Review

Large Language Models in Medical Diagnostics: Scoping Review With Bibliometric Analysis

Hankun Su[1,2,3*], MS; Yuanyuan Sun[1,2*], MD; Ruiting Li[4], MD; Aozhe Zhang[3], MS; Yuemeng Yang[1,2,3], MS; Fen Xiao[5], MD, PhD; Zhiying Duan[1,2], MD; Jingjing Chen[1,2], MD; Qin Hu[1,2], MD; Tianli Yang[1,2], MD; Bin Xu[1,2], MD, PhD; Qiong Zhang[1,2], MD, PhD; Jing Zhao[1,2], MD, PhD; Yanping Li[1,2], MD, PhD; Hui Li[1,2], MD, PhD

[1]Department of Reproductive Medicine, Xiangya Hospital Central South University, Changsha, China

[2]Clinical Research Center for Women's Reproductive Health in Hunan Province, Changsha, China

[3]Xiangya School of Medicine, Central South University, Changsha, China

[4]School of Biomedical Sciences and Engineering, South China University of Technology, Guangzhou, China

[5]Department of Metabolism and Endocrinology, Second Xiangya Hospital of Central South University, Changsha, China

[*]these authors contributed equally

Corresponding Author:
Hui Li, MD, PhD
Department of Reproductive Medicine
Xiangya Hospital Central South University
87 Xiangya Road
Changsha, 410008
China
Phone: 86 13272047403
Email: huili257@csu.edu.cn

Abstract

Background: The integration of large language models (LLMs) into medical diagnostics has garnered substantial attention due to their potential to enhance diagnostic accuracy, streamline clinical workflows, and address health care disparities. However, the rapid evolution of LLM research necessitates a comprehensive synthesis of their applications, challenges, and future directions.

Objective: This scoping review aimed to provide an overview of the current state of research regarding the use of LLMs in medical diagnostics. The study sought to answer four primary subquestions, as follows: (1) Which LLMs are commonly used? (2) How are LLMs assessed in diagnosis? (3) What is the current performance of LLMs in diagnosing diseases? (4) Which medical domains are investigating the application of LLMs?

Methods: This scoping review was conducted according to the Joanna Briggs Institute Manual for Evidence Synthesis and adheres to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews). Relevant literature was searched from the Web of Science, PubMed, Embase, IEEE Xplore, and ACM Digital Library databases from 2022 to 2025. Articles were screened and selected based on predefined inclusion and exclusion criteria. Bibliometric analysis was performed using VOSviewer to identify major research clusters and trends. Data extraction included details on LLM types, application domains, and performance metrics.

Results: The field is rapidly expanding, with a surge in publications after 2023. GPT-4 and its variants dominated research (70/95, 74% of studies), followed by GPT-3.5 (34/95, 36%). Key applications included disease classification (text or image-based), medical question answering, and diagnostic content generation. LLMs demonstrated high accuracy in specialties like radiology, psychiatry, and neurology but exhibited biases in race, gender, and cost predictions. Ethical concerns, including privacy risks and model hallucination, alongside regulatory fragmentation, were critical barriers to clinical adoption.

Conclusions: LLMs hold transformative potential for medical diagnostics but require rigorous validation, bias mitigation, and multimodal integration to address real-world complexities. Future research should prioritize explainable artificial intelligence frameworks, specialty-specific optimization, and international regulatory harmonization to ensure equitable and safe clinical deployment.

(J Med Internet Res 2025;27:e72062) doi: 10.2196/72062

KEYWORDS

large language model; scoping review; medical diagnosis; bibliometric analysis; artificial intelligence

Introduction

In the critical domain of health care, the efficacy of medical decision-making and diagnostic accuracy is essential for managing medical conditions effectively. To bolster these processes, artificial intelligence (AI) models have been increasingly used, demonstrating the potential to rival the diagnostic prowess of seasoned clinicians [1,2].

Large language models (LLMs) are sophisticated AI systems that undergo extensive pretraining, absorbing statistical laws and discernible patterns from extensive datasets [3]. Consequently, they possess the remarkable ability to autonomously generate responses to inquiries and engage in interactive dialogues with users [4]. This capability has also raised the interest of the medical community, who see in LLMs a tool that could significantly enhance various facets of health care, especially in diagnostics [5]. LLMs have presented promising performance in undertaking medical tasks [4,6]; for instance, models such as OpenAI's ChatGPT and Google's Bard have showcased the potential to enhance clinical decision-making processes [7], refine diagnostic accuracy [8], and facilitate the synthesis of complex medical literature [9].

Despite the growing interest in LLMs in diagnostic application, there is a noticeable absence of comprehensive analysis that examines the evolution and analytic appraisal of LLMs in medical diagnosis. Current scoping reviews on LLMs are predominantly conceptual and focused mainly on the entire area of biomedical health [10], highlighting an urgent need for a targeted, domain-specific review to guide future research directions [4,11,12]. With this scoping review, we therefore intend to answer the following research question (RQ): What is the state of research regarding medical diagnosis based on LLMs?

However, traditional scoping review alone may struggle to capture the rapid expansion of LLM research, which has grown exponentially in volume and complexity. To address this, we integrated bibliometric analysis—a quantitative and visualized method for evaluating research landscape and trend-within the scoping review [13]. This hybrid approach enables a dual perspective: the bibliometric analysis serves as an indicator for scoping review, mapping the fields growth, interest, and future trend, while the scoping reviews dig deeper into synthesizing qualitative and quantitative insights on LLMs' diagnostic applications. Together, the hybrid approach provides a holistic view of the state of research regarding medical diagnosis. Therefore, complemented by bibliometric clustering, we broke the large research question into 4 smaller subjects, each representing a major research focus that has been extensively studied in the field, as follows:

1. Which LLMs are commonly used in medical diagnosis?
2. How to examine the performance of LLM in medical diagnosis?
3. What is the current performance of LLMs in diagnosing diseases?
4. Which medical domains are investigating the application of LLMs?

With a combination of scoping review and bibliometric method, we hope to enhance readers' comprehension of LLMs and provide a guideline for prospective collaborative pursuits and clinical implementations.

Methods

Overview

This scoping review was conducted according to the Joanna Briggs Institute Manual for Evidence Synthesis and adheres to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) [14].

Search Strategy and Study Selection

Web of Science, PubMed, Embase, IEEE Xplore, and ACM Digital Library databases were searched as the source of relevant literature from 2022 to January 2025. The first search was conducted on September 27, 2024, and the final search on January 12, 2025. Multimedia Appendix 1 provides the detailed search strategies. Raw data were first imported into the Rayyan platform. Subsequently, 4 authors (HS, YS, AZ, and YY) independently screened for potentially relevant articles in a blind mode with the following procedure: duplicates were first removed, followed by a screening of browsing titles and abstracts. The remaining articles were screened for the second time to evaluate the full text of all eligible articles. At any point in the process, conflicting articles were resolved by RL. The eligibility criteria are listed in subsequent section.

Inclusion and Exclusion Criteria

The inclusion criteria were as follows: (1) articles that focused on reporting the outcome of LLMs in medical diagnosis–related field and (2) articles that described the application of LLMs for medical diagnosis. LLMs here are defined as deep learning models with more than one million parameters, trained on unlabeled text data

To obtain only necessary data, we excluded articles that were conference abstracts or abstracts only and (2) comments and retracted articles.

Bibliometric Analysis

After the selection of included articles, the data from the included articles from Web of Science were first exported in tab-delimited format and then imported into Microsoft Excel 2019 for preliminary collation. Non–Web of Science articles were manually entered into Microsoft Excel 2019 by 2 independent researchers (AZ and RL) to ensure accuracy. The combined dataset was converted to tab-delimited format for compatibility with a bibliometric tool. VoSviewer (Leiden University) was used to map research trends and frontiers in LLM applications for medical diagnosis. This analysis identified 3 dominant research clusters reflecting key thematic foci, which

subsequently informed the design of data extraction categories. The bibliometric analysis objectively identified field-level hot spots and trends, while the scoping review contextualized these patterns through qualitative appraisal of study designs and clinical validations. This dual approach ensured both macroscopic trend detection (via VOSviewer keyword clustering) and microscopic evidence evaluation (via structured data extraction).

Data Extraction

The bibliometric clusters guided the development of an extraction template to systematically capture bibliometric attributes (eg, authors, publication years, and countries), thematic focus (eg, specialty task and LLM type), and clinical validation metrics (eg, primary end point, comparator, and primary result). Two authors (HS and YS) independently extracted data using this template, resolving discrepancies through consensus discussions. A pilot extraction on 10% of articles preceded full-text review to refine category definitions.

Results

Study Selection and Characteristics

Figure 1 displays the study selection process. A total of 17,757 records were identified from Web of Science, PubMed, Embase, IEEE Xplore, and ACM Digital Library databases. After removing 2445 duplicates, we screened titles and abstracts, and 6708 records were assessed for inclusion, resulting in 95 articles in the final review. The overall characteristic of included articles is provided in Multimedia Appendix 2 [15-108]. Figure 2A illustrates the distribution of articles by year, which indicates that this research area is still emerging, with a significant increase in publications occurring in 2024. This trend suggests that the field is likely to continue experiencing positive growth over the coming years. Figure 2B illustrates that the United States and China are the most productive countries in the research of LLM in diagnosis. In addition, the United States has the highest number of interconnected targets in country cooperation. It is suggested for other countries' researchers to strengthen international cooperation and communication for this research topic.

Figure 1. Illustration of study selection process. LLM: large language model.

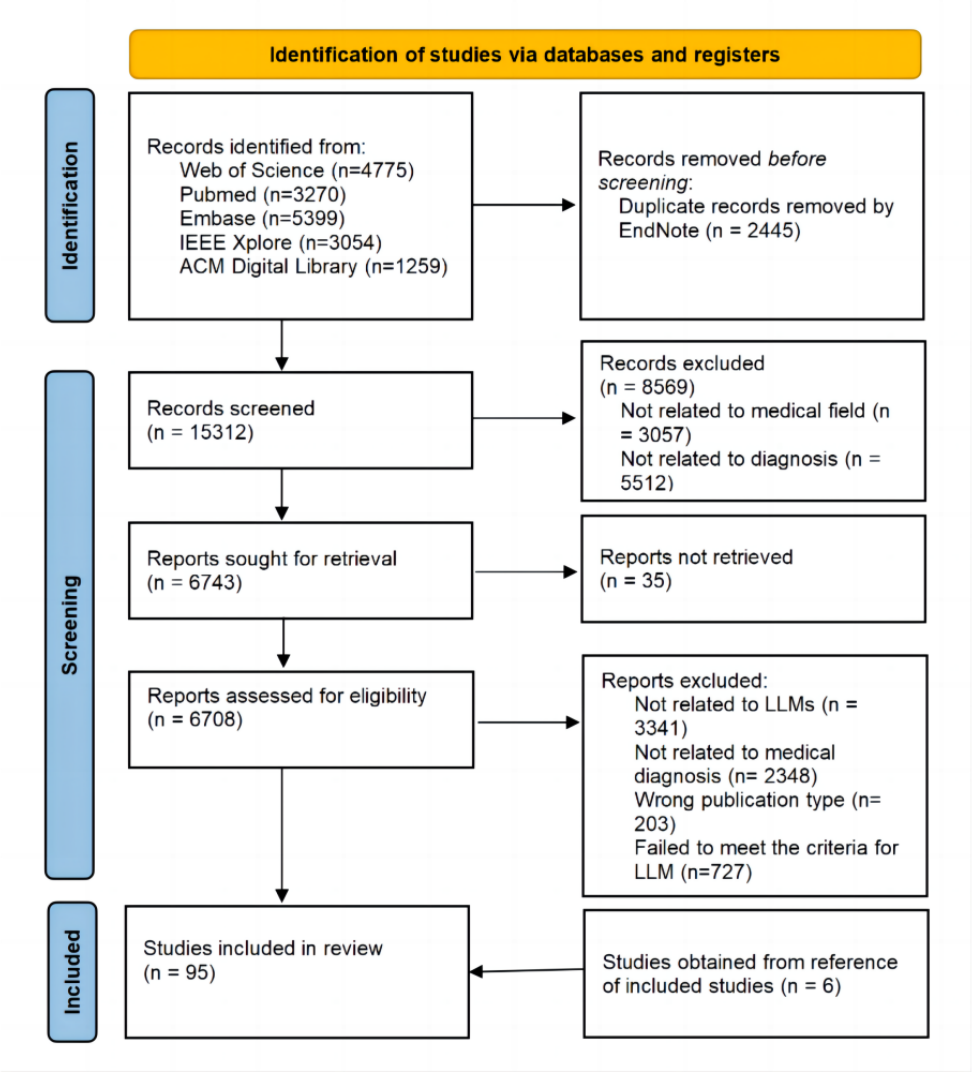
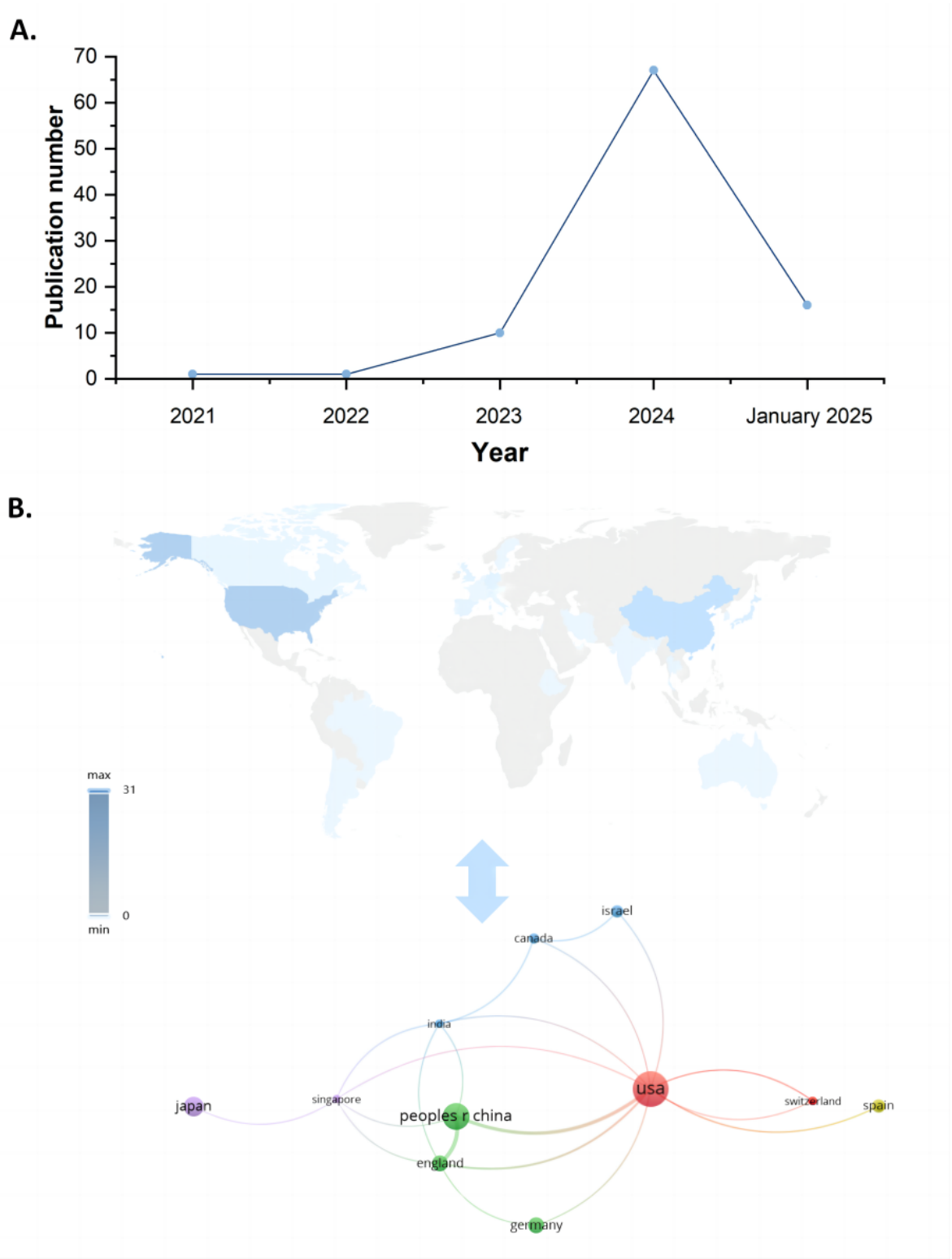


