Review

Comparing Diagnostic Accuracy of Clinical Professionals and Large Language Models: Systematic Review and Meta-Analysis

Guxue Shan[1], BSc; Xiaonan Chen[1], BSc; Chen Wang[1], BSc; Li Liu[2], BMed; Yuanjing Gu[3], MNS; Huiping Jiang[4], BSc; Tingqi Shi[5], MPH

[1]Nanjing Drum Tower Hospital Clinical College of Nanjing University of Chinese Medicine, Nanjing, China
[2]Jiangsu Province Hospital of Chinese Medicine, Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China
[3]Department of Emergency, Nanjing Drum Tower Hospital, Nanjing, China
[4]Department of Nursing, Nanjing Drum Tower Hospital, Nanjing, China
[5]Department of Quality Management, Nanjing Drum Tower Hospital, Affiliated Hospital of Medical School, Nanjing University, Nanjing, China

Corresponding Author:

Tingqi Shi, MPH
Department of Quality Management
Nanjing Drum Tower Hospital, Affiliated Hospital of Medical School, Nanjing University
321 Zhongshan Road, Gulou District
Nanjing, 210008
China
Phone: 86 1-391-299-6998
Email: 13912996998@163.com

Abstract

Background: With the rapid development of artificial intelligence (AI) technology, especially generative AI, large language models (LLMs) have shown great potential in the medical field. Through massive medical data training, it can understand complex medical texts and can quickly analyze medical records and provide health counseling and diagnostic advice directly, especially in rare diseases. However, no study has yet compared and extensively discussed the diagnostic performance of LLMs with that of physicians.

Objective: This study systematically reviewed the accuracy of LLMs in clinical diagnosis and provided reference for further clinical application.

Methods: We conducted searches in CNKI (China National Knowledge Infrastructure), VIP Database, SinoMed, PubMed, Web of Science, Embase, and CINAHL (Cumulative Index to Nursing and Allied Health Literature) from January 1, 2017, to the present. A total of 2 reviewers independently screened the literature and extracted relevant information. The risk of bias was assessed using the Prediction Model Risk of Bias Assessment Tool (PROBAST), which evaluates both the risk of bias and the applicability of included studies.

Results: A total of 30 studies involving 19 LLMs and a total of 4762 cases were included. The quality assessment indicated a high risk of bias in the majority of studies, primary cause is known case diagnosis. For the optimal model, the accuracy of the primary diagnosis ranged from 25% to 97.8%, while the triage accuracy ranged from 66.5% to 98%.

Conclusions: LLMs have demonstrated considerable diagnostic capabilities and significant potential for application across various clinical cases. Although their accuracy still falls short of that of clinical professionals, if used cautiously, they have the potential to become one of the best intelligent assistants in the field of human health care.

JMIR Med Inform 2025;13:e64963; doi: 10.2196/64963

Keywords: machine learning; ML; artificial intelligence; AI; large language model; LLM; natural language processing; algorithm; model; analytics; NLP; deep learning; clinical diagnosis; diagnosis; diagnostic accuracy; accuracy; systematic review

Introduction

The Google Brain research team has consistently aimed to push the boundaries of recurrent language models and encoder-decoder architectures. In 2017, Vaswani et al [1] introduced a novel and simple network architecture known as the Transformer. This architecture uses a new mechanism called "self-attention," leading to significant advancements in the development and training of large language models (LLMs). These models possess advanced capabilities beyond extraction or summarization tasks and include natural language generation. Although there is no official definition of LLM, based on the literature [2,3], we define LLM as a model with over a billion parameters, designed for typical artificial intelligence (AI) applications.

Accurate clinical diagnosis is essential for patient treatment outcomes and survival rates. However, even when health care professionals gather extensive information and conduct numerous observations and tests, absolute diagnostic accuracy cannot be guaranteed. Minimizing diagnostic uncertainty and making the most appropriate treatment decisions remain persistent clinical challenges [4,5]. As of May 2024, the US Food and Drug Administration has approved 882 medical devices that use AI or machine learning assistance. By June 2024, the National Medical Products Administration of China has approved 17 AI-assisted diagnostic devices. In the era of big data in health care, the integration of AI with clinical decision support is a developing trend [6]. Numerous experts and scholars have explored the application of specialized AI and software tools in clinical diagnosis, yet there is limited knowledge about the performance of LLMs in this context. Therefore, this study aims to comprehensively evaluate the performance and accuracy of LLMs in clinical diagnosis, providing references for their clinical application.

Methods

Overview

This systematic review was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analysis of Diagnostic Test Accuracy Studies (PRISMA-DTA) statement [7]. Specific details can be found in Checklist 1.

Data Sources

A computer-assisted literature search of PubMed, Web of Science, Embase, CINAHL (Cumulative Index to Nursing and Allied Health Literature), CNKI (China National Knowledge Infrastructure), VIP, and SinoMed databases was performed from January 1, 2017, to the present. Search terms included controlled terms (MeSH [Medical Subject Heading] in PubMed and Emtree in Embase) and free-text terms. The following terms were used (including synonyms and closely related words) as index terms or free-text words: "large language model," "medicine," "diagnosis," and "accuracy." A search filter was applied to limit the results to humans. Only peer-reviewed cross-sectional studies and cohort studies were included. Multimedia Appendix 1 provides more details of the search strategy and study selection.

Selection Criteria

This review included studies meeting the following criteria: (1) investigated the application of LLMs in the initial diagnosis of human cases, (2) published between January 1, 2017, and the date of the search, (3) study type was either cross-sectional or cohort, (4) a primary source, and (5) written in English or Chinese.

An article was excluded if it (1) was a nonprimary source such as theses, conference papers, etc, (2) did not compare the diagnostic accuracy of clinical professionals in relevant departments with that of LLMs, (3) did not specify the type or scale of the LLM used for diagnosis, (4) did not have LLM independently conduct clinical case diagnoses, (5) was a duplicate publication, and (6) did not provide complete data or the full text could not be obtained.

Data Selection and Extraction

A total of 2 reviewers (GS and XC) independently reviewed the full texts of the eligible articles and extracted data. Any disagreements between the reviewers were discussed until a consensus was reached. The detailed characteristics extracted from each included study were: the first author and publication year, the country where the research was conducted, the study type, the study population, the source of cases, sample size, the LLMs used, control groups, and outcome measures.

Quality of Evidence and Risk of Bias

Due to the significant heterogeneity often present in the design and implementation of diagnostic accuracy studies, it is crucial to carefully assess the quality of the included studies. The Prediction Model Risk of Bias Assessment Tool (PROBAST) was used to evaluate the risk of bias and applicability of all included studies [8]. PROBAST assesses risk of bias across 4 domains: study participants, predictors, outcomes, and statistical analysis, while applicability is evaluated through the first 3 domains.

Given the complex structure and vast number of parameters in LLMs, they can be considered a "black box" to some extent, meaning that their internal workings and decision-making processes may not be entirely transparent or easily understood by humans [6]. Consequently, during the quality assessment, certain signal issues were excluded as they were unrelated to generative AI models [9].

