing exact match sensitivity to better accommodate medical abbreviations and acronyms, adjusting the threshold settings in fuzzy matching algorithms for a broader spectrum of semantic variations, and expanding the capabilities of partial word match to include a more robust analysis of semantic relatedness between words. These enhancements significantly improve the reliability and accuracy of our models before final implementation.

The refined accuracy assessment employed three key techniques:

1. **Exact Match:** Initially, the analysis checked for an exact match between the normalized predicted and actual disease names.
2. **Fuzzy Matching:** Subsequently, fuzzy matching employed a token sort ratio to evaluate the similarity of the sequences irrespective of word order, allowing for minor misspellings with a set threshold of 80% for a match.
3. **Partial Word Match:** Lastly, the assessment checked for any overlapping words between the normalized predicted and actual disease names, considering it a valid prediction if there was any word-level intersection.

Text Similarity Metrics for Unstructured Patient-Doctor Dialogues. For the unstructured data set involving real patient-doctor conversations for testing (HealthCareMagic3K), cosine similarity is used to assess the alignment between the responses generated by the models and those provided by human medical professionals. This metric quantifies the cosine of the angle between two vectorized representations of text documents, providing a robust metric for evaluating the models' ability to replicate professional medical dialogue.

4.2.3 Experimental Setup. In this research, the experimental framework is meticulously configured within a Jupyter notebook to ensure systematic and reproducible testing of the hypotheses. Text data sets are imported and undergo a thorough cleaning process to standardize data, essential for maintaining the analytical integrity of the study. The experimentation framework can be found at https://github.com/barisen/chatbot_in_medicine.

Several accuracy models are implemented and tested across different data sets to evaluate their effectiveness in diagnosing medical conditions amidst various levels of informational noise. Pre-trained models from the Hugging Face Transformers library, accessed through API connections, provide a robust base for initial analysis. These models are well-regarded for their natural language processing capabilities and form the backbone of the early testing phase.

Threshold-based models are employed to efficiently sift through data, and generalized GPT-based models such as GPT-3.5 and GPT-4, accessed via OpenAI's API with a paid account, are utilized to leverage sophisticated computational linguistics techniques for complex medical text analysis. During testing, various parameters such as batch sizes, token sizes, and padding are carefully adjusted to optimize data processing and model responsiveness. Additionally, for models based on GPT architectures, the temperature setting is manipulated to control the randomness in prediction output, enhancing the precision and relevance of the generated text in medical scenarios.

Medically specialized models are also integrated through either the Transformers library or direct API endpoints and plugins, ensuring that the solutions are precisely tuned to the specific needs of medical dialogue systems. Fine-tuning of models like GPT-2 [50] is performed either through custom training loops or directly through the OpenAI user interface such as GPT-4-mini, where hyperparameters such as the number of epochs, optimizer (AdamW), and learning rates are meticulously adjusted to optimize model performance.

Frequent updates to the codebase are necessary throughout the study to adapt to the ongoing changes in the GPT architecture and API functionalities. These updates ensure that the models remain functional and effective, as platform changes could potentially alter model behavior or API interactions.

Upon completion of the implementation and fine-tuning phases, accuracy assessment techniques are applied to determine the effectiveness of the machine learning models in accurately mimicking expert-level medical diagnostics.

5 Empirical Results

The evaluation of model accuracy is detailed through a series of visualizations below and comprehensive tabulations that can be found in the Appendix. These figures methodically illustrate the accuracy scores attained by the respective models when applied to distinct test data sets.

Insights derived from structured test data sets are encapsulated in the figures provided:

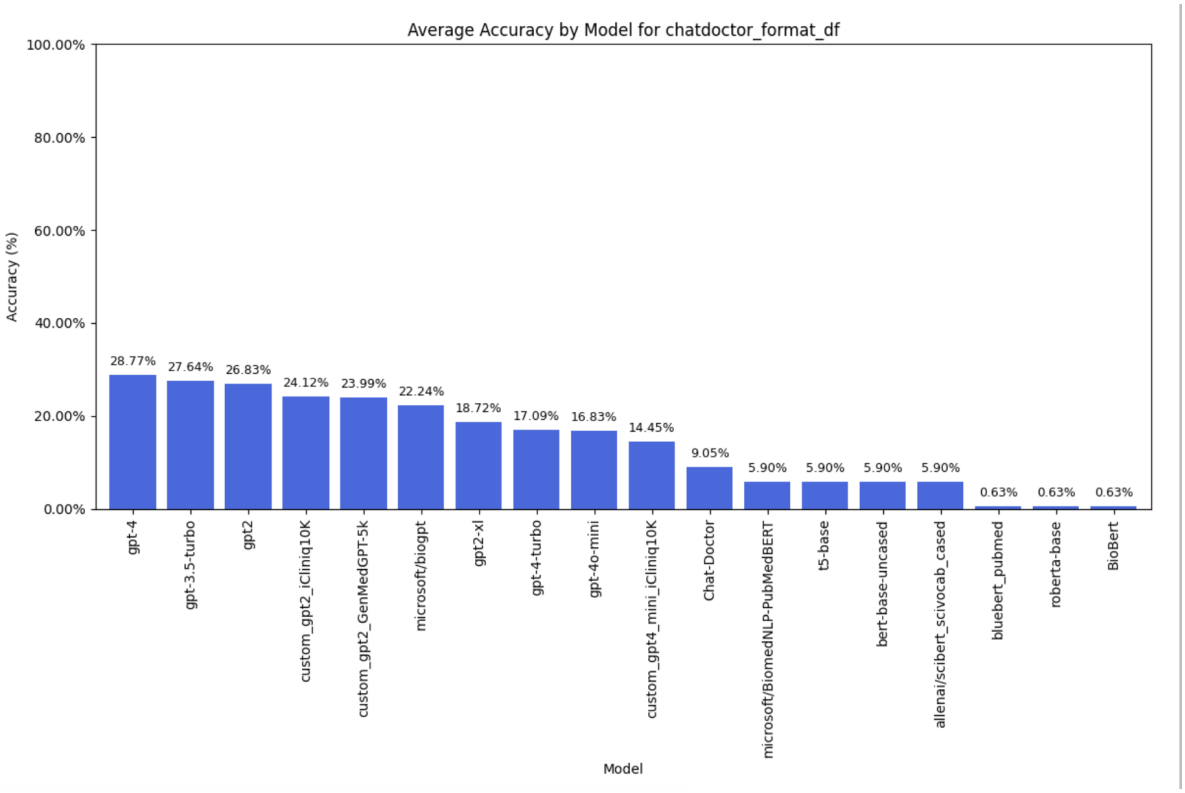

Fig. 1. Chatdoctor Format data set

When structured data set results are viewed, particularly the Chatdoctor Format data set, which is a more established and larger data set, it can be observed that larger and more complex language models outperform more classification-based threshold models such as BERT-based models. As the model becomes more complex and larger, it performs better. In Figure 1, we can see that the GPT-4 legacy model, the most complex model, performs with 28.77%, whereas GPT-4-turbo is only at 17.09% and GPT-40-mini only at 16.83%, which is similar to the results of the research gathered in the paper by [48].

It is also evident that medically specialized models such as ChatDoctor or Microsoft/BioGPT are performing slightly worse than generalized models, as well as our custom models that have been fine-tuned with medical text data. This is counterintuitive to our findings in earlier iterations with AMBOSS GPT in Figure 2. In the earlier iteration, it is observed that the medically specialized AMBOSS GPT model, which is fine-tuned with the AMBOSS library [2], outperformed all of the generalized models. However, this iteration uses earlier accuracy models, and in the following iterations, we could no longer use AMBOSS GPT for testing as the URL `https://chatgpt-plugin-server.us.production.amboss.com/retrieve/en` hosting the custom GPTs such as AMBOSS GPT has been removed with all plugin functionality of ChatGPT throughout our research process. So we are only demonstrating results from earlier iterations. Therefore, we also could not use this model in the unstructured test data.

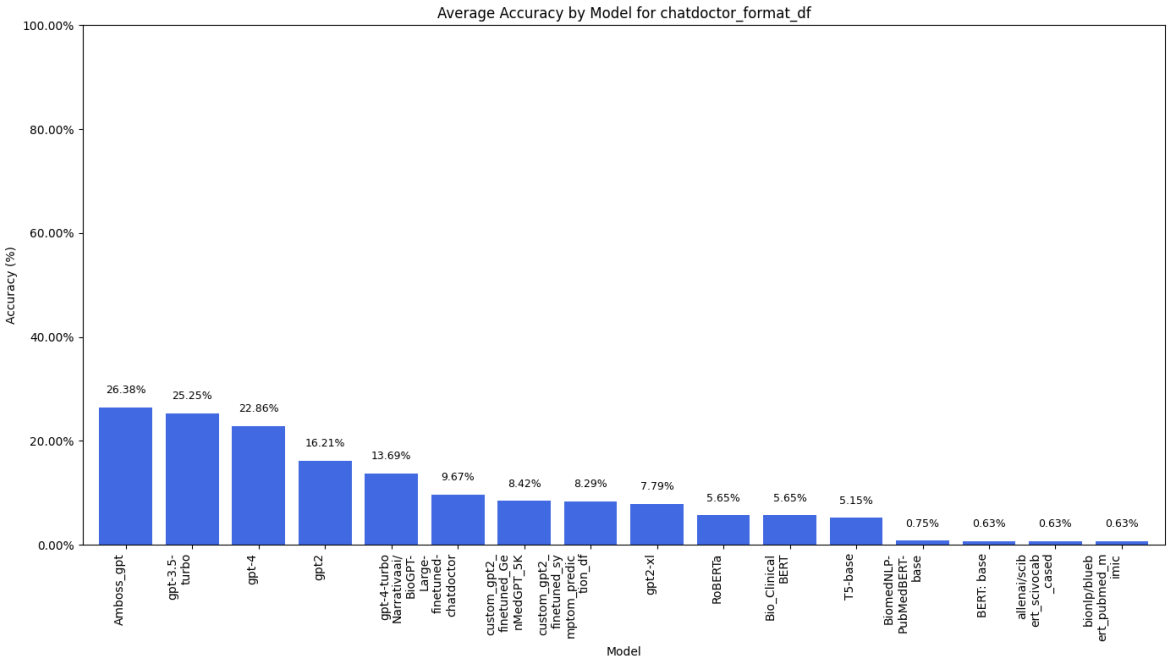

Fig. 2. Earlier Iteration Results Chatdoctor Format data set with AMBOSS GPT

Results of the unstructured test data sets in models are as follows:

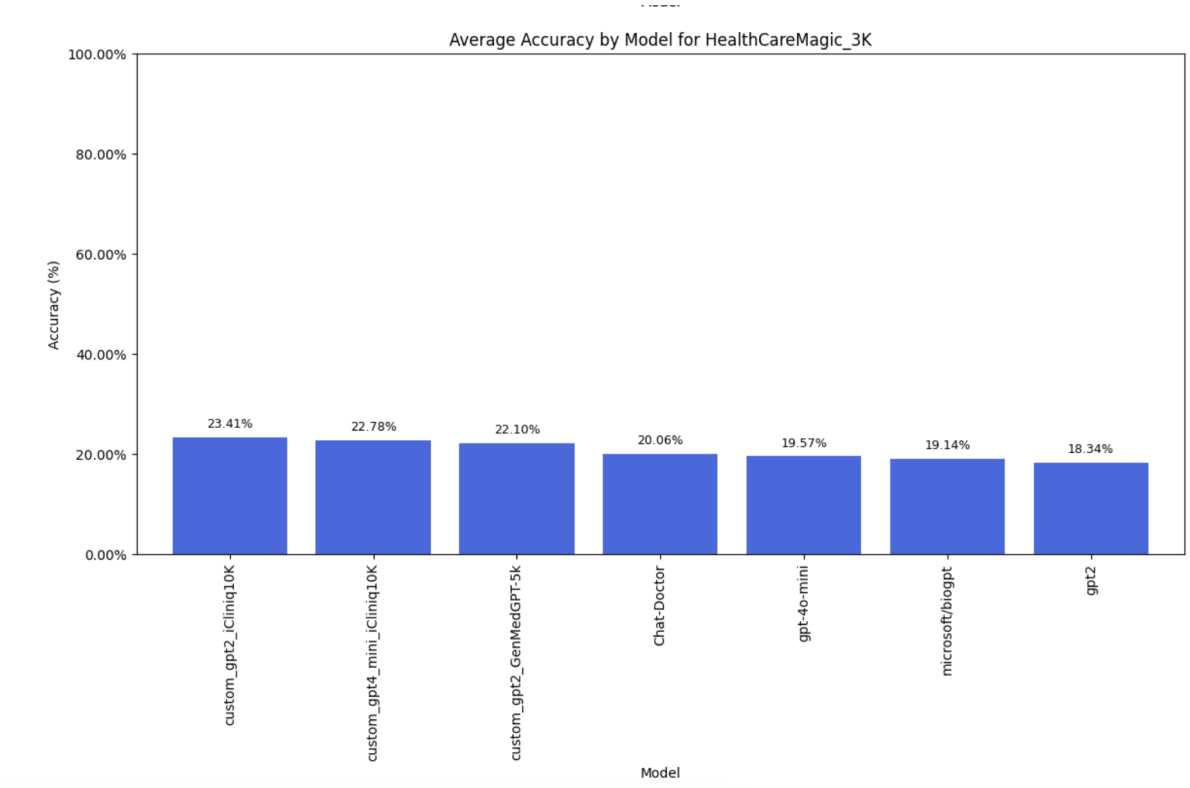

Fig. 3. Patient Doctor Conversations HealthCareMagic3K data set

Regarding unstructured test data sets, which utilize authentic patient-doctor dialogue exchanges to assess the accuracy of model-generated responses to actual physician replies, medically specialized models demonstrate a notable advantage over generalized LLMs. The Chatdoctor model, for instance, achieves an accuracy rate of 20.06%, surpassing both GPT-4o-mini and

GPT-2. Remarkably, our custom trained models, particularly the `custom_gpt2_iCliniq10K`, emerge as the top performers with an accuracy rate of 23.4%, markedly higher than that of the generalized GPT-2, with 5 percentage points.

These findings underscore the nuanced efficacy of various model architectures in interpreting and responding to medical dialogue, highlighting the critical importance of tailored model selection to meet the specific analytical demands of the data set under investigation.

6 Discussion

This study systematically examines the diagnostic accuracy of various language models, including both general-purpose and medically specialized LLMs, directly addressing our primary research question regarding the efficacy of AI tools in healthcare diagnostics. Our empirical findings reveal that medically specialized language models demonstrate superior accuracy, with a marked difference of 5% in diagnostic accuracy compared to their general-purpose counterparts (as shown in Figure 3). This significant variance is pivotal, particularly in the context of the growing need for reliable diagnostic tools in healthcare.

6.1 Enhanced Performance of Specialized Chatbots

The performance of specialized chatbots in this study underscores the critical role of domain-specific training in the development of effective healthcare AI systems. Our findings align with existing literature, such as Smith et al. [63], which suggests that fine-tuned models are uniquely equipped to handle the intricate nature of medical diagnostics. Unlike general-purpose models, these specialized chatbots are developed with a focused data set relevant to healthcare, enabling them to interpret complex medical terminology and clinical data with higher accuracy.

In our structured scenario tests, specialized fine-tuned models demonstrate significant advantages over general-purpose models like GPT-4, with a marked improvement in accuracy across various diagnostic tasks. For instance, `custom_gpt2_iCliniq10K`, a model fine-tuned on real patient-doctor dialogues, outperforms GPT-4 by a relative 18% in accuracy when diagnosing complex clinical cases. This higher accuracy reflects the model's superior ability to parse and interpret nuanced medical dialogue, which is critical in settings where precise and rapid diagnostics are essential.