Figure 2. (A) Study distribution by final publication year and (B) study distribution across countries and bibliometric visualization of country collaborations.



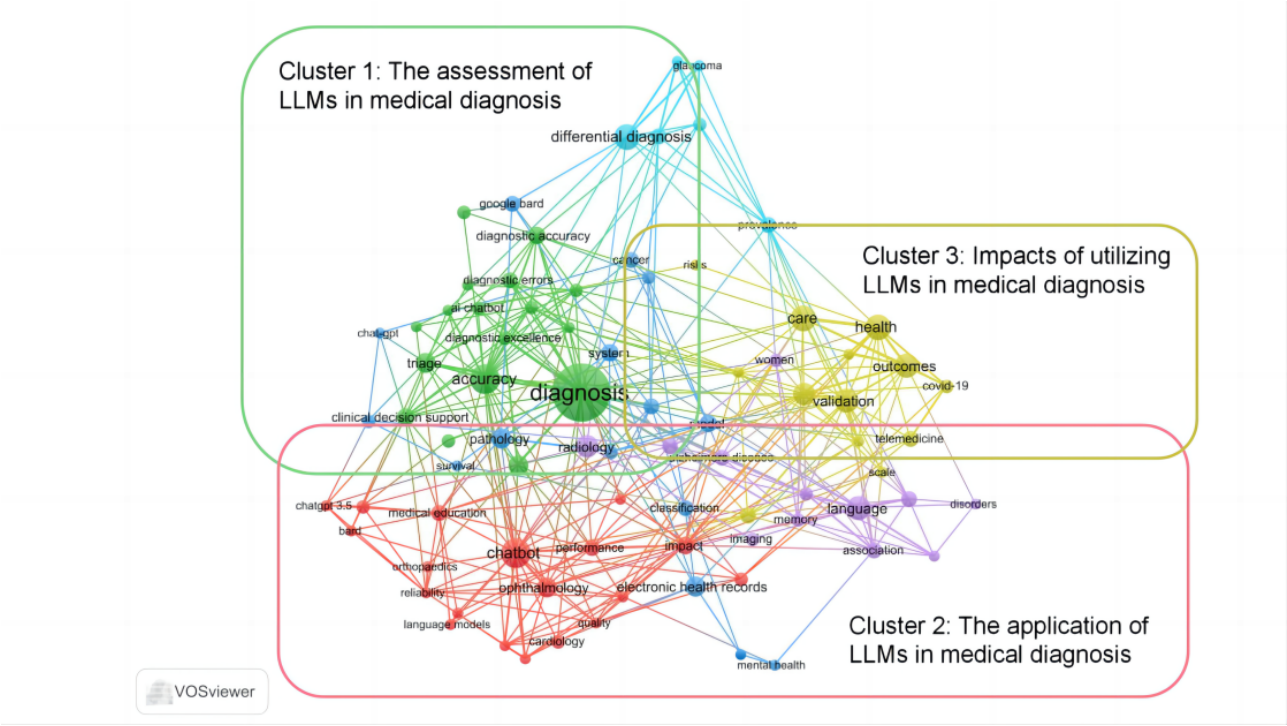
Bibliometric Analysis of Articles

We used VOSviewer software for bibliometric labeling of included studies, and a total of 65 keywords and 3 clusters were obtained (Figure 3). Cluster 1, "The assessment of LLMs in medical diagnosis," is noted in green, with keywords including "accuracy" and "diagnostic accuracy." This cluster focuses on the accuracy and reliability of LLMs in medical diagnosis and how they support clinical decision-making and triage. Cluster 2, "The application of LLMs in medical diagnosis," is represented by red, with keywords including "orthopedics," "cardiology," and "ophthalmology," among others. This cluster illustrates the application of LLMs in diagnosis across various medical specialties, including orthopedics, cardiology, and ophthalmology. Cluster 3, "Impacts of using LLMs in medical diagnosis," is indicated in yellow, with keywords including

"health," and "outcomes," among others. This cluster focuses on the practical outcomes and impacts of using LLMs in medical diagnosis, including their impact on care, health outcomes, and telemedicine. These clusters are in line with other reported analysis and reviews, suggesting a high reliability [109,110].

We subsequently labeled the articles into the 3 clusters (overlapping articles remained in all mother clusters) for further analysis. Results are provided in Multimedia Appendix 2.

Figure 3. Bibliometric labeling of articles' keywords. LLM: large language model.



Quantitative Analysis of LLM Models Used in Current Research

As shown in Table 1, among the 95 included articles, GPT-4 and its variant models were studied in 70 (74%) studies, followed by GPT-3.5 and its variant models in 34 (36%) studies, specific developed GPTs in 9 (9%) studies, Claude and its variant models in 9 (9%) studies, Gemini and its variant models in 8 (8%) studies, and Llama and its variant models in 8 (8%) of the included studies. This suggests that ChatGPT is currently the most popular LLM used in medical diagnosis.

It is also worth noting that specifically developed GPT versions have also surged in numbers in recent years, indicating a translation from general to specific LLMs. Among the specifically developed GPT versions, we extracted 3 major paths for developing a specific GPT version, as described in Table 2 and Textbox 1.

Table 1. Qualitative analysis of large language models (LLMs) used in studies.

LLMs	Frequency, n
ChatGPT-4 (variant model included)	70
ChatGPT-3 (variant model included)	34
Specific developed GPT	9
Gemini	8
Llama	8
Claude	9
Others	5

Table 2. Quantitative analysis of ways of developing specific GPT versions in studies.

Ways of developing specifically developed GPT versions	Frequency, n
Prompt tuning	8
Fine-tuning	7
Pretrained from scratch	2

Textbox 1. Major paths in developing specific GPT versions.

Prompt tuning

- This involves using carefully crafted prompts to guide the model's output, often without modifying the model's weights.
- This method is often used to enhance the performance of existing large language models (LLMs) or to mitigate bias in LLMs [15,111,112].

Fine-tuning

- This involves using a subset of labeled data to adjust the model's weights, allowing it to better adapt to a specific task or domain [113].
- A typical example of this type of LLM is Med-PaLM 2, which is trained based on fine-tuning of Google PaLM [16].

Pretrained from scratch

- This involves using a large corpus of unlabeled text to train the model from the beginning, learning general language patterns and representations [114].
- A typical example of this type of LLM is BioGPT, which was pretrained on a corpus of PubMed articles from scratch [115].

Application and Evaluation of LLMs in Diagnosing Diseases

Overview

To assist our understanding of the overall scope of how performance of LLMs were examined, we further conducted a clustering analysis within the articles in cluster 1 and cluster 2 (nested bibliometric analysis). As shown in Figure 4, this suggested that the application and evaluation of LLMs can be mainly categorized into three aspects: (1) disease classification, (2) medical question answering, and (3) quality of generated diagnostic content (Tables 3 and 4).

Figure 4. (A) Nested bibliometric analysis of performance evaluation in studies; (B) quantitative analysis of the major assessment aspects; and (C) existing large language models' performance on accuracy. USMLE: United States Medical Licensing Examination.

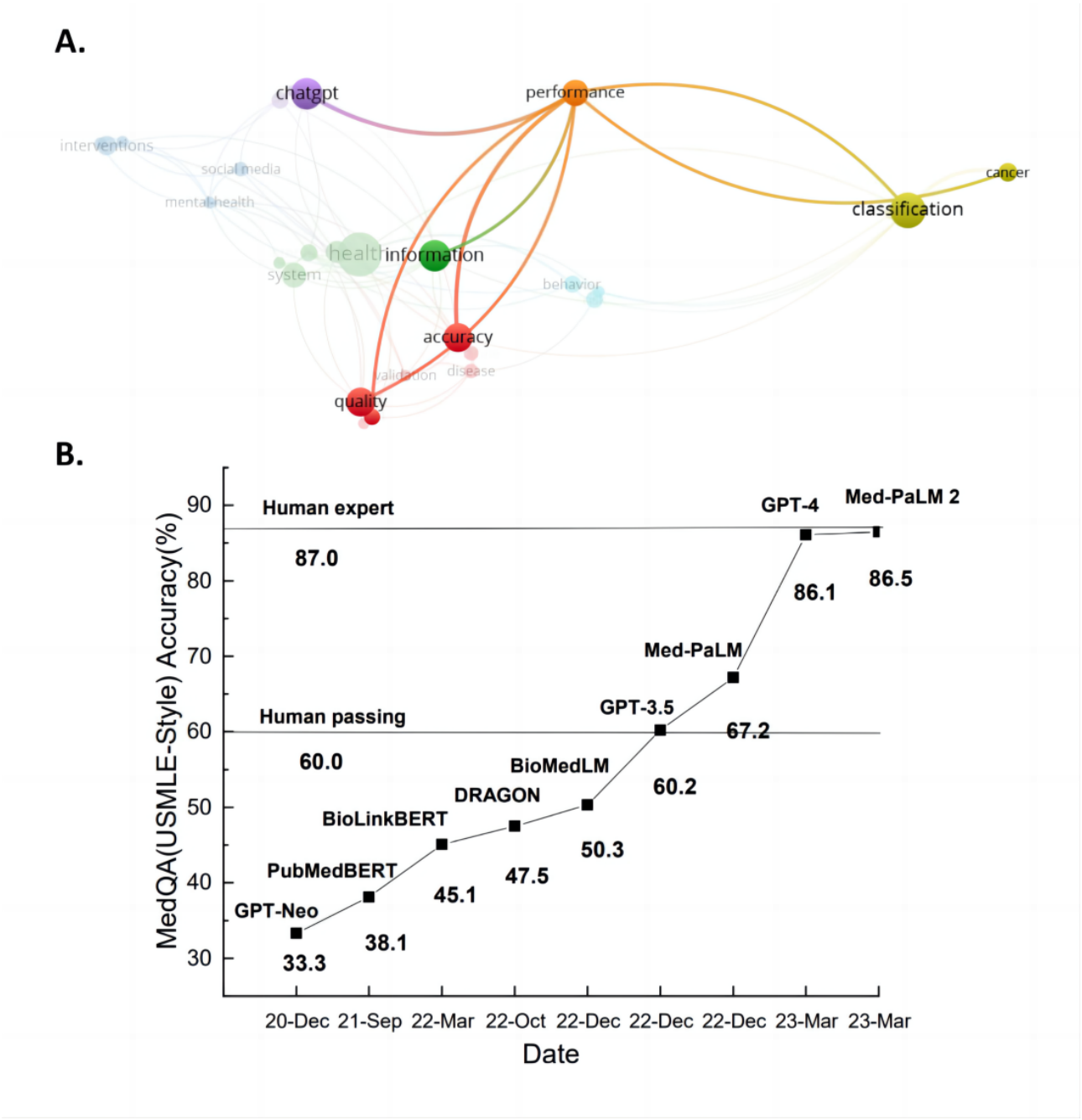


Table 3. Quantitative analysis of the major assessment aspects.

