IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 December 2024, accepted 5 January 2025, date of publication 8 January 2025, date of current version 14 January 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3527025

SURVEY

Exploring the Impact of Large Language Models on Disease Diagnosis

IBRAHIM ALMUBARK
Department of Information Technology, College of Computer, Qassim University, Buraydah 52571, Saudi Arabia
e-mail: imbark@qu.edu.sa

This work was supported by the Deanship of Graduate Studies and Scientific Research at Qassim University under Grant QU-APC-2024.

ABSTRACT The emergence of large language models (LLMs) has revolutionized various fields, including education, finance, marketing, healthcare, and medicine. In this review, we aim to explore the application of LLMs in the healthcare sector, with a specific focus on disease diagnostics. The review highlighted the widespread use of LLMs, such as GPT-4, ChatGPT, GPT-3.5, and LLaMA, with GPT-4 being the most frequently used in disease diagnostics due to its diverse applications, improved accuracy, and efficiency. This review shows that LLMs have utilized a variety of medical data sources, including general medical databases, specialized documents, medical images, and genomic data. Moreover, the focus of these LLMs spans a broad spectrum of healthcare fields, addressing chronic conditions, respiratory diseases, cancer, and rare diseases. The performance evaluation of LLMs involves both qualitative and quantitative measures assessing their diagnostic accuracy. The findings highlight the evolving nature of LLMs in improving diagnostic accuracy.

INDEX TERMS Large language models, health care applications, ChatGPT, GPT-4, GPT-3.5, BARD, BERT, ChatGLM, LLaMA, rare diseases, PaLM.

I. INTRODUCTION

Large Language Models (LLMs), represent an important advancement in Artificial Intelligence (AI), designed to process and generate human-like text. These models mainly support transformer-based encoder-decoder frameworks, enabling them to excel in natural language tasks such as translation, summarization, and content generation [1], [2]. LLMs may have just stacks of only encoders or only decoders, using the mechanism for considering the importance of different words in a sentence. This self-attention is imperative for the capturing of context, meaning, and dependencies in the text. Figure 1 was created to demonstrate the general architecture of an encoder-decoder-based LLM.

This figure shows the architecture of a transformer model, containing both the encoder and decoder components. The input embedding layer gets the input, which is then passed to Positional Encoding. Positional Encoding maintains the original sequence of words in input sentences. Both encoder and decoder have multi-head attention units, which are responsible for conserving the contextual meaning of input vectors through normalization functions and feed-forward neural networks. After encoding, input embeddings are moved to the decoder, a linear layer, and a softmax function to generate output probabilities [3].

These models require substantial computing power often utilizing specialized hardware like Graphic Processing Unit (GPU) or Tensor Processing Unit (TPU). Despite their complexity, they can generalize to a wide range of natural language tasks such as translation and summarization. Pre-training involves training on a general corpus, followed by fine-tuning on a specific task, enhancing their versatility [4], [5]. They use numerous layers of neurons that capture diverse levels of abstraction ranging from simple syntactic structure to complex semantics. The output layer of the LLMs predicts the next word in a sentence or produces a response to a query. In the training phase, the weights of LLMs are adjusted through backpropagation to reduce the gap in the predicted and actual output [6].

In healthcare, LLMs are powerful AI systems, such as GPT-4; that have been trained on enormous volumes of textual data, covering patient records, clinical recommendations, medical literature, and more. These models are capable of producing, comprehending, and processing human language in ways that are highly relevant to the field of medicine [7], [8]. Their capacity to both produce and interpret medical language allows them to assist medical professionals in making well-informed decisions, resulting in

The associate editor coordinating the review of this manuscript and approving it for publication was Chun-Wei Tsai.

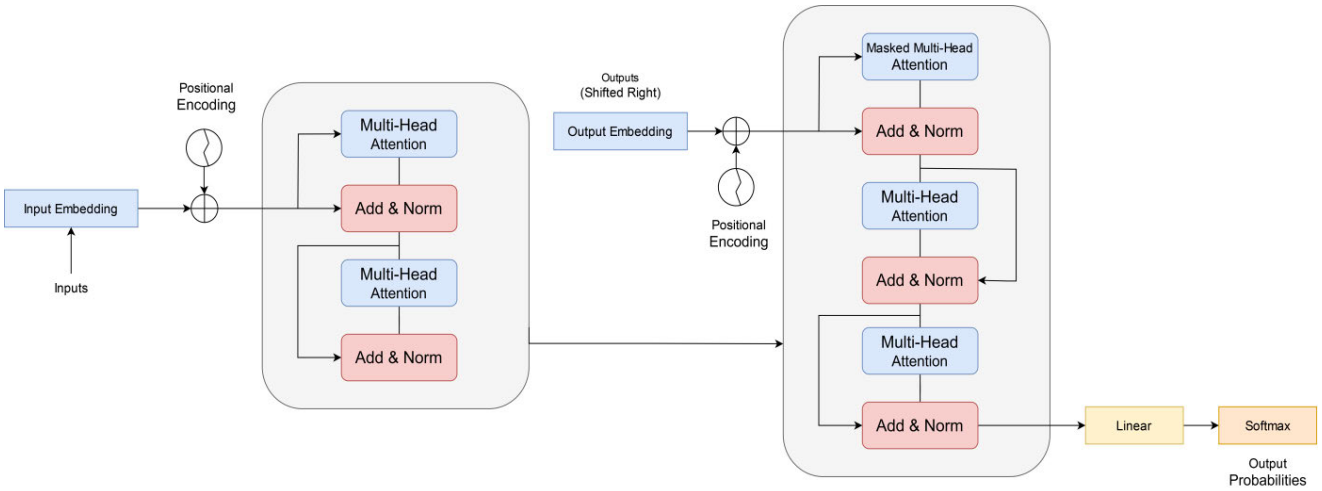

FIGURE 1. Architecture of transformer model.

improved patient outcomes and enhanced healthcare service efficiency [9].

LLMs have transformed the medical and healthcare field by providing modern solutions to natural language processing. For healthcare professionals, they are invaluable tools that can assist in examining patient records and clinical notes for disease diagnosis [9]. They can also contribute to advanced informed decision-making through the formulation of evidence-based recommendations that are based on the patient's relevant medical history [10], [11]. Further, through the understanding and interpretation of subtle descriptions of symptoms and a patient's medical history, they can propose differential diagnoses, offer a customized treatment plan, and suggest medicines based on clinical diagnostics [12]. The support of LLMs in various areas within the medical sector is shown by the authors in Figure 2.

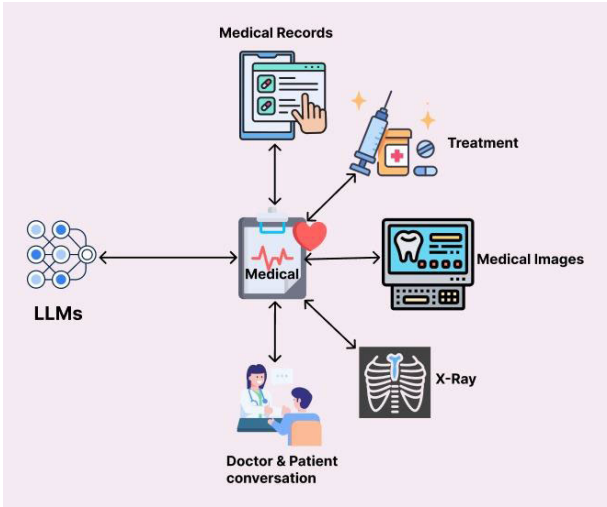

FIGURE 2. Support of LLMs in healthcare.

Moreover, they can assist in detecting rare diseases by identifying complex patterns that may escape the eye of the human clinician [13]. Based on their predictive analysis, LLMs can perform risk assessments to estimate the course of the diseases and the condition of the patients. Additionally, LLMs can summarize the advanced research of medical science, hence keeping clinicians updated with advancing treatment protocols [14]. The processes involved in disease diagnosis differs significantly between the traditional healthcare practices and LLMs. Traditional approaches depend on heavily on clinician expertise and manual analysis of patient records, while LLMs integrate advanced computational methods to automate and enhance decision-making [15]. The key steps involved in both traditional and LLM-based diagnostic approaches are shown by the authors in Figure 3.

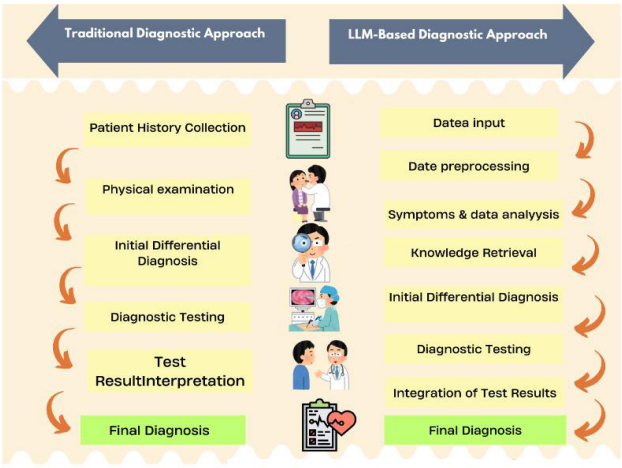

FIGURE 3. Traditional and LLM-based disease diagnostic process.

LLMs can diagnose a variety of diseases including cancer, cardiovascular disease, neurological conditions, and respiratory diseases; as well as dermatological, metabolic, psychiatric, and musculoskeletal disorders [16], [17]. The diagnostic support provided by LLMs for various diseases is shown by the authors in Figure 4.

Various literature reviews have been conducted on the implementation of LLMs in the field of healthcare.

FIGURE 4. Disease diagnostic support is provided by LLMs.