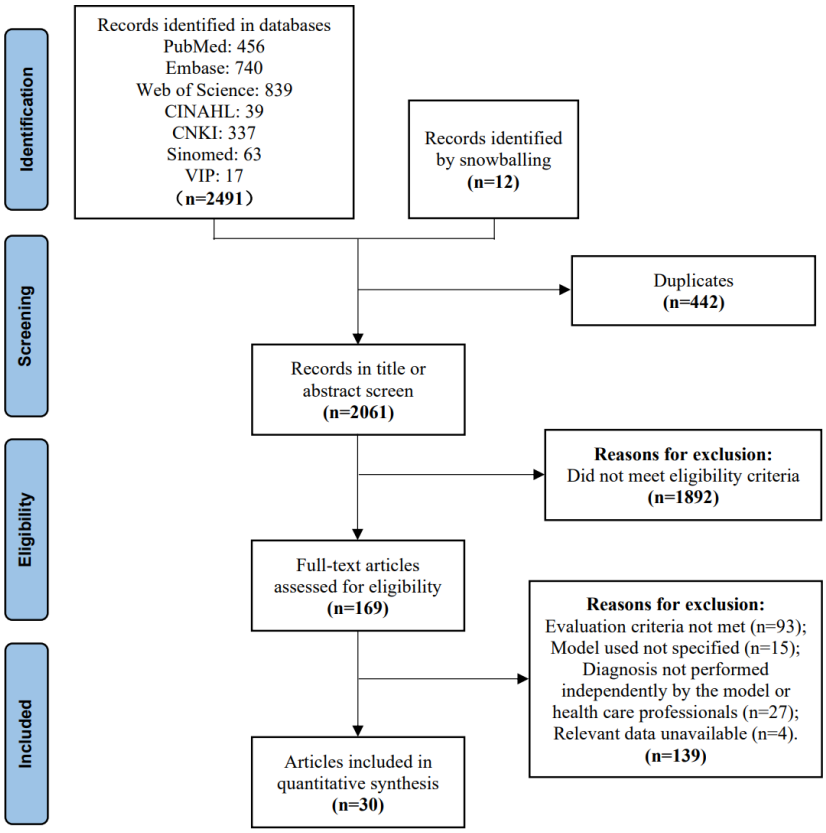
Results

Selection of Studies

A total of 2491 studies were found in the databases by 2 researchers independently following the predefined search strategies and data collection methods. An additional 12 articles were identified through reference tracing, bringing the total number of studies screened to 2503. Among these, 169

studies were read in full, resulting in 30 studies that met the inclusion criteria for synthesis. Reasons for exclusion at this stage were recorded and can be found in the flow diagram (see Figure 1).

Figure 1. Flow diagram. Papers identified in databases, title or abstract screened, read full text, and included in the synthesis. Reasons for exclusion are listed. CINAHL: Cumulative Index to Nursing and Allied Health Literature; CNKI: China National Knowledge Infrastructure.



Studies Characteristics

The 30 included studies [10-39] were concentrated within the past 3 years, with 12 published in 2023, 16 in 2024, and 2 in 2025. These studies cover a wide range of countries, primarily from Japan, the United States, and China. A total of 4762 cases were analyzed, involving 19 LLMs. The studies predominantly focused on GPT-3.5 (n=14) and GPT-4 (n=20) versions (OpenAI), extensively applied in assessing clinical diagnostic accuracy. In contrast, fewer studies addressed Google Bard (n=3), Bing (n=3), GPT-4o (n=2), and GPT-4V (n=2). The case diagnoses encompassed various fields, including ophthalmology (n=9), internal medicine (n=6), emergency medicine (n=3), and general medicine (n=3), among others. The control groups included at least 193 clinical professionals, ranging from resident doctors to medical experts with over 30 years of clinical experience, to compare their diagnostic capabilities with those of the LLMs. All included studies used LLMs for data testing purposes only and were not used for real-time diagnosis of clinical patients. Table 1 shows the basic characteristics of the included studies.

Table 1. Characteristics and results of the eligible studies.

Study	Year	Country	Study type	Subjects	Case source	Sample size	LLM[a]	Comparison group	Outcome measures
Zhang et al [10]	2024	China	Prospective study	Ophthalmology cases	Patient visit records	26	GPT-4o	Ophthalmologists	c[b], g[c]
Makhoul et al [11]	2024	Lebanon	Cross-sectional study	Otolaryngology cases	Published case reports	32	GPT-3.5	ENT[d] physicians, FM[e] specialists	a[f], b[g]
Pillai et al [12]	2023	The United States	Cross-sectional study	Autoimmune diseases cases	Published case reports	40	GPT-3.5 GPT-4 LLaMa 2	A certified internist	a[f], b[g]

Study	Year	Country	Study type	Subjects	Case source	Sample size	LLM[a]	Comparison group	Outcome measures
Levin et al [13]	2024	Israel	Cross-sectional study	Neonatal cases	Developed by researchers	6	GPT-4 Claude-2.0	Certified neonatal nurse practitioners	c[b], g[c]
Lyons et al [14]	2023	The United States	Cross-sectional study	Ophthalmology cases	Developed by researchers	44	GPT-4 Bing	Ophthalmology physicians	b[g], d[h]
Sarangi et al [15]	2023	India	Cross-sectional study	General cases	Developed by researchers	120	GPT-3.5 Bard Bing	Radiology residents	a[f]
Paslı et al [16]	2024	Turkey	Prospective study	Emergency cases	Patient visit records	758	GPT-4	The ED[i] triage team	d[h]
Wang et al [17]	2024	China	Retrospective cohort study	Thyroid cases	Patient visit records	109	GPT-4	Thyroid doctors	c[b]
Huang et al [18]	2024	The United States	Cross-sectional study	Ophthalmology cases	Patient visit records	20	GPT-4	Subspecialists (in glaucoma or retina)	c[b], g[c]
Stoneham et al [19]	2023	UK	Retrospective study	Dermatology cases	Patient visit records	36	GPT-4	A dermatologist	a[f]
Hirosawa et al [20]	2023	Japan	Cross-sectional study	Internal medicine cases	Published case reports	52	GPT-3.5 GPT-4	GIM[j] physicians	a[f], b[g]
Horiuchi et al [21]	2025	Japan	Retrospective study	Musculoskeletal cases	Published case reports	106	GPT-4 GPT-4V	Radiologists	a[f], b[g]
Mitsuyama et al [22]	2024	Japan	Retrospective study	Brain tumors cases	Patient visit records	150	GPT-4	Radiologists	a[f], b[g]
Hirosawa et al [23]	2023	Japan	Retrospective cohort study	Internal medicine cases	Published case reports and developed by researchers	82	Bard	GIM[j] physicians	a[f], b[g]
Suh et al [24]	2024	Korea	Retrospective study	General cases	Published case reports	190	GPT-4V Gemini Pro Vision	Radiologists	b[g]
Fraser et al [25]	2023	The United States	Cross-sectional study	Emergency cases	Patient visit records	40	GPT-3.5 GPT-4	ED physician	a[f], b[g], d[h]
Hirosawa et al [26]	2023	Japan	Prospective study	Internal medicine cases	Developed by researchers	30	GPT-3.5	GIM[j] physicians	a[f], b[g]
Shemer et al [27]	2024	Israel	Retrospective cohort study	Ophthalmology cases	Patient visit records	63	GPT-3.5	Ophthalmology residents and ophthalmologists	a[f], g[c]
Mohammadi et al [28]	2024	Iran	Retrospective study	Tibial plateau fracture cases	Retrospective study	111	GPT-4 GPT-4o	An ED physician and radiologist	f[k]
Arslan et al [29]	2025	Turkey	Prospective study	Emergency cases	Patient visit records	468	GPT-4 Copilot Pro	Triage nurses	d[h]
Rojas-Carabali et al [30]	2023	Singapore	Cross-sectional study	Ophthalmology cases	Developed by researchers	25	GPT-3.5 GPT-4	Ophthalmologists	a[f], b[g]
Kaya et al [31]	2024	Germany	Retrospective study	Myocarditis cases	Patient visit records	396	GPT-4	Radiologists	a[f], e[l]
Delsoz et al [32]	2024	The United States	Cross-sectional study	Ophthalmology cases	Published case reports	20	GPT-3.5 GPT-4	Cornea specialists	a[f]
Ming et al [33]	2024	China	Cross-sectional study	Ophthalmology cases	Published case reports	104	GPT-3.5	Ophthalmic residents	a[f], b[g]