Assessment aspects	Frequency, n
Diagnosis accuracy	77
Disease classification	14
Quality of generated diagnosis content	6

Table 4. Detailed description of application.

Performance and evaluation domain and examples	Description	Reference
Disease classification		
Primary and secondary glaucoma	A total of 26 glaucoma cases were selected as a convenience sample; researchers compared the diagnostic performance of GPT-4o with that of ophthalmologists of varying experience levels, focusing on both primary and differential diagnoses of glaucoma.	[17]
Melanoma	Prompts were designed to either involve conditioning of asymmetry, border irregularity, color variation, diameter >6 mm, and evolution melanoma features or to assess effects of background skin color on predictions.	[18,116,117]
Diagnosis QA[a]		
MedQA answering	The MedQA dataset is a QA dataset oriented toward the medical field, emulating the style of the United States Medical Licensing Examination. It contains questions in English, simplified Chinese, and traditional Chinese, aiming to assess the model's understanding and reasoning abilities in medical knowledge.	[118]
PubMedQA answering	The PubMedQA dataset is a novel biomedical QA dataset collected from PubMed abstracts. It requires models to be capable of understanding and reasoning biomedical research texts, especially their quantitative content, to answer research questions.	[119]
MedMCQA answering	MedMCQA is a large-scale multiple-choice QA dataset specifically designed to address practical medical entrance examination problems. It contains over 194,000 high-quality multiple-choice questions from AIIMS[b] and NEET[c] PG[d] entrance exams, covering 2400 health care topics and 21 medical subjects.	[120]
DSM-5[e] Clinical Cases book	The book clarifies and discusses psychiatric diagnosis with a particular focus on how diagnoses have evolved from the *DSM-5*. It is commonly used to determine accuracy of mental health diagnosis.	[121]
NEJM[f] image challenge dataset	This dataset includes medical image challenges from the *NEJM*. It is often used to determine accuracy of vision diagnosis.	[122]
Quality of generated diagnostic content		
Responsiveness	Test to see if LLMs[g] can respond to every question, instead of giving "I'm sorry, I cannot provide medical diagnoses or interpret medical images."	[122]
Ethical bias	Critically evaluate LLMs in their generated diagnostic content for potential bias in gender and race	[19]
Cognitive bias	Use specifically developed QA sets that include clinically biased questions as compared to unbiased ones to test LLMs for their generated diagnostic content.	[15]

[a]QA: question and answer.

[b]AIIMS: All India Institute of Medical Sciences.

[c]NEET: National Eligibility cum Entrance Test.

[d]PG: postgraduate.

[e]DSM-5: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition.

[f]NEJM: New England Journal of Medicine.

[g]LLM: large language model.

Disease Classification

LLMs such as GPT-4 have shown significant potential in disease classification. Their ability to process and analyze vast amounts of textual data makes them valuable tools for classifying diseases. They can be categorized into text-based classification and image-based classification.

For text-based classification, LLMs have been successfully applied to detect and classify diseases, such as gout and calcium pyrophosphate deposition disease, from electronic health records (EHRs), outperforming traditional methods like regex-based approaches [123]. They achieved high classification accuracy and predictive values, demonstrating their capability to handle non-English medical documents effectively. In addition, in clinicopathological conferences, LLMs such as ChatGPT and Google Bard have been used to generate differential diagnoses for neurodegenerative disorders. While they correctly identified primary diagnoses in a significant number of cases, their inclusion of correct diagnoses in broader lists was even higher, indicating their utility in supporting clinical discussions [124]. All of these have shown LLMs' potential in analyzing unstructured text (eg, patient symptoms and medical histories) to suggest potential diagnoses, which could be used in public health sections for early detection and classification of diseases to reduce financial and physician's burden.

In addition to text-based classification, recent research have also explored the possibility of LLMs in classifying diseases based on image data. With the introduction of vision-language models, LLMs exhibit powerful representational learning capabilities, enabling them to comprehend, generate, and process various image data types, showcasing their application potential [125,126]. For example, in dermatology, LLMs are being trained to recognize and classify skin conditions from simple phone pictures, which can assist in early detection of diseases like melanoma. Alternatively, LLMs can be used to interpret professional medical images in radiology, such as X-rays [127,128] and MRIs [129]. These LLMs are being trained to identify abnormalities and assist in diagnosing conditions. The integration of LLMs in these areas offers several potential benefits. It can enhance the speed and accuracy of diagnosis, reduce the workload on health care professionals, and provide consistent interpretations that can be especially valuable in remote or underserved areas where access to specialized care is limited. Furthermore, LLMs can be trained to recognize subtle patterns and features that may be challenging for the human eye to detect, potentially leading to improved diagnostic performance. Furthermore, a noticeable trend has emerged within this subtopic, which is to combine other mechanisms with LLMs. For example, Wang et al [130] combined computer-aided diagnosis network with LLMs, which have shown amazing ability in reading chest X-rays. Similarly, Selivanov et al [127] combined a preset-attention mechanism with LLMs, which showcased efficient applicability to the chest X-ray image captioning task. This combination trend could be a promising future direction of LLMs development in medical image understanding.

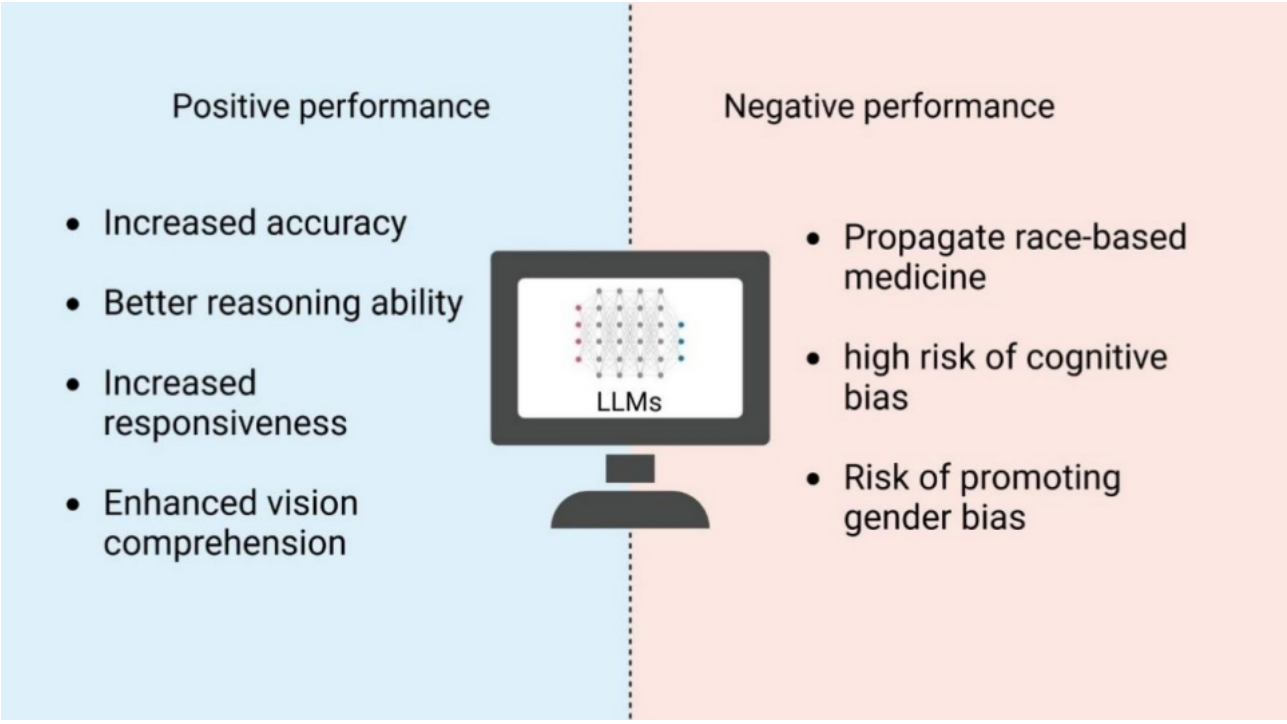
Medical Question Answering

On the basis of our results, it is clear that diagnosis accuracy dominates a crucial part of research. This consists of enhancing diagnosis accuracy of LLMs and exploring the diagnosis accuracy of LLMs in various medical domains.

One major indicator of diagnosis accuracy is medical-related questions and answers (QAs). Here, we listed the QA accuracy of LLM performance in Figure 4 and Multimedia Appendix 3 [20,131-138]. Models such as MedPaLM [20], GPT-4, and MedPaLM2 clearly indicate a trend that suggests a substantial increase in the accuracy of LLMs in answering questions in the MedQA dataset in a short period. In addition to the increased accuracy, a noticeable trend in developing medical task–specific LLMs for QAs was observed when combining the results of Figures 4 and 5. As previously discovered, the application of LLMs is constrained by their limited medical domain knowledge and the complexities of clinical tasks. For example, the performance of ChatGPT with human respondents in answering genetic questions was not significantly different from human respondents [139]. Given such observed limitations in medical knowledge, Li et al [140] introduced ChatDoctor, using the Llama model with an autonomous information retrieval mechanism. This allows real-time access and use of Wikipedia web-based resources, leading to a substantial enhancement in the quality of patient-physician interactive dialogue. The system has demonstrated notable progress in comprehending patient needs and offering precise treatment options. Yasunaga et al [141] developed DRAGON, a deep bidirectional language-knowledge graph pretraining method that enhances the model's reasoning ability by providing knowledge graph that complements text data, offering structured background knowledge. Therefore, DRAGON achieved a higher score in QA across general and biomedical domains.

However, as most LLMs are trained on English corpora, advanced LLMs often struggle to perform effectively in non–English language settings, such as Chinese medical question answering systems. To address this limitation, researchers have focused on developing and applying Chinese-specific LLMs and datasets. For instance, Xiong et al [142] created DoctorGLM, a large-scale language model trained on a comprehensive Chinese health care database. DoctorGLM features a prompt designer module that extracts relevant keywords from user input, uses potential disease names as labels, and generates detailed descriptions based on a disease knowledge library. Therefore, DoctorGLM can provide users with accurate and reliable information, including disease symptoms, diagnoses, treatment options, and preventive measures. In addition, Li et al [21] introduced AcupunctureGPT, a model that not only supports the Chinese language but also integrates data from traditional Chinese medicine (TCM). By combining TCM principles with modern technology, AcupunctureGPT enhances diagnostic accuracy through objective methods and offers valuable scientific insights into the efficacy of TCM practices. This fusion of traditional knowledge and advanced AI represents a significant step forward in bridging the gap between ancient medical practices and contemporary health care technologies.

Figure 5. Positive and negative performance of current large language models (LLMs). Increased accuracy means LLMs have evolved to provide more accurate diagnosis results. Better reasoning ability suggests LLMs can analyze complex symptoms and make logical deductions. Increased responsiveness shows LLMs can respond to a wider range of queries. Enhanced vision comprehension suggests LLMs can interpret and understand visual data, which is beneficial for tasks like image analysis. Propagating race-based medicine suggests that there is a risk that these models could reinforce or propagate biased practices in medicine, leading to unequal treatment. High risk of cognitive bias reveals that the models might inherit and amplify biases present in the data they were given, affecting decision-making processes. The risk of promoting gender bias is similar to racial bias, which could also be perpetuated, affecting how information is processed and decisions are made.



Quality of Generated Diagnostic Content

One major concern hindering LLMs' use in medical diagnosis is closely related to the technical and ethical concerns that emerged in the diagnoses content.

Different from laboratory questions or QAs that give a clear indication and connection between symptoms and diseases, real-world clinical scenarios are often complex and sometimes misleading [143]. Therefore, it not only requires LLMs to have a deep understanding of medical knowledge but also the ability to handle ambiguity and uncertainty. For example, patients may present with atypical symptoms or nonspecific concerns that do not directly point to a particular disease. LLMs need to be capable of considering a broad differential diagnosis. It has been found that LLMs produce less accurate responses when faced with clinically biased questions as compared to unbiased ones, suggesting their vulnerability in dealing with the complexity of real patient-physician interactions [15].

More importantly, ethical problems shown in the diagnosis content of LLMs have caused more concerns in the field. This includes possible bias and privacy leakage problems during the diagnosing process. The importance of fairness in LLMs has become increasingly recognized for ensuring stable performance and unbiased diagnoses decisions. Language models, including LLMs, are known to reflect and sometimes amplify biases present in the historical data they learn from, which can exacerbate existing inequalities in health care [19]. For example, Omiye et al [144] revealed that every LLM model has instances of promoting race-based medicine or racist tropes or repeating unsubstantiated claims around race. Similarly, Yang et al [145] showed race-based tendencies in both GPT-3.5-turbo and GPT-4 models, including generating biased patient backgrounds, associating diseases with specific demographic groups, favoring White patients in treatment recommendations, and showing disparities in projected cost, hospitalization duration, and prognosis [145]. Such biased models could negatively impact the quality of diagnoses that patients receive. In addition, the diverse data sources used to train LLMs, including biomedical and clinical texts, may contain sensitive personal information, posing serious privacy risks [146]. Research has shown that language models can inadvertently leak personal information in their generated content [147]. Therefore, it is crucial to develop and deploy LLMs in medical diagnosis with robust mechanisms to prevent bias and protect patient privacy to ensure equitable and ethical health care outcomes.

Quantitative Analysis of LLMs in Diagnostic Fields

The applications of LLMs in medical diagnosis varied and showed promise across several medical specialties. The frequency of LLM applications in different medical domains, as depicted in Table 5, indicates a high interest and potential in using these models for diagnostic purposes.

Table 5. Large language models' use by diagnostic domains.