Hart et al., the researchers primarily focused on pathology informatics [18]. The authors explored the infrastructure and organizational changes required to support LLMs' implementation, along with the considerations for education, data management, and security. Other aspects of healthcare such as oncology, cardiology, or mental health were not covered in the review. Another systemic survey was conducted by Wang et al., [19], where researchers explored pre-trained language models in the biomedical field. The study focused upon pre-trained language models and their usage in performing various tasks involving natural language processing, such as text classification, information extraction, and question-answering. However, the study did not specifically categorize the fields of healthcare where LLMs were being employed for disease diagnostics. Moreover, it lacked a comprehensive assessment of performance measures explicitly focused upon disease diagnosis.

Similarly, the study by Yuan et al., [20] was conducted to explore the potential of LLMs in knowledge retrieval, research support, clinical workflow automation, and diagnostic assistance. This study did not extensively break down the types of medical data employed or their distribution across various studies. Moreover, the specific fields of health care and the explicit categorization of performance measures across targeted studies were missing. Likewise, the study by Nazi and Peng [21] explored the implementation of LLMs in the healthcare sector. The paper highlighted the transformative role of LLMs in improving clinical decision support, patient care, medical literature analysis, drug discovery, and virtual medical assistants. Their research included a general discussion and exploration of the broader applications of LLMs rather than listing specific models, data sources, or fields of healthcare targeted.

This current review aims to analyze the application of LLMs in the field of healthcare, specifically focusing on the disease diagnosis process. The primary studies have been collected from four renowned repositories including IEEE Xplore, ACM Digital Library, SpringerLink, and Science Direct. After a careful review, twenty studies have been selected for inclusion in this review, as these studies closely match our research domain and criteria. A summary of these studies regarding which digital repositories they were located in is included in Table 1.

TABLE 1. Summary of primary studies.

Repository	Final selection
IEEE Xplore	5
ACM Digital Library	2
SpringerLink	9
Science Direct	4

II. RESEARCH OBJECTIVES

The key objectives of this review are to:

1. Identify the specific LLMs that have been employed in the disease diagnosis process.
2. Investigate and classify the data sources employed by the LLMs for disease diagnosis.
3. Determine the specific fields of healthcare targeted by LLMs for disease diagnosis.
4. Examine the performance measures applied to assess the performance of LLMs in disease diagnosis.

To the best of our knowledge, this is the first review that explores the use of diverse LLMs in disease diagnostics and has a particular focus on elucidating specific fields of healthcare, particular data sources, and explicit performance measures.

The research questions addressed by this review, along with the motivation behind each question are presented in Table 2.

TABLE 2. Research questions and motivation.

Research question	Motivation
Which LLMs have been employed in the disease diagnostic process?	To identify advanced LLMs suited for medical diagnosis
What types of medical data are utilized by LLMs for disease diagnosis?	To assess the capability of models to handle diverse diagnostic scenarios based on various types of medical data
Which fields of healthcare are focused on by LLMs within disease diagnosis?	To identify the areas of healthcare within which LLMs are utilized
How is the performance of LLMs measured in disease diagnosis?	To measure the utility and effectiveness of LLM-based disease diagnosis

The main contributions of this study are given as:

- Identify the implementation of advance LLMs in the field of medical diagnosis.
- Specify the diverse applications of LLMs to deal with diagnostic scenarios.
- Measure the effectiveness of LLM-based disease diagnosis.

This review is organized as follows: Section V will explore the LLMs identified by the primary sources found in the

literature. Section VI will discuss the answers to the above questions along with the analysis of each research question. Section VI explores the issues and challenges associated with the application of LLMs in healthcare. Finally, section VII concludes the paper along with recommendations and areas of future work and research investigation.

III. PRELIMINARIES

The following section will introduce the LLMs identified within the current study as being employed within primary research.

GPT-4 is a highly advanced multimodal LLM with better computational power and human-like reasoning [22]. Users can send visual or textual queries to interact with GPT-4. This kind of LLM finds its applicability in almost every domain including healthcare, research, problem-solving, education, administration, and other industrial areas [23]. It is more reliable, and able to utilize training from real-world experience, compared with previous variants (GPT-2, GPT-3, GPT-3.5).

GPT-3.5 is a newly advanced large-scale model in the field of natural language processing, with human-understandable text generation abilities. Much more precise, faster, and more capable than the pre-existing models. It houses encyclopedic knowledge and can be applied in many fields including education, healthcare, dataset generation, text creation, and translation [24].

ChatGPT is an advanced form of natural language processing and an application of conversational AI. It responds by generating human-like responses. ChatGPT is popular in multiple platforms, for example chatbots, virtual assistants, problem-solving businesses, and customer service platforms [25]. This model is aligned with user intentions to provide in-context responses. It can understand the tones, styles, and requirements of its users.

GPT-2 is an LLM that was generated to produce human-like text using the potential of natural language processing with unsupervised learning. Though small in size, it is more capable of generating valuable text from prompts. This model finds most of its applications in the field of education, where it can generate examination questions and other study materials. It can also act as a virtual assistant and automation in customer services [26].

BARD (Biomedical Artificial Intelligence Research and Development) is a language model developed within medical and health-oriented niches. The model is trained utilizing large-scale medical data to respond accordingly to the queries raised [17]. In specific terms, it tends to benefit researchers, medical students, and healthcare assistants whenever they require verifiable and valid information. This information can then further be used for the diagnosis and treatment of disease.

PaLM (Pathway Language Model) is an advanced language model of a general nature built for multiple tasks. PaLM has been trained for handling complex and technical terms and is, hence, not domain-specific. Thus, the present model finds wide applicability in diverse fields including law, industry, education, science and technology, and other businesses [16]. It can help professionals in, analyzing, drafting, revisiting, and generating text documents of various kinds.

LLaMA (Large Language Model for Multilingual Applications) has been designed for worldwide applications that span different languages and cultures. It is used in multilingual translation and content generation. It is very important for international organizations as well as in education and research [27].

BERT (Bidirectional Encoder Representations from Transformers), was considered a revolution in natural language processing with the introduction of bidirectional training [15]. It allows the model to understand the full context of a word within a sentence and how it relates to all other words. BERT is widely used in text classification, sentiment analysis, and named entity recognition. It is a powerful tool for voice assistants and search engines.

ChatGLM is a specialized generative language model that dynamically makes the conversation interesting and engaging by providing responses contextually linked to inputs. Such a model applies to a chatbot, virtual assistant, or social media interaction. This model is also applied in medical diagnosis and education, answering questions and providing useful guidelines [28].

GPT-Neo is an open-source LLM similar in capabilities, both in understanding and generating responses, to other LLMs. It has gained lots of fame in research, development, and practical applications in the field of natural language processing as well as a variety of industries. It is developed for study and includes adjustments by the developer to improvise advancements and reduce complications, famous in the field of research, development, and several practical applications including chatbots and virtual assistants [29].

The following section will delve into a discussion focused upon answering the research questions detailed above.

IV. DISCUSSION OF RESEARCH QUESTIONS

A. RQ1: WHICH LLMS HAVE BEEN EMPLOYED IN MEDICAL DIAGNOSTIC APPLICATIONS?

The emergence of LLMs in healthcare is increasing the adaptation of LLM-based diagnosis. These models can play a significant role in clinical decision support, medical record analysis, patient interaction, and medical knowledge synthesis.

L. Caruccio et al., conducted a comparative study between traditional predictive models and LLMs to highlight the significant role of LLMs in medical diagnosis [25]. They introduced an intelligent diagnosis by implementing multiple advanced LLMs namely ChatGPT, Google BARD, and GPT-Neo. The results indicated that ChatGPT-based models, specifically text-davinci-003 and GPT-3.5-turbo-0301, performed better in disease diagnosis. However, Google BARD outperformed ChatGPT and GPT-Neo in some scenarios, particularly those with high variability in handling symptoms.

Z. Wang et al., proposed a novel framework named Radiology Report Generation with Frozen LLMs (R2GenGPT) by implementing the LLaMA2-7B LLM [27]. The authors integrated visual encoder, visual mapper, and Llama2-7B to translate visual features from medical images into coherent textual reports. The framework mainly addressed the challenges associated with medical report generation using three Llama2-7B-based feature visualization methods (shallow alignment, deep alignment, and delta alignment). The authors concluded that the deep alignment variant outperformed state-of-the-art methods to align visual features with LLM.

C. Liu et al., explored the potential of Artificial General Intelligence models (AGI), LLMs, and Large Vision Models (LVMs) in the field of Radiation Oncology [30]. The study particularly implemented LLMs, namely GPT-4 and PaLM2, and LVMs namely Segment Anything Model (SAM), to examine diverse aspects of radiation therapy. The applications of these models were assessed across multiple stages, including initial consultation, simulation, treatment planning, delivery, verification, and follow-up. The study concluded that GPT-4 exhibited remarkable performance compared with other models, specifically for the tasks that demand interpretation and standardization of complex medical data. A novel tool named MoŒil was developed to educate patients in ophthalmology [31]. The researchers utilized the exceptional capabilities of GPT-4 to provide a platform that supported accessible, precise, and comprehensive information about ophthalmology-related topics. GPT4 was further fine-tuned to focus on ophthalmology-related material, educating patients without diagnoses or treatment recommendations. The experiments indicated that MonŒil was a highly effective tool to educate patients in ophthalmology, specifically those with Advanced Muscular Degeneration (AMD).