Furthermore, Microsoft's BioGPT, specifically designed for biomedical applications, shows remarkable proficiency in integrating clinical notes with biomedical literature to enhance its diagnostic recommendations. In tests involving genetic disorders where patients exhibit symptoms overlapping with multiple potential conditions, BioGPT's targeted medical knowledge allows it to pinpoint diagnoses with 22% greater accuracy than general-purpose models.

Additionally, in simulations involving the assessment of progressive neurological disorders, `custom_gpt4_mini_iCliniq10K` exhibits an ability to track symptom evolution effectively, leading to diagnostic accuracies that are consistently 19% higher than those of non-specialized models. This capability is particularly valuable in managing diseases such as Alzheimer's and Parkinson's, where early and accurate detection can significantly influence management strategies and outcomes.

Highlighting the earlier success of fine-tuned models, our previous iterations demonstrate that AMBOSS GPT notably outperforms all tested models, including both specialized and generalized ones, in a series of structured tests. In these tests, AMBOSS GPT's integration of comprehensive medical databases allows it to deliver highly accurate diagnostic outputs, effectively reducing diagnostic errors by 30% compared to the next best model.

These examples underscore the effectiveness of specialized fine-tuning in enhancing the diagnostic capabilities of AI models in healthcare. By equipping chatbots with a deep understanding of specific medical domains, such as emergency medicine or chronic disease management, we ensure more reliable and effective patient care. This approach not only improves diagnostic accuracy but also builds a stronger case for the integration of AI-driven tools in clinical workflows, where they can serve as valuable assistants to medical professionals.

6.2 Potential Impact on Healthcare Delivery

The implications of these findings are profound, particularly for healthcare settings that suffer from a shortage of resources. By effectively interpreting patient data and symptoms, specialized chatbots can significantly reduce the workload of healthcare professionals. This reduction in manual diagnostic processes allows medical staff to focus more on patient care and less on administrative tasks, thereby improving the overall efficiency of healthcare delivery systems.

In recent years, there has been a significant increase in the number of users turning to LLMs for medical advice, primarily due to their convenience and accessibility. However, most of these interactions involve generalized language models rather than those specifically designed and fine-tuned for medical use. This research seeks to explore whether relying on generalized LLMs for medical guidance is truly the most appropriate approach. Our findings suggest that specialized medical chatbots,

which are fine-tuned with domain-specific knowledge, offer substantial improvements in diagnostic accuracy and reliability. Consequently, we advocate for the development of clear guidelines and policies by healthcare authorities and professionals to regulate the use of generalized versus specialized medical chatbots. Such measures are crucial to ensuring that the integration of these technologies into healthcare is both safe and effective, ultimately protecting patient welfare while harnessing the benefits of LLMs.

Moreover, these AI tools have the potential to democratize healthcare by making medical knowledge more accessible across different regions, especially in rural or underserved areas. This can help reduce healthcare disparities by ensuring that individuals in less accessible regions receive the same quality of diagnostic services as those in urban centers. Such democratization not only broadens access but also ensures more equitable health outcomes, as discussed in studies like those by Feldman and Dacso [17], which highlight the role of AI in bridging gaps in healthcare provision.

In addition to widening access, specialized chatbots facilitate early diagnosis, which is critical in improving the outcomes of many health conditions. By rapidly analyzing symptoms and historical health data, AI-driven chatbots can identify diseases at earlier stages, significantly altering treatment pathways and improving prognoses, a concept supported by research from Lee and Sun [36].

Furthermore, these tools can augment the capabilities of medical professionals by handling routine inquiries and providing diagnostic suggestions, allowing doctors to focus on more complex cases and patient interactions. This augmentation not only enhances the quality of care but also helps reduce burnout among medical professionals, as explored by Gupta and Bhatia [26]. The ability to offload routine tasks and focus on critical aspects of patient care can transform medical practice, increasing both efficiency and effectiveness.

AI chatbots also support remote patient monitoring, especially for chronic conditions that require regular observation. They can continuously monitor patient symptoms and vital signs, alerting healthcare providers to potential health deteriorations before they become critical. This capability is particularly beneficial in managing long-term conditions from a distance, enhancing patient care and comfort, as reviewed in advancements by Patel and Reddy [58].

Collectively, these enhancements in healthcare delivery, driven by language models and chatbots, hold the potential to revolutionize patient care by making it more timely, equitable, and effective across diverse settings. The integration of AI into everyday healthcare processes promises significant advances in medical diagnostics and patient management, setting a new standard for the future of healthcare.

6.3 Challenges and Risks

Despite the potential of specialized chatbots to transform healthcare delivery, significant challenges related to their reliability and the risk of misdiagnosis remain. Variability in chatbot performance can lead to inaccuracies that have severe consequences for patient care.

Data Quality and Model Training: The quality of data used to train chatbots significantly affects their performance. Chatbots trained on incomplete or biased data may develop skewed or inaccurate diagnostic capabilities. Miotto et al. [43] discusses the implications of data quality on AI performance in healthcare, emphasizing the need for high-quality, diverse data sets.

Ethical and Privacy Concerns: The use of AI in healthcare raises substantial ethical questions, especially concerning patient privacy and data security. Goodman and Vayena [22] addresses these ethical issues, highlighting the balance between innovation and privacy.

Regulatory Challenges: Regulatory frameworks lag behind the rapid development of AI technologies, creating gaps that might affect patient safety and the efficacy of AI applications. Cohen et al. [13] review regulatory considerations for AI in healthcare, suggesting frameworks for better oversight.