Study	Year	Country	Study type	Subjects	Case source	Sample size	LLM[a]	Comparison group	Outcome measures
Nakaura et al [34]	2024	Japan	Retrospective study	Internal medicine cases	Patient visit records	28	GPT-4 GPT-2 GPT-3.5 GPT-4	Radiologists	a[f], b[g]
Ito et al [35]	2023	Japan	Cross-sectional study	General cases	Published case reports	45	GPT-4	Emergency physicians	a[f], d[h]
Gunes et al [36]	2024	Turkey	Cross-sectional study	thoracic cases	Published case reports	124	10 LLMs including GPT-3.5/4 Claude 3 Opus…	Published case reports	a[f]
Delsoz et al [37]	2023	The United States	Cross-sectional study	Ophthalmology cases	Published case reports	11	GPT-3.5	Ophthalmology residents	a[f]
Liu et al [38]	2023	China	Prospective study	Ophthalmology cases	Patient visit records	1226	GPT-3.5	Ophthalmologists	e[l]
Li et al [39]	2024	China	Retrospective study	Abdominal cases	Patient visit records	300	ERNie, 4.0 Claude 3.5 Sonnet	Radiologists	c[b]

[a]LLM: large language model.
[b]Accuracy score.
[c]Other auxiliary indicators (such as diagnostic completeness, diagnostic time, number of answers, etc).
[d]ENT: ear, nose, and throat.
[e]FM: family medicine.
[f]Frequency of correct primary diagnosis (answer).
[g] Frequency of correct diagnosis in a differential diagnosis list.
[h]Triage accuracy.
[i]ED: emergency department.
[j]GIM: general internal medicine.
[k]AUC: area under the curve.
[l]F_1-score

Quality of Evidence and Risk of Bias

The included articles were evaluated using the PROBAST tool, with the results presented in Multimedia Appendix 2. Overall, 10/30 (33.3%) studies had a low risk of bias, while 20/30 (66.6%) exhibited a high risk of bias. Regarding applicability, majority of study had low applicability concerns. Due to ethical concerns and patient privacy issues associated with the use of LLMs in clinical settings, most of the studies consist of retrospective studies with deidentified data and are limited to data testing. A total of 14 studies evaluated the diagnostic accuracy of models using small test sets. In addition, the "black box" nature of LLMs, whose training data are often undisclosed, complicates external evaluation and verification.

LLM Feature Analysis

Although a total of 19 different LLMs were used in the included studies, extracting the LLM with the best diagnostic performance in studies tested with multiple large models simultaneously, we found that the optimal LLM did not belong to the GPT series in only 6 studies. In 80% (24/30) of the studies, the researchers chose to obtain and use the corresponding LLMs directly on the official website by online access, which somewhat lowered the threshold for the use of the LLMs in the medical field and made it more accessible to the public. In total, 18 of the included studies specified the date of access or version of the LLM used. Retrieval-augmented generation (RAG) is a technique that combines information retrieval and generation to enhance task performance by incorporating relevant information into LLMs [40]. RAG was mentioned in 2 of the studies by further training of pretrained models specific to task datasets, and although RAG has been widely used in large model studies, it needs to be strengthened in the medical field. Specific details can be found in Table 2.

Table 2. Characteristics of the large language models (LLMs) in eligible studies.

Study	Optimal LLM[a] in research	Issuing company	Access mode	Date accessed (version)	Parameter settings	RAG[b]
Zhang et al [10]	GPT-4o	Open AI	—[c]	—	—	Unused
Makhoul et al [11]	GPT-3.5	—	Application-based ChatGPT 3.5	—	—	Unused
Pillai et al [12]	GPT-4	Open AI	Online access	August 12, 2023	—	Unused
Levin et al [13]	Claude-2.0	Anthropic	Platform developed by Anthropic (@Poe)	—	—	Unused
Lyons et al [14]	GPT-4	Open AI	Online access	March 19-24, 2023	—	Unused
Sarangi et al [15]	Bing	Microsoft	Search engine-based GPT-4	June 2023	—	Unused
Paslı et al [16]	GPT-4	Open AI	Online access	September 25, 2023	—	RAG
Wang et al [17]	GPT-4	—	Platform-based GPT-4 developed by researchers (ThyroAIGuide)	—	—	Unused
Huang et al [18]	GPT-4	Open AI	Online access	May 12, 2023	—	Unused
Stoneham et al [19]	GPT-4	Open AI	Online access	—	—	Unused
Hirosawa et al [20]	GPT-4	Open AI	Online access	April 10, 2023	—	Unused
Horiuchi et al [21]	GPT-4	Open AI	Online access	September 25, 2023	—	Unused
Mitsuyama et al [22]	GPT-4	Open AI	Online access	May 24, 2024	—	Unused
Hirosawa et al [23]	Bard	Google	Online access	June 8, 2023	—	Unused
Suh et al [24]	GPT-4V	Open AI	Online access	—	Temperature=1	Unused
Fraser et al [25]	GPT-3.5	Open AI	Online access	July 2023	—	Unused
Hirosawa et al [26]	GPT-3.5	Open AI	Online access	January 5, 2023	—	Unused
Shemer et al [27]	GPT-3.5	Open AI	Online access	March 2023	—	Unused
Mohammadi et al [28]	GPT-4o	Open AI	Online access	December 2023	—	Unused
Arslan et al [29]	GPT-4	Open AI	Online access	—	—	Unused
Rojas-Carabali et al [30]	GPT-4	Open AI	Online access	—	—	Unused
Kaya et al [31]	GPT-4	Open AI	Online access	March to July 2023	—	Unused
Delsoz et al [32]	GPT-4	Open AI	Online access	—	—	Unused
Ming et al [33]	GPT-4	Open AI	Online access	March 5-18, 2024	—	Unused
Nakaura et al [34]	Bing	Microsoft	Search engine-based GPT-4	—	—	Unused
Ito et al [35]	GPT-4	Open AI	Online access	March 15, 2023	—	Unused
Gunes et al [36]	Claude 3 Opus	Anthropic	Online access	May 2024	—	Unused
Delsoz et al [37]	GPT-3.5	Open AI	Online access	—	—	Unused
Liu et al [38]	GPT-3.5	Open AI	Online access	—	Temperature=0	Unused
Li et al [39]	Claude 3.5 Sonnet	Anthropic	Online access	June 13 to July 5, 2024	Temperature=1×10^{-10}	RAG

[a]LLM: large language model.
[b]RAG: retrieval-augmented generation.
[c]Not available.