H. Zhang et al., introduced a novel LLMs to enhance the accuracy of automated medical diagnostics [32]. They integrated Markov Logic Networks (MLNs) with external knowledge extracted using multiple LLMs - including ChatGPT-3.5-turbo for summarizing disease knowledge, GPT-4 to formalize disease knowledge, and text-embedding-ada-002 for document integrations. The proposed approach comprised three stages: knowledge acquisition, knowledge formalization, and iterative optimization. LLMs were employed in combination with a search engine which provided structured external medical knowledge. This knowledge was further interpreted into first-order logic rules which were passed to the MLN-based diagnostic system that produced final predictions. The statistical results indicated that the proposed approach outperformed several baseline methods in terms of enhanced accuracy and interpretability. A further study developed and utilized a novel advanced Java-based Android application to support medical diagnosis using GPT-3.5 [33]. The application required disease symptoms through input and provided diagnosis and advice through a user-friendly interface. The study concluded that the application offered users satisfactory results with significant accuracy and informed healthcare decisions.

D. P. Panagoulias et al., proposed a rule-augmented-based patient-doctor communication system using ChatGPT [34]. They streamlined diagnostic procedures using various external machine learning and analytical Application Programming Interfaces (APIs) to offer diagnostic suggestions. The system aimed to improve healthcare diagnosis and reduce costs by leveraging the enhanced capabilities of ChatGPT. The proposed approach was demonstrated through various cases which revealed that ChatGPT-based patient-doctor system could perform precise disease diagnosis with further diagnostic exams. A. E. Saddik et al., integrated ChatGPT with Metaverse to provide enhanced medical consultancy [35]. They proposed a model named ChatGPT-Metaverse-Medical (CMM) that combined the metaverse environment, ChatGPT, and the healthcare sector to design a novel approach for digital medical consultancy. The proposed model could visualize organ anatomy, examine body morphology, and offer remote surgery. CMM also supported patient privacy and data security while providing healthcare services at economical prices. The study concluded that the proposed ChatGPT-based model could deliver real-time consultancy with efficient medical advice. J. Kim et al., examined the use of both commercial and non-commercial LLMs to support doctors in the medical field [36]. The commercial LLMs included BARD and a series of GPT-3.5 series comprising text-davinci-003, GPT-3.5-turbo, and davinci-002. While non-commercial LLMs included two fine-tuned versions of LLaMA, named Alpaca-7B and Alpaca-7B LoRA. To assess the effectiveness of the models, the authors employed a list of synonyms. The study concluded that the disease prediction by these models was correct if it was the synonym of the disease belonging to broad categories. The experimental results showed that GPT-3.5-turbo, achieved the highest accuracy among all the models while other models struggled to produce accurate diagnoses.

D. P. Panagoulias et al., integrated a rule-based decision approach, external APIs, and GPT-4 for improved medical diagnosis [37]. They also implemented natural language processing-based algorithms to extract domain-specific knowledge. The study focused on user interactions with LLM-based systems to provide precise medical advice. The system was evaluated using pathology-based multiple-choice questions. The statistical analysis showed that the approach achieved remarkable accuracy and offered reliable diagnostic services. An open-source LLM, named ChatGLM-6B, was employed for fine-tuning medical applications [28].The study proposed an innovative framework named MOELoRA (Mixture-of-Experts and Low-Rank Adaptation) which could handle various medical tasks. These tasks included named entity recognition, diagnosis prediction, clinical report generation, and doctor recommendation. The study conducted a comparative analysis of MOELoRA with other baseline methods named LoRAHub and MoLoRA.

The results revealed that MOELoRA outperformed these other methods. A Chinese medication system named ShennongMGS was presented by Y. Dou et al., [38]. They employed a Chinese language expert LLM, called ChatGLM-6B LLM, to design ShennongMGS. It was pre-trained on preprocessed data to build the medical knowledge base. This knowledge base was continuously updated using a web crawler to ensure the latest medication guidance. ShennongMGS was fine-tuned using doctor-patient dialogues and various medical cases to produce efficient results. The study concluded that ShennongMGS could offer rational advice based on user communication.

A. Ríos-Hoyo *et el.,* evaluated the effectiveness of two LLMs, namely GPT-3.5 and GPT-4, to diagnose complex medical cases [39]. These cases were published between 2022 and 2023 and were selected based on their exclusion from training data. The models were examined using three distinct prompts. The experiments demonstrated that GPT-3.5 produced more diagnoses with less accuracy, while GPT-4 produced fewer diagnoses with precise results with persistent accuracy. The study concluded that GPT-4 outperformed GPT-3.5 in precisely diagnosing intricate medical conditions. A novel system was designed and named SkinGPT-4 with the aim of enhancing dermatological diagnosis using LLaMA-2-13b-chat as a LLM [40].The system was pre-trained using a vision transformer for image encoding. SkinGPT-4 was fine-tuned using comprehensive clinical notes and doctors' remarks to improve the system's diagnostic abilities. The system allowed users to upload images of skin diseases, with the result that it could identify skin conditions along with the medical recommendations. The assessment ensured that SkinGPT-4 offered a reliable diagnosis in comparison with traditional dermatologists. S. Zhang et al., designed a question-answer-based system, named Chat Ella, using the GPT-2 model [41]. Chat Ella offered a user-friendly interface to provide patients with an interactive environment. The users could provide disease symptoms through a conversational interface and Chat Ella was able to produce remote and efficient healthcare consultancy. The front-end system was developed using React while the back-end system was designed through the flask framework. The study revealed that Chat Ella was an efficient tool for diagnosing chronic diseases with accessible remote services. C. Shyr et al., focused on analyzing and phenotyping rare diseases using two advanced LLMs (ChatGPT and BioClinicalBERT) [42]. They extracted disease phenotypes hidden in unstructured text, which is a critical part of rare disease treatment. The study employed two approaches, 1) training the system using ChatGPT and 2) fine-tuning it using a BERT variant LLM named BioClinicalBERT. The assessment showed that BioClinicalBERT outperformed ChatGPT in extracting phenotypes.

X. Hu *et al.,* analyzed the diagnostic capabilities of GPT-4 to identify rare eye diseases [43]. The researchers focused on three user groups patients, family physicians, and junior ophthalmologists. GPT-4 was provided with diverse inputs including only main complaints for patients, main complaints, and disease history for family. The input for junior ophthalmologists comprised main complaints, disease history, and descriptions of ophthalmic assessments. The analysis found that GPT-4 could effectively diagnose eye disease based on the detailed disease description specifically for junior ophthalmologist cases. S. Rau et al., examined the diagnostic accuracy of a vision-enabled LLM named ChatGPT-4V [44]. The study focused on three diagnostic categories–chest CT scans for COVID-19, non-small cell lung cancer, and control cases. The study evaluated ChatGPT-4V using sixty CT scans extracted from a cancer imaging archive. The results suggested that ChatGPT-4V could extract useful insights from radiographic features.

S. Bushuven et al., assessed the diagnostic capabilities of ChatGPT and GPT-4 to deal with pediatric emergencies and acute medical emergency scenarios [45]. An analytical study was conducted using Cross-sectional investigative evaluation. The content was validated by five emergency physicians who examined multiple diseases including airway obstructions, anaphylaxis, asthma, bronchiolitis, pneumonia, shock types, and cardiac problems. The investigation revealed that both models effectively diagnosed most of the diseases, except septic shock and pulmonary embolism. A further study performed a comparative analysis between a conversational LLM, ChatGPT, and a traditional diagnosis tool, named Isabel pro differential diagnosis generator, to diagnose ophthalmic diseases [46]. Both models were provided with detailed ophthalmic cases having multiple ophthalmic conditions. Each case was fed into both models where ChatGPT was inquired about the most likely diagnosis and differential diagnosis. Isabel was analyzed based on the list and free text comprising disease symptoms. The findings indicated that ChatGPT was more efficient at diagnosing ophthalmic conditions.

A graphical representation of the frequency of LLMs found in the primary research is presented in Figure 5. The summary of primary studies along with the LLM employed by each study is presented in Table 3.

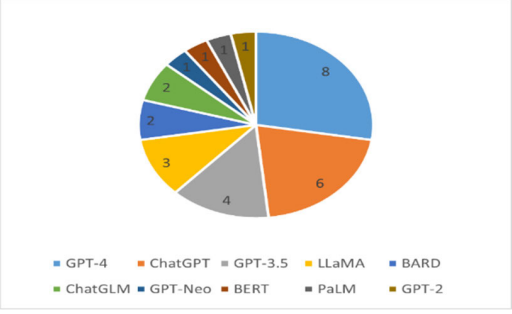

FIGURE 5. Frequency of LLMs employed by primary studies.

Analysis: This literature review found a variety of LLMs being employed in the field of healthcare for disease diagnostic purposes. GPT-4 was found to be the most popular LLM, highlighting its exceptional capabilities in medical

TABLE 3. Summary of primary studies.

Reference	Publication Year	LLM Used	Type of Study
3	2024	ChatGPT, Google Bard, GPT-Neo	Journal
7	2023	Llama2-7B	Journal
8	2023	GPT-4, PaLM 2	Journal
9	2024	GPT-4	Journal
12	2023	ChatGPT-3.5-turbo, text-embedding-ada-002, GPT-4	Conference
13	2023	ChatGPT	Conference
14	2024	ChatGPT	Journal
15	2023	Bard, GPT-3.5, LLaMA	Conference
16	2024	GPT-4, PaLM 2	Journal
18	2024	ChatGLM-6B	Conference
19	2024	ChatGLM-6B	Journal
20	2024	GPT-3.5, GPT-4	Journal
21	2024	Llama2-13b-chat	Journal
22	2024	GPT-2	Journal
23	2024	ChatGPT, BioClinicalBERT	Journal
27	2023	GPT-4	Journal
24	2024	ChatGPT-4V	Journal
25	2023	ChatGPT, GPT-4	Journal
26	2023	ChatGPT	Journal
11	2023	GPT-3.5	Conference

diagnosis. Other LLMs, including ChatGPT, GPT-3.5, and LLaMA, are also widely adopted by researchers indicating their key role in diagnosing diverse diseases. Additional LLMs, such as BARD, ChatGLM, GOT-Neo, and PaLM, are employed less commonly. This, perhaps, suggests a shift towards utilization of newer LLMs within the disease diagnostic process. The implementation of GPT-2 reveals that there is still research going ongoing in healthcare using legacy models. These experiments with legacy models provide a baseline for state-of-the-art models showing the evolution of diagnostic capabilities within LLMs over time. Some studies also employed a combination of LLMs, providing a comparative analysis of disease diagnostic abilities of multiple LLMs and offering insights into their strengths and weaknesses. This diverse implementation of LLMs highlights the evolving nature of the field of healthcare, alongside continued efforts to achieve improved accuracy for medical diagnosis.