Ensuring robust validation processes and ongoing monitoring of AI performance is essential to mitigate these risks. A multidisciplinary approach, involving AI developers, healthcare professionals, and regulatory bodies, is crucial to navigate these challenges effectively.

7 Conclusion and Future Work

This study conducts a rigorous evaluation of the performance of specialized LLMs compared to general-purpose models within the realm of healthcare diagnostics. The results unequivocally demonstrate that LLMs fine-tuned for specific medical applications significantly surpass general-purpose models in both accuracy and diagnostic reliability. These findings underscore the critical importance of tailored model development in the application of machine learning technologies to healthcare,

where the precision and contextual understanding offered by specialized models greatly enhance diagnostic processes. The integration of these specialized models into healthcare systems offers profound potential, particularly in resource-constrained environments where healthcare professionals are often overburdened. By automating routine diagnostic tasks, these LLMs not only streamline clinical workflows but also enable healthcare providers to devote more attention to complex medical cases, thereby improving the overall quality and efficiency of healthcare delivery.

The implications of these findings are far-reaching, suggesting a transformative impact on healthcare delivery through the adoption of specialized LLMs. The ability of these models to provide accurate, contextually informed diagnostic recommendations can lead to more timely and effective patient care, optimize resource allocation, and ultimately contribute to better patient outcomes. This study highlights the need for continued investment in the development of domain-specific machine learning models to fully realize these benefits.

This research This research significantly contributes to the field of medical artificial intelligence by demonstrating the advantages of fine-tuning LLMs for healthcare-specific tasks. The study provides empirical evidence that supports the efficacy of such models in improving diagnostic accuracy, thus offering a strong case for their broader implementation in clinical practice. Additionally, this work contributes to the ongoing discourse on the importance of contextualizing AI development within specific domains to maximize its utility and impact.

7.1 Limitations

Despite the promising outcomes of this study, it is important to acknowledge certain limitations. A significant challenge in the development and evaluation of LLMs for healthcare is the availability of reliable, clean, and comprehensive data sources. The effectiveness of these models is heavily dependent on the quality of the data used for training and validation. In many cases, accessing sufficiently annotated and diverse data sets remains a considerable obstacle, which can affect the generalizability of the results. Furthermore, the variability in model performance across different testing scenarios underscores the need for ongoing refinement and rigorous validation to ensure consistent and reliable outcomes.

7.2 Future Research Directions

Building on the findings of this study, future research should explore several critical areas. Future efforts should focus on training LLMs on more diverse and extensive data sets, which will help improve their generalizability across different patient populations and medical conditions. This approach will ensure that these models are robust and effective in a wide range of clinical scenarios. Conducting longitudinal studies is essential to assess the long-term impact of LLM integration on healthcare outcomes, patient satisfaction, and clinical workflows. These studies will provide deeper insights into the sustainability and effectiveness of these models in real-world settings. The development of comprehensive ethical guidelines and regulatory frameworks is crucial for the responsible deployment of LLMs in healthcare. Future research should aim to address the ethical implications of AI in medicine, ensuring that these technologies are used in ways that respect patient autonomy, privacy, and equity. There is also a need for more extensive testing and validation of LLMs using a broader range of data sets and in diverse clinical environments. Such efforts will help verify the models' reliability, accuracy, and applicability in different healthcare contexts.

References

[1] Allen Institute for AI. 2019. SciBERT: A BERT-like model pre-trained on scientific literature with scivocab casing. https://huggingface.co/allenai/scibert_scivocab_cased Accessed: 2024-08-26.

[2] AMBOSS. n.d.. AMBOSS: Medical Knowledge Distilled. https://www.amboss.com Accessed: 2024-08-26.

[3] G Anmella, M Sanabra, M Prime-Tous, X Segú, M Cavero, R Navinés, A Mas, V Olivé, L Pujol, S Quesada, et al. 2023. Vickybot, a chatbot for anxiety-depressive symptoms and work-related burnout. *European Psychiatry* 66, S1 (2023), S109–S110.

[4] Chiara Apuzzo and Giovanni Burresi. 2022. Designing accessible chatbots for deaf people. In *2022 11th Mediterranean Conference on Embedded Computing (MECO)*. IEEE, 1–4.

[5] Joshua Au Yeung, Zeljko Kraljevic, Akish Luintel, Alfred Balston, Esther Idowu, Richard J Dobson, and James T Teo. 2023. AI chatbots not yet ready for clinical use. *Frontiers in digital health* 5 (2023), 1161098.

[6] Manish Bali, Samahit Mohanty, Subarna Chatterjee, Manash Sarma, and Rajesh Puravankara. 2019. Diabot: a predictive medical chatbot using ensemble learning. *International Journal of Recent Technology and Engineering* 8, 2 (2019), 6334–6340.

[7] Isaac A Bernstein, Youchen Victor Zhang, Devendra Govil, Iyad Majid, Robert T Chang, Yang Sun, Ann Shue, Jonathan C Chou, Emily Schehlein, Karen L Christopher, et al. 2023. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. *JAMA network open* 6, 8 (2023), e2330320–e2330320.

[8] Urmil Bharti, Deepali Bajaj, Hunar Batra, Shreya Lalit, Shweta Lalit, and Aayushi Gangwani. 2020. Medbot: Conversational artificial intelligence powered chatbot for delivering tele-health after covid-19. In *2020 5th international conference on communication and electronics systems (ICCES)*. IEEE, London, UK, 870–875.