Results of Diagnosis

The accuracy of the diagnoses made by the LLMs and the clinical professionals in the studies depends on the "standard answer" mentioned in the literature. The comparison is based on how their answers align with this standard. The "standard answer" in the included studies consists of the final diagnoses recorded in patient medical records or case reports, predetermined answers set by case developers, and diagnoses established by experienced clinical experts in the relevant departments.

Application of LLMs in Clinical Diagnosis

The most common model task was the free text task, which appeared in 19 articles, while only 1 article involved a choice task. English was used for input and output in all but 2 articles: one used Hebrew for prompting, and the other used Chinese to compare model diagnostic performance. In LLM, prompt is an input mode that guides the model to specific tasks or generates specific outputs, typically including elements such as instructions (task descriptions), context (background information), examples, input data, output instructions, and roles [41,42]. When LLMs are used for case diagnosis, the most frequently used elements are commands and input data, which primarily include patient basic information, complaints, medical history, physical examination, and laboratory tests. The output content mainly consists of diagnostic lists or triage recommendations. The diagnostic accuracy of health care professionals in each study was evaluated by investigators or experts in relevant fields.

In studies where multiple LLMs were used to diagnose sample cases, only the data for the model with the best diagnostic performance were recorded. Of these studies, 85% (24/30) reported that the ChatGPT series models demonstrated the best diagnostic performance. Several investigators noted that the diagnostic accuracy of GPT models was comparable with that of physicians and did not show significant differences. Specific details can be found in Multimedia Appendix 3.

Comparison of Diagnostic Accuracy Between LLMs and Health Care Professionals

Pooling the data revealed that 70% (21/30) of the studies used the frequency of correct diagnoses in model responses as the primary evaluation indicator of clinical diagnostic accuracy, excluding other auxiliary indicators. All accuracy results were expressed as percentages. For the optimal model, the accuracy of the primary diagnosis ranged from 25% to 97.8%, while triage accuracy ranged from 66.5% to 98%. In medical practice, the diagnostic agreement criterion is usually set at over 80%. The GPT series LLMs achieved diagnostic accuracy greater than 80% in clinical tasks across 3 studies in ophthalmology, 2 studies in general medicine, and 1 study each in radiology, emergency medicine, and general practice. Among the 7 studies focused on ophthalmic case diagnosis, the diagnostic performance was generally high, with 77.8% (7/9) of the large models showing diagnostic accuracy comparable with that of health care professionals.

In these cases, health care professionals received the same prompting words as the LLMs. In 60% (18/30) of the studies, control group participants were blinded to the true nature and goals of the study until it was completed. The diagnostic accuracy of health care professionals was compared with the outcomes of LLMs. The results showed that in 33.7% (20/30) of the studies, professionals had higher diagnostic accuracy than the models. In 33.3% (10/30) of the studies, the LLMs, specifically ChatGPT, had higher diagnostic accuracy than humans. The specific diagnostic accuracy comparisons are detailed in Table 3.

Table 3. Comparison of diagnostic accuracy between large language models (LLMs) and clinical professionals.

Specialty and study	Clinical professionals	Evaluation results (LLMs vs clinical professionals), %					
		a[a]	b[b]	c[c]	d[d]	e[e]	f[f]
Ophthalmology							
Zhang et al [10]	3	—[g]	—	55 vs 74.7	—	—	—
Lyons et al [14]	8	—	93 vs 95[h]	—	98.0 vs 86.0	—	—
Huang et al [18]	15	—	—	50.4 vs 50.3	—	—	—
Shemer et al [27]	6	68 vs 90	—	—	—	—	—
Rojas-Carabali et al [30]	5	64 vs 85.6	72 vs 89.6[h]	—	—	—	—
Delsoz et al [32]	3	85 vs 96.7	—	—	—	—	—
Ming et al [33]	3	59.6 vs 60.6	76 vs 65.4[h]	—	—	—	—
Delsoz et al [37]	3	72.7 vs 66.6	—	—	—	—	—
Liu et al [38]	2	—	—	—	—	80.1 vs 89.4	—
Internal medicine							
Hirosawa et al [20]	3	60 vs 50	81 vs 67[i]; 83 vs 75[j]	—	—	—	—
Mitsuyama et al [22]	5	73 vs 69.4	94 vs 81.6[h]	—	—	—	—
Hirosawa et al [23]	5	40.2 vs 64.6	53.7 vs 78[i]; 56.1 vs 82.9[j]	—	—	—	—
Hirosawa et al [26]	2	53.3 vs 93.3	83.3 vs 98.3[i]	—	—	—	—
Nakaura et al [34]	1	54 vs 100	96 vs 100[i]	—	—	—	—
Li et al [39]	5	—	—	93.8 vs 99.6	—	—	—
Emergency department							
Sinan Paslı et al [16]	Team	—	—	—	95.6 vs 92.8	—	—
Fraser et al [25]	3	40 vs 47	63 vs 69[h]	—	—	—	—
Arslan et al [29]	Team	—	—	—	66.5 vs 65.2	—	—
General medicine							
Sarangi et al [15]	2	53.3 vs 60.4	—	—	—	—	—
Suh et al [24]	8	—	48.9 vs 60.5[h]	—	—	—	—
Ito et al [35]	3	97.8 vs 91.1	—	—	66.7 vs 66.7	—	—
Orthopedics							
Horiuchi et al [21]	2	43 vs 47	58 vs 62.5[h]	—	—	—	—
Mohammadi et al [28]	2	—	—	—	—	—	0.73 vs 0.74
Cardiothoracic							
Kaya et al [31]	3	81 vs 91.3	—	—	—	85 vs 92.7	—
Gunes et al [36]	2	70.3 vs 46.8	—	—	—	—	—
Otolaryngology							
Makhoul et al [11]	20	—	70.8 vs 71.3[h]	—	—	—	—
Immunology							