B. RQ2: WHAT TYPES OF MEDICAL DATA ARE USED TO TRAIN LLMs FOR DIAGNOSTIC PURPOSES?

Training of LLMs requires vast and diverse amounts of medical data, this is most commonly obtained from electronic healthcare records, clinical notes, medical images, and genomic data. LLM-based training data considers data quality, data privacy, data biases, and data annotation. The primary medical data employed by each primary study in this review is mentioned below.

Caruccio., employed two datasets, namely the disease prediction dataset and the medical diagnosis dialogue dataset [25]. The first dataset comprised 132 symptoms and 4,663 symptom combinations, while the second dataset was designed using real-world patient records covering explicit and implicit symptoms related to twelve types of diseases and 118 symptoms. Wang et al., utilized two datasets, IU-Xray and MIMIC-CXR, to assess the performance of R2GenGPT [27]. IU-Xray included 3,955 de-identified radiology reports from the Indiana University Chest X-ray Collection. MIMIC-CXR is the biggest publicly available dataset containing chest X-ray images and respective reports obtained from patients inspected at the Beth Israel Medical Center.

Liu et al., employed multiple datasets, including Medical Information Mart for Intensive Care (MIMIC-III), Amsterdam Open MRI Collection (AOMIC), The Cancer Imaging Archive (TCIA), Cytopathological data (SIPaKMeD), as well as genomic data from the national center for biotechnology information [30]. Xompero et al., performed experiments using patient interactions with the proposed MonŒil system [31]. A. S. et al., employed medical knowledge and diagnostic data to improve knowledge based on medical terminology and symptom descriptions [33]. Zhang et al., combined real-world and synthetic datasets split into training and testing data [32]. The datasets included Muzhi, DXY, and synthetic datasets containing separate cases for training and testing. The researchers employed datasets from blood exams to analyze the diagnostic abilities of ChatGPT [34]. They analyzed patients' blood variables and further processed them using machine learning models to classify metabolic syndrome and patients' weight groups.

Saddik and Ghaboura examined the diagnostic ability of ChatGPT by assessing its performance in the US Medical Licensing Examination (USMLE) [35]. The study also analyzed the responses generated by ChatGPT with medical disease diagnosis. Kim et al., employed PolyMed to train a test dataset that comprised symptoms, diseases, age, gender, departments, and family history [36]. This dataset included patient-doctor conversations obtained from a healthcare platform. GPT-4 was fused with domain-specific rules by Panagoulias et al., [37] for medical diagnosis professionals where it proved its effectiveness in accurate. The performance of the LLM was evaluated using domain-specific knowledge and ground-truth comparisons. The Prompt CBLUE Chinese dataset was employed for training and testing in the work of Liu et al., [28] which comprised eighteen distinct tasks. This dataset included name entity recognition, medical text classification, medical report generation, diagnostic word normalization, as well as several other tasks.

The authors of a 2024 study employed a combination of distilled and real-world data [38]. Distilled datasets include food and drug data, PubMed, DrugBank¸ Drugs.com, UpToDate, PubMedQA, ChatMed, and Med-ChatGLM data. Whereas, real-world data contains patient-doctor conversations and real-world knowledge of questions and

answers. Massachusetts General Hospital Case Records were employed by Ríos-Hoyo et al., to evaluate GPT-3.5 and GPT-4 [39]. The included records were medical cases reported between 2022-2023. Zhou et al., employed skin disease images to train LLM [40]. The included dataset had both public and proprietary dermatological data. The proposed model, SkinGPT-4, was tested using real-life cases examined by board-certified dermatologists. Publicly available raw data was acquired for training and testing from Kaggle [41]. The dataset was pre-processed to organize chronic disease symptoms. Researchers extracted rare disease data from the RareDis corpus for training and testing [42]. The dataset had descriptions of rare diseases in textual form.

In another study, differential diagnosis was performed using training data from radiographic documents [44]. While self-created gastrointestinal pathological cases were used to test the diagnostic approach. Basic life support and pediatric advanced life support cases were utilized to examine the diagnostic capabilities of ChatGPT and GPT-4 [45]. The study by Balas and Ing randomly selected cases from the EyeRounds service provided by the University of Iowa's department of ophthalmology and visual science [46]. Hu *et al.,* used ophthalmic case descriptions obtained from the EyeRounds service [43]. These case descriptions were provided by the University of Iowa's Department of Ophthalmology and Visual Sciences.

The results of the current investigation show that data sources utilized by primary studies can be classified into specific categories-as shown in Table 4. The distribution of data sources across various categories is graphically presented in Figure 6.

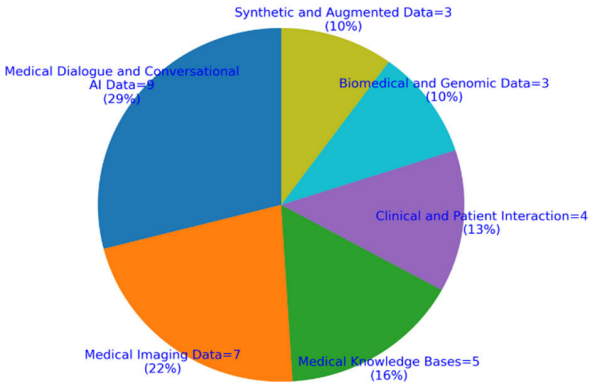

FIGURE 6. Distribution of data sources across various categories.

Analysis: This review of primary studies revealed that researchers have employed data from diverse sources to perform LLM-based disease diagnosis. The included primary sources were found across general medical knowledge databases i.e., PubMed, DrugBank, and Drugs.com to specialized sources, such as radiographic documents and gastrointestinal pathological cases. The nature of data used varies across the included studies, comprising medical images, doctor-patient conversations, blood samples, and genomic data. Moreover, researchers have also employed synthetic and real-world datasets, including Muzhi and DXY, which allow the LLMs to learn from both genuine clinical cases and controlled, hypothetical, scenarios. These insights are helpful to both researchers and practitioners wishing to integrate artificial intelligence into the medical field. The detailed breakdown of data sources used in each primary study highlights the importance of varied data sources. The researchers are encouraged to combine both structured and unstructured medical data to train the LLMs for efficient

TABLE 4. Classification of data sources across diverse categories.

Main Category	Sub-category
Medical Imaging Data	IU-Xray (https://openi.nlm.nih.gov/)
	MIMIC-CXR (https://openi.nlm.nih.gov/)
	TCIA (https://pubmed.ncbi.nlm.nih.gov/21115458/)
	SIPaKMeD (https://paperswithcode.com/dataset/sipakmed)
	EyeRounds service data (EYEROUNDS.ORG)
	Radiographic documents (https://pubs.rsna.org/action/doSearch?AllField=dataset&SeriesKey=radiographics)
	Self-created gastrointestinal pathological cases
Clinical and Patient Interaction Data	MIMIC-III (https://physionet.org/content/mimiciii/1.4/)
	AOMIC (https://pubmed.ncbi.nlm.nih.gov/21115458)
	Patient interactions
	Kaggle healthcare data (https://www.kaggle.com/datasets/shubhamgoel27/dermnet)
Medical Dialogue and Conversational AI Data	Disease prediction dataset (https://muzhi.baidu.com/#/)
	Medical diagnosis dialogue dataset (http://www.sdspeople.fudan.edu.cn/zywei/data/acl2018-mds.zip)
	US medical licensing exams (https://arxiv.org/abs/2304.08247)
	PolyMed (https://zenodo.org/records/7866103)
	PromptCBLUE Chinese dataset (https://tianchi.aliyun.com/competition/entrance/532084/information)
	PubMedQA (https://github.com/vgupta123/sumpubmed)
	ChatMed (https://github.com/michael-wzhu/ChatMed)
	Med-ChatGLM data (https://github.com/SCIR-HI/Huatuo-Llama-Med-Chinese)
	Real-world medical conversation
Biomedical and Genomic Data	Genomic data (https://pubmed.ncbi.nlm.nih.gov/21115458/)
	RareDis corpus data (https://github.com/isegura/NLP4RARE-CM-UC3M)
Medical Knowledge Bases	PubMed (https://github.com/vgupta123/sumpubmed)
	DrugBank (https://go.drugbank.com/)
	Drugs.com (https://www.emerald.com/insight/content/doi/10.1108/09504120410552697/full/html)
	UpToDate (https://www.mdedge.com/familymedicine/article/65481/practice-management/uptodate-comprehensive-clinical-database?sso=true)
	Massachusetts General Hospital Case Records (https://www.nejm.org/doi/full/10.1056/NEJMe030079)
Synthetic and Augmented Data	Muzhi (https://muzhi.baidu.com/#/)
	DXY (https://dxy.com/)
	Synthetic datasets (http://www.symcat.com/)

disease predictions. They are also provided with future guidance to explore the latest data sources and generate synthetic data sources.