Diagnostic domains	Frequency, n
Radiology	16
General medicine	14
Neurology	10
Ophthalmology	10
Psychiatry	7
Emergency medicine	5
Cardiology	4
Dermatology	4
Dentistry	4
Surgery	3
Others	17

Among all the medical domains, research has focused on radiology, which accounts for 16 (17%) studies. In radiology, the application of LLMs is mainly focused on image recognition, interpretation, and diagnosis. LLMs can effectively identify pathological areas and improve diagnostic accuracy [148]. At the same time, LLMs can also conduct risk prediction and formulate personalized treatment plans based on patients' medical imaging data [149]. In addition, through natural language generation technology, LLMs can also provide patients with understandable diagnostic reports and health guidance, which would aid the diagnostic process [150].

It is also worth noting that the application of LLMs on psychiatry diagnosis, including depression and attention-deficit/hyperactivity disorder are very promising [151-153]. Similarly, diseases of the nervous system also attract considerable attention, with studies covering disorders from predicting seizure to determining Alzheimer disease [154-156]. This might be because decision-making in psychiatry and neurology is particularly unique, as they have fewer objective tools that can be used to either confirm or refute a diagnosis [156]. Therefore, the emergence of LLMs can be an objective assessment tool that can help experts in decision-making process for diagnoses [121].

Clinical Statistical Analysis

Clinical trials related to LLMs were searched in ClinicalTrials.gov (March 17, 2025) as indicators for LLMs' real-world integration (Tables 6-8). After removing unrelated trials, 23 clinical trials were selected (Multimedia Appendix 4 provides detailed information on trials). These trials spanned across 11 medical specialties, with oncology (6/23, 26%) and ophthalmology (5/23, 22%) being the most extensively studied. All trials were funded by public sponsors and conducted in publicly available hospitals. None of the trials had reported their results by the date of search. Most trials (15/23, 65%) used LLMs as support in physicians whole diagnostic process, while some (6/23, 26%) investigated LLMs' potential in direct diagnosis of diseases conducted by a physician. We observed that LLMs used for direct diagnosis were centered in emergency medicine and radiology. These areas showed shortages in physicians, where LLMs served as aid to reduce the workload of physicians [157,158]. For example, LLMs' quick reasoning and responding makes them an ideal choice for grading emergency patients, and their enhanced vision understanding capabilities are valuable for generating radiology reports. It was found that ChatGPT and its variants still dominated the use in clinical trials (10/23, 43%), which is consistent with our previous findings. However, we observed a rise in using specifically develop LLMs in clinical settings compared with our previous findings (6/23, 26% vs. 9/95, 10%); this might be because the clinical use of LLMs was specialized in particular subject. Therefore, applying fine-tuning or prompt engineering method could enhance the usability of LLMs in expertized clinical scenario. In addition, it should be noted that certain trials did not report the LLMs used in the study or reported them without proper disclosure of their type (eg, only mentioned ChatGPT without its model types). Privacy concerns have driven innovations, such as locally deployed LLMs (ie, the ongoing trial registered as NCT06865534). These models minimize data transmission risks by processing information on-site, aligning with regulations like General Data Protection Regulation (GDPR). However, only this trial explicitly mentioned privacy safeguards, indicating a need for standardized reporting on data security measures in LLM research.

Table 6. Quantitative analysis of large language models (LLMs) used in clinical trials.

LLMs	Frequency (n=32), n (%)
ChatGPT	11 (34)
Gemini Pro	1 (3)
Marvin	2 (6)
Deepseek-Janus-Pro	1 (3)
Claude	4 (12)
Specifically developed LLM	6 (19)
Not reported	7 (22)

Table 7. The function of large language models (LLMs) in clinical trials. Note: the reason for having an intervention group is because one trial uses LLM both for diagnosis and intervention.

LLMs' function	Frequency (n=23), n (%)
Diagnosis support	16 (70)
Direct diagnosis	6 (26)
Model comparison	1 (4)

Table 8. Large language models' use in clinical trials by diagnostic domains.

Diagnostic domains	Frequency, n
Oncology	6
Ophthalmology	5
Cardiology	4
Pulmonology	3
General medicine	3
Gastroenterology	2
Neurology	2
Emergency medicine	2
Rheumatology	1
Radiology	1
Psychiatry	1

Discussion

Principal Findings

The field of LLMs in diagnosis has witnessed tremendous development in recent years, which is evident from our findings on study distribution across years. The results of our study contribute to a growing body of literature on the application of LLMs in health care. Previous scoping reviews have demonstrated that considerable research in the field of LLMs in health care is related to medical diagnosis [10]. In addition, there are existing literatures examining the current application of LLMs in diagnosis [159,160]. However, there is still a noticeable absence of a comprehensive analysis of LLMs in diagnosis to identify the gap between research and mapping the current research landscape of LLMs in medical diagnosis.

This is a pioneering study to quantitatively and comprehensively chart the integration of LLMs into the medical diagnosis domain. In this scoping review, we noticed a surge in publications regarding LLMs in diagnosis, signifying growth potential in this research field. This surge in research activity underscores the growing interest and potential impact of LLMs in transforming diagnostic processes in health care.

It is also worth noting that we performed a bibliometric analysis before our data extraction process. This was partly inspired by the bibliometric-systematic literature review approach proposed by Marzi et al [161]. In short, instead of gathering data directly from the 95 included articles, we performed a bibliometric analysis to cluster their keywords as indicators for determining research questions and data to extract. By using this technique, we were able to extract our needed information from a vast amount of data to support our findings. In addition, by performing a nested bibliometric analysis, this study was able to map the current landscape of LLM diagnostic performance evaluation, which provides valuable insight and is compatible with existing literature recommending a framework for future LLM performance evaluation [162].

Our review reveals that LLMs, particularly ChatGPT and its variants, show substantial promise in enhancing diagnostic accuracy and facilitating clinical decision-making. Traditional clinical decision support systems using decision trees increase users' accuracy in diagnosing diseases [163] but struggle with complex presentations requiring probabilistic reasoning [164]. In contrast, LLMs demonstrate a superior diagnostic accuracy rate with a clearer indication of reasoning. However, the hallucination problem in LLMs creates regulatory challenges for high-stakes applications [165]. Nevertheless, the ability of LLMs to process vast amounts of textual and visual data enables them to classify diseases effectively, as demonstrated in various medical specialties such as radiology, psychiatry, neurology, and dermatology. For instance, GPT-4 has been successfully applied to detect and classify diseases from EHRs and medical images. Currently, the integration of LLMs with other technologies, such as computer-aided diagnosis networks and attention mechanisms, creates hybrid systems that combine the strengths of both approaches [130], which further enhances their diagnostic capabilities and suggests a promising direction for LLM development.

The rapid advancement in the accuracy of LLMs in medical QA tasks is another notable achievement. Models like MedPaLM, GPT-4, and MedPaLM2 have shown significant improvements in their ability to provide accurate and relevant answers to medical queries. This progress is partly attributed to the development of medical task–specific LLMs, which are fine-tuned on specialized datasets to enhance their domain knowledge and reasoning abilities. For example, ChatDoctor and AcupunctureGPT, which incorporate real-time information retrieval and traditional Chinese medicine principles, respectively, demonstrate the potential of LLMs to adapt to diverse medical contexts and improve diagnostic outcomes.

Challenges Ahead

Despite promising advancements, several critical challenges remain. One of the most significant concerns is the potential for bias in LLMs, which can lead to inequitable and unethical health care outcomes. Our review highlights that LLMs are susceptible to reflecting and amplifying biases present in their training data, resulting in biased diagnostic recommendations and disparities in patient care. For instance, studies have shown that LLMs can exhibit race-based biases in medical diagnosis and patient information processing. Specifically, LLMs have been found to generate biased patient backgrounds and associate diseases with specific racial or ethnic groups. For example, GPT-3.5-turbo has been shown to attribute unwarranted details to patients based on their race or ethnicity, such as associating Black male patients with a safari trip in South Africa, and varying its diagnoses for different racial and ethnic groups even under identical conditions (eg, diagnosing HIV in Black patients, tuberculosis in Asian patients, and cysts in White patients) [145]. In addition, LLMs may project higher costs and longer hospitalizations for certain racial groups. For example, GPT-3.5-turbo predicts higher costs and longer hospital stays for White patients compared to other racial and ethnic groups, potentially reflecting real-world health care disparities. Furthermore, LLMs have been found to show overly optimistic outcomes in challenging medical scenarios with higher survival rates for some groups compared to others, which may negatively impact the quality of diagnoses and treatment decisions [144]. It should also be noted that apart from the race and gender bias that were directly reported by researchers studying medical diagnosis, LLMs had also exhibited bias in patient care regarding age, social status, and income level. These are also possible hidden bias that might affect the decision-making of LLMs in medical diagnosis.

Addressing these biases is crucial to ensure that LLMs provide fair and accurate diagnostic support to all patients. Currently, the most widely adopted method for bias mitigation is through prompt tuning. By carefully designing and adjusting the prompts used to query the LLMs, it is possible to guide the models toward more equitable and accurate responses. For example, researchers have found that using prompts that explicitly emphasize fairness and inclusivity can reduce the likelihood of biased outputs [166]. In addition, incorporating counternarratives or examples that challenge stereotypes within the prompts can help counteract the biases that may be present in the training data [167]. However, such methods only provide a shield to prevent the occurrence of bias in diagnosis without changing its core. Reinforcement learning from human feedback is another strategy that allows humans to grade the model's responses, which helps correct some model outputs, particularly on sensitive questions with known online misinformation. However, such a method costs additional computing and could result in a more ambiguous answer in diagnosis. For example, the racial bias in cost prediction for GPT-4 has been reduced from 18% in GPT-3.5 to 5%, but the "uncertainty rate" increased from 16.25% to 29.46% [145]. Therefore, future research in finding a suitable way fully eradicate biased medical diagnosis, and constant monitoring of LLMs bias performance are important for the application of LLM in medical diagnosis.

In addition, the complexity of real-world clinical scenarios presents a challenge for LLMs. From a technical perspective, the first hurdle for LLMs deployment in real-world clinical settings is the ambiguity and uncertainty in unstructured inputs [15]. Unlike structured QA tasks, patient concerns (eg, "chest pain") may correspond to dozens of potential diseases. LLMs must generate reasonable differential diagnoses despite lacking definitive information. Additional hurdles include the need for dynamic interaction and real-time feedback. LLMs in clinical consultations require interactive information gathering in follow-up history and test results, among others. Current LLMs' limited context windows (eg, GPT-4's 32,000 tokens) may lead to critical information loss. In addition, dynamically synthesized laboratory results, imaging reports, and other multimodal data are required for an accurate and reasonable diagnosis, but cross-modal alignment techniques remain underdeveloped in LLMs. Therefore, ensuring that LLMs can effectively navigate these complexities without compromising diagnostic accuracy is a critical area for future research.

Beyond technical challenges, the clinical application of LLMs faces multifaceted nontechnological barriers, including fragmented regulatory frameworks, physician skepticism, data privacy risks, ethical dilemmas, and regulatory inconsistencies across regions, such as the United States. The requirement of software as a medical device for consistent and reproducible

results for medical software contradicts LLMs' nature as products generating differentiated content for each answer. In addition, the United States' requirement for predefined change control plans for adaptive algorithms, the European Union's stringent transparency and data traceability mandates under the AI Act and the GDPR, and China's life cycle quality control protocols create complex compliance landscapes that delay global deployment [168,169]. Clinician trust is further eroded by the black box nature of LLMs, concerns over data representativeness (eg, biases arising from existing studies), and ambiguous liability frameworks for errors [170]. Data privacy remains a critical hurdle, as LLMs rely on sensitive patient information while navigating strict anonymization rules (eg, the GDPR) and cross-border data flow restrictions. Current LLMs in medical diagnosis are still mainly commercially available models, with only a few reports on local deployment of LLMs, which raises the concern of patient data leakage. Ethical risks, such as algorithmic biases in gender or race and inadequate patient consent processes, compound public distrust, particularly among vulnerable populations [19]. These challenges underscore the need for international regulatory harmonization, enhanced algorithmic transparency through adversarial testing, and collaborative ethical governance to balance innovation with patient safety and trust in real-world clinical settings.

Future Directions

Overview

Building on the current advancements and challenges identified in this review, we propose a structured research road map to guide the next phase of LLM development in medical diagnosis. In total, 4 critical domains warrant prioritized investigation.

Multimodal AI Integration for Holistic Diagnostics

Future studies should focus on integrating text, imaging, and structured clinical data (eg, laboratory results and genomic profiles) into unified multimodal frameworks. Although LLMs like GPT-4 have shown some success in processing EHRs and radiology reports, their ability to dynamically synthesize diverse data streams is still underdeveloped. Key priorities include developing cross-modal alignment techniques to harmonize semantic relationships between textual symptoms, imaging findings (eg, MRI anomalies), and biomarker patterns; creating benchmark datasets that reflect real-world clinical complexity, such as longitudinal patient records with asynchronous laboratory and imaging updates; and investigating hybrid architectures that combine LLMs with computer vision systems (eg, vision transformers) for simultaneous analysis of dermatology images and symptom narratives.

Enhancing Trust Through Explainable AI and Clinical Validation

To address physician skepticism and regulatory concerns, research must bridge the interpretability gap by developing chain-of-reasoning frameworks that visualize LLM diagnostic pathways, explicitly linking symptom inputs to disease hypotheses through intermediate evidence (eg, "Elevated CRP → infection likelihood → differential weighting of tuberculosis vs lymphoma"); implementing adversarial testing protocols to quantify model uncertainty in ambiguous presentations (eg, differentiating psychosomatic vs organic causes of chest pain); and conducting large-scale prospective trials comparing LLM-assisted versus conventional diagnostic workflows across institutions, with metrics including time-to-diagnosis, cost efficiency, and diagnostic error rates.