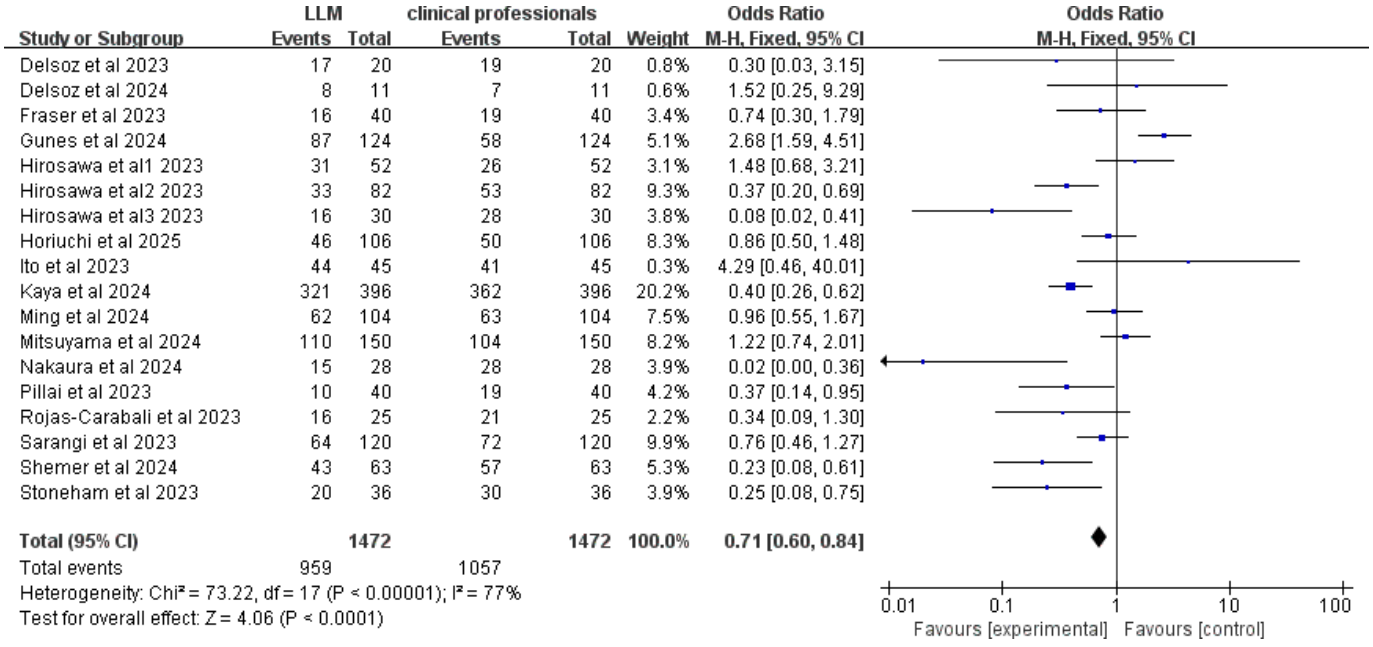
Specialty and study	Clinical professionals	Evaluation results (LLMs vs clinical professionals), %					
		a[a]	b[b]	c[c]	d[d]	e[e]	f[f]
Pillai et al [12]	1	25 vs 47.5	45 vs 60[i]; 47.5 vs 75[j]	—	—	—	—
Neonatology							
Levin et al [13]	32	—	—	70.8 vs 82.5	—	—	—
Thyroid							
Wang et al [17]	40	—	—	73.6 vs 87.4	—	—	—
Dermatology							
Stoneham et al [19]	1	56 vs 83	—	—	—	—	—

[a]Frequency of correct primary diagnosis (answer)
[b]Frequency of correct diagnosis in the 3, 5, or 10 differential diagnoses.
[c]Accuracy score
[d]Triage accuracy
[e]F_1-score.
[f]AUC: area under the curve.
[g]Not available.
[h]Frequency of correct diagnosis in the 3 differential diagnoses.
[i]Frequency of correct diagnosis in the 5 differential diagnoses.
[j]Frequency of correct diagnosis in the 10 differential diagnoses.

Meta-Analysis

Although this paper synthesizes over 4000 clinical cases, these cases exhibit significant heterogeneity in terms of clinical departments, diagnostic methodologies, and evaluation metrics. Due to these inherent differences, only 18 studies that used primary diagnostic accuracy as the evaluation metric were included in a meta-analysis. The analysis revealed that clinical professionals generally outperformed LLMs in diagnostic accuracy across various conditions, as shown in Figure 2. The P value was less than 0.05, and the I^2 value was 77%, indicating significant heterogeneity among the studies. Sensitivity analysis did not significantly improve the heterogeneity. Subgroup analyses by clinical department showed reduced heterogeneity in ophthalmology-related research, yet results still favored the diagnostic accuracy of ophthalmology professionals over LLMs.

Figure 2. Forest plot comparing diagnostic accuracy of large language models (LLMs) and clinical professionals [37,32,25,36,20,23,26,21,35,31,33,22,34,12,30,15,27,19].



Discussion

Principal Findings

In this systematic review, we analyzed the diagnostic accuracy of LLMs compared with clinical professionals, encompassing various LLMs and common medical specialties. Although the results typically indicated superior diagnostic accuracy among professionals, this study compiled the methodologies, functionalities, and outcomes of using LLMs in medical diagnostics. It affirmed the diagnostic capabilities of generic LLMs, providing evidence for their potential as healthcare assistants.

Application of LLMs in Clinical Diagnosis Still in Exploratory Stage

This review includes only peer-reviewed and published literature, so the models examined in the included studies primarily use text-based input and output for diagnostic tasks. However, with the advancement of large models, multimodal capabilities have also been integrated [43]. Some preprint studies [44,45] have explored using GPT-4V, incorporating imaging data into input prompts. Notably, adding images to LLM did not improve diagnostic performance. In a study by Horiuchi et al [44], ChatGPT-4, which relied solely on text prompts, achieved higher diagnostic accuracy compared to GPT-4V, which combined text and images. Without few-shot learning, LLMs may struggle with image recognition and interpretation, sometimes leading to counterproductive outcomes.

Currently, the performance of general LLMs continues to improve, showing strong results in health care question answering, text classification, and clinical concept extraction [46]. However, these studies remain experimental and laboratory-based. Issues such as the interpretability of model responses and medical ethics pose significant challenges to applying these models in real clinical settings. Furthermore, the trust and acceptance of AI models by clinicians directly affect their adoption and implementation. Therefore, education and training programs are crucial for enhancing physicians' AI literacy [47].

Evolution of Artificial Intelligence in Clinical Diagnosis

The evolution of AI in clinical diagnosis has progressed from simple specialized systems to complex deep learning models. Early AI systems were based on fixed rules and expert knowledge bases. While these systems achieved some success in specific tasks, they had limited scalability and flexibility. The advancement of deep learning technologies, particularly the emergence of LLMs, has ushered AI applications in the health care sector into a new era [48,49].

LLM can learn from vast amounts of medical data to autonomously discover and summarize diagnostic rules, significantly enhancing diagnostic accuracy and reliability. The development of RAG technology and fine-tuning techniques has further enabled LLM to acquire advanced domain expertise and effectively perform specialized tasks.

Ethics of Artificial Intelligence in Clinical Diagnosis

Although the pace of artificial intelligence development is swift, its broad implementation in clinical settings continues to encounter numerous obstacles, including concerns over data privacy, accountability, and ethics. Consequently, numerous scholars [50-53] underscore the imperative of utmost caution in using these technologies. Advances in the future will necessitate not only technological innovations but also comprehensive enhancements in legal and ethical frameworks to ensure that AI technology is safely and effectively woven into clinical diagnostic processes. In deploying LLMs within actual clinical workflows, it is crucial to first guarantee the transparency of all used data and secure patients' informed consent. In addition, to tackle potential biases within AI models, periodic audits are advised to identify and amend any discrepancies. Furthermore, to safeguard patient safety and adhere to regulatory demands, medical institutions should work alongside legal and ethical experts to establish stringent guidelines and oversight mechanisms for AI use. For instance, forming an ethics committee to assess and monitor AI applications could ensure compliance with ethical standards and legal requirements. These targeted measures are essential to surmount existing challenges and foster the successful incorporation of AI technologies in clinical diagnostics.

Application of LLMs in Specific Medical Fields

The application of LLMs in the medical field is gradually expanding, especially in imaging diagnosis, clinical decision support, and personalized treatment planning. Due to their specific needs and challenges, each medical field shows different ways and effects of LLMs' application.

Ophthalmology is one of the pioneers of LLMs' applications. In ophthalmic diagnosis, imaging data such as fundus images, retinal scans are typically complex, but LLMs excel in processing and analyzing these types of data [45,54]. Research has shown that LLMs can identify minor lesions in fundus images and diagnose conditions such as glaucoma and macular degeneration [55,56]. Ophthalmic diagnostics rely not only on imaging but also on additional data such as patients' genetic information and blood sugar levels. In the future, LLMs could integrate these multimodal data to achieve more accurate disease predictions through personalized treatment. Particularly in resource-limited areas, easily accessible LLMs with low usage thresholds could replace some ophthalmologists in preliminary screenings, further providing efficient diagnostic support in remote regions.