Moreover, practitioners get insight into the inclusion of diverse datasets from various medical sources, cultural backgrounds, and real-world interactions. Thus, giving them confidence that LLM-based disease prediction systems are better prepared to support informed decision-making in the field of healthcare.

C. RQ3: WHICH SPECIFIC FIELD OF HEALTHCARE IS FOCUSED ON LLMs-BASED DISEASE DIAGNOSTIC APPROACHES?

LLMs can be trained on a vast variety of medical healthcare data. They can diagnose a wide range of diseases efficiently, saving both time and money. The primary studies found within the current review found LLM utilization across various healthcare fields. These are highlighted below.

The study by L. Caruccio et al., focused on various low- and medium-risk diseases for diagnosis by LLMs [25]. The low-risk diseases included jaundice, hepatitis, fungal infection, and dermatitis, but were not limited to this list. The medium-risk diseases included asthma, coronary heart disease, pneumonia, thyroiditis, and traumatic brain injury, among others. The researchers targeted thoracic diseases to be diagnosed by LLMs [27]. The dataset employed by the study comprised distinct categories related to thoracic diseases and support devices. A further study performed cancer-based experiments with LLMs to analyze various cases in radiation oncology [30]. The researchers specifically focused on neck cancer, hepatocellular carcinoma, prostate cancer, and pulmonary neoplasm. Macular degeneration was diagnosed by LLMs in the work of Xompero et al., [31]. The researchers designed an advanced system, named MonŒil), which was customized for ophthalmology-related queries.

The study by A. S et al., focused on diagnosing various chronic diseases through the use of GPT-3.5 [33]. The researchers particularly analyzed the response of LLM for promoting healthy lifestyles, for example, smoking cessation and medication adherence. Zhang et al., were primarily focused on diagnosing three diseases: pulmonary neoplasm, tuberculosis, and influenza through utilizing GPT-3.5 [32]. The study by Panagoulias et al., was dedicated to diagnosing multiple diseases, including alcoholic liver disease, metabolic syndrome, gout, and hyperlipidemia [34]. Alcoholic liver disease is characterized by liver damage due to chronic heavy alcohol consumption. Similarly, metabolic syndrome increases the risk of heart disease, stroke, and type 2 diabetes. Gout is a type of arthritis caused by uric acid crystals; while hyperlipidemia is a major risk for developing cardiovascular disease and is caused by elevated cholesterol levels.

The study by Kim et al., focused on diagnosing multiple diseases, these included respiratory infections, allergic diseases, along with chronic and acute conditions [36]. These diseases were diagnosed based on patient medical records, including symptoms, age, sex, family history, and underlying diseases?. Various medical cases were addressed by the researchers in a 2024 study [35]. They aimed at providing mental health support, rehabilitation, and gynecological consultations using ChatGPT along with the Metaverse. The study by Panagoulias et al., targeted various medical cases for effective diagnosis using GPT-4 [37]. These cases belonged to cardiology, neurology, psychiatry and psychology, dermatology, endocrinology, and general pathology. Liu et al., focused on various healthcare tasks to be performed by ChatGLM [28]. These tasks included doctor recommendation, diagnosis prediction, medicine recommendation, medical named entity recognition, and clinical report generation.

In the study by Dou et al., adverse drug reactions were predicted using ChatGLM [38]. The LLM also provided personalized guidance, and treatment plans utilizing multiple drugs. The researchers diagnosed various complex medical diseases across various groups of patients [39]. The first group included neurology and psychiatry, while the second group included oncology and hematology. The third group comprised infectious diseases, internal medicine, endocrinology, and toxicology. The fourth group involved rheumatology, allergy, and autoimmune diseases. The fifth group belonged to the category of 'others' where multiple diseases were included such as cardiology, gastroenterology, genetic diseases, dermatology, nephrology, and pediatrics. The study by Zhou et al., was focused on dermatological conditions [40]. These conditions included skin cancer, onychomycosis, alopecia areata, mpox virus infection, actinic keratosis, and eczema.

The work of Zhang et al., was devoted to diagnosing 24 distinct diseases using GPT-2 [41]. These diseases were grouped around similar diseases; including cardiovascular diseases, respiratory diseases, metabolic and endocrine disorders, neurological and psychiatric disorders, musculoskeletal disorders, gastrointestinal disorders, and chronic organ diseases. A further study by Shyr et al., focused on diagnosing rare diseases using ChatGPT and BioClinicalBERT [42]. The rare diseases included neurofibromatosis type I (also called Von Recklinghausen's disease) and primary antiphospholipid syndrome. Rau et al., diagnosed a variety of abdominal pathologies with the study investigating various gastrointestinal cases including malignancies, inflammatory disorders, obstructive disorders, benign neoplasms, and vascular pathologies [44]. Basic and advanced life support cases were evaluated in the study by Bushuven et al., [45]. The study focused on diagnosing respiratory conditions, cardiovascular and shock conditions, neurological and muscular conditions, and toxicological conditions.

Researchers focused on diagnosing diseases in a study published in 2023 [46]. Multiple ophthalmic diseases were diagnosed within the study, including optic neuritis, adenoviral conjunctivitis, orbital cellulitis, tuberculosis uveitis, and corneal and external eye conditions. Rare eye diseases were diagnosed by LLMs in the work of Hu et al., [43]. The study specifically focused on inflammatory and autoimmune

disorders, genetic and congenital disorders, and oncological disorders.

The diseases diagnosed by LLMs in the primary literature can be classified into diverse categories - shown in Table 5. The graphical distribution of the diseases against each category is shown in Figure 7.

TABLE 5. Disease classification into distinct categories.

Primary category	Sub-category
Chronic conditions	Chronic diseases
	Cardiovascular conditions
	Abdominal pathologies
	Endocrine disorders
	Alcoholic liver disease
	Metabolic syndrome,
	Gout
	Hyperlipidemia
Respiratory-pulmonary diseases	Thoracic diseases
	Tuberculosis
	Influenza
	Respiratory infections
	Allergic diseases
	Respiratory conditions
Cance-tumors	Neck cancer
	Hepatocellular carcinoma
	Prostate cancer
	Pulmonary neoplasm
Neuropsychiatric disorders	Neurological disorders
	Psychiatric disorder
	Muscular conditions
Dermatological diseases	Eczema
	Alopecia areata
	Skin cancer
Ophthalmic conditions	Macular degeneration
	Autoimmune eye disorders
	Infectious eye conditions
	Oncological eye disorders

Analysis: The analysis of primary studies with a focus on specific diseases diagnosed by LLMs shows that researchers are working on a broad spectrum of diseases. The prime attention is drawn towards chronic conditions, respiratory and pulmonary diseases, and cancer. These disease groups constitute significant public health challenges, so it is, perhaps, unsurprising that research is focused on these areas and the development of treatment plans for them. Both common and less prevalent conditions are addressed by LLMs, including categories such as mental health support, rehabilitation, and rare diseases. This trend highlights a commitment to

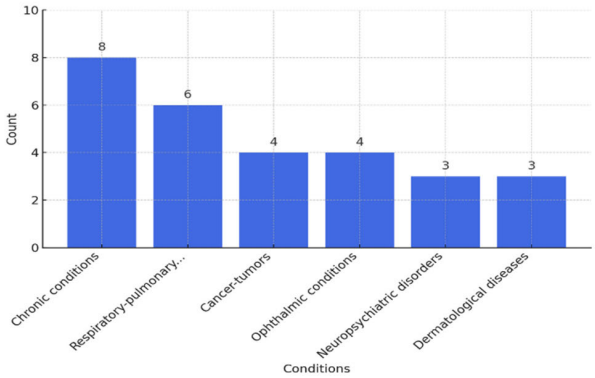

FIGURE 7. Distribution of diseases against each category.

general and specialized healthcare services provided by LLMs. Adverse drug reactions are also considered by LLMs ensuring patient safety and recommending personalized medication. Additionally, the presence of diseases, including metabolic syndrome and alcoholic liver disease shows how lifestyle and environmental factors affect public health and are also a focus of LLM-focused research. This analysis also reflects an integrated method of health care, which involves cardiology, neurology, and endocrinology. Overall, the current study suggests that LLMs can prove helpful in shaping a well-rounded healthcare system that prioritizes a broad spectrum of diseases while considering emerging trends in healthcare.

D. RQ4: HOW IS THE PERFORMANCE OF LLMs MEASURED IN DISEASE DIAGNOSIS?

Measuring the effectiveness of LLMs in diagnosing various diseases across diverse medical cases is a crucial aspect of current research. The studies included herein apply a variety of measures to evaluate the effectiveness of LLMs in the disease diagnostic process. The performance measures implemented by primary research to examine the efficiency of LLMs are explored in the below section.

In the study by Caruccio et al., the performance of the ChatGPT-based disease diagnostic process was measured using precision, recall, accuracy, and F-measure [25]. Several performance measures have been adopted in further studies, for example in the work of Wang et al., [27], including bleu scores, rouge-l, cider, meteor, precision, recall, and f-measure. Bleu scores, rouge-l, cider, and meteor provide a detailed assessment of LLMs by considering the overlap in generated and reference text over linguistic variety.While precision, recall, and f-measures examine the clinical efficiency of LLMs in producing relevant reports. A variety of qualitative measures have also been employed to assess the diagnostic process [30]. These measures include patient outcome prediction, clinical decision support post-treatment analysis, standardization, and data labeling. Patient outcome prediction further involves measures such as tumor control, toxicity levels, and overall survival rate.