Specialty-Specific Optimization and Cross-Disciplinary Comparisons

The heterogeneous performance of LLMs across medical domains necessitates targeted investigations, including establishing specialty-specific evaluation benchmarks (eg, psychiatry: symptom trajectory modeling and oncology: rare cancer detection in pathology reports); conducting comparative studies analyzing performance variance across disciplines; and exploring transfer learning paradigms where diagnostic patterns learned in data-rich specialties (eg, radiology) inform models for underserved domains (eg, tropical medicine).

Ethical and Regulatory Harmonization

Parallel technical efforts must address systemic barriers to clinical implementation by developing international standards for bias auditing that require LLMs to demonstrate less demographic variance in diagnostic accuracy across protected attributes, such as race, gender, and socioeconomic status; creating federated learning infrastructures that enable secure multi-institutional training while complying with regulations like the GDPR and the Health Insurance Portability and Accountability Act through techniques such as differential privacy; and proposing adaptive regulatory frameworks that balance the probabilistic nature of LLM outputs with medical device safety requirements, potentially introducing real-time clinician oversight protocols for high-stakes diagnoses.

This road map emphasizes translational research that bridges technical innovation with clinical pragmatism. By focusing on multimodal integration, explainable reasoning, specialty-specific validation, and ethical governance, the field can transition from demonstrating diagnostic potential to delivering measurable improvements in patient outcomes. Collaborative efforts between AI researchers, clinicians, and policy makers will be essential to realize LLMs' transformative potential while maintaining rigorous standards of care.

Limitations of the Scoping Review Process

The scoping review process, while comprehensive, has inherent limitations. First, our reliance on selected databases might have excluded pertinent publications from nonindexed sources, for example, arXiv. Therefore, the source for this scoping review is limited. Despite that, we have tried to include articles from other sources by screening the citation of included studies; however, certain valuable gray literature or ongoing research may be omitted. In addition, this review lacked longitudinal studies evaluating the long-term clinical impact of LLMs, such as sustained diagnostic accuracy, clinician reliance, and patient outcomes over extended periods. This gap restricts conclusions about the durability of LLM performance and its integration into dynamic health care workflows. Moreover, because research into LLMs in medical diagnosis is only beginning, the rapid evolution of LLM technology introduces a temporal limitation:

studies published after January 2025 or advancements in multimodal architectures may already outpace findings in this review. In addition, due to the expected heterogeneity in tasks and end points, we did not conduct formal meta-analyses or other statistical analysis. Instead, we presented simple descriptive statistics to provide an overview of the features of the LLMs' landscape in medical diagnosis. Future updates with statistical analysis are critical to maintaining relevance in this fast-moving field.

Conclusions

In conclusion, this scoping review highlights the significant potential of LLMs to revolutionize medical diagnostics, while also emphasizing the critical need to address biases, privacy concerns, and the complexities of real-world clinical scenarios. By focusing on these areas, future research can pave the way for the successful integration of LLMs into health care systems, ultimately improving diagnostic accuracy and patient outcomes.

Acknowledgments

This work was supported by the Science Foundation For Excellent Young Scholars of Hunan Province, China (grant 2024JJ4091); Young Scientists Fund of the National Natural Science Foundation of China (grant 82301835); Science Foundation for Young Scholars of Hunan Province, China (grant 2023JJ40956); China Postdoctoral Science Foundation (general program; grant 2022M713522); China Postdoctoral Foundation (grant 2021TQ0372); Young Investigator Grant of Xiangya Hospital, Central South University (grant 2021Q03); Fundamental Research Funds for the Central Universities of Central South University (grant 105332022249); National Natural Science Foundation of China (general program; grant 82371682); and Innovation and Entrepreneurship Training Program for College Students (grant S202410533307).

Data Availability

The datasets used and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

HS was responsible for conceptualization, writing—original draft, and software. YS was responsible for conceptualization, writing—original draft, writing—review and editing, and methodology. RL was responsible for data curation and formal analysis. AZ was responsible for visualization. YY was responsible for data curation. FX was responsible for writing—review and editing. ZD was responsible for writing—review and editing. JC was responsible for writing—review and editing. QH was responsible for writing—review and editing. YL was responsible for writing—original draft. BX was responsible for writing—review and editing. QZ was responsible for writing—review and editing. JZ was responsible for writing—review and editing. YL was responsible for resources. HL was responsible for funding acquisition, project administration supervision, and validation.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Search strategy.
[DOCX File , 16 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Characteristics of included studies.
[DOCX File , 43 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Detailed data of large language models included in the accuracy results.
[DOCX File , 15 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Characteristics of included clinical trials.
[DOCX File , 21 KB-Multimedia Appendix 4]

Multimedia Appendix 5

PRISMA-ScR checklist.
[DOCX File , 25 KB-Multimedia Appendix 5]

References

1. Kaur S, Singla J, Nkenyereye L, Jha S, Prashar D, Joshi GP, et al. Medical diagnostic systems using artificial intelligence (AI) algorithms: principles and perspectives. IEEE Access. 2020;8:228049-228069. [doi: 10.1109/access.2020.3042273]
2. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology. Mar 2018;286(3):800-809. [doi: 10.1148/radiol.2017171920] [Medline: 29309734]
3. Raiaan MA, Mukta MS, Fatema K, Fahad NM, Sakib S, Mim MM. A review on large language models: architectures, applications, taxonomies, open issues and challenges. IEEE Access. 2024;12:26839-26874. [doi: 10.1109/access.2024.3365742]
4. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med. Aug 17, 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
5. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med. Jan 7, 2019;25(1):24-29. [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
6. Liu X, Liu H, Yang G, Jiang Z, Cui S, Zhang Z, et al. A generalist medical language model for disease diagnosis assistance. Nat Med. Mar 08, 2025;31(3):932-942. [doi: 10.1038/s41591-024-03416-6] [Medline: 39779927]
7. Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: a systematic review and meta-analysis. J Biomed Inform. Mar 2024;151:104620. [FREE Full text] [doi: 10.1016/j.jbi.2024.104620] [Medline: 38462064]
8. Sandmann S, Riepenhausen S, Plagwitz L, Varghese J. Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. Nat Commun. Mar 06, 2024;15(1):2050. [FREE Full text] [doi: 10.1038/s41467-024-46411-8] [Medline: 38448475]
9. Tang L, Sun Z, Idnay B, Nestor JG, Soroush A, Elias PA, et al. Evaluating large language models on medical evidence summarization. NPJ Digit Med. Aug 24, 2023;6(1):158. [FREE Full text] [doi: 10.1038/s41746-023-00896-7] [Medline: 37620423]
10. Meng X, Yan X, Zhang K, Liu D, Cui X, Yang Y, et al. The application of large language models in medicine: a scoping review. iScience. May 17, 2024;27(5):109713. [FREE Full text] [doi: 10.1016/j.isci.2024.109713] [Medline: 38746668]
11. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel). Mar 19, 2023;11(6):887. [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
12. Gödde D, Nöhl S, Wolf C, Rupert Y, Rimkus L, Ehlers J, et al. A SWOT (strengths, weaknesses, opportunities, and threats) analysis of ChatGPT in the medical literature: concise review. J Med Internet Res. Nov 16, 2023;25:e49368. [FREE Full text] [doi: 10.2196/49368] [Medline: 37865883]
13. Ninkov A, Frank JR, Maggio LA. Bibliometrics: methods for studying academic publishing. Perspect Med Educ. Jun 16, 2022;11(3):173-176. [FREE Full text] [doi: 10.1007/s40037-021-00695-4] [Medline: 34914027]
14. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. Oct 02, 2018;169(7):467-473. [doi: 10.7326/m18-0850]
15. Schmidgall S, Harris C, Essien I, Olshvang D, Rahman T, Kim JW, et al. Evaluation and mitigation of cognitive biases in medical language models. NPJ Digit Med. Oct 21, 2024;7(1):295. [FREE Full text] [doi: 10.1038/s41746-024-01283-6] [Medline: 39433945]
16. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, et al. Toward expert-level medical question answering with large language models. Nat Med. Mar 2025;31(3):943-950. [doi: 10.1038/s41591-024-03423-7] [Medline: 39779926]
17. Zhang J, Ma Y, Zhang R, Chen Y, Xu M, Rina S, et al. A comparative study of GPT-4o and human ophthalmologists in glaucoma diagnosis. Sci Rep. Dec 05, 2024;14(1):30385. [FREE Full text] [doi: 10.1038/s41598-024-80917-x] [Medline: 39639068]
18. Liu X, Duan C, Kim MK, Zhang L, Jee E, Maharjan B, et al. Claude 3 Opus and ChatGPT with GPT-4 in dermoscopic image analysis for melanoma diagnosis: comparative performance analysis. JMIR Med Inform. Aug 06, 2024;12:e59273. [FREE Full text] [doi: 10.2196/59273] [Medline: 39106482]
19. Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health. Jan 2024;6(1):e12-e22. [doi: 10.1016/s2589-7500(23)00225-x]
20. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. Aug 12, 2023;620(7972):172-180. [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
21. Li S, Tan W, Zhang C, Li J, Ren H, Guo Y, et al. Taming large language models to implement diagnosis and evaluating the generation of LLMs at the semantic similarity level in acupuncture and moxibustion. Expert Syst Appl. Mar 2025;264:125920. [doi: 10.1016/j.eswa.2024.125920]
22. Beaulieu-Jones BR, Berrigan MT, Shah S, Marwaha JS, Lai SL, Brat GA. Evaluating capabilities of large language models: performance of GPT-4 on surgical knowledge assessments. Surgery. Apr 2024;175(4):936-942. [doi: 10.1016/j.surg.2023.12.014] [Medline: 38246839]

23. Liu J, Liang X, Fang D, Zheng J, Yin C, Xie H, et al. The diagnostic ability of GPT-3.5 and GPT-4.0 in surgery: comparative analysis. J Med Internet Res. Sep 10, 2024;26:e54985. [FREE Full text] [doi: 10.2196/54985] [Medline: 39255016]
24. Palenzuela DL, Mullen JT, Phitayakorn R. AI versus MD: evaluating the surgical decision-making accuracy of ChatGPT-4. Surgery. Aug 2024;176(2):241-245. [doi: 10.1016/j.surg.2024.04.003] [Medline: 38769038]
25. Collado-Montañez J, Martín-Valdivia MT, Martínez-Cámara E. Data augmentation based on large language models for radiological report classification. Knowl Based Syst. Jan 2025;308:112745. [doi: 10.1016/j.knosys.2024.112745]
26. Horiuchi D, Tatekawa H, Oura T, Shimono T, Walston SL, Takita H, et al. ChatGPT's diagnostic performance based on textual vs. visual information compared to radiologists' diagnostic performance in musculoskeletal radiology. Eur Radiol. Jan 2025;35(1):506-516. [doi: 10.1007/s00330-024-10902-5] [Medline: 38995378]
27. Huang J, Yang R, Huang X, Zeng K, Liu Y, Luo J, et al. Feasibility of large language models for CEUS LI-RADS categorization of small liver nodules in patients at risk for hepatocellular carcinoma. Front Oncol. 2024;14:1513608. [FREE Full text] [doi: 10.3389/fonc.2024.1513608] [Medline: 39744002]
28. Huppertz MS, Siepmann R, Topp D, Nikoubashman O, Yüksel C, Kuhl CK, et al. Revolution or risk?-Assessing the potential and challenges of GPT-4V in radiologic image interpretation. Eur Radiol. Mar 2025;35(3):1111-1121. [doi: 10.1007/s00330-024-11115-6] [Medline: 39422726]
29. Kikuchi T, Nakao T, Nakamura Y, Hanaoka S, Mori H, Yoshikawa T. Toward improved radiologic diagnostics: investigating the utility and limitations of GPT-3.5 turbo and GPT-4 with quiz cases. AJNR Am J Neuroradiol. Oct 03, 2024;45(10):1506-1511. [doi: 10.3174/ajnr.A8332] [Medline: 38719605]
30. Liu C, Wei M, Qin Y, Zhang M, Jiang H, Xu J, et al. Harnessing large language models for structured reporting in breast ultrasound: a comparative study of Open AI (GPT-4.0) and Microsoft Bing (GPT-4). Ultrasound Med Biol. Nov 2024;50(11):1697-1703. [doi: 10.1016/j.ultrasmedbio.2024.07.007] [Medline: 39138026]
31. Mitsuyama Y, Tatekawa H, Takita H, Sasaki F, Tashiro A, Oue S, et al. Comparative analysis of GPT-4-based ChatGPT's diagnostic performance with radiologists using real-world radiology reports of brain tumors. Eur Radiol. Apr 2025;35(4):1938-1947. [doi: 10.1007/s00330-024-11032-8] [Medline: 39198333]
32. Mori Y, Izumiyama T, Kanabuchi R, Mori N, Aizawa T. Large language model may assist diagnosis of SAPHO syndrome by bone scintigraphy. Mod Rheumatol. Aug 20, 2024;34(5):1043-1046. [doi: 10.1093/mr/road115] [Medline: 38153762]
33. Nakaura T, Yoshida N, Kobayashi N, Shiraishi K, Nagayama Y, Uetani H, et al. Preliminary assessment of automated radiology report generation with generative pre-trained transformers: comparing results to radiologist-generated reports. Jpn J Radiol. Feb 2024;42(2):190-200. [FREE Full text] [doi: 10.1007/s11604-023-01487-y] [Medline: 37713022]
34. Reith TP, D'Alessandro DM, D'Alessandro MP. Capability of multimodal large language models to interpret pediatric radiological images. Pediatr Radiol. Sep 2024;54(10):1729-1737. [doi: 10.1007/s00247-024-06025-0] [Medline: 39133401]
35. Silva TP, Andrade-Bortoletto MF, Ocampo TS, Alencar-Palha C, Bornstein MM, Oliveira-Santos C, et al. Performance of a commercially available Generative Pre-trained Transformer (GPT) in describing radiolucent lesions in panoramic radiographs and establishing differential diagnoses. Clin Oral Investig. Mar 09, 2024;28(3):204. [FREE Full text] [doi: 10.1007/s00784-024-05587-5] [Medline: 38459362]
36. Song M, Wang J, Yu Z, Wang J, Yang L, Lu Y, et al. PneumoLLM: harnessing the power of large language model for pneumoconiosis diagnosis. Med Image Anal. Oct 2024;97:103248. [doi: 10.1016/j.media.2024.103248] [Medline: 38941859]
37. Sonoda Y, Kurokawa R, Nakamura Y, Kanzawa J, Kurokawa M, Ohizumi Y, et al. Diagnostic performances of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro in "Diagnosis Please" cases. Jpn J Radiol. Nov 2024;42(11):1231-1235. [doi: 10.1007/s11604-024-01619-y] [Medline: 38954192]
38. Suh PS, Shim WH, Suh CH, Heo H, Park CR, Eom HJ, et al. Comparing diagnostic accuracy of radiologists versus GPT-4V and Gemini Pro vision using image inputs from diagnosis please cases. Radiology. Jul 2024;312(1):e240273. [doi: 10.1148/radiol.240273] [Medline: 38980179]
39. Sun SH, Huynh K, Cortes G, Hill R, Tran J, Yeh L, et al. Testing the ability and limitations of ChatGPT to generate differential diagnoses from transcribed radiologic findings. Radiology. Oct 2024;313(1):e232346. [doi: 10.1148/radiol.232346] [Medline: 39404623]
40. Valdez D, Bunnell A, Lim SY, Sadowski P, Shepherd JA. Performance of progressive generations of GPT on an exam designed for certifying physicians as certified clinical densitometrists. J Clin Densitom. 2024;27(2):101480. [doi: 10.1016/j.jocd.2024.101480] [Medline: 38401238]
41. Santos W, Yoon S, Paraboni I. Mental health prediction from social media text using mixture of experts. IEEE Latin Am Trans. Jun 2023;21(6):723-729. [doi: 10.1109/TLA.2023.10172137]
42. Franco D'Souza R, Amanullah S, Mathew M, Surapaneni KM. Appraising the performance of ChatGPT in psychiatry using 100 clinical case vignettes. Asian J Psychiatr. Nov 2023;89:103770. [doi: 10.1016/j.ajp.2023.103770] [Medline: 37812998]
43. Gargari OK, Fatehi F, Mohammadi I, Firouzabadi SR, Shafiee A, Habibi G. Diagnostic accuracy of large language models in psychiatry. Asian J Psychiatr. Oct 2024;100:104168. [doi: 10.1016/j.ajp.2024.104168] [Medline: 39111087]
44. Kim J, Leonte KG, Chen ML, Torous JB, Linos E, Pinto A, et al. Large language models outperform mental and medical health care professionals in identifying obsessive-compulsive disorder. NPJ Digit Med. Jul 19, 2024;7(1):193. [FREE Full text] [doi: 10.1038/s41746-024-01181-x] [Medline: 39030292]