There are many internal medicine diseases that require long-term follow-up and monitoring. LLMs can process all historical data of patients simultaneously and updating personalized treatment plans, assisting clinical professionals in making more beneficial decisions [57-59]. In the future, LLMs will be paired with wearable devices to monitor patients' health in real time, predict potential risks through data analysis, and provide early intervention for patients

with medical diseases, thereby reducing the incidence and recurrence of the disease.

In the fields of otolaryngology [60,61] and dermatology [62,63], LLMs have been used to analyze imaging data for detecting lesions in respective areas. The latest models now offer voice input features, allowing patients to use the model anytime and anywhere to help in the early detection of speech disorders and vocal cord issues. In the future, integrating voice recognition with physiological data can also assist physicians in more accurately locating lesion areas during otolaryngological surgeries, thus improving treatment efficacy. Furthermore, by combining images of skin lesions with patients' genetic data, LLMs can help predict the risk of dermatological conditions and provide early warnings.

Exploration of the Use of LLMs in Various Clinical Departments

Currently, extensive research in fields such as ophthalmology, internal medicine, and radiology has demonstrated the substantial potential of LLMs in clinical diagnostics and pathological analysis. These models have even been implemented in some hospitals. Many clinical professionals are actively exploring how to integrate these technologies into their daily diagnostic and treatment routines.

However, the application of LLMs in other specialized areas remains limited, and research in these fields appears to be lacking. Several reasons account for this disparity: First, the departments mentioned above primarily focus on diagnostic issues, providing rich training data for large models, especially in terms of imaging and case data. Second, the main challenges these departments face in clinical practice, such as accurate diagnosis and disease prediction, are areas where LLMs can excel. In contrast, other departments such as surgery, although also using imaging data, face complexities in surgical and procedural tasks that hinder the maturity of AI applications. Gynecology has seen some applications of image recognition, but lacks depth in research and sufficient data accumulation, making model training challenging. In addition, real-world factors such as data privacy protection and technology dissemination also restrict the application of large models in certain departments.

Future Directions

The "human-AI collaboration" model involves an initial diagnosis provided by AI, which is then reviewed and confirmed by clinicians. AI's capability to analyze clinical data in real time enables it to offer personalized monitoring plans based on the specific conditions of patients. This continuous tracking of patient health and treatment outcomes helps achieve the goals of personalized medicine and precise diagnosis [64,65]. In addition, AI can provide customized services and recommendations based on user preferences and backgrounds, enhancing user experience and effectiveness. This model combines the rapid processing capabilities of AI with the expert judgment of clinicians, improving the efficiency and reliability of clinical trials. It also enhances data analysis and patient management, offering significant advantages in cost reduction, resource use, and ensuring the reliability of trial results.

Although LLMs are not inherently designed for clinical diagnostic tasks, advancements in technology and data accumulation are expected to improve their performance in clinical settings. Techniques such as large-scale medical literature analysis, specific clinical data training, task-specific fine-tuning, personalized training for particular scenarios, and integration with APIs or other supplementary software tools are anticipated to enhance the diagnostic support and treatment recommendations provided by these models [66,67]. Hybrid models could be developed by combining rule-based clinical decision support systems with the pattern recognition capabilities of LLMs. For example, Vision China 2023 introduced Eye GPT [68], a system that integrates ophthalmic medical knowledge with LLM. This system aims to assist clinicians in disease diagnosis and improve medical efficiency by combining extensive ophthalmic information with powerful computational capabilities. This innovation in integrating large models with specialized clinical fields is expected to play a crucial role in future clinical applications and provide research directions for other medical specialties.

Limitations

This study has several limitations. First, the inclusion criteria restricted the review to studies comparing the diagnostic accuracy of LLMs with that of clinical health care professionals using case groups. This limitation may affect the comprehensiveness of the review and introduce selection bias. In addition, there is no specialized tool for assessing the risk of bias in literature related to LLMs. Although PROBAST was used to evaluate the quality of the included studies, its focus on diagnostic accuracy may influence the evaluation results. Finally, significant heterogeneity among the studies was observed, with variations in outcome measures potentially related to differences in intervention subjects, prompt inputs, and information modalities. Further exploration of LLM diagnostic performance is needed through large-scale, multicenter, and high-quality cross-sectional and cohort studies.

Conclusions

This systematic review included 20 studies comparing the diagnostic accuracy of LLMs with that of health care professionals, encompassing various generative AI models and medical specialties. The findings indicate that while LLMs still have a long way to go in accurately diagnosing real-world clinical scenarios and currently lack the level of understanding of human experts, they undeniably possess significant potential as health care assistants. With ongoing advancements and optimizations in technology, it is anticipated that LLMs will play an increasingly important role in future clinical diagnostics.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Details of the search strategy in PubMed.
[DOCX File (Microsoft Word File), 16 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Quality assessment of included studies.
[DOCX File (Microsoft Word File), 23 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Characteristics of large language models (LLMs) applied in clinical diagnostic studies.
[DOCX File (Microsoft Word File), 29 KB-Multimedia Appendix 3]

Checklist 1

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) checklist.
[PDF File (Adobe File), 93 KB-Checklist 1]

References

1. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Presented at: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; Dec 4-9, 2017; Long Beach California USA.
2. Wentao S, Ruixiao L, Tianxiang S, et al. Large language models: principles, implementation, and progress. J Comp Res Dev. (2):351-361. [doi: 10.7544/issn1000-1239.202330303]
3. Guo H, Liu P, Lu R, et al. Research on a massively large artificial intelligence model and its application in medicine. Sci Sin-Vitae. Jan 1, 2024;54. [doi: 10.1360/SSV-2022-0298]
4. Our stubborn quest for diagnostic certainty. N Engl J Med. Nov 2, 1989;321(18):1272-1273. [doi: 10.1056/NEJM198911023211820]
5. Nour M, Senturk U, Polat K. Diagnosis and classification of Parkinson's disease using ensemble learning and 1D-PDCovNN. Comput Biol Med. Jul 2023;161:107031. [doi: 10.1016/j.compbiomed.2023.107031] [Medline: 37211002]
6. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. Aug 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
7. McInnes MDF, Moher D, Thombs BD, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: the PRISMA-DTA statement. JAMA. Jan 23, 2018;319(4):388-396. [doi: 10.1001/jama.2017.19163] [Medline: 29362800]
8. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. Jan 1, 2019;170(1):51-58. [doi: 10.7326/M18-1376] [Medline: 30596875]
9. Takita H, Kabata D, Walston SL, et al. Diagnostic performance comparison between generative AI and physicians: a systematic review and meta-analysis. medRxiv. Preprint posted online on Mar 18, 2024. [doi: 10.1101/2024.01.20.24301563]
10. Zhang J, Ma Y, Zhang R, et al. A comparative study of GPT-4o and human ophthalmologists in glaucoma diagnosis. Sci Rep. 2024;14(1):30385. [doi: 10.1038/s41598-024-80917-x]
11. Makhoul M, Melkane AE, Khoury PE, Hadi CE, Matar N. A cross-sectional comparative study: ChatGPT 3.5 versus diverse levels of medical experts in the diagnosis of ENT diseases. Eur Arch Otorhinolaryngol. May 2024;281(5):2717-2721. [doi: 10.1007/s00405-024-08509-z] [Medline: 38365990]
12. Pillai J, Pillai K. Accuracy of generative artificial intelligence models in differential diagnoses of familial Mediterranean fever and deficiency of Interleukin-1 receptor antagonist. J Transl Autoimmun. Dec 2023;7:100213. [doi: 10.1016/j.jtauto.2023.100213] [Medline: 37927888]
13. Levin C, Kagan T, Rosen S, Saban M. An evaluation of the capabilities of language models and nurses in providing neonatal clinical decision support. Int J Nurs Stud. Jul 2024;155:104771. [doi: 10.1016/j.ijnurstu.2024.104771]
14. Lyons RJ, Arepalli SR, Fromal O, Choi JD, Jain N. Artificial intelligence chatbot performance in triage of ophthalmic conditions. Can J Ophthalmol. Aug 2024;59(4):e301-e308. [doi: 10.1016/j.jcjo.2023.07.016] [Medline: 37572695]
15. Sarangi PK, Narayan RK, Mohakud S, Vats A, Sahani D, Mondal H. Assessing the capability of ChatGPT, Google Bard, and Microsoft Bing in solving radiology case vignettes. Indian J Radiol Imaging. Apr 2024;34(2):276-282. [doi: 10.1055/s-0043-1777746] [Medline: 38549897]
16. Paslı S, Şahin AS, Beşer MF, Topçuoğlu H, Yadigaroğlu M, İmamoğlu M. Assessing the precision of artificial intelligence in ED triage decisions: insights from a study with ChatGPT. Am J Emerg Med. Apr 2024;78:170-175. [doi: 10.1016/j.ajem.2024.01.037] [Medline: 38295466]