A system usability scale questionnaire was employed by Xompero et al., to examine ophthalmology patient outcomes [31]. The scores achieved from this questionnaire revealed the

patients' usability experience. The medical diagnosis process was evaluated in the study by A. S. et al., using diagnostic accuracy, adaptability to rareness, handling incomplete responses, and interpretations of symptom descriptions [33]. Moreover, a statistical analysis was also performed using response accuracy rate, and response time. The performance of the diagnostic process was measured using accuracy and interpretability by Zhang et al., [32]. The interpretability is derived using Markov logic networks to get first-order logic rules. The study by Panagoulias et al., employed reliability and precision measures to examine the diagnostic process [34]. The response produced by ChatGPT was assessed based on the accuracy of the answer, usefulness for doctor and patient, and economic value produced by LLM.

The study by Saddik and Ghaboura, measured the effectiveness of ChatGPT using accuracy and reliability [35]. The study mainly employed qualitative measures, such as user understanding and patient engagement. A further study employed both qualitative and quantitative measures to examine the response of LLMs [36]. The qualitative measures included the overall quality of responses along with consistency and format of responses, while the quantitative analysis was performed using top-1, top-3, and top-5 Accuracy. This top-n accuracy analysis presents whether the LLMs predicted correct diagnosis in the first 1, 3, or 5 responses. The study by Panagoulias et al. employed a variety of performance measures, including response correctness and action ability [37]. Moreover, the study performed a holistic assessment to examine the medical reasoning capabilities? of GPT-4. The response was also measured using precision categories including precise, generic, or misleading?. A scoring scale was employed to assess performance measures including correctness, action ability, and precision. The LLM was also examined using a multiple-choice questions-based quiz comprising pathology questions.

Researchers employed micro and macro-F1 scores, average scores, and Rouge-L in a further study [28]. The average score measured overall performance against all tasks performed by ChatGLM. The performance criteria utilized in the study by Dou et al., were primarily focused on qualitative analysis [38]. The performance measures involved the comprehension of the query, situation analysis,rationality of medication advice, flagging potential adverse reactions, and comprehensive description. The performance of LLMs was also examined using the average score obtained against fifty questions regarding pertinent medical guidance provided by LLMs. Both qualitative and quantitative measures were implemented in the work of Ríos-Hoyo et al., to assess the performance of LLMs [39]. The qualitative measures included the correctness of diagnosis order and list overlap with case discussants. The quantitative measures involved the Jaccard similarity index, the inclusion of correct diagnosis, accuracy of top diagnoses, intraclass correlation coefficients, fisher's exact test, Mann–Whitney U test, and odds ratios.

Two tests were conducted to examine the performance of LLaMA to perform statistical analysis between groups in the study by Zhou et al., [40]. These tests included a two-tailed student's t-test and a consistency test. The LLM went through multiple trials to assess the response consistency. Moreover, the researchers also performed qualitative evaluation by dermatologists using a Likert scale. The study mainly focused on statistical analysis of GPT-2 using multiple performance measures, such as precision, recall, accuracy, F-measure, and Area Under the Curve (AUC). The researchers also conducted a user satisfaction survey to highlight the medical guidance provided by LLM.

A variety of performance measures were employed by Shyr I including precision, recall, F-measure, exact match, and relaxed match [42]. Additionally, an error analysis was also conducted to understand the nature of the error produced by LLM. This analysis divided error into five categories: incorrect boundary, incorrect entity type, incorrect boundary and entity type, spurious entities, and missed entities.

The differential diagnosis by LLM was analyzed using the accuracy of the main diagnosis, top-3 differential diagnosis, response time, and cost per case in the work of Rau et al., [44]. The explanations and clinical soundness were also assessed by experienced radiologists. The study by Bushuven et al., conducted both quantitative and qualitative analysis to examine the diagnostic capabilities of ChatGPT and GPT-4 [45]. The quantitative measures included diagnostic accuracy, inter-rater reliability,and Fleiss' kappa, while the qualitative measures included patient safety and advice quality. The primary metrics used to analyze performance in a 2023 study included diagnostic accuracy, differential diagnosis inclusion, and rank of correct diagnosis [46]. The study also employed statistical measures - means, standard deviations, medians, and interquartile ranges - to summarize the data. The researchers analyzed the diagnostic abilities of ChatGPT using accuracy and suitability. Responses generated by ChatGPT were labeled 'appropriate' if they contained no misconceptions while accuracy measured the correctness of responses against ground truth. The performance measures applied for effective disease diagnosis are further classified into sub-categories as shown in Table 6 The distribution of performance measures across various categories is shown in Figure 8.

Analysis: This review of primary studies with a focus on performance measures reveals that a comprehensive range of measures have been employed within the research on LLMs for efficient disease diagnostics. These measures include both qualitative and quantitative measures showing a diverse assessment of LLMs with multiple perspectives of healthcare. Quantitative assessment has been performed mostly using precision, recall, accuracy, and F-measure. These measures represent LLMs' effectiveness in making correct predictions and identifying the factual extent of medical conditions. Statistical measures, including AUC, BLEU scores, ROUGE-L, CIDEr, and METEOR, were applied to examine the coherence and relevance of the generated text. The Jaccard similarity index was implemented to assess the similarity between actual and predicted diagnoses. Similarly, Fisher's

TABLE 6. Classification of performance measures across distinct categories.

Main Category	Sub-category
General accuracy measures	Reliability
	Consistency
	Precision
	Recall
	Accuracy
	F1 score
Text Generation Evaluation	Rouge-L
	BLEU
	CIDEr
	METEOR
Statistical measures	AUC
	Jaccard Similarity Index
	Fleiss' Kappa
	Odds Ratios
	Mann–Whitney U Test
	Fisher's Exact Test
	Error Analysis
Usability and Practicality	System Usability Scale
	Response Time
	Cost
	Actionability
Qualitative and Expert Evaluations	Expert Evaluations
	Survey-Based Analysis
	Holistic Assessment
Contextual Assessments	Patient Outcome Prediction
	Clinical Decision Support
	Post-Treatment Analysis
	Interpretations of Symptom Descriptions
	Comprehension of Query
	Rationality of Medication Advice

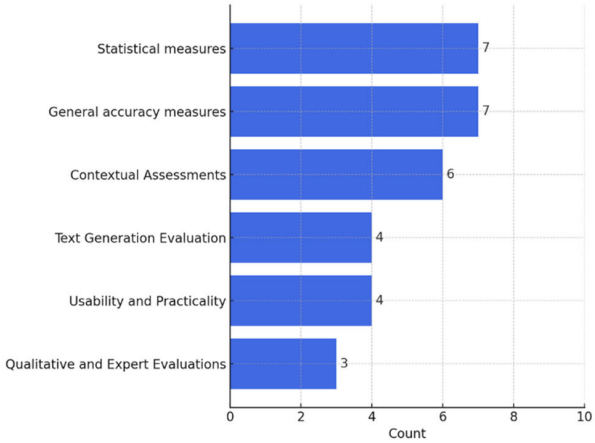

FIGURE 8. Distribution of performance measure across specific categories.

exact test, Mann–Whitney U test, odds ratios,intraclass correlation coefficients, and Fleiss' kappa were also implemented to examine the efficacy, consistency, and clinical utility of LLM-based disease diagnosis. Descriptive measures like mean, standard deviation, median, and interquartile range were employed to summarize the statistics offering insights into the patterns and variability in the generated results.

Additionally, the analysis of primary literature undertaken here indicates that numerous qualitative measures have also been implemented to highlight the dimensions that surpass quantitative measures. These measures include system usability scale, overall quality of responses, consistency and format of responses, interpretations of symptom descriptions, comprehension of the query, situation analysis, and rationality of medication advice. These are particularly key metrics for assessing the ability of a model to make clinical decisions and ensure that the provided recommendations are safe to apply.

Holistic assessment and expert evaluation conducted by radiologists or dermatologists using a Likert scale, provide invaluable insights into the performance of LLMs in terms of reliability and relevance to healthcare. Measures such as patient safety, as well as the quality of advice, add an extra dimension by considering non-adherence to ethical and practical standards in healthcare.

V. ISSUES AND CHALLENGES

The utilization of LLMs in medical diagnostics poses several difficulties and challenges. The bias in the training data is the most crucial concern. LLMs are designed using enormous amounts of data that may fail to accurately represent all populations, or individuals within populations. Thus, biased predictions may have a negative impact on under-represented communities [47], [48]. These biases may lead to incorrect diagnoses and inadequate access to medical care. Completing training data with a diverse range of demographic, geographic, and socioeconomic backgrounds further complicates the process of creating unbiased LLMs in the healthcare industry [49].

The LLMs' interpretability and transparency are the additional significant issues. They are called black box models, as in most cases experts find it difficult to understand or justify the reasoning behind the decisions they make. Furthermore, transparency in diagnosis is critical to the medical sciences to establish patient trust and follow ethical medical standards [50], [51]. Since LLMs are unable to provide a justification for the findings of the model, clinicians may be reluctant to depend upon their results. In addition, it may be difficult to identify and fix problems with the models as a result of the absence of this transparency, leading to incorrect diagnosis and treatment suggestions [52].

Risks about data security and privacy are also brought up by integrating LLMs in therapeutic environments [53]. Sensitive patient information may be found in the majority of medical data used for LLM training and implementation. Hence, it might be challenging for professionals to efficiently protect and maintain data confidentiality in an era where cyber threats are constantly evolving. Moreover, there are ethical concerns associated with the utilization of data with patient consent to uphold patient trust [54]. To balance the benefits of advanced diagnostic tools, such as GPT-4, ChatGPT, and LLaMA, there is also a need to protect patient privacy and maintain legal standards.

Moreover, preserving the accuracy and reliability of LLMs while integrating them in clinical workflows is also a

challenging scenario raising concerns for both researchers and practitioners.

An important challenge in the present literature is the lack of discussion on the application of LLMs in specific areas, such as disease diagnosis. Although prior surveys [1], [2], [12], [18], [19], [20], have provided valuable insights into general LLM applications in healthcare, they do not sufficiently address their role in diagnostic processes or their use across various medical data sources. Our survey links these gaps by offering a detailed analysis of LLMs applications in disease diagnostics, offering novel views on their performance and challenges in real-world healthcare scenarios.

VI. CONCLUSION AND FUTURE WORK

This review highlights the transformative role of LLMs in disease diagnostics, highlighting their adaptability across diverse healthcare fields such as chronic diseases, mental health, and rare conditions. Key models, including GPT-4, ChatGPT, and LLaMA, utilized in disease diagnosis in healthcare. Medical data from diverse data sources, from general to specialized datasets, support these models' training and application. LLMs were found to be implemented across various healthcare fields, including chronic diseases, mental health, and rare conditions. These LLMs were assessed using various measures comprising both quantitative and qualitative measures, highlighting the complexity of evaluating the LLMs. Our findings emphasize the potential of LLMs to improve diagnostic accuracy and contribute to a more comprehensive healthcare system. Continued research and exploration of new models and data sources are essential for advancing medical diagnostics. The analysis of primary studies reveals that researchers aim to enhance the utilization of LLMs by monitoring patients' health through wearable or home monitoring devices to diagnose various diseases. In the future, the scope of this literature review can be enhanced by exploring the application of LLMs in other areas of healthcare, for example, treatment planning and patient monitoring.

ACKNOWLEDGMENT

The researcher would like to thank the Deanship of Graduate Studies and Scientific Research at Qassim University for their financial support (QU-APC-2025).

REFERENCES

[1] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature Med.*, vol. 29, no. 8, pp. 1930–1940, Aug. 2023, doi: 10.1038/s41591-023-02448-8.

[2] S. Harrer, "Attention is not all you need: The complicated case of ethically using large language models in healthcare and medicine," *eBioMedicine*, vol. 90, Apr. 2023, Art. no. 104512, doi: 10.1016/j.ebiom.2023.104512.

[3] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly," *High-Confidence Comput.*, vol. 4, no. 2, Jun. 2024, Art. no. 100211, doi: 10.1016/j.hcc.2024.100211.

[4] M. Mizrahi, G. Kaplan, D. Malkin, R. Dror, D. Shahaf, and G. Stanovsky, "State of what art? A call for multi-prompt LLM evaluation," *Trans. Assoc. Comput. Linguistics*, vol. 12, pp. 933–949, Aug. 2024, doi: 10.1162/tacl_a_00681.

[5] A. Bhat, D. Shrivastava, and J. L. C. Guo, "Do LLMs meet the needs of software tutorial writers? Opportunities and design implications," in *Proc. Designing Interact. Syst. Conf.* Copenhagen, Denmark: IT Univ. Copenhagen Denmark, Jul. 2024, pp. 1760–1773, doi: 10.1145/3643834.3660692.

[6] M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, J. Ahmad, M. E. Ali, and S. Azam, "A review on large language models: Architectures, applications, taxonomies, open issues and challenges," *IEEE Access*, vol. 12, pp. 26839–26874, 2024, doi: 10.1109/ACCESS.2024.3365742.

[7] A. Mosavi, F. Imre, and V. T. Hung, "ChatGPT and large language models in healthcare; a bibliometrics analysis and review," in *Proc. IEEE 11th Int. Conf. Comput. Cybern. Cyber-Medical Syst. (ICCC)*, Hanoi, Vietnam, Apr. 2024, pp. 1–6, doi: 10.1109/iccc62278.2024.10582937.

[8] M. J. Boonstra, D. Weissenbacher, J. H. Moore, G. Gonzalez-Hernandez, and F. W. Asselbergs, "Artificial intelligence: Revolutionizing cardiology with large language models," *Eur. Heart J.*, vol. 45, no. 5, pp. 332–345, Jan. 2024, doi: 10.1093/eurheartj/ehad838.

[9] H. Y. Kwan, "User-focused telehealth powered by LLMs: Bridging the gap between technology and human-centric care delivery," in *Proc. 4th Int. Conf. Comput. Commun. Artif. Intell. (CCAI)*, Xi'an, China, May 2024, pp. 187–191, doi: 10.1109/ccai61966.2024.10603150.

[10] R. Yang, T. F. Tan, W. Lu, A. J. Thirunavukarasu, D. S. W. Ting, and N. Liu, "Large language models in health care: Development, applications, and challenges," *Health Care Sci.*, vol. 2, no. 4, pp. 255–263, Aug. 2023, doi: 10.1002/hcs2.61.

[11] Y. Lee, T. Shin, L. Tessier, A. Javidan, J. Jung, D. Hong, A. T. Strong, T. McKechnie, S. Malone, D. Jin, M. Kroh, and J. T. Dang, "Harnessing artificial intelligence in bariatric surgery: Comparative analysis of ChatGPT-4, bing, and bard in generating clinician-level bariatric surgery recommendations," *Surg. Obesity Rel. Diseases*, vol. 20, no. 7, pp. 603–608, Jul. 2024, doi: 10.1016/j.soard.2024.03.011.

[12] S. Sai, A. Gaur, R. Sai, V. Chamola, M. Guizani, and J. J. P. C. Rodrigues, "Generative AI for transformative healthcare: A comprehensive study of emerging models, applications, case studies, and limitations," *IEEE Access*, vol. 12, pp. 31078–31106, 2024, doi: 10.1109/ACCESS.2024.3367715.

[13] S. Reddy, "Evaluating large language models for use in healthcare: A framework for translational value assessment," *Informat. Med. Unlocked*, vol. 41, Jul. 2023, Art. no. 101304, doi: 10.1016/j.imu.2023.101304.

[14] B. Meskó, "The impact of multimodal large language models on health care's future," *J. Med. Internet Res.*, vol. 25, Nov. 2023, Art. no. e52865, doi: 10.2196/52865.

[15] Á. García-Barragán, A. G. Calatayud, L. Prieto-Santamaría, V. Robles, E. Menasalvas, and A. Rodríguez, "Step-forward structuring disease phenotypic entities with LLMs for disease understanding," in *Proc. IEEE 37th Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Guadalajara, Mexico, Jun. 2024, pp. 213–218, doi: 10.1109/cbms61543.2024.00043.

[16] A. Nolin-Lapalme, P. Theriault-Lauzier, D. Corbin, O. Tastet, A. Sharma, J. G. Hussin, S. Kadoury, R. Jiang, A. D. Krahn, R. Gallo, and R. Avram, "Maximising large language model utility in cardiovascular care: A practical guide," *Can. J. Cardiol.*, vol. 40, no. 10, pp. 1774–1787, 2024, doi: 10.1016/j.cjca.2024.05.024.

[17] Y. Lee, L. Tessier, K. Brar, S. Malone, D. Jin, T. McKechnie, J. J. Jung, M. Kroh, and J. T. Dang, "Performance of artificial intelligence in bariatric surgery: Comparative analysis of ChatGPT-4, bing, and bard in the American society for metabolic and bariatric surgery textbook of bariatric surgery questions," *Surg. Obesity Rel. Diseases*, vol. 20, no. 7, pp. 609–613, Jul. 2024, doi: 10.1016/j.soard.2024.04.014.

[18] S. N. Hart, N. G. Hoffman, P. Gershkovich, C. Christenson, D. S. McClintock, L. J. Miller, R. Jackups, V. Azimi, N. Spies, and V. Brodsky, "Organizational preparedness for the use of large language models in pathology informatics," *J. Pathol. Informat.*, vol. 14, Oct. 2023, Art. no. 100338, doi: 10.1016/j.jpi.2023.100338.

[19] B. Wang, Q. Xie, J. Pei, Z. Chen, P. Tiwari, Z. Li, and J. Fu, "Pre-trained language models in biomedical domain: A systematic survey," *ACM Comput. Surv.*, vol. 56, no. 3, pp. 1–52, Mar. 2024, doi: 10.1145/3611651.

[20] M. Yuan, P. Bao, J. Yuan, Y. Shen, Z. Chen, Y. Xie, J. Zhao, Q. Li, Y. Chen, L. Zhang, L. Shen, and B. Dong, "Large language models illuminate a progressive pathway to artificial intelligent healthcare assistant," *Med. Plus*, vol. 1, no. 2, Jun. 2024, Art. no. 100030, doi: 10.1016/j.medp.2024.100030.

[21] Z. Al Nazi and W. Peng, "Large language models in healthcare and medical domain: A review," 2023, *arXiv:2401.06775*.

[22] E. Waisberg, J. Ong, M. Masalkhi, S. A. Kamran, N. Zaman, P. Sarker, A. G. Lee, and A. Tavakkoli, "GPT-4: A new era of artificial intelligence in medicine," *Irish J. Med. Sci.*, vol. 192, no. 6, pp. 3197–3200, Dec. 2023, doi: 10.1007/s11845-023-03377-8.
[23] S. Grassini, "Shaping the future of education: Exploring the potential and consequences of AI and ChatGPT in educational settings," *Educ. Sci.*, vol. 13, no. 7, p. 692, Jul. 2023, doi: 10.3390/educsci13070692.
[24] A. Koubaa, "GPT-4 vs. GPT-3.5: A concise showdown," *Engineering*, Mar. 2023, doi: 10.20944/preprints202303.0422.v1.
[25] L. Caruccio, S. Cirillo, G. Polese, G. Solimando, S. Sundaramurthy, and G. Tortora, "Can ChatGPT provide intelligent diagnoses? A comparative study between predictive models and ChatGPT to define a new medical diagnostic bot," *Exp. Syst. Appl.*, vol. 235, Jan. 2024, Art. no. 121186, doi: 10.1016/j.eswa.2023.121186.
[26] M. Bahani, A. E. Ouaazizi, and K. Maalmi, "The effectiveness of t5, GPT-2, and BERT on text-to-image generation task," *Pattern Recognit. Lett.*, vol. 173, pp. 57–63, Sep. 2023, doi: 10.1016/j.patrec.2023.08.001.
[27] Z. Wang, L. Liu, L. Wang, and L. Zhou, "R2GenGPT: Radiology report generation with frozen LLMs," *Meta-Radiol.*, vol. 1, no. 3, Nov. 2023, Art. no. 100033, doi: 10.1016/j.metrad.2023.100033.
[28] Q. Liu, X. Wu, X. Zhao, Y. Zhu, D. Xu, F. Tian, and Y. Zheng, "When MOE meets LLMs: Parameter efficient fine-tuning for multi-task medical applications," in *Proc. 47th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Washington, DC, USA, Jul. 2024, pp. 1104–1114, doi: 10.1145/3626772.3657722.
[29] S. Huang and C. Chen, "Combining LoRa to GPT-neo to reduce large language model hallucination," *Review*, Jun. 2024, doi: 10.21203/rs.3.rs-4515250/v1.
[30] C. Liu, Z. Liu, J. Holmes, L. Zhang, L. Zhang, Y. Ding, P. Shu, Z. Wu, H. Dai, Y. Li, D. Shen, N. Liu, Q. Li, X. Li, D. Zhu, T. Liu, and W. Liu, "Artificial general intelligence for radiation oncology," *Meta-Radiol.*, vol. 1, no. 3, Nov. 2023, Art. no. 100045, doi: 10.1016/j.metrad.2023.100045.
[31] C. Xompero, W. Benettayeb, E. H. Souied, and C.-J. Mehanna, "Pilot study evaluating the usability of Monæil, a ChatGPT-based education tool in ophthalmology," *AJO Int.*, vol. 1, no. 2, Jul. 2024, Art. no. 100032, doi: 10.1016/j.ajoint.2024.100032.
[32] H. Zhang, J. Li, Y. Wang, and Y. Song, "Integrating automated knowledge extraction with large language models for explainable medical decision-making," in *Proc. IEEE Int. Conf. Bioinf. Biomedicine (BIBM)*, Istanbul, Turkiye, Dec. 2023, pp. 1710–1717, doi: 10.1109/BIBM58861.2023.10385557.
[33] S. Akilesh, R. Abinaya, S. Dhanushkodi, and R. Sekar, "A novel AI-based chatbot application for personalized medical diagnosis and review using large language models," in *Proc. Int. Conf. Res. Methodologies Knowl. Manage., Artif. Intell. Telecommun. Eng. (RMKMATE)*, Nov. 2023, pp. 1–5, doi: 10.1109/rmkmate59243.2023.10368616.
[34] D. P. Panagoulias, F. A. Palamidas, M. Virvou, and G. A. Tsihrintzis, "Rule-augmented artificial intelligence-empowered systems for medical diagnosis using large language models," in *Proc. IEEE 35th Int. Conf. Tools Artif. Intell. (ICTAI)*, Nov. 2023, pp. 70–77, doi: 10.1109/ictai59109.2023.00018.
[35] A. E. Saddik and S. Ghaboura, "The integration of ChatGPT with the metaverse for medical consultations," *IEEE Consum. Electron. Mag.*, vol. 13, no. 3, pp. 6–15, May 2024, doi: 10.1109/MCE.2023.3324978.
[36] J. Kim, C.-Y. Ju, and D.-H. Lee, "Who can be your AI doctor? Evaluation for disease diagnosis on large language models," in *Proc. 14th Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Oct. 2023, pp. 154–159, doi: 10.1109/ICTC58733.2023.10392305.
[37] D. P. Panagoulias, M. Virvou, and G. A. Tsihrintzis, "Augmenting large language models with rules for enhanced domain-specific interactions: The case of medical diagnosis," *Electronics*, vol. 13, no. 2, p. 320, Jan. 2024, doi: 10.3390/electronics13020320.
[38] Y. Dou, Y. Huang, X. Zhao, H. Zou, J. Shang, Y. Lu, X. Yang, J. Xiao, and S. Peng, "ShennongMGS: An LLM-based Chinese medication guidance system," *ACM Trans. Manage. Inf. Syst.*, Apr. 2024, Art. no. 3658451, doi: 10.1145/3658451.
[39] A. Ríos-Hoyo, N. L. Shan, A. Li, A. T. Pearson, L. Pusztai, and F. M. Howard, "Evaluation of large language models as a diagnostic aid for complex medical cases," *Frontiers Med.*, vol. 11, Jun. 2024, Art. no. 1380148, doi: 10.3389/fmed.2024.1380148.
[40] J. Zhou, X. He, L. Sun, J. Xu, X. Chen, Y. Chu, L. Zhou, X. Liao, B. Zhang, S. Afvari, and X. Gao, "Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4," *Nature Commun.*, vol. 15, no. 1, p. 5649, Jul. 2024, doi: 10.1038/s41467-024-50043-3.
[41] S. Zhang and J. Song, "A chatbot based question and answer system for the auxiliary diagnosis of chronic diseases based on large language model," *Sci. Rep.*, vol. 14, no. 1, p. 17118, Jul. 2024, doi: 10.1038/s41598-024-67429-4.
[42] C. Shyr, Y. Hu, L. Bastarache, A. Cheng, R. Hamid, P. Harris, and H. Xu, "Identifying and extracting rare diseases and their phenotypes with large language models," *J. Healthcare Informat. Res.*, vol. 8, no. 2, pp. 438–461, Jun. 2024, doi: 10.1007/s41666-023-00155-0.
[43] X. Hu, A. R. Ran, T. X. Nguyen, S. Szeto, J. C. Yam, C. K. M. Chan, and C. Y. Cheung, "What can GPT-4 do for diagnosing rare eye diseases? A pilot study," *Ophthalmol. Therapy*, vol. 12, no. 6, pp. 3395–3402, Dec. 2023, doi: 10.1007/s40123-023-00789-8.
[44] S. Rau, A. Rau, J. Nattenmüller, A. Fink, F. Bamberg, M. Reisert, and M. F. Russe, "A retrieval-augmented chatbot based on GPT-4 provides appropriate differential diagnosis in gastrointestinal radiology: A proof of concept study," *Eur. Radiol. Experim.*, vol. 8, no. 1, p. 60, May 2024, doi: 10.1186/s41747-024-00457-x.
[45] S. Bushuven, M. Bentele, S. Bentele, B. Gerber, J. Bansbach, J. Ganter, M. Trifunovic-Koenig, and R. Ranisch, "'ChatGPT, can you help me save my child's life?'—Diagnostic accuracy and supportive capabilities to lay rescuers by ChatGPT in prehospital basic life support and paediatric advanced life support cases—An in-silico analysis," *J. Med. Syst.*, vol. 47, no. 1, p. 123, Nov. 2023, doi: 10.1007/s10916-023-02019-x.
[46] M. Balas and E. B. Ing, "Conversational AI models for ophthalmic diagnosis: Comparison of ChatGPT and the isabel pro differential diagnosis generator," *JFO Open Ophthalmol.*, vol. 1, Mar. 2023, Art. no. 100005, doi: 10.1016/j.jfop.2023.100005.
[47] B. Meskó and E. J. Topol, "The imperative for regulatory oversight of large language models (or generative AI) in healthcare," *Npj Digit. Med.*, vol. 6, no. 1, p. 120, Jul. 2023, doi: 10.1038/s41746-023-00873-0.
[48] D. Saha, S. Tarek, K. Yahyaei, S. K. Saha, J. Zhou, M. Tehranipoor, and F. Farahmandi, "LLM for SoC security: A paradigm shift," *IEEE Access*, vol. 12, pp. 155498–155521, 2024, doi: 10.1109/access.2024.3427369.
[49] R. Patil and V. Gudivada, "A review of current trends, techniques, and challenges in large language models (LLMs)," *Appl. Sci.*, vol. 14, no. 5, p. 2074, Mar. 2024, doi: 10.3390/app14052074.
[50] A. Arora and A. Arora, "The promise of large language models in health care," *Lancet*, vol. 401, no. 10377, p. 641, Feb. 2023, doi: 10.1016/s0140-6736(23)00216-7.
[51] S. Ebrahimi, N. Shahbazi, and A. Asudeh, "REQUAL-LM: Reliability and equity through aggregation in large language models," in *Proc. Findings Assoc. Comput. Linguistics*. Mexico City, Mexico: Association for Computational Linguistics, 2024, pp. 549–560, doi: 10.18653/v1/2024.findings-naacl.37.
[52] S. Abumusab, "Introduction to the special issue: Large language models and teaching writing," *Teaching Philosophy*, vol. 47, no. 2, pp. 139–142, 2024, doi: 10.5840/teachphil2024472195.
[53] E. Ullah, A. Parwani, M. M. Baig, and R. Singh, "Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology—A recent scoping review," *Diagnostic Pathol.*, vol. 19, no. 1, p. 43, Feb. 2024, doi: 10.1186/s13000-024-01464-7.
[54] F. Wu, N. Zhang, S. Jha, P. McDaniel, and C. Xiao, "A new era in LLM security: Exploring security concerns in real-world LLM-based systems," 2024, *arXiv:2402.18649*.

IBRAHIM ALMUBARK received the B.Sc. degree in computer science from King Saud University, Saudi Arabia, the M.Sc. degree in computer science, along with a Graduate Certificate in Computer Security and Information Assurance, from The George Washington University, Washington, DC, USA, and the Ph.D. degree from The Catholic University of America, Washington, DC, USA. He is currently an Assistant Professor with the Department of Information Technology, Qassim University, Saudi Arabia. His research and teaching interests include artificial intelligence, machine learning, data science, big data analytics, databases, and data privacy and security.