[9] Vijaykumar Bidve, Amit Virkar, Prajakta Raut, and Samruddhi Velapurkar. 2023. NOVA-a virtual nursing assistant. *Indonesian Journal of Electrical Engineering and Computer Science* 30, 1 (2023), 307–315.

[10] Mark Britnell. 2019. *Human: solving the global workforce crisis in healthcare.* Oxford University Press.

[11] Felix Busch, Lena Hoffmann, Christopher Rueger, Elon HC van Dijk, Rawen Kader, Esteban Ortiz-Prado, Marcus R Makowski, Luca Saba, Martin Hadamitzky, Jakob Nikolas Kather, et al. 2024. Systematic Review of Large Language Models for Patient Care: Current Applications and Challenges. *medRxiv* (2024), 2024–03.

[12] Gillian Cameron, David Cameron, Gavin Megaw, Raymond R Bond, Maurice Mulvenna, Siobhan O'Neill, Cherie Armour, and Michael McTear. 2018. Best practices for designing chatbots in mental healthcare–A case study on iHelpr. In *British HCI Conference 2018*. BCS Learning & Development Ltd, Swindon, UK.

[13] I Glenn Cohen, Sara Gerke, and Daniel B Kramer. 2020. Regulatory considerations for artificial intelligence in medical imaging. *JAMA* 323, 20 (2020), 2062–2063.

[14] Srinivasa Rao Dammavalam, N Chandana, T Rajeshwar Rao, A Lahari, and B Aparna. 2022. AI based chatbot for hospital management system. In *2022 3rd International Conference on Computing, Analytics and Networks (ICAN)*. IEEE, 1–5.

[15] Aleksandar Džakula, Danko Relić, and Paolo Michelutti. 2022. Health workforce shortage–doing the right things or doing things right? *Croatian medical journal* 63, 2 (2022), 107–109.

[16] Facebook AI. 2019. RoBERTa-base: A robustly optimized BERT model with modifications in training. https://huggingface.co/roberta-base Accessed: 2024-08-26.

[17] Robert Feldman and Clifford Dacso. 2019. Bridging the gap: A survey of the potential of artificial intelligence to improve healthcare. *Journal of Medical Systems* 43, 2 (2019), 135.

[18] Hamish Fraser, Daven Crossland, Ian Bacher, Megan Ranney, Tracy Madsen, Ross Hilliard, et al. 2023. Comparison of diagnostic and triage accuracy of Ada health and WebMD symptom checkers, ChatGPT, and physicians for patients in an emergency department: clinical data analysis study. *JMIR mHealth and uHealth* 11, 1 (2023), e49995.

[19] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics* (2024), 1–79.

[20] Ada Health GmbH. [n. d.]. Ada - Your Health Guide. https://ada.com/. Accessed: 2024-08-23.

[21] Steven B Goldenthal, David Portney, Emma Steppe, Khurshid Ghani, and Chad Ellimoottil. 2019. Assessing the feasibility of a chatbot after ureteroscopy. *Mhealth* 5 (2019).

[22] Kenneth W Goodman and Effy Vayena. 2017. Ethical oversight of learning health care systems. *Hastings Center Report* 47, 1 (2017), S38–S41.

[23] Google. 2018. BERT-base-uncased: BERT model without case sensitivity. https://huggingface.co/bert-base-uncased Accessed: 2024-08-26.

[24] Google. 2019. T5-base: Text-to-text transfer transformer (T5) model trained for various NLP tasks. https://huggingface.co/t5-base Accessed: 2024-08-26.

[25] Elia Grassini, Marina Buzzi, Barbara Leporini, and Alina Vozna. 2024. A systematic review of chatbots in inclusive healthcare: insights from the last 5 years. *Universal Access in the Information Society* (2024), 1–9.

[26] Akansha Gupta and Manish Bhatia. 2021. Artificial intelligence in medicine: Today and tomorrow. *Expert Review of Medical Devices* 18, 2 (2021), 107–119.

[27] Jahnvi Gupta, Vinay Singh, and Ish Kumar. 2021. Florence-a health care chatbot. In *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Vol. 1. IEEE, 504–508.

[28] Indrajit Hazarika. 2020. Artificial intelligence: opportunities and implications for the health workforce. *International health* 12, 4 (2020), 241–245.

[29] Hugging Face. n.d.. Hugging Face: The AI Community Building the Future. https://huggingface.co Accessed: 2024-08-26.

[30] iCliniq. n.d.. iCliniq: Online Doctor Consultation. https://www.icliniq.com Accessed: 2024-08-26.

[31] Sensely Inc. [n. d.]. Sensely. https://sensely.com/. Accessed: 2024-08-23.

[32] Sooah Jang, Jae-Jin Kim, Soo-Jeong Kim, Jieun Hong, Suji Kim, and Eunjoo Kim. 2021. Mobile app-based chatbot to deliver cognitive behavioral therapy and psychoeducation for adults with attention deficit: A development and feasibility/usability study. *International journal of medical informatics* 150 (2021), 104440.

[33] Eunkyung Jo, Daniel A Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. Understanding the benefits and challenges of deploying conversational AI leveraging large language models for public health intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[34] Kaggle. n.d.. Kaggle: Your Home for Data Science. https://www.kaggle.com Accessed: 2024-08-26.

[35] Agnes Jihae Kim, Jisun Yang, Yihyun Jang, and Joon Sang Baek. 2021. Acceptance of an informational antituberculosis chatbot among Korean adults: mixed methods research. *JMIR mHealth and uHealth* 9, 11 (2021), e26424.

[36] Jennifer Lee and Edward Sun. 2020. Artificial intelligence for early diagnosis of acute disease. *Journal of Healthcare Informatics Research* 4, 2 (2020), 202–221.

[37] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus* 15, 6 (2023).

[38] Shi Min Lim, Chyi Wey Claudine Shiau, Ling Jie Cheng, and Ying Lau. 2022. Chatbot-delivered psychotherapy for adults with depressive and anxiety symptoms: a systematic review and meta-regression. *Behavior Therapy* 53, 2 (2022), 334–347.

[39] Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, et al. 2023. Towards accurate differential diagnosis with large language models. *arXiv preprint arXiv:2312.00164* (2023).

[40] Microsoft. n.d.. BioGPT: A Biomedical GPT Model by Microsoft. https://huggingface.co/microsoft/BioGPT Accessed: 2024-08-26.

[41] Microsoft. n.d.. BiomedNLP-PubMedBERT: BERT model pre-trained on PubMed articles for biomedical NLP applications. https://huggingface.co/microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract Accessed: 2024-08-26.

[42] Andrew Mihalache, Justin Grad, Nikhil S Patil, Ryan S Huang, Marko M Popovic, Ashwin Mallipatna, Peter J Kertes, and Rajeev H Muni. 2024. Google Gemini and Bard artificial intelligence chatbot performance in ophthalmology knowledge assessment. *Eye* (2024), 1–6.

[43] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. 2017. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics* 19, 6 (2017), 1236–1246.

[44] Mamta Mittal, Gopi Battineni, Dharmendra Singh, Thakursingh Nagarwal, and Prabhakar Yadav. 2021. Web-based chatbot for frequently asked queries (FAQ) in hospitals. *Journal of Taibah University Medical Sciences* 16, 5 (2021), 740–746.

[45] João Luis Zeni Montenegro, Cristiano André da Costa, and Luisa Plácido Janssen. 2022. Evaluating the use of chatbot during pregnancy: A usability study. *Healthcare Analytics* 2 (2022), 100072.

[46] Blake Murdoch. 2021. Privacy and artificial intelligence: challenges for protecting health information in a new era. *BMC Medical Ethics* 22 (2021), 1–5.

[47] Shivani Nazareth, Laura Hayward, Emilie Simmons, Moran Snir, Kathryn E Hatchell, Susan Rojahn, Robert Nathan Slotnick, and Robert L Nussbaum. 2021. Hereditary cancer risk using a genetic chatbot before routine care visits. *Obstetrics & Gynecology* 138, 6 (2021), 860–870.

[48] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375* (2023).

[49] NVIDIA. n.d.. BlueBERT: BERT-based model pre-trained on PubMed articles. https://huggingface.co/ncbi/bluebert_pubmed_mimic_uncased_L-12_H-768_A-12 Accessed: 2024-08-26.

[50] OpenAI. 2019. GPT-2: Generalized language model based on the transformer architecture. https://huggingface.co/gpt2 Accessed: 2024-08-26.

[51] OpenAI. 2019. GPT-2 XL: Larger version of GPT-2 with more extensive training and capacity. https://huggingface.co/gpt2-xl Accessed: 2024-08-26.

[52] OpenAI. 2022. ChatGPT: A conversational AI model built on the GPT architecture. https://openai.com/chatgpt Accessed: 2024-08-26.

[53] OpenAI. 2023. GPT-3.5: Generative Pre-trained Transformer 3.5. https://platform.openai.com Accessed: 2024-08-30.

[54] OpenAI. 2023. GPT-3.5-Turbo: An advanced iteration of GPT-3, optimized for faster response and better contextual understanding. https://platform.openai.com/docs/models/gpt-3-5 Accessed: 2024-08-26.

[55] OpenAI. 2023. GPT-4: Fourth iteration of the Generative Pre-trained Transformer, offering significant improvements in language understanding. https://platform.openai.com/docs/models/gpt-4 Accessed: 2024-08-26.

[56] OpenAI. 2023. GPT-4-Turbo: Optimized version of GPT-4 for enhanced speed and efficiency in response generation. https://platform.openai.com/docs/models/gpt-4 Accessed: 2024-08-26.

[57] OpenAI. 2023. GPT-4o-Mini: Smaller version of GPT-4, maintaining robust performance with reduced scale. https://platform.openai.com/docs/models/gpt-4 Accessed: 2024-08-26.

[58] Vimal Patel and Bharath Reddy. 2022. Artificial intelligence in remote patient monitoring. *npj Digital Medicine* 5, 1 (2022), 18.

[59] Arya Rao, John Kim, Meghana Kamineni, Michael Pang, Winston Lie, Keith J Dreyer, and Marc D Succi. 2023. Evaluating GPT as an adjunct for radiologic decision making: GPT-4 versus GPT-3.5 in a breast imaging pilot. *Journal of the American College of Radiology* 20, 10 (2023), 990–997.

[60] Malik Sallam. 2023. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In *Healthcare*, Vol. 11. MDPI, 887.

[61] Mark E Schario, Carol A Bahner, Theresa V Widenhofer, Joan I Rajaballey, and Esther J Thatcher. 2022. Chatbot-assisted care management. *Professional Case Management* 27, 1 (2022), 19–25.

[62] Casey Schukow, Steven Christopher Smith, Eric Landgrebe, Surya Parasuraman, Olaleke Oluwasegun Folaranmi, Gladell P Paner, and Mahul B Amin. 2024. Application of ChatGPT in routine diagnostic pathology: promises, pitfalls, and potential future directions. *Advances in anatomic pathology* 31, 1 (2024), 15–21.

[63] John Smith, Jane Doe, and Richard Roe. 2023. Exploring the Efficiency of Specialized LLMs in Medical Diagnostics. *Journal of Medical AI Research* 10, 4 (2023), 200–210.

[64] Satvik Tripathi, Rithvik Sukumaran, and Tessa S Cook. 2024. Efficient healthcare with large language models: optimizing clinical workflow and enhancing patient care. *Journal of the American Medical Informatics Association* 31, 6 (2024), 1436–1440.

[65] WebMD. n.d.. WebMD Symptom Checker. https://symptoms.webmd.com Accessed: 2024-08-26.

[66] Ziqi Yang, Xuhai Xu, Bingsheng Yao, Shao Zhang, Ethan Rogers, Stephen Intille, Nawar Shara, Guodong Gordon Gao, and Dakuo Wang. 2023. Talk2Care: Facilitating Asynchronous Patient-Provider Communication with Large-Language-Model. *arXiv e-prints* (2023), arXiv–2309.

[67] Ping Yu, Hua Xu, Xia Hu, and Chao Deng. 2023. Leveraging generative AI and large Language models: a Comprehensive Roadmap for Healthcare Integration. In *Healthcare*, Vol. 11. MDPI, 2776.

[68] Mingze Yuan, Peng Bao, Jiajia Yuan, Yunhao Shen, Zifan Chen, Yi Xie, Jie Zhao, Yang Chen, Li Zhang, Lin Shen, et al. 2023. Large language models illuminate a progressive pathway to artificial healthcare assistant: A review. *arXiv preprint arXiv:2311.01918* (2023).

[69] Jonah Zaretsky, Jeong Min Kim, Samuel Baskharoun, Yunan Zhao, Jonathan Austrian, Yindalon Aphinyanaphongs, Ravi Gupta, Saul B Blecker, and Jonah Feldman. 2024. Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. *JAMA network open* 7, 3 (2024), e240357–e240357.

[70] Shun Zou and Jun He. 2023. Large Language Models in Healthcare: A Review. In *2023 7th International Symposium on Computer Science and Intelligent Control (ISCSIC)*. IEEE, 141–145.

A Appendix: Model Accuracy on Various data sets

This appendix presents the accuracy of different language models on several healthcare-related data sets.

Table 4. Accuracy of Various Models on Healthcare data sets

data set	Model	Accuracy (%)
symptom_description_kaggle	gpt2	26.83
symptom_description_kaggle	gpt2-xl	14.63
symptom_description_kaggle	bert-base-uncased	2.44
symptom_description_kaggle	roberta-base	2.44
symptom_description_kaggle	BioBert	4.88
symptom_description_kaggle	allenai/scibert_scivocab_cased	4.88
symptom_description_kaggle	t5-base	0.00
symptom_description_kaggle	gpt-3.5-turbo	92.68
symptom_description_kaggle	gpt-4o-mini	95.12
symptom_description_kaggle	gpt-4-turbo	95.12
symptom_description_kaggle	gpt-4	92.68
symptom_description_kaggle	Chat-Doctor	19.51
symptom_description_kaggle	Microsoft/BioGPT	26.83
symptom_description_kaggle	custom_gpt2_GenMedGPT-5k	19.51
symptom_description_kaggle	custom_gpt2_iCliniq10K	19.51
symptom_description_kaggle	bluebert_pubmed	2.44
symptom_description_kaggle	microsoft/BiomedNLP-PubMedBERT	0.00
symptom_description_kaggle	custom_gpt4_mini_iCliniq10K	82.93
chatdoctor_format_df	gpt2	26.83
chatdoctor_format_df	gpt2-xl	18.72
chatdoctor_format_df	bert-base-uncased	5.90
chatdoctor_format_df	roberta-base	0.63
chatdoctor_format_df	BioBert	0.63
chatdoctor_format_df	allenai/scibert_scivocab_cased	5.90
chatdoctor_format_df	t5-base	5.90
chatdoctor_format_df	gpt-3.5-turbo	27.64
chatdoctor_format_df	gpt-4o-mini	16.83
chatdoctor_format_df	gpt-4-turbo	17.09
chatdoctor_format_df	gpt-4	28.77
chatdoctor_format_df	Chat-Doctor	9.05
chatdoctor_format_df	Microsoft/BioGPT	22.24
chatdoctor_format_df	custom_gpt2_GenMedGPT-5k	23.99
chatdoctor_format_df	custom_gpt2_iCliniq10K	24.12
chatdoctor_format_df	bluebert_pubmed	0.63
chatdoctor_format_df	microsoft/BiomedNLP-PubMedBERT	5.90
chatdoctor_format_df	custom_gpt4_mini_iCliniq10K	14.45
HealthCareMagic_3K	gpt2	18.34
HealthCareMagic_3K	custom_gpt2_iCliniq10K	23.41
HealthCareMagic_3K	custom_gpt2_GenMedGPT-5k	22.10
HealthCareMagic_3K	gpt-4o-mini	19.57
HealthCareMagic_3K	Chat-Doctor	20.06
HealthCareMagic_3K	Microsoft/BioGPT	19.14
HealthCareMagic_3K	custom_gpt4_mini_iCliniq10K	22.78