17. Wang Z, Zhang Z, Traverso A, Dekker A, Qian L, Sun P. Assessing the role of GPT-4 in thyroid ultrasound diagnosis and treatment recommendations: enhancing interpretability with a chain of thought approach. Quant IMAGING Med Surg. Feb 1, 2024;14(2):1602-1615. [doi: 10.21037/qims-23-1180] [Medline: 38415150]
18. Huang AS, Hirabayashi K, Barna L, Parikh D, Pasquale LR. Assessment of a large language model's responses to questions and cases about glaucoma and retina management. JAMA Ophthalmol. Apr 1, 2024;142(4):371-375. [doi: 10.1001/jamaophthalmol.2023.6917] [Medline: 38386351]
19. Stoneham S, Livesey A, Cooper H, Mitchell C. ChatGPT versus clinician: challenging the diagnostic capabilities of artificial intelligence in dermatology. Clin Exp Dermatol. Jun 25, 2024;49(7):707-710. [doi: 10.1093/ced/llad402] [Medline: 37979201]
20. Hirosawa T, Kawamura R, Harada Y, et al. ChatGPT-generated differential diagnosis lists for complex case-derived clinical vignettes: diagnostic accuracy evaluation. JMIR Med Inform. Oct 9, 2023;11:e48808. [doi: 10.2196/48808] [Medline: 37812468]
21. Horiuchi D, Tatekawa H, Oura T, 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]
22. Mitsuyama Y, Tatekawa H, Takita H, 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]
23. Hirosawa T, Mizuta K, Harada Y, Shimizu T. Comparative evaluation of diagnostic accuracy between Google Bard and physicians. Am J Med. Nov 2023;136(11):1119-1123. [doi: 10.1016/j.amjmed.2023.08.003] [Medline: 37643659]
24. Suh PS, Shim WH, Suh CH, 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]
25. Fraser H, Crossland D, Bacher I, Ranney M, Madsen T, Hilliard R. Comparison of diagnostic and triage accuracy of Ada Health and WebMD symptom checkers, ChatGPT, and physicians for patients in an emergency department: clinical data analysis study. JMIR MHealth UHealth. Oct 3, 2023;11:e49995. [doi: 10.2196/49995] [Medline: 37788063]
26. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. Int J Environ Res Public Health. Feb 15, 2023;20(4):3378. [doi: 10.3390/ijerph20043378] [Medline: 36834073]
27. Shemer A, Cohen M, Altarescu A, et al. Diagnostic capabilities of ChatGPT in ophthalmology. GRAEFES Arch Clin Exp Ophthalmol. Jul 2024;262(7):2345-2352. [doi: 10.1007/s00417-023-06363-z] [Medline: 38183467]
28. Mohammadi M, Parviz S, Parvaz P, Pirmoradi MM, Afzalimoghaddam M, Mirfazaelian H. Diagnostic performance of ChatGPT in tibial plateau fracture in knee X-ray. Emerg Radiol. Feb 2025;32(1):59-64. [doi: 10.1007/s10140-024-02298-y] [Medline: 39613920]
29. Arslan B, Nuhoglu C, Satici MO, Altinbilek E. Evaluating LLM-based generative AI tools in emergency triage: a comparative study of ChatGPT Plus, Copilot Pro, and triage nurses. Am J Emerg Med. Mar 2025;89:174-181. [doi: 10.1016/j.ajem.2024.12.024] [Medline: 39731895]
30. Rojas-Carabali W, Cifuentes-González C, Wei X, et al. Evaluating the diagnostic accuracy and management recommendations of ChatGPT in uveitis. Ocul Immunol Inflamm. Oct 2024;32(8):1526-1531. [doi: 10.1080/09273948.2023.2253471] [Medline: 37722842]
31. Kaya K, Gietzen C, Hahnfeldt R, 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. [doi: 10.1016/j.jocmr.2024.101068] [Medline: 39079602]
32. Delsoz M, Madadi Y, Raja H, et al. Performance of ChatGPT in diagnosis of corneal eye diseases. Cornea. May 1, 2024;43(5):664-670. [doi: 10.1097/ICO.0000000000003492] [Medline: 38391243]
33. Ming S, Yao X, Guo X, et al. Performance of ChatGPT in ophthalmic registration and clinical diagnosis: cross-sectional study. J Med Internet Res. 2024;26:e60226. [doi: 10.2196/60226]
34. Nakaura T, Yoshida N, Kobayashi N, 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. [doi: 10.1007/s11604-023-01487-y] [Medline: 37713022]
35. Ito N, Kadomatsu S, Fujisawa M, et al. The accuracy and potential racial and ethnic biases of GPT-4 in the diagnosis and triage of health conditions: evaluation study. JMIR Med Educ. Nov 2, 2023;9:e47532. [doi: 10.2196/47532] [Medline: 37917120]

36. Gunes YC, Cesur T. The diagnostic performance of large language models and general radiologists in thoracic radiology cases: a comparative study. J Thorac Imaging. Sep 13, 2024. [doi: 10.1097/RTI.0000000000000805] [Medline: 39269227]
37. Delsoz M, Raja H, Madadi Y, et al. The use of ChatGPT to assist in diagnosing glaucoma based on clinical case reports. Ophthalmol Ther. Dec 2023;12(6):3121-3132. [doi: 10.1007/s40123-023-00805-x] [Medline: 37707707]
38. Liu X, Wu J, Shao A, et al. Uncovering language disparity of ChatGPT in healthcare: non-English clinical environment for retinal vascular disease classification. medRxiv. Preprint posted online on Jul 14, 2023. [doi: 10.1101/2023.06.28.23291931]
39. Chao LI, Youmei C, Yani D, Yaoping C, Xiuzhen C, Jie Q. Evaluation of the performance of generative artificial intelligence in generating radiology reports. Journal of New Medicine. 2024;55(11):853-860. [doi: 10.3969/j.issn.0253-9802.2024.11.001]
40. Zhao P, Zhang H, Yu Q, et al. Retrieval-augmented generation for ai-generated content: a survey. arXiv. Preprint posted online on Feb 29, 2024. [doi: 10.48550/arXiv.2402.19473]
41. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res. Oct 4, 2023;25:e50638. [doi: 10.2196/50638] [Medline: 37792434]
42. Giray L. Prompt engineering with ChatGPT: a guide for academic writers. Ann Biomed Eng. Dec 2023;51(12):2629-2633. [doi: 10.1007/s10439-023-03272-4] [Medline: 37284994]
43. Han T, Adams LC, Bressem KK, Busch F, Nebelung S, Truhn D. Comparative analysis of multimodal large language model performance on clinical vignette questions. JAMA. Apr 16, 2024;331(15):1320-1321. [doi: 10.1001/jama.2023.27861] [Medline: 38497956]
44. Horiuchi D, Tatekawa H, Oura T, et al. Comparison of the diagnostic accuracy among GPT-4 based ChatGPT, GPT-4V based ChatGPT, and radiologists in musculoskeletal radiology. medRxiv. Preprint posted online on Dec 9, 2023. [doi: 10.1101/2023.12.07.23299707]
45. Sorin V, Kapelushnik N, Hecht I, et al. Integrated visual and text-based analysis of ophthalmology clinical cases using a large language model. Sci Rep. Feb 10, 2025;15(1):4999. [doi: 10.1038/s41598-025-88948-8] [Medline: 39930078]
46. He K, Mao R, Lin Q, et al. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. Information Fusion. Jun 2025;118:102963. [doi: 10.1016/j.inffus.2025.102963]
47. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. Jan 2019;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
48. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nat Med. Jan 2019;25(1):24-29. [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
49. Reddy S, Fox J, Purohit MP. Artificial intelligence-enabled healthcare delivery. J R Soc Med. Jan 2019;112(1):22-28. [doi: 10.1177/0141076818815510] [Medline: 30507284]
50. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res. Aug 11, 2023;25:e48009. [doi: 10.2196/48009] [Medline: 37566454]
51. Meng X, Yan X, Zhang K, et al. The application of large language models in medicine: a scoping review. iScience. May 17, 2024;27(5):109713. [doi: 10.1016/j.isci.2024.109713] [Medline: 38746668]
52. Zhang J, Sun K, Jagadeesh A, 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 1, 2024;31(9):1884-1891. [doi: 10.1093/jamia/ocae184] [Medline: 39018498]
53. Lalmuanawma S, Hussain J, Chhakchhuak L. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: a review. Chaos Solitons Fractals. Oct 2020;139:110059. [doi: 10.1016/j.chaos.2020.110059] [Medline: 32834612]
54. Carlà MM, Crincoli E, Rizzo S. Retinal imaging analysis performed by ChatGPT-4O and Gemini Advanced: the turning point of the revolution? Retina (Philadelphia, Pa). Apr 1, 2025;45(4):694-702. [doi: 10.1097/IAE.0000000000004351] [Medline: 39715322]
55. Ghalibafan S, Taylor Gonzalez DJ, Cai LZ, et al. Applications of multimodal generative artificial intelligence in a real-world retina clinic setting. Retina (Philadelphia, Pa). Oct 1, 2024;44(10):1732-1740. [doi: 10.1097/IAE.0000000000004204] [Medline: 39287535]
56. Raja H, Huang X, Delsoz M, et al. Diagnosing glaucoma based on the ocular hypertension treatment study dataset using chat generative pre-trained transformer as a large language model. Ophthalmol Sci. 2025;5(1):100599. [doi: 10.1016/j.xops.2024.100599] [Medline: 39346574]
57. Kuzan BN, Meşe İ, Yaşar S, Kuzan TY. A retrospective evaluation of the potential of ChatGPT in the accurate diagnosis of acute stroke. Diagn Interv Radiol. Sep 2, 2024. [doi: 10.4274/dir.2024.242892] [Medline: 39221691]

58. Chiang KL, Chou YC, Tung H, 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]
59. Ding JE, Thao PNM, Peng WC, et al. Large language multimodal models for new-onset type 2 diabetes prediction using five-year cohort electronic health records. Sci Rep. Sep 6, 2024;14(1):20774. [doi: 10.1038/s41598-024-71020-2] [Medline: 39237580]
60. 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]
61. Lechien JR, Naunheim MR, Maniaci A, et al. Performance and consistency of ChatGPT-4 versus otolaryngologists: a clinical case series. Otolaryngol Head Neck Surg. Jun 2024;170(6):1519-1526. [doi: 10.1002/ohn.759] [Medline: 38591726]
62. Gabashvili IS. ChatGPT in dermatology: a comprehensive systematic review. medRxiv. Preprint posted online on Jun 12, 2023. [doi: 10.1101/2023.06.11.23291252]
63. Pillai A, Joseph SP, Hardin J. Evaluating the diagnostic and treatment recommendation capabilities of GPT-4 vision in dermatology. medRxiv. Preprint posted online on Jan 26, 2024. [doi: 10.1101/2024.01.24.24301743]
64. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA. Apr 3, 2018;319(13):1317-1318. [doi: 10.1001/jama.2017.18391] [Medline: 29532063]
65. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med. 2018;1(1):18. [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]
66. Zhou J, He X, Sun L, et al. SkinGPT-4: an interactive dermatology diagnostic system with visual large language model. medRxiv. Preprint posted online on Jun 13, 2023. [doi: 10.1101/2023.06.10.23291127]
67. Betzler BK, Chen H, Cheng CY, et al. Large language models and their impact in ophthalmology. Lancet Digit Health. Dec 2023;5(12):e917-e924. [doi: 10.1016/S2589-7500(23)00201-7] [Medline: 38000875]
68. Chen X, Zhao Z, Zhang W, et al. EyeGPT: ophthalmic assistant with large language models. arXiv. Preprint posted online on Feb 29, 2024. [doi: 10.48550/arXiv.2403.00840]

Abbreviations

AI: artificial intelligence
CINAHL: Cumulative Index to Nursing and Allied Health Literature
CNKI: China National Knowledge Infrastructure
LLM: large language model
MeSH: Medical Subject Heading
PRISMA-DTA: Preferred Reporting Items for Systematic Reviews and Meta-analysis of Diagnostic Test Accuracy Studies
PROBAST: Prediction Model Risk of Bias Assessment Tool
RAG: retrieval-augmented generation

Edited by Alexandre Castonguay; peer-reviewed by Ali Jafarizadeh, Simon Laplante, Soroosh Tayebi Arasteh; submitted 31.07.2024; final revised version received 19.03.2025; accepted 25.03.2025; published 25.04.2025

Please cite as:
Shan G, Chen X, Wang C, Liu L, Gu Y, Jiang H, Shi T
Comparing Diagnostic Accuracy of Clinical Professionals and Large Language Models: Systematic Review and Meta-Analysis
JMIR Med Inform 2025;13:e64963
URL: https://medinform.jmir.org/2025/1/e64963
doi: 10.2196/64963