45. Levkovich I, Rabin E, Brann M, Elyoseph Z. Large language models outperform general practitioners in identifying complex cases of childhood anxiety. Digit Health. 2024;10:20552076241294182. [FREE Full text] [doi: 10.1177/20552076241294182] [Medline: 39687523]
46. Pavez J, Allende H. A hybrid system based on bayesian networks and deep learning for explainable mental health diagnosis. Appl Sci. Sep 14, 2024;14(18):8283. [doi: 10.3390/app14188283]
47. Shin D, Kim H, Lee S, Cho Y, Jung W. Using large language models to detect depression from user-generated diary text data as a novel approach in digital mental health screening: instrument validation study. J Med Internet Res. Sep 18, 2024;26:e54617. [FREE Full text] [doi: 10.2196/54617] [Medline: 39292502]
48. Gianola S, Bargeri S, Castellini G, Cook C, Palese A, Pillastrini P, et al. Performance of ChatGPT compared to clinical practice guidelines in making informed decisions for lumbosacral radicular pain: a cross-sectional study. J Orthop Sports Phys Ther. Mar 2024;54(3):222-228. [doi: 10.2519/jospt.2024.12151] [Medline: 38284363]
49. Young CC, Enichen E, Rivera C, Auger CA, Grant N, Rao A, et al. Diagnostic accuracy of a custom large language model on rare pediatric disease case reports. Am J Med Genet A. Feb 2025;197(2):e63878. [doi: 10.1002/ajmg.a.63878] [Medline: 39268988]
50. Akhondi-Asl A, Yang Y, Luchette M, Burns JP, Mehta NM, Geva A. Comparing the quality of domain-specific versus general language models for artificial intelligence-generated differential diagnoses in PICU patients. Pediatr Crit Care Med. Jun 01, 2024;25(6):e273-e282. [doi: 10.1097/PCC.0000000000003468] [Medline: 38329382]
51. Apornvirat S, Thinpanja W, Damrongkiet K, Benjakul N, Laohawetwanit T. Comparing customized ChatGPT and pathology residents in histopathologic description and diagnosis of common diseases. Ann Diagn Pathol. Dec 2024;73:152359. [doi: 10.1016/j.anndiagpath.2024.152359] [Medline: 38972166]
52. Teixeira-Marques F, Medeiros N, Nazaré F, Alves S, Lima N, Ribeiro L, et al. Exploring the role of ChatGPT in clinical decision-making in otorhinolaryngology: a ChatGPT designed study. Eur Arch Otorhinolaryngol. Apr 2024;281(4):2023-2030. [doi: 10.1007/s00405-024-08498-z] [Medline: 38345613]
53. Mayo-Yáñez M, González-Torres L, Saibene AM, Allevi F, Vaira LA, Maniaci A, et al. Application of ChatGPT as a support tool in the diagnosis and management of acute bacterial tonsillitis. Health Technol. Apr 11, 2024;14(4):773-779. [doi: 10.1007/s12553-024-00858-3]
54. Maniaci A, Chiesa-Estomba CM, Lechien JR. ChatGPT-4 consistency in interpreting laryngeal clinical images of common lesions and disorders. Otolaryngol Head Neck Surg. Oct 2024;171(4):1106-1113. [doi: 10.1002/ohn.897] [Medline: 39045737]
55. Li J, Gao X, Dou T, Gao Y, Li X, Zhu W. Quantitative evaluation of GPT-4's performance on US and Chinese osteoarthritis treatment guideline interpretation and orthopaedic case consultation. BMJ Open. Dec 30, 2024;14(12):e082344. [FREE Full text] [doi: 10.1136/bmjopen-2023-082344] [Medline: 39806703]
56. Pagano S, Holzapfel S, Kappenschneider T, Meyer M, Maderbacher G, Grifka J, et al. Arthrosis diagnosis and treatment recommendations in clinical practice: an exploratory investigation with the generative AI model GPT-4. J Orthop Traumatol. Nov 28, 2023;24(1):61. [FREE Full text] [doi: 10.1186/s10195-023-00740-4] [Medline: 38015298]
57. Wilhelm TI, Roos J, Kaczmarczyk R. Large language models for therapy recommendations across 3 clinical specialties: comparative study. J Med Internet Res. Oct 30, 2023;25:e49324. [FREE Full text] [doi: 10.2196/49324] [Medline: 37902826]
58. Chen X, Zhang W, Zhao Z, Xu P, Zheng Y, Shi D, et al. ICGA-GPT: report generation and question answering for indocyanine green angiography images. Br J Ophthalmol. Sep 20, 2024;108(10):1450-1456. [doi: 10.1136/bjo-2023-324446] [Medline: 38508675]
59. Delsoz M, Madadi Y, Raja H, Munir WM, Tamm B, Mehravaran S, et al. Performance of ChatGPT in diagnosis of corneal eye diseases. Cornea. May 01, 2024;43(5):664-670. [doi: 10.1097/ICO.0000000000003492] [Medline: 38391243]
60. Hu X, Ran AR, Nguyen TX, Szeto S, Yam JC, Chan CK, et al. What can GPT-4 do for diagnosing rare eye diseases? a pilot study. Ophthalmol Ther. Dec 2023;12(6):3395-3402. [FREE Full text] [doi: 10.1007/s40123-023-00789-8] [Medline: 37656399]
61. Mikhail D, Mihalache A, Huang RS, Khairy T, Popovic MM, Milad D, et al. Performance of ChatGPT in French language analysis of multimodal retinal cases. J Fr Ophtalmol. Mar 2025;48(3):104391. [doi: 10.1016/j.jfo.2024.104391] [Medline: 39708623]
62. Ming S, Yao X, Guo X, Guo Q, Xie K, Chen D, et al. Performance of ChatGPT in ophthalmic registration and clinical diagnosis: cross-sectional study. J Med Internet Res. Nov 14, 2024;26:e60226. [FREE Full text] [doi: 10.2196/60226] [Medline: 39541581]
63. Rojas-Carabali W, Sen A, Agarwal A, Tan G, Cheung CY, Rousselot A, et al. Chatbots vs. Human experts: evaluating diagnostic performance of chatbots in uveitis and the perspectives on AI adoption in ophthalmology. Ocul Immunol Inflamm. Oct 2024;32(8):1591-1598. [doi: 10.1080/09273948.2023.2266730] [Medline: 37831553]
64. Xu P, Chen X, Zhao Z, Shi D. Unveiling the clinical incapabilities: a benchmarking study of GPT-4V(ision) for ophthalmic multimodal image analysis. Br J Ophthalmol. Sep 20, 2024;108(10):1384-1389. [doi: 10.1136/bjo-2023-325054] [Medline: 38789133]

65. Yang Z, Wang D, Zhou F, Song D, Zhang Y, Jiang J, et al. Understanding natural language: potential application of large language models to ophthalmology. Asia Pac J Ophthalmol (Phila). 2024;13(4):100085. [FREE Full text] [doi: 10.1016/j.apjo.2024.100085] [Medline: 39059558]
66. Zandi R, Fahey JD, Drakopoulos M, Bryan JM, Dong S, Bryar PJ, et al. Exploring diagnostic precision and triage proficiency: a comparative study of GPT-4 and bard in addressing common ophthalmic complaints. Bioengineering (Basel). Jan 26, 2024;11(2):120. [FREE Full text] [doi: 10.3390/bioengineering11020120] [Medline: 38391606]
67. Kaiser P, Yang S, Bach M, Breit C, Mertz K, Stieltjes B, et al. The interaction of structured data using openEHR and large Language models for clinical decision support in prostate cancer. World J Urol. Jan 13, 2025;43(1):67. [doi: 10.1007/s00345-024-05423-1] [Medline: 39804478]
68. Kozel G, Gurses ME, Gecici NN, Gökalp E, Bahadir S, Merenzon MA, et al. Chat-GPT on brain tumors: an examination of Artificial Intelligence/Machine Learning's ability to provide diagnoses and treatment plans for example neuro-oncology cases. Clin Neurol Neurosurg. Apr 2024;239:108238. [doi: 10.1016/j.clineuro.2024.108238] [Medline: 38507989]
69. Kumar RP, Sivan V, Bachir H, Sarwar SA, Ruzicka F, O'Malley GR, et al. Can artificial intelligence mitigate missed diagnoses by generating differential diagnoses for neurosurgeons? World Neurosurg. Jul 2024;187:e1083-e1088. [doi: 10.1016/j.wneu.2024.05.052] [Medline: 38759788]
70. Ward M, Unadkat P, Toscano D, Kashanian A, Lynch DG, Horn AC, et al. A quantitative assessment of ChatGPT as a neurosurgical triaging tool. Neurosurgery. Aug 01, 2024;95(2):487-495. [doi: 10.1227/neu.0000000000002867] [Medline: 38353523]
71. Sorin V, Klang E, Sobeh T, Konen E, Shrot S, Livne A, et al. Generative pre-trained transformer (GPT)-4 support for differential diagnosis in neuroradiology. Quant Imaging Med Surg. Oct 01, 2024;14(10):7551-7560. [FREE Full text] [doi: 10.21037/qims-24-200] [Medline: 39429611]
72. Hewitt KJ, Wiest IC, Carrero ZI, Bejan L, Millner TO, Brandner S, et al. Large language models as a diagnostic support tool in neuropathology. J Pathol Clin Res. Nov 2024;10(6):e70009. [FREE Full text] [doi: 10.1002/2056-4538.70009] [Medline: 39505569]
73. Chiang KL, Chou YC, Tung H, Huang CY, Hsieh LP, Chang KP, et al. Customized GPT model largely increases surgery decision accuracy for pharmaco-resistant epilepsy. J Clin Neurosci. Dec 2024;130:110918. [doi: 10.1016/j.jocn.2024.110918] [Medline: 39541652]
74. de Arriba-Pérez F, García-Méndez S, Otero-Mosquera J, González-Castaño FJ. Explainable cognitive decline detection in free dialogues with a machine learning approach based on pre-trained large language models. Appl Intell. Sep 24, 2024;54:12613-12628. [doi: 10.1007/s10489-024-05808-0]
75. Koga S, Martin NB, Dickson DW. Evaluating the performance of large language models: ChatGPT and Google Bard in generating differential diagnoses in clinicopathological conferences of neurodegenerative disorders. Brain Pathol. May 2024;34(3):e13207. [FREE Full text] [doi: 10.1111/bpa.13207] [Medline: 37553205]
76. Rezaii N, Hochberg D, Quimby M, Wong B, Brickhouse M, Touroutoglou A, et al. Artificial intelligence classifies primary progressive aphasia from connected speech. Brain. Sep 03, 2024;147(9):3070-3082. [doi: 10.1093/brain/awae196] [Medline: 38912855]
77. Wang C, Liu S, Li A, Liu J. Text dialogue analysis for primary screening of mild cognitive impairment: development and validation study. J Med Internet Res. Dec 29, 2023;25:e51501. [FREE Full text] [doi: 10.2196/51501] [Medline: 38157230]
78. Miao J, Thongprayoon C, Craici IM, Cheungpasitporn W. How to improve ChatGPT performance for nephrologists: a technique guide. J Nephrol. Jun 2024;37(5):1397-1403. [doi: 10.1007/s40620-024-01974-z] [Medline: 38771519]
79. Li KC, Bu ZJ, Shahjalal M, He BX, Zhuang ZF, Li C, et al. Performance of ChatGPT on Chinese master's degree entrance examination in clinical medicine. PLoS One. Apr 04, 2024;19(4):e0301702. [FREE Full text] [doi: 10.1371/journal.pone.0301702] [Medline: 38573944]
80. Leypold T, Lingens LF, Beier JP, Boos AM. Integrating AI in lipedema management: assessing the efficacy of GPT-4 as a consultation assistant. Life (Basel). May 20, 2024;14(5):646. [FREE Full text] [doi: 10.3390/life14050646] [Medline: 38792666]
81. Ríos-Hoyo A, Shan NL, Li A, Pearson AT, Pusztai L, Howard FM. Evaluation of large language models as a diagnostic aid for complex medical cases. Front Med (Lausanne). Jun 20, 2024;11:1380148. [FREE Full text] [doi: 10.3389/fmed.2024.1380148] [Medline: 38966538]
82. Wu W, Guo Y, Li Q, Jia C. Exploring the potential of large language models in identifying metabolic dysfunction-associated steatotic liver disease: a comparative study of non-invasive tests and artificial intelligence-generated responses. Liver Int. Apr 2025;45(4):e16112. [doi: 10.1111/liv.16112] [Medline: 39526465]
83. Afshar M, Gao Y, Gupta D, Croxford E, Demner-Fushman D. On the role of the UMLS in supporting diagnosis generation proposed by Large Language Models. J Biomed Inform. Sep 2024;157:104707. [doi: 10.1016/j.jbi.2024.104707] [Medline: 39142598]
84. Andreadis K, Newman DR, Twan C, Shunk A, Mann DM, Stevens ER. Mixed methods assessment of the influence of demographics on medical advice of ChatGPT. J Am Med Inform Assoc. Sep 01, 2024;31(9):2002-2009. [doi: 10.1093/jamia/ocae086] [Medline: 38679900]

85. Ho CN, Tian T, Ayers AT, Aaron RE, Phillips V, Wolf RM, et al. Qualitative metrics from the biomedical literature for evaluating large language models in clinical decision-making: a narrative review. BMC Med Inform Decis Mak. Nov 26, 2024;24(1):357. [FREE Full text] [doi: 10.1186/s12911-024-02757-z] [Medline: 39593074]
86. Niu S, Ma J, Bai L, Wang Z, Guo L, Yang X. EHR-KnowGen: knowledge-enhanced multimodal learning for disease diagnosis generation. Inf Fusion. Feb 1, 2024;102(C):102069. [doi: 10.1016/j.inffus.2023.102069]
87. Taub-Tabib H, Shamay Y, Shlain M, Pinhasov M, Polak M, Tiktinsky A, et al. Identifying symptom etiologies using syntactic patterns and large language models. Sci Rep. Jul 13, 2024;14(1):16190. [FREE Full text] [doi: 10.1038/s41598-024-65645-6] [Medline: 39003296]
88. Zhang J, Sun K, Jagadeesh A, Falakaflaki P, Kayayan E, Tao G, et al. The potential and pitfalls of using a large language model such as ChatGPT, GPT-4, or LLaMA as a clinical assistant. J Am Med Inform Assoc. Sep 01, 2024;31(9):1884-1891. [doi: 10.1093/jamia/ocae184] [Medline: 39018498]
89. Chen J, Liu L, Ruan S, Li M, Yin C. Are different versions of ChatGPT's ability comparable to the clinical diagnosis presented in case reports? A descriptive study. J Multidiscip Healthc. Dec 6, 2023;16:3825-3831. [FREE Full text] [doi: 10.2147/JMDH.S441790] [Medline: 38084123]
90. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthc. Oct 15, 2021;3(1):1-23. [doi: 10.1145/3458754]
91. Liévin V, Egeberg Hother C, Geert Motzfeldt A, Winther O. Can large language models reason about medical questions? arXiv. Preprint posted online on July 17, 2022. [doi: 10.48550/arXiv.2207.08143]
92. Hirosawa T, Kawamura R, Harada Y, Mizuta K, Tokumasu K, Kaji Y, et al. ChatGPT-generated differential diagnosis lists for complex case-derived clinical vignettes: diagnostic accuracy evaluation. JMIR Med Inform. Oct 09, 2023;11:e48808. [FREE Full text] [doi: 10.2196/48808] [Medline: 37812468]
93. Hirosawa T, Harada Y, Tokumasu K, Ito T, Suzuki T, Shimizu T. Comparative study to evaluate the accuracy of differential diagnosis lists generated by Gemini Advanced, Gemini, and Bard for a case report series analysis: cross-sectional study. JMIR Med Inform. Oct 02, 2024;12:e63010. [FREE Full text] [doi: 10.2196/63010] [Medline: 39357052]
94. Günay S, Öztürk A, Özerol H, Yiğit Y, Erenler AK. Comparison of emergency medicine specialist, cardiologist, and chat-GPT in electrocardiography assessment. Am J Emerg Med. Jun 2024;80:51-60. [doi: 10.1016/j.ajem.2024.03.017] [Medline: 38507847]
95. Hoppe JM, Auer MK, Strüven A, Massberg S, Stremmel C. ChatGPT with GPT-4 outperforms emergency department physicians in diagnostic accuracy: retrospective analysis. J Med Internet Res. Jul 08, 2024;26:e56110. [FREE Full text] [doi: 10.2196/56110] [Medline: 38976865]
96. Lee S, Lee J, Park J, Park J, Kim D, Lee J, et al. Deep learning-based natural language processing for detecting medical symptoms and histories in emergency patient triage. Am J Emerg Med. Mar 2024;77:29-38. [doi: 10.1016/j.ajem.2023.11.063] [Medline: 38096637]
97. Levine DM, Tuwani R, Kompa B, Varma A, Finlayson SG, Mehrotra A, et al. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model: an observational study. Lancet Digit Health. Aug 2024;6(8):e555-e561. [FREE Full text] [doi: 10.1016/S2589-7500(24)00097-9] [Medline: 39059888]
98. Shah-Mohammadi F, Finkelstein J. Accuracy evaluation of GPT-assisted differential diagnosis in emergency department. Diagnostics (Basel). Aug 15, 2024;14(16):1779. [FREE Full text] [doi: 10.3390/diagnostics14161779] [Medline: 39202267]
99. Shiraishi M, Kanayama K, Kurita D, Moriwaki Y, Okazaki M. Performance of artificial intelligence chatbots in interpreting clinical images of pressure injuries. Wound Repair Regen. 2024;32(5):652-654. [doi: 10.1111/wrr.13189] [Medline: 38747443]
100. Zhuang S, Zeng Y, Lin S, Chen X, Xin Y, Li H, et al. Evaluation of the ability of large language models to self-diagnose oral diseases. iScience. Nov 29, 2024;27(12):111495. [FREE Full text] [doi: 10.1016/j.isci.2024.111495] [Medline: 39758998]
101. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. J Med Internet Res. Aug 22, 2023;25:e48659. [FREE Full text] [doi: 10.2196/48659] [Medline: 37606976]
102. Ke Y, Yang R, Lie SA, Lim TX, Ning Y, Li I, et al. Mitigating cognitive biases in clinical decision-making through multi-agent conversations using large language models: simulation study. J Med Internet Res. Nov 19, 2024;26:e59439. [FREE Full text] [doi: 10.2196/59439] [Medline: 39561363]
103. Avidan Y, Tabachnikov V, Court OB, Khoury R, Aker A. In the face of confounders: atrial fibrillation detection - practitioners vs. ChatGPT. J Electrocardiol. 2025;88:153851. [doi: 10.1016/j.jelectrocard.2024.153851] [Medline: 39667153]
104. Delaunay J, Cusido J. Evaluating the performance of large language models in predicting diagnostics for Spanish clinical cases in cardiology. Appl Sci. Dec 25, 2024;15(1):61. [doi: 10.3390/app15010061]
105. Kaya K, Gietzen C, Hahnfeldt R, Zoubi M, Emrich T, Halfmann MC, et al. Generative pre-trained transformer 4 analysis of cardiovascular magnetic resonance reports in suspected myocarditis: a multicenter study. J Cardiovasc Magn Reson. 2024;26(2):101068. [FREE Full text] [doi: 10.1016/j.jocmr.2024.101068] [Medline: 39079602]
106. Laohawetwanit T, Apornvirat S, Namboonlue C. Thinking like a pathologist: morphologic approach to hepatobiliary tumors by ChatGPT. Am J Clin Pathol. Jan 28, 2025;163(1):3-11. [doi: 10.1093/ajcp/aqae087] [Medline: 39030695]

107. Laohawetwanit T, Namboonlue C, Apornvirat S. Accuracy of GPT-4 in histopathological image detection and classification of colorectal adenomas. J Clin Pathol. Feb 18, 2025;78(3):202-207. [doi: 10.1136/jcp-2023-209304] [Medline: 38199797]
108. Traini DO, Palmisano G, Peris K. Large language models and dermoscopy: assessing the potential of task-specific GPT-4 vision in diagnosing basal cell carcinoma. J Eur Acad Dermatol Venereol. Dec 2024;38(12):2320-2322. [doi: 10.1111/jdv.20333] [Medline: 39258894]
109. Mosavi A, Imre F, Hung VT. ChatGPT and large language models in healthcare; a bibliometrics analysis and review. In: Proceedings of the IEEE 11th International Conference on Computational Cybernetics and Cyber-Medical Systems. 2024. Presented at: ICCC 2024; April 4-6, 2024; Hanoi, Vietnam. [doi: 10.1109/iccc62278.2024.10582937]
110. Gencer G, Gencer K. Large language models in healthcare: a bibliometric analysis and examination of research trends. J Multidiscip Healthc. Jan 2025;Volume 18:223-238. [doi: 10.2147/jmdh.s502351]
111. Wada A, Akashi T, Shih G, Hagiwara A, Nishizawa M, Hayakawa Y, et al. Optimizing GPT-4 turbo diagnostic accuracy in neuroradiology through prompt engineering and confidence thresholds. Diagnostics (Basel). Jul 17, 2024;14(14):1541. [FREE Full text] [doi: 10.3390/diagnostics14141541] [Medline: 39061677]
112. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res. Oct 04, 2023;25:e50638. [FREE Full text] [doi: 10.2196/50638] [Medline: 37792434]
113. Anisuzzaman DM, Malins JG, Friedman PA, Attia ZI. Fine-tuning large language models for specialized use cases. Mayo Clin Proc Digit Health. Mar 2025;3(1):100184. [FREE Full text] [doi: 10.1016/j.mcpdig.2024.11.005] [Medline: 40206998]
114. Han X, Zhang Z, Ding N, Gu Y, Liu X, Huo Y, et al. Pre-trained models: past, present and future. AI Open. 2021;2:225-250. [doi: 10.1016/j.aiopen.2021.08.002]
115. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform. Nov 19, 2022;23(6):bbac409. [doi: 10.1093/bib/bbac409] [Medline: 36156661]
116. Cirone K, Akrout M, Abid L, Oakley A. Assessing the utility of multimodal large language models (GPT-4 vision and large language and vision assistant) in identifying melanoma across different skin tones. JMIR Dermatol. Mar 13, 2024;7:e55508. [FREE Full text] [doi: 10.2196/55508] [Medline: 38477960]
117. Shifai N, van Doorn R, Malvehy J, Sangers TE. Can ChatGPT vision diagnose melanoma? An exploratory diagnostic accuracy study. J Am Acad Dermatol. May 2024;90(5):1057-1059. [FREE Full text] [doi: 10.1016/j.jaad.2023.12.062] [Medline: 38244612]
118. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Appl Sci. Jul 12, 2021;11(14):6421. [doi: 10.3390/app11146421]
119. Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. PubMedQA: a dataset for biomedical research question answering. arXiv. Preprint posted online on September 13, 2019. [FREE Full text] [doi: 10.18653/v1/d19-1259]
120. Pal A, Umapathi LK, Sankarasubbu M. MedMCQA: a large-scale multi-subject multi-choice dataset for medical domain question answering. In: Proceedings of the Conference on Health, Inference, and Learning. 2022. Presented at: PMLR 2022; April 7-8, 2022; Virtual Event.
121. Gargari OK, Fatehi F, Mohammadi I, Firouzabadi SR, Shafiee A, Habibi G. Diagnostic accuracy of large language models in psychiatry. Asian J Psychiatr. Oct 2024;100:104168. [doi: 10.1016/j.ajp.2024.104168] [Medline: 39111087]
122. Kaczmarczyk R, Wilhelm TI, Martin R, Roos J. Evaluating multimodal AI in medical diagnostics. NPJ Digit Med. Aug 07, 2024;7(1):205. [FREE Full text] [doi: 10.1038/s41746-024-01208-3] [Medline: 39112822]
123. Bürgisser N, Chalot E, Mehouachi S, Buclin CP, Lauper K, Courvoisier DS, et al. Large language models for accurate disease detection in electronic health records: the examples of crystal arthropathies. RMD Open. Dec 20, 2024;10(4):e005003. [FREE Full text] [doi: 10.1136/rmdopen-2024-005003] [Medline: 39794274]
124. Koga S, Martin NB, Dickson DW. Evaluating the performance of large language models: ChatGPT and Google Bard in generating differential diagnoses in clinicopathological conferences of neurodegenerative disorders. Brain Pathol. May 08, 2024;34(3):e13207. [FREE Full text] [doi: 10.1111/bpa.13207] [Medline: 37553205]
125. Tian D, Jiang S, Zhang L, Lu X, Xu Y. The role of large language models in medical image processing: a narrative review. Quant Imaging Med Surg. Jan 03, 2024;14(1):1108-1121. [FREE Full text] [doi: 10.21037/qims-23-892] [Medline: 38223123]
126. Waisberg E, Ong J, Masalkhi M, Zaman N, Sarker P, Lee AG, et al. GPT-4 and medical image analysis: strengths, weaknesses and future directions. J Med Artif Intell. Dec 2023;6:29. [doi: 10.21037/jmai-23-94]
127. Selivanov A, Rogov OY, Chesakov D, Shelmanov A, Fedulova I, Dylov DV. Medical image captioning via generative pretrained transformers. Sci Rep. Mar 13, 2023;13(1):4171. [FREE Full text] [doi: 10.1038/s41598-023-31223-5] [Medline: 36914733]
128. Tiu E, Talius E, Patel P, Langlotz CP, Ng AY, Rajpurkar P. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. Nat Biomed Eng. Dec 15, 2022;6(12):1399-1406. [FREE Full text] [doi: 10.1038/s41551-022-00936-9] [Medline: 36109605]
129. Dong R, Cheng X, Kang M, Qu Y. Classification of lumbar spine disorders using large language models and MRI segmentation. BMC Med Inform Decis Mak. Nov 18, 2024;24(1):343. [FREE Full text] [doi: 10.1186/s12911-024-02740-8] [Medline: 39558285]
130. Wang S, Zhao Z, Ouyang X, Liu T, Wang Q, Shen D. Interactive computer-aided diagnosis on medical image using large language models. Commun Eng. Sep 17, 2024;3(1):133. [doi: 10.1038/s44172-024-00271-8] [Medline: 39284899]

131. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, et al. Toward expert-level medical question answering with large language models. Nat Med. Mar 2025;31(3):943-950. [doi: 10.1038/s41591-024-03423-7] [Medline: 39779926]
132. Yasunaga M, Leskovec J, Liang P. LinkBERT: pretraining language models with document links. arXiv. Preprint posted online on March 29, 2022. [doi: 10.48550/arXiv.2203.15827]
133. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthc. Oct 15, 2021;3(1):1-23. [doi: 10.1145/3458754]
134. Liévin V, Egeberg Hother C, Geert Motzfeldt A, Winther O. Can large language models reason about medical questions? arXiv. Preprint posted online on July 17, 2022. [doi: 10.48550/arXiv.2207.08143]
135. Nori H, King N, Mayer McKinney S, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv. Preprint posted online on March 20, 2023. [doi: 10.48550/arXiv.2303.13375]
136. Yasunaga M, Bosselut A, Ren H, Zhang X, Manning CD, Liang P, et al. Deep bidirectional language-knowledge graph pretraining. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. 2022. Presented at: NIPS'22; November 28-December 9, 2022; New Orleans, LA.
137. Bolton E, Venigalla A, Yasunaga M, Hall D, Xiong B, Lee T, et al. BioMedLM: a 2.7B parameter language model trained on biomedical text. arXiv. Preprint posted online on March 27, 2024. [doi: 10.48550/arXiv.2403.18421]
138. Taylor R, Kardas M, Cucurull G, Scialom T, Hartshorn A, Saravia E, et al. Galactica: a large language model for science. arXiv. Preprint posted online on November 16, 2022. [doi: 10.48550/arXiv.2211.09085]
139. Duong D, Solomon BD. Analysis of large-language model versus human performance for genetics questions. Eur J Hum Genet. Apr 29, 2024;32(4):466-468. [doi: 10.1038/s41431-023-01396-8] [Medline: 37246194]
140. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: a medical chat model fine-tuned on a Large Language Model Meta-AI (LLaMA) using medical domain knowledge. Cureus. Jun 2023;15(6):e40895. [FREE Full text] [doi: 10.7759/cureus.40895] [Medline: 37492832]
141. Yasunaga M, Bosselut A, Ren H, Zhang X, Manning CD, Liang P, et al. Deep bidirectional language-knowledge graph pretraining. arXiv. Preprint posted online on October 17, 2022. [FREE Full text]
142. Xiong H, Wang S, Zhu Y, Zhao Z, Liu Y, Huang L, et al. DoctorGLM: fine-tuning your Chinese doctor is not a herculean task. arXiv. Preprint posted online on April 3, 2023. [FREE Full text]
143. Seibert K, Domhoff D, Bruch D, Schulte-Althoff M, Fürstenau D, Biessmann F, et al. Application scenarios for artificial intelligence in nursing care: rapid review. J Med Internet Res. Nov 29, 2021;23(11):e26522. [FREE Full text] [doi: 10.2196/26522] [Medline: 34847057]
144. Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. NPJ Digit Med. Oct 20, 2023;6(1):195. [FREE Full text] [doi: 10.1038/s41746-023-00939-z] [Medline: 37864012]
145. Yang Y, Liu X, Jin Q, Huang F, Lu Z. Unmasking and quantifying racial bias of large language models in medical report generation. Commun Med (Lond). Sep 10, 2024;4(1):176. [FREE Full text] [doi: 10.1038/s43856-024-00601-z] [Medline: 39256622]
146. Tian S, Jin Q, Yeganova L, Lai PT, Zhu Q, Chen X, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Brief Bioinform. Nov 22, 2023;25(1):bbad493. [FREE Full text] [doi: 10.1093/bib/bbad493] [Medline: 38168838]
147. Huang J, Shao H, Chang KC. Are large pre-trained language models leaking your personal information? arXiv. Preprint posted online on May 25, 2022. [FREE Full text] [doi: 10.18653/v1/2022.findings-emnlp.148]
148. Wu SH, Tong WJ, Li MD, Hu HT, Lu XZ, Huang ZR, et al. Collaborative enhancement of consistency and accuracy in US diagnosis of thyroid nodules using large language models. Radiology. Mar 01, 2024;310(3):e232255. [doi: 10.1148/radiol.232255] [Medline: 38470237]
149. Sun D, Hadjiiski L, Gormley J, Chan HP, Caoili E, Cohan R, et al. Outcome prediction using multi-modal information: integrating large language model-extracted clinical information and image analysis. Cancers (Basel). Jun 29, 2024;16(13):2402. [FREE Full text] [doi: 10.3390/cancers16132402] [Medline: 39001463]
150. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. NPJ Digit Med. Jan 24, 2024;7(1):20. [FREE Full text] [doi: 10.1038/s41746-024-01010-1] [Medline: 38267608]
151. Guo Z, Lai A, Thygesen JH, Farrington J, Keen T, Li K. Large language models for mental health applications: systematic review. JMIR Ment Health. Oct 18, 2024;11:e57400. [FREE Full text] [doi: 10.2196/57400] [Medline: 39423368]
152. Lawrence HR, Schneider RA, Rubin SB, Matarić MJ, McDuff DJ, Jones Bell M. The opportunities and risks of large language models in mental health. JMIR Ment Health. Jul 29, 2024;11:e59479. [FREE Full text] [doi: 10.2196/59479] [Medline: 39105570]
153. Obradovich N, Khalsa SS, Khan WU, Suh J, Perlis RH, Ajilore O, et al. Opportunities and risks of large language models in psychiatry. NPP Digit Psychiatry Neurosci. May 24, 2024;2(1):8. [doi: 10.1038/s44277-024-00010-z] [Medline: 39554888]
154. Bouchouras G, Bitilis P, Kotis K, Vouros GA. LLMs for the engineering of a Parkinson disease monitoring and alerting ontology. In: Proceedings of the First International Workshop on Generative Neuro-Symbolic Artificial Intelligence. 2024. Presented at: GeNeSy'24; May 26, 2024; Crete, Greece.

155. Zeng J, Zou X, Li S, Tang Y, Teng S, Li H, et al. Assessing the role of the generative pretrained transformer (GPT) in Alzheimer's disease management: comparative study of neurologist- and artificial intelligence-generated responses. J Med Internet Res. Oct 31, 2024;26:e51095. [FREE Full text] [doi: 10.2196/51095] [Medline: 39481104]
156. Mbizvo GK, Buchan I. Predicting seizure recurrence from medical records using large language models. Lancet Digit Health. Dec 2023;5(12):e851-e852. [doi: 10.1016/s2589-7500(23)00205-4]
157. Susiku E, Hewitt-Taylor J, Akudjedu TN. Graduate competencies, employability and the transnational Radiography workforce shortage: a systematic literature review of current pre-registration Radiography education and training models. Radiography (Lond). Mar 2024;30(2):457-467. [FREE Full text] [doi: 10.1016/j.radi.2024.01.001] [Medline: 38211453]
158. Rimmer A. Radiologist shortage leaves patient care at risk, warns royal college. BMJ. Oct 11, 2017;359:j4683. [doi: 10.1136/bmj.j4683] [Medline: 29021184]
159. Bellanda VC, Santos ML, Ferraz DA, Jorge R, Melo GB. Applications of ChatGPT in the diagnosis, management, education, and research of retinal diseases: a scoping review. Int J Retina Vitreous. Oct 17, 2024;10(1):79. [FREE Full text] [doi: 10.1186/s40942-024-00595-9] [Medline: 39420407]
160. Wu J, Ma Y, Wang J, Xiao M. The application of ChatGPT in medicine: a scoping review and bibliometric analysis. J Multidiscip Healthc. Apr 2024;Volume 17:1681-1692. [doi: 10.2147/jmdh.s463128]
161. Marzi G, Balzano M, Caputo A, Pellegrini MM. Guidelines for bibliometric‐systematic literature reviews: 10 steps to combine analysis, synthesis and theory development. Int J Manag Rev. Oct 07, 2024;27(1):81-103. [doi: 10.1111/ijmr.12381]
162. Johri S, Jeong J, Tran BA, Schlessinger DI, Wongvibulsin S, Barnes LA, et al. An evaluation framework for clinical use of large language models in patient interaction tasks. Nat Med. Jan 02, 2025;31(1):77-86. [doi: 10.1038/s41591-024-03328-5] [Medline: 39747685]
163. Diogo RC, Gengo And Silva Butcher RC, Peres HH. Evaluation of the accuracy of nursing diagnoses determined by users of a clinical decision support system. J Nurs Scholarsh. Jul 15, 2021;53(4):519-526. [doi: 10.1111/jnu.12659] [Medline: 33860621]
164. Böhnke J, Varghese J, ELISE Study Group, Karch A, Rübsamen N. Systematic review identifies deficiencies in reporting of diagnostic test accuracy among clinical decision support systems. J Clin Epidemiol. Nov 2022;151:171-184. [doi: 10.1016/j.jclinepi.2022.08.003] [Medline: 35987404]
165. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans Inf Syst. Jan 24, 2025;43(2):1-55. [doi: 10.1145/3703155]
166. Omar M, Sorin V, Agbareia R, Apakama DU, Soroush A, Sakhuja A, et al. Evaluating and addressing demographic disparities in medical large language models: a systematic review. Int J Equity Health. Mar 26, 2025;24(1):57. [FREE Full text] [doi: 10.1186/s12939-025-02419-0] [Medline: 40011901]
167. Sheng E, Chang KW, Natarajan P, Peng N. The woman worked as a babysitter: on biases in language generation. arXiv. Preprint posted online on September 3, 2019. [FREE Full text] [doi: 10.18653/v1/d19-1339]
168. van Kolfschooten H, van Oirschot J. The EU Artificial Intelligence Act (2024): implications for healthcare. Health Policy. Nov 2024;149:105152. [FREE Full text] [doi: 10.1016/j.healthpol.2024.105152] [Medline: 39244818]
169. Tang D, Xi X, Li Y, Hu M. Regulatory approaches towards AI medical devices: a comparative study of the United States, the European Union and China. Health Policy. Mar 2025;153:105260. [doi: 10.1016/j.healthpol.2025.105260] [Medline: 39951854]
170. Choudhury A, Chaudhry Z. Large language models and user trust: consequence of self-referential learning loop and the deskilling of health care professionals. J Med Internet Res. Apr 25, 2024;26:e56764. [FREE Full text] [doi: 10.2196/56764] [Medline: 38662419]

Abbreviations

AI: artificial intelligence
EHR: electronic health record
GDPR: General Data Protection Regulation
LLM: large language model
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
QA: question and answer
TCM: traditional Chinese medicine

Edited by J Sarvestan; submitted 02.02.25; peer-reviewed by R Singh, H Maheshwari, O Ibikunle; comments to author 13.03.25; revised version received 24.03.25; accepted 21.04.25; published 09.06.25

Please cite as:
Su H, Sun Y, Li R, Zhang A, Yang Y, Xiao F, Duan Z, Chen J, Hu Q, Yang T, Xu B, Zhang Q, Zhao J, Li Y, Li H
Large Language Models in Medical Diagnostics: Scoping Review With Bibliometric Analysis
J Med Internet Res 2025;27:e72062
URL: https://www.jmir.org/2025/1/e72062
doi: 10.2196/72062
PMID: