Review

Accuracy of Large Language Models When Answering Clinical Research Questions: Systematic Review and Network Meta-Analysis

Ling Wang[1,2*], MD; Jinglin Li[2*], BD; Boyang Zhuang[3*], MD; Shasha Huang[4*], BD; Meilin Fang[2*], BD; Cunze Wang[2], BD; Wen Li[1], BD; Mohan Zhang[2], BD; Shurong Gong[5], MD

[1]Fuzhou University Affiliated Provincial Hospital, Shengli Clinical Medical College, Fujian Medical University, Fuzhou, China

[2]School of Pharmacy, Fujian Medical University, Fuzhou, China

[3]Fujian Center For Drug Evaluation and Monitoring, Fuzhou, China

[4]School of Pharmacy, Fujian University of Traditional Chinese Medicine, Fuzhou, China

[5]The Third Department of Critical Care Medicine, Fuzhou University Affiliated Provincial Hospital, Shengli Clinical Medical College, Fujian Medical University, Fuzhou, Fujian, China

[*]these authors contributed equally

Corresponding Author:
Shurong Gong, MD
The Third Department of Critical Care Medicine
Fuzhou University Affiliated Provincial Hospital
Shengli Clinical Medical College, Fujian Medical University
No.134 Dongjie Road
Fuzhou, Fujian, 350001
China
Phone: 86 15060677447
Email: shurong_gong@fjmu.edu.cn

Abstract

Background: Large language models (LLMs) have flourished and gradually become an important research and application direction in the medical field. However, due to the high degree of specialization, complexity, and specificity of medicine, which results in extremely high accuracy requirements, controversy remains about whether LLMs can be used in the medical field. More studies have evaluated the performance of various types of LLMs in medicine, but the conclusions are inconsistent.

Objective: This study uses a network meta-analysis (NMA) to assess the accuracy of LLMs when answering clinical research questions to provide high-level evidence-based evidence for its future development and application in the medical field.

Methods: In this systematic review and NMA, we searched PubMed, Embase, Web of Science, and Scopus from inception until October 14, 2024. Studies on the accuracy of LLMs when answering clinical research questions were included and screened by reading published reports. The systematic review and NMA were conducted to compare the accuracy of different LLMs when answering clinical research questions, including objective questions, open-ended questions, top 1 diagnosis, top 3 diagnosis, top 5 diagnosis, and triage and classification. The NMA was performed using Bayesian frequency theory methods. Indirect intercomparisons between programs were performed using a grading scale. A larger surface under the cumulative ranking curve (SUCRA) value indicates a higher ranking of the corresponding LLM accuracy.

Results: The systematic review and NMA examined 168 articles encompassing 35,896 questions and 3063 clinical cases. Of the 168 studies, 40 (23.8%) were considered to have a low risk of bias, 128 (76.2%) had a moderate risk, and none were rated as having a high risk. ChatGPT-4o (SUCRA=0.9207) demonstrated strong performance in terms of accuracy for objective questions, followed by Aeyeconsult (SUCRA=0.9187) and ChatGPT-4 (SUCRA=0.8087). ChatGPT-4 (SUCRA=0.8708) excelled at answering open-ended questions. In terms of accuracy for top 1 diagnosis and top 3 diagnosis of clinical cases, human experts (SUCRA=0.9001 and SUCRA=0.7126, respectively) ranked the highest, while Claude 3 Opus (SUCRA=0.9672) performed well at the top 5 diagnosis. Gemini (SUCRA=0.9649) had the highest rated SUCRA value for accuracy in the area of triage and classification.

Conclusions: Our study indicates that ChatGPT-4o has an advantage when answering objective questions. For open-ended questions, ChatGPT-4 may be more credible. Humans are more accurate at the top 1 diagnosis and top 3 diagnosis. Claude 3 Opus performs better at the top 5 diagnosis, while for triage and classification, Gemini is more advantageous. This analysis offers valuable insights for clinicians and medical practitioners, empowering them to effectively leverage LLMs for improved decision-making in learning, diagnosis, and management of various clinical scenarios.

Trial Registration: PROSPERO CRD42024558245; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024558245

(J Med Internet Res 2025;27:e64486) doi: 10.2196/64486

KEYWORDS

large language models; LLM; clinical research questions; accuracy; network meta-analysis; PRISMA

Introduction

Recent research has demonstrated the considerable success of large language models (LLMs) in a multitude of natural language tasks, including automatic summarization (the generation of a condensed version of a passage of text), machine translation (the automatic translation of text from one language to another), and question-and-answer systems (the construction of a system to automatically answer questions based on a passage of text) [1]. In this context, with the development of big biomedical data and artificial intelligence, the emergence of flexible natural language processing models such as ChatGPT provides a number of new possibilities for health care and biomedical research and has the potential to be a turning point in the field [2-4].

Although LLMs have shown great potential in the medical field, medicine is a demanding field, it is associated with life, and its complexity as well as specificity mean that any application must meet extremely high standards of accuracy. Controversy remains about whether LLMs can be applied to the medical field. Mu and He [5] reviewed the potential applications and challenges of ChatGPT in health care, noting that a lack of understanding of medical knowledge and specialized medical backgrounds hinder the ability of ChatGPT to delve into the complexity of medical concepts and terminology. Consequently, the capacity of ChatGPT to address specific medical queries, diagnose ailments, or furnish precise medical recommendations is restricted. Another study noted that the role of LLMs in health care may be limited by the presence of bias in training materials, their tendency to "hallucinate," and ethical and legal considerations when LLMs provide inaccurate advice that leads to patient harm, as well as patient privacy issues [6].

Given the controversy over the application of LLMs in medicine and the continuous emergence and versioning of LLMs, more research has been devoted to evaluating the performance of various LLMs in medicine to provide stronger evidence. In addition to ChatGPT developed by OpenAI, the performance of many other LLMs such as Microsoft (eg, Copilot [7]), Google (eg, Gemini [8]), and Meta (eg, LLaMA [9]) in the medical domain has also been compared. Many aspects of assessment have been included, such as medical exams [10], case text diagnosis [11], and disease classification or grading [12].

Unfortunately, there are differences in the performance of different LLMs in different studies. For example, in a study by Vaishya et al [13] that explored the performance of ChatGPT-3.5, ChatGPT-4, and Google Bard when answering 120 multiple-choice questions, the results showed that Google Bard had 100% accuracy and was significantly more accurate than both ChatGPT-3.5 and ChatGPT-4 (P<.001). Another study showed that ChatGPT-4 was more accurate than Google Bard (83% vs 76%) [14]. At present, most related research is limited to a single type of LLM [15,16] or a specific domain area [17,18], and there is no high-level evidence comparing the accuracy rankings of different LLMs when responding to clinical research questions.

Therefore, this study aimed to compare the accuracy of different LLMs when answering clinical research questions, including objective questions, open-ended questions, top 1 diagnosis, top 3 diagnosis, top 5 diagnosis, and triage and classification. This study aimed to provide high-level evidence-based support for future clinical applications, enabling clinical workers to better use LLMs to make more accurate and informed decisions for future learning, diagnosis, and different clinical scenarios.

Methods

Network Meta-Analysis

The network meta-analysis (NMA) was based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) reporting guidelines. The PRISMA checklist is shown in Multimedia Appendix 1. The Bayesian approach permits the indirect comparison of performance between a range of LLMs that were not explicitly articulated throughout the experiment. The study protocol was defined and registered in the PROSPERO database prior to the commencement of the study.

Search Strategy and Selection Criteria

A computer search of the PubMed, Embase, Web of Science, and Scopus databases was conducted to identify relevant studies on the accuracy of different LLMs when answering questions in the medical field. The last search was updated to October 14, 2024, to identify studies published since the first search, with no restrictions on the type of study. When the results of a study were reported in multiple publications, we included the study with the richest and most recent findings. We also searched the list of literature on LLMs in medicine-related systematic reviews and manually searched the references included in the reviews for additional access to relevant literature. The search subject terms were "LLM," "generative AI," "open AI," "Large language model," "ChatGPT-3.5," "ChatGPT-4," "Google

Bard," and "Bing," without any language restriction. The complete search strategies for all databases are shown in Multimedia Appendix 2.

A combination of EndNote X9 deduplication and manual deduplication was used to screen the literature in accordance with the developed inclusion criteria. The results of the literature searches conducted in different databases were then combined to create a new information database, which could be downloaded in full text. Independent review and assessment of the titles, abstracts, and full texts of the relevant literature were undertaken by 4 authors (LW, JL, BZ, and SH). The review encompassed studies using disparate LLMs systems to respond to medical queries. Letters, conference abstracts, editorials, reviews, and expert opinions for which no information was available were excluded from the review. In addition, the following studies were excluded: those that evaluated the performance of only 1 LLM; those that assessed the performance of 2 or more LLMs without specifying the LLM versions used (eg, the article only mentioned evaluating ChatGPT without mentioning ChatGPT-3.5, ChatGPT-4, or other versions), with the updated versions and timelines of various LLMs so far shown in Multimedia Appendix 3; those that assessed the performance of 2 or more LLMs but did not provide data isolating their accuracy when answering different types of questions; and the questions included in the study contained images. In addition, to reduce bias, we excluded research on accessing LLMs through an application programming interface (API).

Assessment of Results

The primary outcomes were the accuracy of LLMs when answering medical questions. These included objective questions, open-ended questions, top 1 diagnosis, top 3 diagnosis, top 5 diagnosis, and triage and classification accuracy. Objective questions are exam questions with a clear, quantifiable answer that is usually predetermined, unique, or with a limited number of options. Open-ended questions are a type of question that does not have a fixed answer nor standardized answer. Diagnosis and triage and classification are open-ended questions, but most diagnostic questions end with "What is the most probable diagnosis?" whereas triage and classification questions end with "How would you classify this disease?" Corresponding examples are shown in Multimedia Appendix 4.

Accuracy for objective questions was calculated as the number of correctly answered questions divided by the total number of questions. For diagnosis and classification, accuracy was defined as the number of cases correctly diagnosed or triaged divided by the total number of cases. Specifically for open-ended questions, accuracy was determined based on the number of questions rated "good" or "accurate" on the accuracy scale divided by the total number of questions.

Data Extraction

The 4 researchers jointly extracted and verified the following data: (1) basic information about the included studies, such as study title and first author; (2) baseline characteristics and interventions of the study population; (3) key elements evaluated for risk of bias; and (4) outcome indicators and relevant outcome measure data. Our study involved extracting raw data from each study. In cases of disagreement, these were resolved through discussion and consultation with a third party.

Quality Assessment

Because they were cross-sectional studies, the quality of the included studies was evaluated using the Newcastle-Ottawa Scale [19]. The quality assessment was conducted by 3 independent researchers (LW, JL, and BZ), with a fourth researcher (SH) resolving any disagreements. A low overall risk of bias was determined when the Newcastle-Ottawa Scale score ranged from 7 to 9, moderate risk was determined when the score was between 4 and 6, and high risk was determined when the score was 0 to 3.

Statistical Analyses

Statistical analyses were performed using Stata 18.0 and R (version 4.3.1), with the odds ratio (OR) as the analytical statistic. Accuracy was assessed using 95% CIs and the credible interval. NMA analyses were performed on different types of LLMs.

The confidence of the NMA results estimates was assessed according to the Confidence in Network Meta-Analysis (CINEMA) methodology, which is broadly based on the Grading of Recommendations Assessment, Development, and Evaluation (GRADE). An NMA was conducted within a Bayesian framework using Markov chain Monte Carlo methods and was computed using the BUGSnet and GeMTC packages in R (V.4.3.1) software. A network graph was constructed for each LLM included in the experiment in order to facilitate a comparison of the performance of multiple LLMs. The consistency between direct and indirect evidence was evaluated using a node-splitting method when there was a closed loop. If the *P* value between the direct, indirect, and network comparisons of the 2 interventions was >.05, we concluded that there was no statistical difference and consistency was good. The convergence of the network models derived from the Markov chain Monte Carlo simulations was assessed using trace and density plots. We used noninformative priors for all parameters and assumed common heterogeneity. Furthermore, for all LLMs, we determined the ranking probabilities, which were articulated as the surface under the cumulative ranking curve (SUCRA). Higher SUCRA values suggest superior accuracy in model ranking.

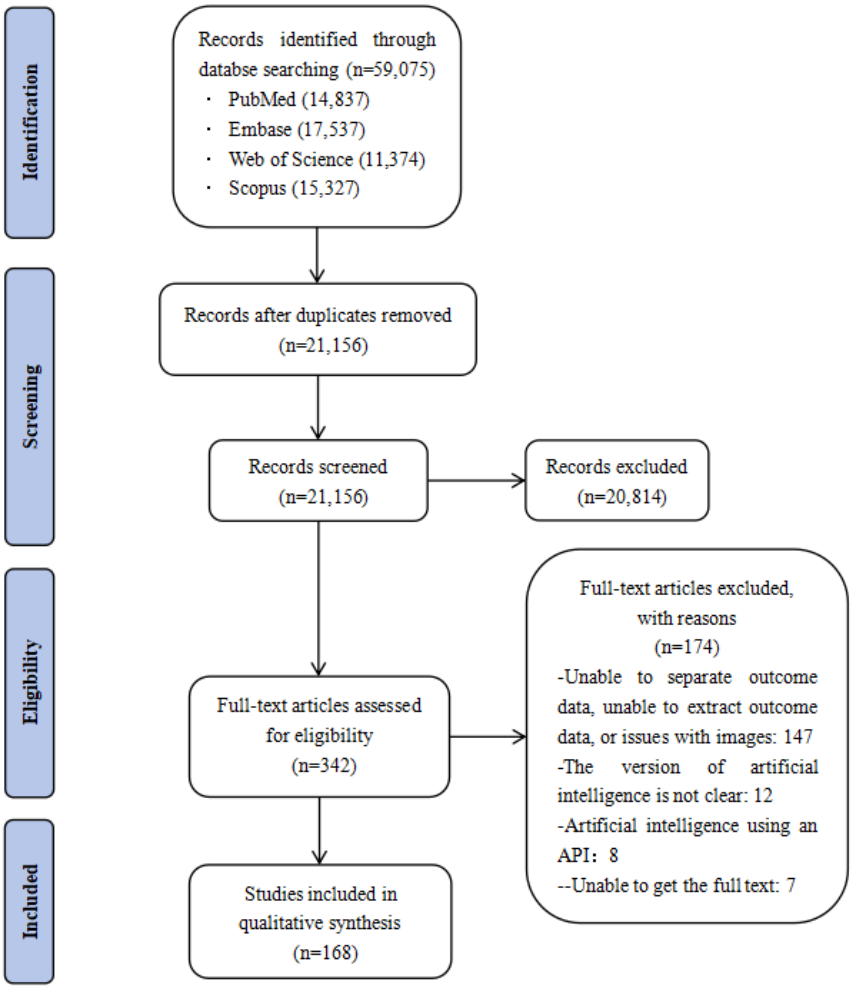
Results

Literature Search and Selection

A bibliographic search yielded 59,075 citations, of which 21,156 studies were identified as potential conditions based on abstract screening and retrieved for full text evaluation. Manual reading of the titles and abstracts of the remaining literature excluded 20,814 papers whose topics and interventions did not match the inclusion criteria for this study. Further reading of the full texts excluded the following: 174 articles that could not be separated nor extracted from the ending; 147 articles in which we were unable to separate outcome data, unable to extract outcome data, or detected issues related to images; 12 articles with unclear versions of the LLMs; and 8 articles that used an API to access

LLMs. In addition, the full text of 7 articles was not available, resulting in the final inclusion of 168 articles from the literature. The literature screening process is shown in Figure 1.

Figure 1. Literature screening flowchart. API: application programming interface.



Basic Characteristics of the Incorporated Literature

To assess the accuracy of different LLMs when answering medical questions, a total of 168 studies underwent a screening process to determine their suitability for inclusion. A total of 35,896 questions and 3063 clinical cases were included in the study. The basic information of the 168 studies is presented in Multimedia Appendix 5.

Quality Assessment of the Included Studies

In the quality assessment, 40 (40/168, 23.8%) studies were assessed as having a low overall risk of bias, while 128 (128/168, 76.2%) had a moderate overall risk of bias. No studies were identified as having a high overall risk of bias. The detailed quality assessment results for each study can be found in Multimedia Appendix 6.

Network Meta-Analysis

Objective Questions

The accuracy of LLMs when answering objective questions was reported in 105 studies [10,13,14,20-121]. The evidence network relationships are plotted in Figure 2A and involve 30 LLMs and a total of 33,838 multiple choice questions. Direct and indirect comparisons were formed for each LLM, partially forming a closed loop. The results of the indirect comparison are shown in Figure 3 and Multimedia Appendix 7. The red cells indicate there are statistically significant differences between the column-defining regimen and the row-defining regimen. The values in the green and blue cells are the logOR and 95% CI, respectively, from the comparison of the LLMs represented in the columns with the LLMs represented in the rows. A logOR value <0 indicates that the accuracy of the LLM corresponding to a column is lower than the LLM corresponding to a row. A value >0 indicates a higher accuracy. There was no evidence of statistically significant inconsistency (all $P>.05$) in the node-splitting test for NMA, except for Claude 2 versus ChatGPT-4 ($P=.04$), Bing chat versus people ($P=.004$), and Perplexity versus people ($P=.04$; Multimedia Appendix 8). The convergence of iterations was evaluated as good in trace and density plots, with the bandwidth tending toward 0 and reaching stability (Multimedia Appendix 9). The best probability ranking showed that ChatGPT-4o (SUCRA=0.9207) ranked first in terms of accuracy when answering objective questions, Aeyeconsult (SUCRA=0.9187) ranked second, and ChatGPT-4 (SUCRA=0.8087) ranked third (Table 1, Figure 4A).

Figure 2. Comparison network diagram of different outcomes, where larger nodes indicate more questions and thicker line segments indicate more questions between 2 types of large language models (LLMs) when answering (A) objective questions, (B) open-ended questions, (C) a top 1 diagnosis, (D) a top 3 diagnosis, (E) a top 5 diagnosis, and (F) triage and classification questions.

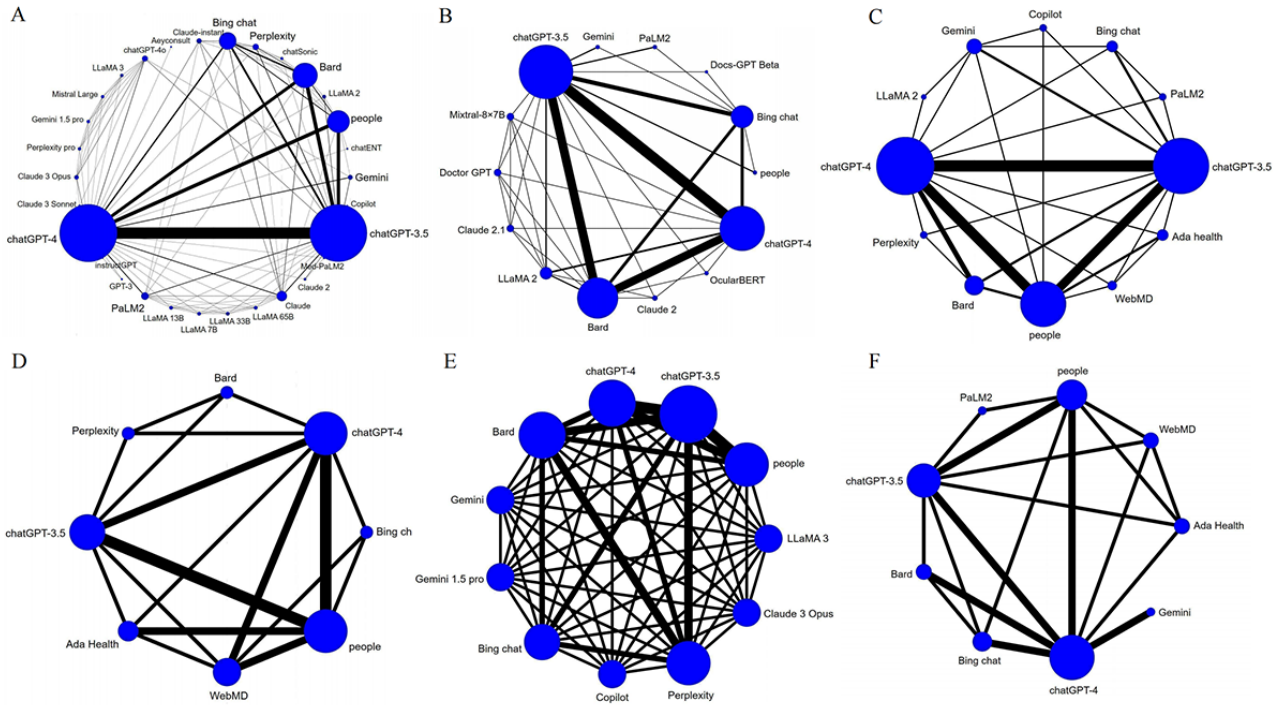


Figure 3. Indirect comparison of the accuracy of large language models (LLMs) when answering objective questions: A: instructGPT; A1: LLaMA 2; B: GTP-3; B1: LLaMA 3; C: ChatGPT-3.5; D: ChatGPT-4; D1: Mistral Large; E: ChatGPT-4o; E1: people; F1: chatENT; G: Bard; G1: ChatSonic; H: PaLM2; H1: Aeyeconsult; I: Gemini; I1: Med-PaLM 2; K: Gemini 1.5 pro; L: Bing chat; M: Copilot; N: Perplexity; O: Perplexity Pro; P: Claude; Q: Claude-instant; R: Claude 2; T: Claude 3 Opus; U: Claude 3 Sonnet; W: LLaMA 7B; X: LLaMA 13B; Y: LLaMA 33B; Z: LLaMA 65B.

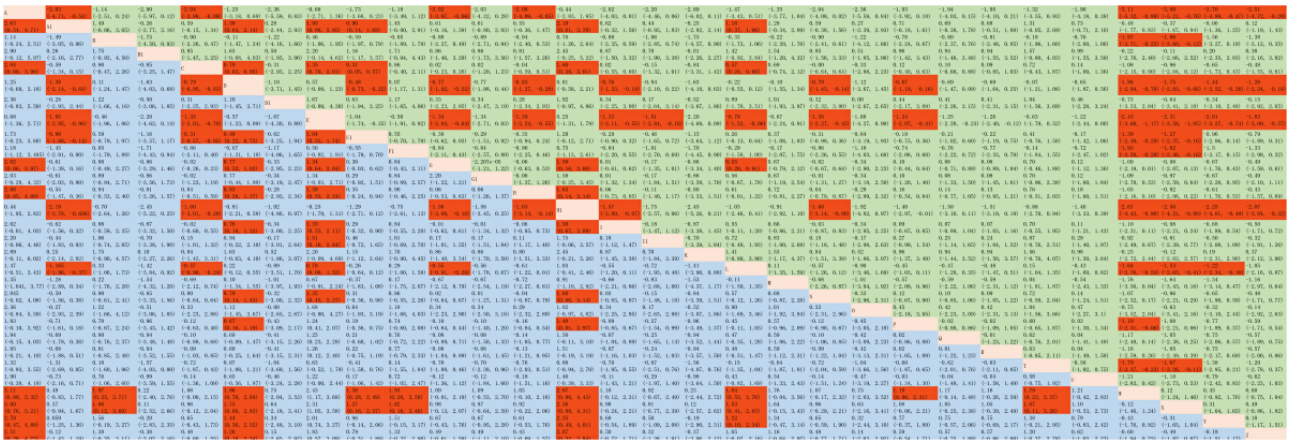
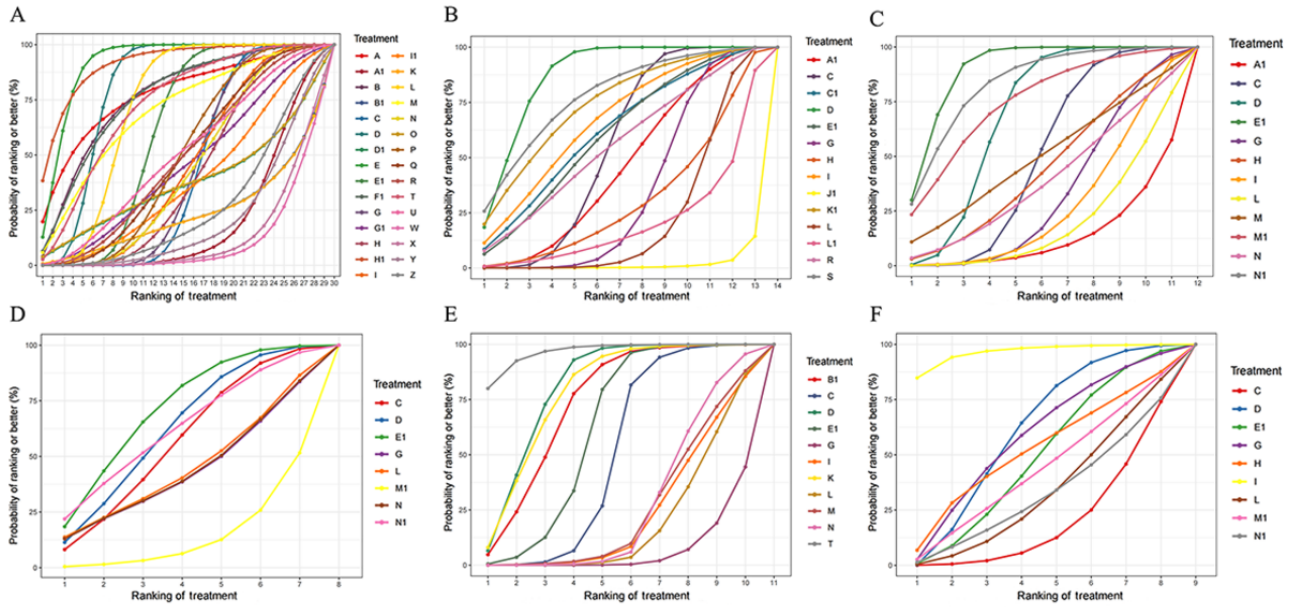


Table 1. Bayesian ranking results (surface under the cumulative ranking curve [SUCRA] value) of the network meta-analysis for each large language model (LLM).

LLM	SUCRA					
	Objective questions	Open-ended questions	Top 1 diagnosis	Top 3 diagnosis	Top 5 diagnosis	Triage and classification
instructGPT (A)	0.7805	__[a]	—	—	—	—
LLaMA 2 (A1)	0.2086	0.4629	0.1395	—	—	—
GTP-3 (B)	0.7704	—	—	—	—	—
LLaMA 3 (B1)	0.239	—	—	—	0.7405	—
ChatGPT-3.5 (C)	0.4343	0.5548	0.5039	0.565	0.5084	0.2093
Mixtral-8x7B (C1)	—	0.6224	—	—	—	—
ChatGPT-4 (D)	0.8087	0.8708	0.693	0.6302	0.8089	0.6185
Mistral Large (D1)	0.3842	—	—	—	—	—
ChatGPT-4o (E)	0.9207	—	—	—	—	—
People (E1)	0.6172	0.6067	0.9001	0.7126	0.6241	0.4934
chatENT (F1)	0.7687	—	—	—	—	—
Bard (G)	0.4443	0.3512	0.3353	0.4329	0.0722	0.5885
ChatSonic (G1)	0.4617	—	—	—	—	—
PaLM2 (H)	0.421	0.312	0.4496	—	—	0.5197
Aeyeconsult (H1)	0.9187	—	—	—	—	—
Gemini (I)	0.4543	0.6703	0.2812	—	0.2405	0.9649
Med-PaLM 2 (I1)	0.3919	—	—	—	—	—
OcularBERT (J1)	—	0.0176	—	—	—	—
Gemini 1.5 pro (K)	0.2449	—	—	—	0.7905	—
Doctor GPT (K1)	—	0.745	—	—	—	—
Bing chat (L)	0.728	0.23	0.2073	0.4499	0.2042	0.3391
Docs-GPT Beta (L1)	—	0.212	—	—	—	—
Copilot (M)	0.7038	—	0.5048	—	0.2633	—
WebMD (M1)	—	—	0.7511	0.1452	—	0.4348
Perplexity (N)	0.4424	—	0.3980	0.4367	0.2801	—
Ada Health (N1)	—	—	0.8363	0.6273	—	0.3319
Perplexity Pro (O)	0.3821	—	—	—	—	—
Claude (P)	0.5048	—	—	—	—	—
Claude-instant (Q)	0.4949	—	—	—	—	—
Claude 2 (R)	0.4928	0.5647	—	—	—	—
Claude 3 Opus (T)	0.7365	—	—	—	0.9672	—
Claude 3 Sonnet (U)	0.5094	—	—	—	—	—
LLaMA 7B (W)	0.1131	—	—	—	—	—
LLaMA 13B (X)	0.1365	—	—	—	—	—
LLaMA 33B (Y)	0.2147	—	—	—	—	—
LLaMA 65B (Z)	0.2721	—	—	—	—	—

[a]Not applicable because the LLM was not in the network.

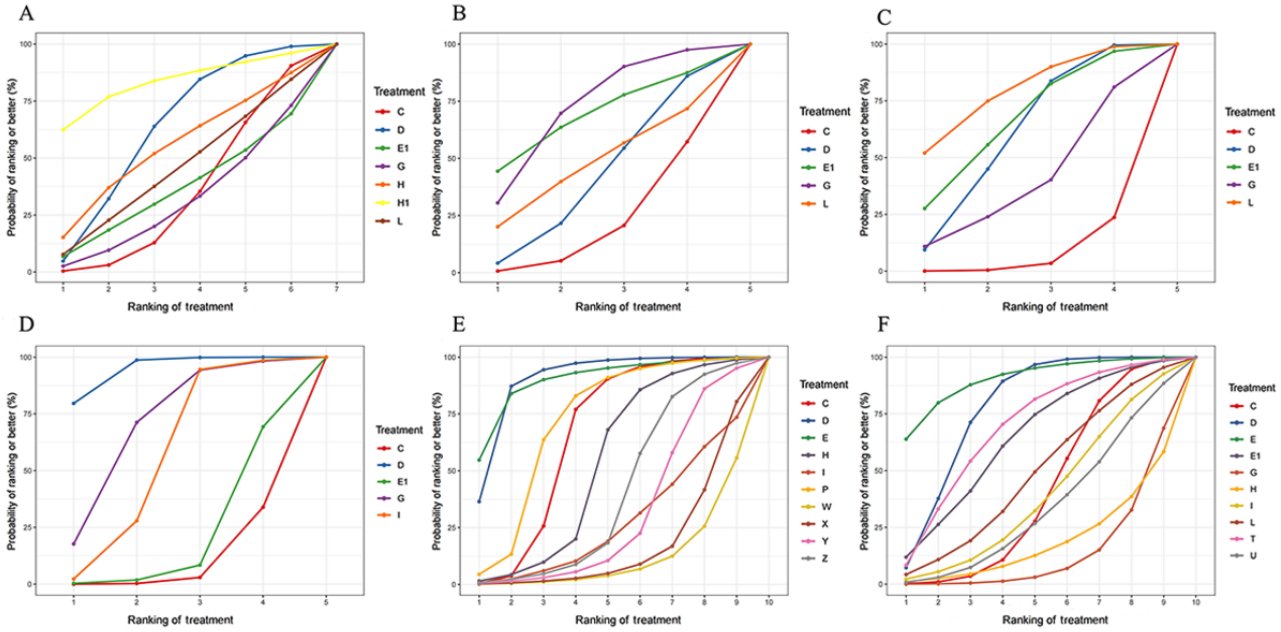
Figure 4. Surface under the cumulative ranking curve (SUCRAs) for the accuracy, with higher rankings associated with larger outcome values, of different large language models (LLMs) when answering (A) objective questions, (B) open-ended questions, (C) the top 1 diagnosis, (D) the top 3 diagnosis, (E) the top 5 diagnosis, and (F) triage and classification questions. The letters in the keys indicate the following LLMs: A: instructGPT; A1: LLaMA 2; B: GTP-3; B1: LLaMA 3; C: ChatGPT-3.5; C1: Mixtral-8x7B; D: ChatGPT-4; D1: Mistral Large; E: ChatGPT-4o; E1: people; F1: chatENT; G: Bard; G1: ChatSonic; H: PaLM2; H1: Aeyeconsult; I: Gemini; I1: Med-PaLM 2; J1: OcularBERT; K: Gemini 1.5 pro; K1: Doctor GPT; L: Bing chat; L1: Docs-GPT Beta; M: Copilot; M1: WebMD; N: Perplexity; N1: Ada Health; O: Perplexity Pro; P: Claude; Q: Claude-instant; R: Claude 2; S: Claude 2.1; T: Claude 3 Opus; U: Claude 3 Sonnet; W: LLaMA 7B; X: LLaMA 13B; Y: LLaMA 33B; Z: LLaMA 65B.



Subgroup Analysis

We stratified the results based on the fields of the problem (Multimedia Appendix 10). Based on the results, we compared the accuracy of LLMs in 6 fields: ophthalmology, orthopedics, urology, dentistry, oncology, and radiology. In ophthalmology, the LLM with the highest accuracy was Aeyeconsult (SUCRA=0.8334), followed by ChatGPT-4 (SUCRA=0.6331) and PaLM2 (SUCRA=0.5517). In the field of orthopedics, the LLM accuracy rates, from highest to lowest, were for Bard (SUCRA=0.7219), people (SUCRA=0.6802), and Bing chat (SUCRA=0.4732). For urology, Bing chat (SUCRA=0.7905) was the most accurate, followed by people (SUCRA=0.6587) and ChatGPT-4 (SUCRA=0.5941). In dentistry, ChatGPT-4 (SUCRA=0.9473) was the most accurate, followed by Bard (SUCRA=0.7068) and Gemini (SUCRA=0.5535). ChatGPT-4 (SUCRA=0.9002) performed the best in oncology, followed by ChatGPT-4o (SUCRA=0.8998) and Claude (SUCRA=0.7159). In radiology, ChatGPT-4o (SUCRA=0.9053) performed the best, ChatGPT-4 (SUCRA=0.7777) was second, and Claude 3 Opus (SUCRA=0.6935) ranked third. The SUCRAs are shown in Figure 5.

Figure 5. Surface under the cumulative ranking curve (SUCRAs) for the accuracy, with higher rankings associated with larger outcome values, of different large language models (LLMs) in (A) ophthalmology, (B) orthopedics, (C) urology, (D) dentistry, (E) oncology, and (F) radiology. The letters in the keys indicate the following LLMs: C: ChatGPT-3.5; D: ChatGPT-4; E: ChatGPT-4o; E1=people; G: Bard; H: PaLM2; H1: Aeyeconsult; I: Gemini; L: Bing chat; P: Claude; T: Claude 3 Opus; U: Claude 3 Sonnet; W: LLaMA 7B; X: LLaMA 13B; Y: LLaMA 33B; Z: LLaMA 65B.



Open-Ended Questions

The accuracy of the LLMs when responding to open-ended questions was examined in 34 studies [122-155]. The relationships within the evidence network are plotted in Figure 2B and include 14 LLMs and a total of 2026 open-ended questions. Direct and indirect comparisons were formed for each LLM, partially forming a closed loop. The results of the indirect comparison are presented in Multimedia Appendix 10, where red cells indicate statistically significant differences between the column-defining regimen and the row-defining regimen (Multimedia Appendix 7). There was no evidence of a statistically significant inconsistency (all P>.05) in the node-splitting test for the NMA, except for Bard versus ChatGPT-3.5 (P=.02; Multimedia Appendix 8). The trace and density plots are shown in Multimedia Appendix 9**,** and from the results, the iterative convergence was good. The best probability ranking indicated that ChatGPT-4 (SUCRA=0.8708) exhibited the highest accuracy when answering open-ended questions, followed by Claude 2.1 (SUCRA=0.7796) and Doctor GPT (SUCRA=0.7450; Table 1, Figure 4B).

Top 1 Diagnosis, Top 3 Diagnosis, and Top 5 Diagnosis

The accuracy of the top 1 diagnosis in clinical cases by LLMs was reported in 19 studies [11,156-173]. The evidence network relationship diagram is shown in Figure 2C and involves 12 LLMs and a total of 1266 clinical cases. The accuracy of LLMs for the top 3 diagnosis was reported in 7 studies [158,161,169,171,174-176]. The evidence network relationships are plotted in Figure 2D and involve 8 LLMs and a total of 453 clinical cases. The accuracy of LLMs for the top 5 diagnosis in clinical cases was reported in 7 studies [158,167,168,173,177-179]. The evidence network relationships are plotted in Figure 2E and involve 11 LLMs and a total of 443 clinical cases. Each LLM formed direct and indirect comparisons, partially closing the loop.

In terms of the top 1 diagnosis and top 5 diagnosis, the results of the indirect comparison are presented in Multimedia Appendix 7, where red cells indicate statistically significant differences between the column-defining regimen and the row-defining regimen. For the top 3 diagnosis, there was no statistical difference (all P>.05) in the comparisons between the LLMs (Multimedia Appendix 7). There was no evidence of a statistically significant inconsistency (all P>.05) for the top 1 diagnosis, except for Ada Health versus ChatGPT-3.5 (P=.04). For the top 3 diagnosis and top 5 diagnosis, all P were >.05 in the node-splitting test for the NMA (Multimedia Appendix 8). Iterative convergence was good, as shown by the trace and density plots (Multimedia Appendix 9). The best probability ranking showed that, in terms of accuracy of the top 1 diagnosis in clinical cases, people ranked first (SUCRA=0.9001), Ada Health ranked second (SUCRA=0.8363), and WebMD ranked third (SUCRA=0.7511; Table 1, Figure 4C). In terms of the accuracy of the top 3 diagnosis, people ranked first (SUCRA=0.7126), ChatGPT-4 ranked second (SUCRA=0.6302), and Ada Health ranked third (SUCRA=0.6273; Table 1, Figure 4D). For the accuracy of the top 5 diagnosis, Claude 3 Opus ranked first (SUCRA=0.9672), ChatGPT-4 ranked second (SUCRA=0.8089), and Gemini 1.5 pro ranked third (SUCRA=0.7905; Table 1, Figure 4E).

Triage and Classification

The accuracy of LLMs in triage and classification was reported in 7 studies [12,167,169,174,180-182]. The evidence network relationships are plotted in Figure 2F and involve 9 LLMs and a total of 901 clinical cases. Each LLM formed direct and indirect comparisons, partially closing the loop. The results of the indirect comparison are shown in Multimedia Appendix 7.

There were significant differences between Gemini and ChatGPT-3.5, ChatGPT-4, or Bing chat ($P<.05$). There was no evidence of a statistically significant inconsistency (all $P>.05$) in the node-splitting test for the NMA, except for ChatGPT-3.5 versus ChatGPT-4 ($P=.045$; Multimedia Appendix 8). Iterative convergence was good, as shown by the trace and density plots (Multimedia Appendix 9). The best probability ranking showed that, for the accuracy of triage and classification, Gemini ranked first (SUCRA=0.9649), ChatGPT-4 ranked second (SUCRA=0.6185), and Bard ranked third (SUCRA=0.5885), as shown in Table 1 and Figure 4F.

Discussion

Principal Findings

This study presents the most comprehensive meta-analysis to date on the accuracy of various LLMs when responding to medical queries, encompassing objective questions, open-ended questions, top 1 diagnosis, top 3 diagnosis, top 5 diagnosis, and triage and classification. Variations in accuracy among different LLMs were observed. ChatGPT-4o demonstrated the highest accuracy when answering objective questions, while ChatGPT-4 excelled at open-ended questions. The superior performance of people at the top 1 diagnosis and top 3 diagnosis suggests that human expertise is generally more dependable than LLMs in complex medical scenarios, while Claude 3 Opus seems to perform the best in the top 5 diagnosis. In terms of triage and classification, Gemini appeared to be more reliable.

In addition, we stratified LLMs according to the medical field in which the objective questions were located and explored their accuracy in 6 fields: ophthalmology, orthopedics, urology, dentistry, oncology, and radiology. We found that Aeyeconsult performed the best in ophthalmology, Bard performed the best in orthopedics, Bing chat performed the best in urology, ChatGPT-4 performed the best in both dentistry and oncology, and ChatGPT-4o had the highest accuracy in radiology.

At present, language models based on transformer architecture, whether pretrained or fine-tuned using biomedical corpora, have been proven effective in a series of natural language processing benchmarks in the biomedical field [183]. We attempted to analyze the reasons for the performance differences when different LLMs answer questions. Parameter size is an important factor affecting the accuracy of LLMs when answering questions. Research has found that, when the parameter size of the PaLM model is expanded from 8B to 40B, the accuracy of answering medical questions is doubled [184]. However, the practicality of a model depends not only on its number of parameters but also on many factors such as its training data and architecture, fine-tuning protocols, and overall architecture [185]. Taking GPT-4 as an example, it achieved a higher performance than its predecessor by adopting more advanced training data and architecture. The timeliness and accuracy of training data are also crucial for model performance. Today, models can not only rely on a limited set of pretraining data but also obtain the latest knowledge from the internet in real time. For example, Bing AI and Google Bard already have the ability to obtain real-time updates, and ChatGPT has also begun to follow suit by accepting plugins to expand its capabilities [185,186].

In addition, we found that some models fine-tuned on the backend LLM can achieve higher accuracy and less energy consumption in specific fields. For example, in the field of ophthalmology, Aeyeconsult integrates many ophthalmic data sets based on GPT-4 for training and generation [24]. This targeted training can significantly improve its performance in ophthalmic clinical tasks. Other possible data sources include clinical texts and accurate medical information, such as guidelines and peer-reviewed literature. In fact, there are already some models built or fine-tuned based on clinical text, such as SkinGPT-4 and ChatDoctor, which perform better overall than various general LLMs at biomedical natural language processing tasks [187,188].

Progress on various grand prognostic models has been very rapid, with a newer, more arithmetically powerful version being released every few months. However, our results show that the newer versions do not necessarily outperform the older ones in terms of performance when measured as accuracy, possibly because the newer versions incorporate fewer studies, which may have biased the results somewhat. In addition, updated versions such as ChatGPT-4V provided multimodal models (eg, that can evaluate image problems), and these models may have a greater advantage for image evaluation, for example.

Studies indicate LLMs outperform humans at exams like medical licensing, orthopedics, and pediatrics globally, highlighting LLMs' potential as a study aid. For the top 1 diagnosis and top 3 diagnosis, human accuracy is higher than that of LLMs. Despite the fact that Claude 3 Opus outperformed humans in the top 5 diagnostic results, due to the high level of accuracy required in the medical field and the multifaceted information and complex decision-making involved in medical diagnosis, we still recommend that LLMs should only be used as an auxiliary tool to assist doctors with more efficient data analysis and preliminary diagnostic recommendations.

Several meta-analyses have been conducted to assess the accuracy of LLMs in health care [15,189,190]. However, it is very unfortunate that the LLMs included in these studies included ChatGPT only and that some of the studies simply evaluated its performance on exams. Some studies did not differentiate between the types of questions answered by ChatGPT, which led to a significant amount of heterogeneity between the studies, resulting in biased results.

We acknowledge certain limitations in our study. First, for the top 3 diagnosis, top 5 diagnosis, and triage and classification, this may bias the results due to the number of included studies as well as the sample size, so caution is needed when interpreting these results. Although we minimized the heterogeneity of the research as much as possible, we cannot deny that the inclusion of different fields of study and the complexity of LLMs (such as different instructions and questioning dates) can affect the results of the study and generate heterogeneity. Therefore, caution should be exercised when interpreting the results. In addition, we did not assess the accuracy of multimodal grand prognostic models when solving medical image–related problems; with the development of

artificial intelligence, more multimodal models are being developed, and in the future, these models will become indispensable in the exploration of image-based problems in the medical field.

Conclusion

Existing studies suggest that ChatGPT-4o has an advantage for answering objective questions. For open-ended questions, ChatGPT-4 may be more credible. Humans are more accurate in the top 1 diagnosis and top 3 diagnosis of clinical cases. Claude 3 Opus performs better in the top 5 diagnosis, while for classification accuracy, Gemini is more advantageous. Although some LLMs excel at addressing medical queries, caution is advised due to the critical need for precision and rigor in medicine. Future high-quality studies and trials are necessary to gather more scientific evidence.

Acknowledgments

This work was supported by the Training Program for Young and Middle-aged Backbone Talents of Fujian Provincial Health Commission (grant number 2022GGA001), Natural Science Foundation of Fujian Province (grant number 2021J01395 and 2024J011032), Foundations of Department of Finance of Fujian Province (grant number Min Cai Zhi (2023) 830 and Min Cai Zhi (2024) 881), Joint Funding Projects for Innovation in Science and Technology of Fujian Province (grant number 2023Y9330), and Internal Supporting Project of Fuzhou University Affiliated Provincial Hospital (grant number 0080072220).

Data Availability

The data sets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

All authors were involved in the conceptualization and design of the study and reviewed all documents and materials. LW, JL, BZ, and SH collected the data, performed data analysis, interpreted the results, and wrote the first draft of the manuscript. CW, WL, and MZ were involved in the development of the protocol for the systematic review and critically reviewed the results and the manuscript. MF and SG were involved in the development of the protocol and revised the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

PRISMA checklist.
[DOCX File , 28 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Search strategy.
[DOCX File , 14 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Versions and timelines of LLMs iterations.
[DOCX File , 20 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Examples of the outcomes.
[DOCX File , 1341 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Description of 168 studies included.
[DOCX File , 196 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Quality assessment of observational study.
[DOCX File , 66 KB-Multimedia Appendix 6]

Multimedia Appendix 7

Indirect comparison results.
[DOCX File , 1355 KB-Multimedia Appendix 7]

Multimedia Appendix 8

Node splitting inconsistency test.
[DOCX File , 2606 KB-Multimedia Appendix 8]

Multimedia Appendix 9

Trace and density plots.
[DOCX File , 3634 KB-Multimedia Appendix 9]

Multimedia Appendix 10

Objective questions are stratified according to different fields of the questions.
[DOCX File , 13 KB-Multimedia Appendix 10]

References

1. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. ChatGPT and other large language models are double-edged swords. Radiology. Apr 2023;307(2):e230163. [doi: 10.1148/radiol.230163] [Medline: 36700838]
2. No authors listed. Will ChatGPT transform healthcare? Nat Med. Mar 14, 2023;29(3):505-506. [doi: 10.1038/s41591-023-02289-5] [Medline: 36918736]
3. Park SH, Pinto-Powell R, Thesen T, Lindqwister A, Levy J, Chacko R, et al. Preparing healthcare leaders of the digital age with an integrative artificial intelligence curriculum: a pilot study. Med Educ Online. Dec 31, 2024;29(1):2315684. [FREE Full text] [doi: 10.1080/10872981.2024.2315684] [Medline: 38351737]
4. Sblendorio E, Dentamaro V, Lo Cascio A, Germini F, Piredda M, Cicolini G. Integrating human expertise and automated methods for a dynamic and multi-parametric evaluation of large language models' feasibility in clinical decision-making. Int J Med Inform. Aug 2024;188:105501. [FREE Full text] [doi: 10.1016/j.ijmedinf.2024.105501] [Medline: 38810498]
5. Mu Y, He D. The potential applications and challenges of ChatGPT in the medical field. IJGM. Mar 2024;Volume 17:817-826. [doi: 10.2147/ijgm.s456659]
6. Park Y, Pillai A, Deng J, Guo E, Gupta M, Paget M, et al. Assessing the research landscape and clinical utility of large language models: a scoping review. BMC Med Inform Decis Mak. Mar 12, 2024;24(1):72. [FREE Full text] [doi: 10.1186/s12911-024-02459-6] [Medline: 38475802]
7. Copilot. Microsoft. URL: https://www.microsoft.com/en-us/microsoft-copilot [accessed 2025-04-21]
8. Gemini. Google. URL: https://gemini.google.com/ [accessed 2025-04-21]
9. Llama. Meta. URL: https://llama.meta.com/ [accessed 2025-04-21]
10. Tsoutsanis P, Tsoutsanis A. Evaluation of large language model performance on the Multi-Specialty Recruitment Assessment (MSRA) exam. Comput Biol Med. Jan 2024;168:107794. [doi: 10.1016/j.compbiomed.2023.107794] [Medline: 38043471]
11. Shukla R, Mishra A, Banerjee N, Verma A. The comparison of ChatGPT 3.5, Microsoft Bing, and Google Gemini for diagnosing cases of neuro-ophthalmology. Cureus. Apr 2024;16(4):e58232. [FREE Full text] [doi: 10.7759/cureus.58232] [Medline: 38745784]
12. Pressman SM, Borna S, Gomez-Cabello CA, Haider SA, Forte AJ. AI in hand surgery: assessing large language models in the classification and management of hand injuries. J Clin Med. May 11, 2024;13(10):2832. [FREE Full text] [doi: 10.3390/jcm13102832] [Medline: 38792374]
13. Vaishya R, Iyengar KP, Patralekh MK, Botchu R, Shirodkar K, Jain VK, et al. Effectiveness of AI-powered chatbots in responding to orthopaedic postgraduate exam questions-an observational study. Int Orthop. Aug 15, 2024;48(8):1963-1969. [doi: 10.1007/s00264-024-06182-9] [Medline: 38619565]
14. Lee Y, Tessier L, Brar K, Malone S, Jin D, McKechnie T, et al. ASMBS Artificial Intelligence and Digital Surgery Taskforce. Performance of artificial intelligence in bariatric surgery: comparative analysis of ChatGPT-4, Bing, and Bard in the American Society for Metabolic and Bariatric Surgery textbook of bariatric surgery questions. Surg Obes Relat Dis. Jul 2024;20(7):609-613. [FREE Full text] [doi: 10.1016/j.soard.2024.04.014] [Medline: 38782611]
15. 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]
16. Kaboudi N, Firouzbakht S, Shahir Eftekhar M, Fayazbakhsh F, Joharivarnoosfaderani N, Ghaderi S, et al. Diagnostic accuracy of ChatGPT for patients' triage; a systematic review and meta-analysis. Arch Acad Emerg Med. 2024;12(1):e60. [doi: 10.22037/aaem.v12i1.2384] [Medline: 39290765]

17. Patil A, Serrato P, Chisvo N, Arnaout O, See PA, Huang KT. Large language models in neurosurgery: a systematic review and meta-analysis. Acta Neurochir (Wien). Nov 23, 2024;166(1):475. [doi: 10.1007/s00701-024-06372-9] [Medline: 39579215]
18. Nguyen HC, Dang HP, Nguyen TL, Hoang V, Nguyen VA. Accuracy of latest large language models in answering multiple choice questions in dentistry: a comparative study. PLoS One. Jan 29, 2025;20(1):e0317423. [FREE Full text] [doi: 10.1371/journal.pone.0317423] [Medline: 39879192]
19. Lo CK, Mertz D, Loeb M. Newcastle-Ottawa Scale: comparing reviewers' to authors' assessments. BMC Med Res Methodol. Apr 01, 2014;14:45. [FREE Full text] [doi: 10.1186/1471-2288-14-45] [Medline: 24690082]
20. Long C, Subburam D, Lowe K, Dos Santos A, Zhang J, Hwang S, et al. ChatENT: augmented large language model for expert knowledge retrieval in otolaryngology-head and neck surgery. Otolaryngol Head Neck Surg. Oct 19, 2024;171(4):1042-1051. [doi: 10.1002/ohn.864] [Medline: 38895862]
21. Tao BK, Hua N, Milkovich J, Micieli JA. ChatGPT-3.5 and Bing Chat in ophthalmology: an updated evaluation of performance, readability, and informative sources. Eye (Lond). Jul 20, 2024;38(10):1897-1902. [doi: 10.1038/s41433-024-03037-w] [Medline: 38509182]
22. Shieh A, Tran B, He G, Kumar M, Freed JA, Majety P. Assessing ChatGPT 4.0's test performance and clinical diagnostic accuracy on USMLE STEP 2 CK and clinical case reports. Sci Rep. Apr 23, 2024;14(1):9330. [FREE Full text] [doi: 10.1038/s41598-024-58760-x] [Medline: 38654011]
23. 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 29, 2024;34(2):276-282. [FREE Full text] [doi: 10.1055/s-0043-1777746] [Medline: 38549897]
24. Singer MB, Fu JJ, Chow J, Teng CC. Development and evaluation of Aeyeconsult: a novel ophthalmology chatbot leveraging verified textbook knowledge and GPT-4. J Surg Educ. Mar 2024;81(3):438-443. [doi: 10.1016/j.jsurg.2023.11.019] [Medline: 38135548]
25. Hanna RE, Smith LR, Mhaskar R, Hanna K. Performance of language models on the family medicine in-training exam. Fam Med. Oct 2, 2024;56(9):555-560. [doi: 10.22454/fammed.2024.233738]
26. Kadoya N, Arai K, Tanaka S, Kimura Y, Tozuka R, Yasui K, et al. Assessing knowledge about medical physics in language-generative AI with large language model: using the medical physicist exam. Radiol Phys Technol. Dec 10, 2024;17(4):929-937. [doi: 10.1007/s12194-024-00838-2] [Medline: 39254919]
27. Sallam M, Al-Mahzoum K, Almutawaa RA, Alhashash JA, Dashti RA, AlSafy DR, et al. The performance of OpenAI ChatGPT-4 and Google Gemini in virology multiple-choice questions: a comparative analysis of English and Arabic responses. BMC Res Notes. Sep 03, 2024;17(1):247. [FREE Full text] [doi: 10.1186/s13104-024-06920-7] [Medline: 39228001]
28. Gravina AG, Pellegrino R, Palladino G, Imperio G, Ventura A, Federico A. Charting new AI education in gastroenterology: cross-sectional evaluation of ChatGPT and perplexity AI in medical residency exam. Dig Liver Dis. Aug 2024;56(8):1304-1311. [doi: 10.1016/j.dld.2024.02.019] [Medline: 38503659]
29. Passby L, Jenko N, Wernham A. Performance of ChatGPT on specialty certificate examination in dermatology multiple-choice questions. Clin Exp Dermatol. Jun 25, 2024;49(7):722-727. [doi: 10.1093/ced/llad197] [Medline: 37264670]
30. Sabri H, Saleh MHA, Hazrati P, Merchant K, Misch J, Kumar PS, et al. Performance of three artificial intelligence (AI)-based large language models in standardized testing; implications for AI-assisted dental education. J Periodontal Res. Feb 18, 2025;60(2):121-133. [doi: 10.1111/jre.13323] [Medline: 39030766]
31. Çamur E, Cesur T, Güneş YC. Can large language models be new supportive tools in coronary computed tomography angiography reporting? Clin Imaging. Oct 2024;114:110271. [doi: 10.1016/j.clinimag.2024.110271] [Medline: 39236553]
32. Lubitz M, Latario L. Performance of two artificial intelligence generative language models on the orthopaedic in-training examination. Orthopedics. May 2024;47(3):e146-e150. [doi: 10.3928/01477447-20240304-02] [Medline: 38466827]
33. Gupta R, Hamid A, Jhaveri M, Patel N, Suthar P. Comparative evaluation of AI models such as ChatGPT 3.5, ChatGPT 4.0, and Google Gemini in neuroradiology diagnostics. Cureus. Aug 2024;16(8):e67766. [doi: 10.7759/cureus.67766] [Medline: 39323714]
34. Lee G, Hong D, Kim S, Kim JW, Lee YH, Park SO, et al. Comparison of the problem-solving performance of ChatGPT-3.5, ChatGPT-4, Bing Chat, and Bard for the Korean emergency medicine board examination question bank. Medicine (Baltimore). Mar 01, 2024;103(9):e37325. [FREE Full text] [doi: 10.1097/MD.0000000000037325] [Medline: 38428889]
35. Is EE, Menekseoglu AK. Comparative performance of artificial intelligence models in rheumatology board-level questions: evaluating Google Gemini and ChatGPT-4o. Clin Rheumatol. Nov 28, 2024;43(11):3507-3513. [doi: 10.1007/s10067-024-07154-5] [Medline: 39340572]
36. D'Anna G, Van Cauter S, Thurnher M, Van Goethem J, Haller S. Can large language models pass official high-grade exams of the European Society of Neuroradiology courses? A direct comparison between OpenAI chatGPT 3.5, OpenAI GPT4 and Google Bard. Neuroradiology. Aug 06, 2024;66(8):1245-1250. [doi: 10.1007/s00234-024-03371-6] [Medline: 38705899]
37. Altamimi I, Alhumimidi A, Alshehri S, Alrumayan A, Al-Khlaiwi T, Meo SA, et al. The scientific knowledge of three large language models in cardiology: multiple-choice questions examination-based performance. Ann Med Surg (Lond). Jun 2024;86(6):3261-3266. [FREE Full text] [doi: 10.1097/MS9.0000000000002120] [Medline: 38846858]

38. Schoch J, Schmelz H, Strauch A, Borgmann H, Nestler T. Performance of ChatGPT-3.5 and ChatGPT-4 on the European Board of Urology (EBU) exams: a comparative analysis. World J Urol. Jul 26, 2024;42(1):445. [doi: 10.1007/s00345-024-05137-4] [Medline: 39060792]
39. May M, Körner-Riffard K, Kollitsch L, Burger M, Brookman-May SD, Rauchenwald M, et al. Evaluating the efficacy of AI chatbots as tutors in urology: a comparative analysis of responses to the 2022 In-Service Assessment of the European Board of Urology. Urol Int. Mar 30, 2024;108(4):359-366. [FREE Full text] [doi: 10.1159/000537854] [Medline: 38555637]
40. Sadeq MA, Ghorab RMF, Ashry MH, Abozaid AM, Banihani HA, Salem M, et al. AI chatbots show promise but limitations on UK medical exam questions: a comparative performance study. Sci Rep. Aug 14, 2024;14(1):18859. [FREE Full text] [doi: 10.1038/s41598-024-68996-2] [Medline: 39143077]
41. Khalpey Z, Kumar U, King N, Abraham A, Khalpey A. Large language models take on cardiothoracic surgery: a comparative analysis of the performance of four models on American Board of Thoracic Surgery exam questions in 2023. Cureus. Jul 2024;16(7):e65083. [doi: 10.7759/cureus.65083] [Medline: 39171020]
42. Patel EA, Fleischer L, Filip P, Eggerstedt M, Hutz M, Michaelides E, et al. Comparative performance of ChatGPT 3.5 and GPT4 on rhinology standardized board examination questions. OTO Open. Jun 27, 2024;8(2):e164. [FREE Full text] [doi: 10.1002/oto2.164] [Medline: 38938507]
43. Irmici G, Cozzi A, Della Pepa G, De Berardinis C, D'Ascoli E, Cellina M, et al. How do large language models answer breast cancer quiz questions? A comparative study of GPT-3.5, GPT-4 and Google Gemini. Radiol Med. Oct 13, 2024;129(10):1463-1467. [doi: 10.1007/s11547-024-01872-1] [Medline: 39138732]
44. Kollitsch L, Eredics K, Marszalek M, Rauchenwald M, Brookman-May SD, Burger M, et al. How does artificial intelligence master urological board examinations? A comparative analysis of different large language models' accuracy and reliability in the 2022 In-Service Assessment of the European Board of Urology. World J Urol. Jan 10, 2024;42(1):20. [doi: 10.1007/s00345-023-04749-6] [Medline: 38197996]
45. Morreel S, Verhoeven V, Mathysen D. Microsoft Bing outperforms five other generative artificial intelligence chatbots in the Antwerp University multiple choice medical license exam. PLOS Digit Health. Feb 14, 2024;3(2):e0000349. [FREE Full text] [doi: 10.1371/journal.pdig.0000349] [Medline: 38354127]
46. Bajčetić M, Mirčić A, Rakočević J, Đoković D, Milutinović K, Zaletel I. Comparing the performance of artificial intelligence learning models to medical students in solving histology and embryology multiple choice questions. Ann Anat. Jun 2024;254:152261. [doi: 10.1016/j.aanat.2024.152261] [Medline: 38521363]
47. Canillas Del Rey F, Canillas Arias M. Exploring the potential of artificial intelligence in traumatology: conversational answers to specific questions. Rev Esp Cir Ortop Traumatol. Jan 2025;69(1):38-46. [FREE Full text] [doi: 10.1016/j.recot.2024.05.004] [Medline: 38782358]
48. Meyer A, Riese J, Streichert T. Comparison of the performance of GPT-3.5 and GPT-4 with that of medical students on the written German Medical Licensing Examination: observational study. JMIR Med Educ. Feb 08, 2024;10:e50965. [FREE Full text] [doi: 10.2196/50965] [Medline: 38329802]
49. Toyama Y, Harigai A, Abe M, Nagano M, Kawabata M, Seki Y, et al. Performance evaluation of ChatGPT, GPT-4, and Bard on the official board examination of the Japan Radiology Society. Jpn J Radiol. Feb 04, 2024;42(2):201-207. [FREE Full text] [doi: 10.1007/s11604-023-01491-2] [Medline: 37792149]
50. Touma NJ, Caterini J, Liblk K. Is ChatGPT ready for primetime? Performance of artificial intelligence on a simulated Canadian urology board exam. Can Urol Assoc J. Oct 10, 2024;18(10):329-332. [FREE Full text] [doi: 10.5489/cuaj.8800] [Medline: 38896484]
51. Chan J, Dong T, Angelini G. The performance of large language models in intercollegiate Membership of the Royal College of Surgeons examination. Ann R Coll Surg Engl. Nov 06, 2024;106(8):700-704. [FREE Full text] [doi: 10.1308/rcsann.2024.0023] [Medline: 38445611]
52. Patil NS, Huang RS, van der Pol CB, Larocque N. Comparative performance of ChatGPT and Bard in a text-based radiology knowledge assessment. Can Assoc Radiol J. May 14, 2024;75(2):344-350. [FREE Full text] [doi: 10.1177/08465371231193716] [Medline: 37578849]
53. Hubany S, Scala F, Hashemi K, Kapoor S, Fedorova JR, Vaccaro MJ, et al. ChatGPT-4 surpasses residents: a study of artificial intelligence competency in plastic surgery in-service examinations and its advancements from ChatGPT-3.5. Plast Reconstr Surg Glob Open. Sep 2024;12(9):e6136. [doi: 10.1097/GOX.0000000000006136] [Medline: 39239234]
54. Nakajima N, Fujimori T, Furuya M, Kanie Y, Imai H, Kita K, et al. A comparison between GPT-3.5, GPT-4, and GPT-4V: can the large language model (ChatGPT) pass the Japanese Board of Orthopaedic Surgery examination? Cureus. Mar 2024;16(3):e56402. [FREE Full text] [doi: 10.7759/cureus.56402] [Medline: 38633935]
55. Thibaut G, Dabbagh A, Liverneaux P. Does Google's Bard Chatbot perform better than ChatGPT on the European hand surgery exam? Int Orthop. Jan 15, 2024;48(1):151-158. [doi: 10.1007/s00264-023-06034-y] [Medline: 37968408]
56. Lum Z, Collins D, Dennison S, Guntupalli L, Choudhary S, Saiz AM, et al. Generative artificial intelligence performs at a second-year orthopedic resident level. Cureus. Mar 2024;16(3):e56104. [FREE Full text] [doi: 10.7759/cureus.56104] [Medline: 38618358]

57. Menekşeoğlu AK, İş EE. Comparative performance of artificial ıntelligence models in physical medicine and rehabilitation board-level questions. Rev Assoc Med Bras (1992). 2024;70(7):e20240241. [FREE Full text] [doi: 10.1590/1806-9282.20240241] [Medline: 39045939]
58. Cheong RCT, Pang KP, Unadkat S, Mcneillis V, Williamson A, Joseph J, et al. Performance of artificial intelligence chatbots in sleep medicine certification board exams: ChatGPT versus Google Bard. Eur Arch Otorhinolaryngol. Apr 2024;281(4):2137-2143. [doi: 10.1007/s00405-023-08381-3] [Medline: 38117307]
59. Mesnard B, Schirmann A, Branchereau J, Perrot O, Bogaert G, Neuzillet Y, et al. Artificial intelligence: ready to pass the European Board examinations in urology? Eur Urol Open Sci. Feb 2024;60:44-46. [FREE Full text] [doi: 10.1016/j.euros.2024.01.002] [Medline: 38321995]
60. Ming S, Guo Q, Cheng W, Lei B. Influence of model evolution and system roles on ChatGPT's performance in Chinese medical licensing exams: comparative study. JMIR Med Educ. Aug 13, 2024;10:e52784-e52784. [FREE Full text] [doi: 10.2196/52784] [Medline: 39140269]
61. Chow R, Hasan S, Zheng A, Gao C, Valdes G, Yu F, et al. The accuracy of artificial intelligence ChatGPT in oncology examination questions. J Am Coll Radiol. Nov 2024;21(11):1800-1804. [FREE Full text] [doi: 10.1016/j.jacr.2024.07.011] [Medline: 39098369]
62. Kim SE, Lee JH, Choi BS, Han H, Lee MC, Ro DH. Performance of ChatGPT on solving orthopedic board-style questions: a comparative analysis of ChatGPT 3.5 and ChatGPT 4. Clin Orthop Surg. Aug 2024;16(4):669-673. [FREE Full text] [doi: 10.4055/cios23179] [Medline: 39092297]
63. Oura T, Tatekawa H, Horiuchi D, Matsushita S, Takita H, Atsukawa N, et al. Diagnostic accuracy of vision-language models on Japanese diagnostic radiology, nuclear medicine, and interventional radiology specialty board examinations. Jpn J Radiol. Dec 20, 2024;42(12):1392-1398. [doi: 10.1007/s11604-024-01633-0] [Medline: 39031270]
64. Lewandowski M, Łukowicz P, Świetlik D, Barańska-Rybak W. ChatGPT-3.5 and ChatGPT-4 dermatological knowledge level based on the Specialty Certificate Examination in Dermatology. Clin Exp Dermatol. Jun 25, 2024;49(7):686-691. [doi: 10.1093/ced/llad255] [Medline: 37540015]
65. Knoedler L, Alfertshofer M, Knoedler S, Hoch CC, Funk PF, Cotofana S, et al. Pure wisdom or Potemkin villages? A comparison of ChatGPT 3.5 and ChatGPT 4 on USMLE Step 3 style questions: quantitative analysis. JMIR Med Educ. Jan 05, 2024;10:e51148. [FREE Full text] [doi: 10.2196/51148] [Medline: 38180782]
66. Khan AA, Yunus R, Sohail M, Rehman TA, Saeed S, Bu Y, et al. Artificial intelligence for anesthesiology board-style examination questions: role of large language models. J Cardiothorac Vasc Anesth. May 2024;38(5):1251-1259. [doi: 10.1053/j.jvca.2024.01.032] [Medline: 38423884]
67. Sheikh MS, Thongprayoon C, Qureshi F, Suppadungsuk S, Kashani KB, Miao J, et al. Personalized medicine transformed: ChatGPT's contribution to continuous renal replacement therapy alarm management in intensive care units. J Pers Med. Feb 22, 2024;14(3):233. [FREE Full text] [doi: 10.3390/jpm14030233] [Medline: 38540976]
68. Mayo-Yáñez M, Lechien JR, Maria-Saibene A, Vaira LA, Maniaci A, Chiesa-Estomba CM. Examining the performance of ChatGPT 3.5 and Microsoft Copilot in otolaryngology: a comparative study with otolaryngologists' evaluation. Indian J Otolaryngol Head Neck Surg. Aug 01, 2024;76(4):3465-3469. [doi: 10.1007/s12070-024-04729-1] [Medline: 39130248]
69. Rydzewski NR, Dinakaran D, Zhao SG, Ruppin E, Turkbey B, Citrin DE, et al. Comparative evaluation of LLMs in clinical oncology. NEJM AI. May 25, 2024;1(5):1. [doi: 10.1056/aioa2300151] [Medline: 39131700]
70. Wang T, Mu J, Chen J, Lin C. Comparing ChatGPT and clinical nurses' performances on tracheostomy care: a cross-sectional study. Int J Nurs Stud Adv. Jun 2024;6:100181. [FREE Full text] [doi: 10.1016/j.ijnsa.2024.100181] [Medline: 38746816]
71. Liang R, Zhao A, Peng L, Xu X, Zhong J, Wu F, et al. Enhanced artificial intelligence strategies in renal oncology: iterative optimization and comparative analysis of GPT 3.5 versus 4.0. Ann Surg Oncol. Jun 12, 2024;31(6):3887-3893. [doi: 10.1245/s10434-024-15107-0] [Medline: 38472675]
72. Jaworski A, Jasiński D, Jaworski W, Hop A, Janek A, Sławińska B, et al. Comparison of the performance of artificial intelligence versus medical professionals in the Polish final medical examination. Cureus. Aug 2024;16(8):e66011. [doi: 10.7759/cureus.66011] [Medline: 39221376]
73. Bharatha A, Ojeh N, Fazle Rabbi AM, Campbell M, Krishnamurthy K, Layne-Yarde R, et al. Comparing the performance of ChatGPT-4 and medical students on MCQs at varied levels of Bloom's taxonomy. AMEP. May 2024;Volume 15:393-400. [doi: 10.2147/amep.s457408]
74. Le M, Davis M. ChatGPT yields a passing score on a pediatric board preparatory exam but raises red flags. Global Pediatric Health. Mar 24, 2024;11:1. [doi: 10.1177/2333794x241240327]
75. Arango S, Flynn J, Zeitlin J, Lorenzana DJ, Miller AJ, Wilson MS, et al. The performance of ChatGPT on the American Society for Surgery of the Hand self-assessment examination. Cureus. Apr 2024;16(4):e58950. [FREE Full text] [doi: 10.7759/cureus.58950] [Medline: 38800302]
76. Rojas M, Rojas M, Burgess V, Toro-Pérez J, Salehi S. Exploring the performance of ChatGPT versions 3.5, 4, and 4 with vision in the Chilean medical licensing examination: observational study. JMIR Med Educ. Apr 29, 2024;10:e55048-e55048. [FREE Full text] [doi: 10.2196/55048] [Medline: 38686550]

77. Chau RCW, Thu KM, Yu OY, Hsung RT, Lo ECM, Lam WYH. Performance of generative artificial intelligence in dental licensing examinations. Int Dent J. Jun 2024;74(3):616-621. [FREE Full text] [doi: 10.1016/j.identj.2023.12.007] [Medline: 38242810]
78. Thirunavukarasu AJ, Mahmood S, Malem A, Foster WP, Sanghera R, Hassan R, et al. Large language models approach expert-level clinical knowledge and reasoning in ophthalmology: a head-to-head cross-sectional study. PLOS Digit Health. Apr 17, 2024;3(4):e0000341. [FREE Full text] [doi: 10.1371/journal.pdig.0000341] [Medline: 38630683]
79. Bicknell BT, Butler D, Whalen S, Ricks J, Dixon CJ, Clark AB, et al. ChatGPT-4 Omni performance in USMLE disciplines and clinical skills: comparative analysis. JMIR Med Educ. Nov 06, 2024;10:e63430. [FREE Full text] [doi: 10.2196/63430] [Medline: 39504445]
80. Haddad F, Saade JS. Performance of ChatGPT on ophthalmology-related questions across various examination levels: observational study. JMIR Med Educ. Jan 18, 2024;10:e50842. [FREE Full text] [doi: 10.2196/50842] [Medline: 38236632]
81. Noda R, Izaki Y, Kitano F, Komatsu J, Ichikawa D, Shibagaki Y. Performance of ChatGPT and Bard in self-assessment questions for nephrology board renewal. Clin Exp Nephrol. May 14, 2024;28(5):465-469. [doi: 10.1007/s10157-023-02451-w] [Medline: 38353783]
82. Yudovich MS, Makarova E, Hague CM, Raman JD. Performance of GPT-3.5 and GPT-4 on standardized urology knowledge assessment items in the United States: a descriptive study. J Educ Eval Health Prof. Jul 08, 2024;21:17. [FREE Full text] [doi: 10.3352/jeehp.2024.21.17] [Medline: 38977032]
83. Li D, Kao Y, Tsai S, Bai Y, Yeh T, Chu C, et al. Comparing the performance of ChatGPT GPT-4, Bard, and Llama-2 in the Taiwan Psychiatric Licensing Examination and in differential diagnosis with multi-center psychiatrists. Psychiatry Clin Neurosci. Jun 26, 2024;78(6):347-352. [doi: 10.1111/pcn.13656] [Medline: 38404249]
84. Farhat F, Chaudhry BM, Nadeem M, Sohail SS, Madsen D. Evaluating large language models for the National Premedical Exam in India: comparative analysis of GPT-3.5, GPT-4, and Bard. JMIR Med Educ. Feb 21, 2024;10:e51523. [FREE Full text] [doi: 10.2196/51523] [Medline: 38381486]
85. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. Feb 08, 2023;9:e45312. [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
86. Kung JE, Marshall C, Gauthier C, Gonzalez TA, Jackson JB. Evaluating ChatGPT performance on the orthopaedic in-training examination. JB JS Open Access. 2023;8(3):1. [FREE Full text] [doi: 10.2106/JBJS.OA.23.00056] [Medline: 37693092]
87. Gencer A, Aydin S. Can ChatGPT pass the thoracic surgery exam? Am J Med Sci. Oct 2023;366(4):291-295. [doi: 10.1016/j.amjms.2023.08.001] [Medline: 37549788]
88. Ali R, Tang OY, Connolly ID, Zadnik Sullivan PL, Shin JH, Fridley JS, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. Neurosurgery. Dec 01, 2023;93(6):1353-1365. [doi: 10.1227/neu.0000000000002632] [Medline: 37581444]
89. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT-3.5, ChatGPT-4, and orthopaedic resident performance on orthopaedic assessment examinations. J Am Acad Orthop Surg. Dec 01, 2023;31(23):1173-1179. [FREE Full text] [doi: 10.5435/JAAOS-D-23-00396] [Medline: 37671415]
90. Suchman K, Garg S, Trindade AJ. Chat Generative Pretrained Transformer fails the multiple-choice American College of Gastroenterology self-assessment test. Am J Gastroenterol. Dec 01, 2023;118(12):2280-2282. [doi: 10.14309/ajg.0000000000002320] [Medline: 37212584]
91. Sakai D, Maeda T, Ozaki A, Kanda G, Kurimoto Y, Takahashi M. Performance of ChatGPT in board examinations for specialists in the Japanese Ophthalmology Society. Cureus. Dec 2023;15(12):e49903. [FREE Full text] [doi: 10.7759/cureus.49903] [Medline: 38174202]
92. Huang Y, Gomaa A, Semrau S, Haderlein M, Lettmaier S, Weissmann T, et al. Benchmarking ChatGPT-4 on a radiation oncology in-training exam and Red Journal Gray Zone cases: potentials and challenges for AI-assisted medical education and decision making in radiation oncology. Front Oncol. 2023;13:1265024. [FREE Full text] [doi: 10.3389/fonc.2023.1265024] [Medline: 37790756]
93. Yanagita Y, Yokokawa D, Uchida S, Tawara J, Ikusaka M. Accuracy of ChatGPT on medical questions in the National Medical Licensing Examination in Japan: evaluation study. JMIR Form Res. Oct 13, 2023;7:e48023. [FREE Full text] [doi: 10.2196/48023] [Medline: 37831496]
94. Teebagy S, Colwell L, Wood E, Yaghy A, Faustina M. Improved performance of ChatGPT-4 on the OKAP Examination: a comparative study with ChatGPT-3.5. J Acad Ophthalmol (2017). Jul 11, 2023;15(2):e184-e187. [FREE Full text] [doi: 10.1055/s-0043-1774399] [Medline: 37701862]
95. Kaneda Y, Takahashi R, Kaneda U, Akashima S, Okita H, Misaki S, et al. Assessing the performance of GPT-3.5 and GPT-4 on the 2023 Japanese Nursing Examination. Cureus. Aug 2023;15(8):e42924. [FREE Full text] [doi: 10.7759/cureus.42924] [Medline: 37667724]
96. Flores-Cohaila JA, García-Vicente A, Vizcarra-Jiménez SF, De la Cruz-Galán JP, Gutiérrez-Arratia JD, Quiroga Torres BG, et al. Performance of ChatGPT on the Peruvian National Licensing Medical Examination: cross-sectional study. JMIR Med Educ. Sep 28, 2023;9:e48039. [FREE Full text] [doi: 10.2196/48039] [Medline: 37768724]

97. Fowler T, Pullen S, Birkett L. Performance of ChatGPT and Bard on the official part 1 FRCOphth practice questions. Br J Ophthalmol. Sep 20, 2024;108(10):1379-1383. [doi: 10.1136/bjo-2023-324091] [Medline: 37932006]
98. Moshirfar M, Altaf AW, Stoakes IM, Tuttle JJ, Hoopes PC. Artificial intelligence in ophthalmology: a comparative analysis of GPT-3.5, GPT-4, and human expertise in answering StatPearls questions. Cureus. Jun 2023;15(6):e40822. [FREE Full text] [doi: 10.7759/cureus.40822] [Medline: 37485215]
99. Brin D, Sorin V, Vaid A, Soroush A, Glicksberg BS, Charney AW, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep. Oct 01, 2023;13(1):16492. [FREE Full text] [doi: 10.1038/s41598-023-43436-9] [Medline: 37779171]
100. Miao J, Thongprayoon C, Garcia Valencia OA, Krisanapan P, Sheikh MS, Davis PW, et al. Performance of ChatGPT on nephrology test questions. CJASN. Oct 18, 2023;19(1):35-43. [doi: 10.2215/cjn.0000000000000330]
101. Kaneda Y, Namba M, Kaneda U, Tanimoto T. Artificial intelligence in childcare: assessing the performance and acceptance of ChatGPT responses. Cureus. Aug 2023;15(8):e44484. [FREE Full text] [doi: 10.7759/cureus.44484] [Medline: 37791148]
102. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: comparison study. JMIR Med Educ. Jun 29, 2023;9:e48002. [FREE Full text] [doi: 10.2196/48002] [Medline: 37384388]
103. Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Zadnik Sullivan PL, et al. Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. Neurosurgery. Nov 01, 2023;93(5):1090-1098. [doi: 10.1227/neu.0000000000002551] [Medline: 37306460]
104. Ohta K, Ohta S. The Performance of GPT-3.5, GPT-4, and Bard on the Japanese National Dentist Examination: a comparison study. Cureus. Dec 2023;15(12):e50369. [FREE Full text] [doi: 10.7759/cureus.50369] [Medline: 38213361]
105. Watari T, Takagi S, Sakaguchi K, Nishizaki Y, Shimizu T, Yamamoto Y, et al. Performance comparison of ChatGPT-4 and Japanese medical residents in the General Medicine In-Training Examination: comparison study. JMIR Med Educ. Dec 06, 2023;9:e52202. [FREE Full text] [doi: 10.2196/52202] [Medline: 38055323]
106. Roos J, Kasapovic A, Jansen T, Kaczmarczyk R. Artificial intelligence in medical education: comparative analysis of ChatGPT, Bing, and medical students in Germany. JMIR Med Educ. Sep 04, 2023;9:e46482. [FREE Full text] [doi: 10.2196/46482] [Medline: 37665620]
107. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, Alas-Brun R, Onambele L, Ortega W, et al. Evaluating the efficacy of ChatGPT in navigating the Spanish Medical Residency Entrance Examination (MIR): promising horizons for AI in clinical medicine. Clin Pract. Nov 20, 2023;13(6):1460-1487. [FREE Full text] [doi: 10.3390/clinpract13060130] [Medline: 37987431]
108. Huang RS, Lu KJQ, Meaney C, Kemppainen J, Punnett A, Leung F. Assessment of resident and AI chatbot performance on the University of Toronto family medicine residency progress test: comparative study. JMIR Med Educ. Sep 19, 2023;9:e50514. [FREE Full text] [doi: 10.2196/50514] [Medline: 37725411]
109. Schubert MC, Wick W, Venkataramani V. Performance of large language models on a neurology board-style examination. JAMA Netw Open. Dec 01, 2023;6(12):e2346721. [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.46721] [Medline: 38060223]
110. Torres-Zegarra BC, Rios-Garcia W, Ñaña-Cordova AM, Arteaga-Cisneros KF, Chalco XCB, Ordoñez MAB, et al. Performance of ChatGPT, Bard, Claude, and Bing on the Peruvian National Licensing Medical Examination: a cross-sectional study. J Educ Eval Health Prof. Nov 20, 2023;20:30. [FREE Full text] [doi: 10.3352/jeehp.2023.20.30] [Medline: 37981579]
111. Kirshteyn G, Golan R, Chaet M. Performance of ChatGPT vs. HuggingChat on OB-GYN topics. Cureus. Mar 2024;16(3):e56187. [FREE Full text] [doi: 10.7759/cureus.56187] [Medline: 38618446]
112. van Nuland M, Erdogan A, Açar C, Contrucci R, Hilbrants S, Maanach L, et al. Performance of ChatGPT on factual knowledge questions regarding clinical pharmacy. J Clin Pharmacol. Sep 16, 2024;64(9):1095-1100. [doi: 10.1002/jcph.2443] [Medline: 38623909]
113. Danesh A, Pazouki H, Danesh F, Danesh A, Vardar‐Sengul S. Artificial intelligence in dental education: ChatGPT's performance on the periodontic in‐service examination. Journal of Periodontology. Jan 10, 2024;95(7):682-687. [doi: 10.1002/jper.23-0514]
114. Huang CY, Zhang E, Caussade M, Brown T, Stockton Hogrogian G, Yan AC. Pediatric dermatologists versus AI bots: evaluating the medical knowledge and diagnostic capabilities of ChatGPT. Pediatr Dermatol. May 09, 2024;41(5):831-834. [doi: 10.1111/pde.15649] [Medline: 38721744]
115. Fiedler B, Azua EN, Phillips T, Ahmed AS. ChatGPT performance on the American Shoulder and Elbow Surgeons maintenance of certification exam. J Shoulder Elbow Surg. Sep 2024;33(9):1888-1893. [doi: 10.1016/j.jse.2024.02.029] [Medline: 38580067]
116. Coleman MC, Moore JN. Two artificial intelligence models underperform on examinations in a veterinary curriculum. J Am Vet Med Assoc. May 01, 2024;262(5):692-697. [FREE Full text] [doi: 10.2460/javma.23.12.0666] [Medline: 38382193]
117. Abbas A, Rehman MS, Rehman SS. Comparing the performance of popular large language models on the National Board of Medical Examiners sample questions. Cureus. Mar 2024;16(3):e55991. [FREE Full text] [doi: 10.7759/cureus.55991] [Medline: 38606229]

118. Jarou ZJ, Dakka A, McGuire D, Bunting L. ChatGPT versus human performance on emergency medicine board preparation questions. Ann Emerg Med. Jan 2024;83(1):87-88. [doi: 10.1016/j.annemergmed.2023.08.010] [Medline: 37725017]
119. Sensoy E, Citirik M. Assessing the proficiency of artificial intelligence programs in the diagnosis and treatment of cornea, conjunctiva, and eyelid diseases and exploring the advantages of each other benefits. Cont Lens Anterior Eye. Apr 2024;47(2):102125. [doi: 10.1016/j.clae.2024.102125] [Medline: 38443209]
120. Guerra GA, Hofmann HL, Le JL, Wong AM, Fathi A, Mayfield CK, et al. ChatGPT, Bard, and Bing chat are large language processing models that answered orthopaedic in-training examination questions with similar accuracy to first-year orthopaedic surgery residents. Arthroscopy. Mar 2025;41(3):557-562. [doi: 10.1016/j.arthro.2024.08.023] [Medline: 39209078]
121. Agarwal M, Goswami A, Sharma P. Evaluating ChatGPT-3.5 and Claude-2 in answering and explaining conceptual medical physiology multiple-choice questions. Cureus. Sep 2023;15(9):e46222. [FREE Full text] [doi: 10.7759/cureus.46222] [Medline: 37908959]
122. Cheong KX, Zhang C, Tan T, Fenner BJ, Wong WM, Teo KY, et al. Comparing generative and retrieval-based chatbots in answering patient questions regarding age-related macular degeneration and diabetic retinopathy. Br J Ophthalmol. Sep 20, 2024;108(10):1443-1449. [doi: 10.1136/bjo-2023-324533] [Medline: 38749531]
123. Zhou S, Luo X, Chen C, Jiang H, Yang C, Ran G, et al. The performance of large language model-powered chatbots compared to oncology physicians on colorectal cancer queries. Int J Surg. Oct 01, 2024;110(10):6509-6517. [doi: 10.1097/JS9.0000000000001850] [Medline: 38935100]
124. Kozaily E, Geagea M, Akdogan ER, Atkins J, Elshazly MB, Guglin M, et al. Accuracy and consistency of online large language model-based artificial intelligence chat platforms in answering patients' questions about heart failure. Int J Cardiol. Aug 01, 2024;408:132115. [doi: 10.1016/j.ijcard.2024.132115] [Medline: 38697402]
125. Xia S, Hua Q, Mei Z, Xu W, Lai L, Wei M, et al. Clinical application potential of large language model: a study based on thyroid nodules. Endocrine. Jan 30, 2025;87(1):206-213. [doi: 10.1007/s12020-024-03981-3] [Medline: 39080210]
126. Lee Y, Shin T, Tessier L, Javidan A, Jung J, Hong D, et al. ASMBS Artificial Intelligence and Digital Surgery Task Force. Harnessing artificial intelligence in bariatric surgery: comparative analysis of ChatGPT-4, Bing, and Bard in generating clinician-level bariatric surgery recommendations. Surg Obes Relat Dis. Jul 2024;20(7):603-608. [FREE Full text] [doi: 10.1016/j.soard.2024.03.011] [Medline: 38644078]
127. Doğan L, Özçakmakcı GB, Yılmaz E. The performance of chatbots and the AAPOS website as a tool for amblyopia education. J Pediatr Ophthalmol Strabismus. Apr 25, 2024;61(5):325-331. [doi: 10.3928/01913913-20240409-01] [Medline: 38661309]
128. Lee T, Campbell D, Patel S, Hossain A, Radfar N, Siddiqui E, et al. Unlocking health literacy: the ultimate guide to hypertension education from ChatGPT versus Google Gemini. Cureus. May 2024;16(5):e59898. [FREE Full text] [doi: 10.7759/cureus.59898] [Medline: 38721479]
129. Lang SP, Yoseph ET, Gonzalez-Suarez AD, Kim R, Fatemi P, Wagner K, et al. Analyzing large language models' responses to common lumbar spine fusion surgery questions: a comparison between ChatGPT and Bard. Neurospine. Jun 2024;21(2):633-641. [FREE Full text] [doi: 10.14245/ns.2448098.049] [Medline: 38955533]
130. Iannantuono G, Bracken-Clarke D, Karzai F, Choo-Wosoba H, Gulley J, Floudas C. Comparison of large language models in answering immuno-oncology questions: a cross-sectional study. Oncologist. May 03, 2024;29(5):407-414. [FREE Full text] [doi: 10.1093/oncolo/oyae009] [Medline: 38309720]
131. Anguita R, Downie C, Ferro Desideri L, Sagoo MS. Assessing large language models' accuracy in providing patient support for choroidal melanoma. Eye (Lond). Nov 13, 2024;38(16):3113-3117. [doi: 10.1038/s41433-024-03231-w] [Medline: 39003430]
132. Zhang Y, Dong Y, Mei Z, Hou Y, Wei M, Yeung YH, et al. Performance of large language models on benign prostatic hyperplasia frequently asked questions. Prostate. Jun 2024;84(9):807-813. [doi: 10.1002/pros.24699] [Medline: 38558009]
133. Xue E, Bracken-Clarke D, Iannantuono GM, Choo-Wosoba H, Gulley JL, Floudas CS. Utility of large language models for health care professionals and patients in navigating hematopoietic stem cell transplantation: comparison of the performance of ChatGPT-3.5, ChatGPT-4, and Bard. J Med Internet Res. May 17, 2024;26:e54758. [FREE Full text] [doi: 10.2196/54758] [Medline: 38758582]
134. Cao JJ, Kwon DH, Ghaziani TT, Kwo P, Tse G, Kesselman A, et al. Large language models' responses to liver cancer surveillance, diagnosis, and management questions: accuracy, reliability, readability. Abdom Radiol (NY). Dec 01, 2024;49(12):4286-4294. [doi: 10.1007/s00261-024-04501-7] [Medline: 39088019]
135. Monroe CL, Abdelhafez YG, Atsina K, Aman E, Nardo L, Madani MH. Evaluation of responses to cardiac imaging questions by the artificial intelligence large language model ChatGPT. Clin Imaging. Aug 2024;112:110193. [FREE Full text] [doi: 10.1016/j.clinimag.2024.110193] [Medline: 38820977]
136. Chervonski E, Harish KB, Rockman CB, Sadek M, Teter KA, Jacobowitz GR, et al. Generative artificial intelligence chatbots may provide appropriate informational responses to common vascular surgery questions by patients. Vascular. Feb 18, 2025;33(1):229-237. [doi: 10.1177/17085381241240550] [Medline: 38500300]
137. Kassab J, Hadi El Hajjar A, Wardrop RM, Brateanu A. Accuracy of online artificial intelligence models in primary care settings. Am J Prev Med. Jun 2024;66(6):1054-1059. [doi: 10.1016/j.amepre.2024.02.006] [Medline: 38354991]

138. Al-Sharif E, Penteado R, Dib El Jalbout N, Topilow NJ, Shoji MK, Kikkawa DO, et al. Evaluating the accuracy of ChatGPT and Google BARD in fielding oculoplastic patient queries: a comparative study on artificial versus human intelligence. Ophthalmic Plast Reconstr Surg. 2024;40(3):303-311. [doi: 10.1097/IOP.0000000000002567] [Medline: 38215452]
139. Mejia MR, Arroyave JS, Saturno M, Ndjonko LCM, Zaidat B, Rajjoub R, et al. Use of ChatGPT for determining clinical and surgical treatment of lumbar disc herniation with radiculopathy: a North American Spine Society guideline comparison. Neurospine. Mar 2024;21(1):149-158. [FREE Full text] [doi: 10.14245/ns.2347052.526] [Medline: 38291746]
140. Lee T, Rao A, Campbell D, Radfar N, Dayal M, Khrais A. Evaluating ChatGPT-3.5 and ChatGPT-4.0 responses on hyperlipidemia for patient education. Cureus. May 2024;16(5):e61067. [FREE Full text] [doi: 10.7759/cureus.61067] [Medline: 38803402]
141. Oliveira AL, Coelho M, Guedes LC, Cattoni MB, Carvalho H, Duarte-Batista P. Performance of ChatGPT 3.5 and 4 as a tool for patient support before and after DBS surgery for Parkinson's disease. Neurol Sci. Dec 29, 2024;45(12):5757-5764. [doi: 10.1007/s10072-024-07732-0] [Medline: 39198356]
142. Lim ZW, Pushpanathan K, Yew SME, Lai Y, Sun C, Lam JSH, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine. Sep 2023;95:104770. [FREE Full text] [doi: 10.1016/j.ebiom.2023.104770] [Medline: 37625267]
143. Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI responds to common lung cancer questions: ChatGPT vs Google Bard. Radiology. Jun 2023;307(5):e230922. [doi: 10.1148/radiol.230922] [Medline: 37310252]
144. Pushpanathan K, Lim ZW, Er Yew SM, Chen DZ, Hui'En Lin HA, Lin Goh JH, et al. Popular large language model chatbots' accuracy, comprehensiveness, and self-awareness in answering ocular symptom queries. iScience. Nov 17, 2023;26(11):108163. [FREE Full text] [doi: 10.1016/j.isci.2023.108163] [Medline: 37915603]
145. Coskun BN, Yagiz B, Ocakoglu G, Dalkilic E, Pehlivan Y. Assessing the accuracy and completeness of artificial intelligence language models in providing information on methotrexate use. Rheumatol Int. Mar 25, 2024;44(3):509-515. [doi: 10.1007/s00296-023-05473-5] [Medline: 37747564]
146. King RC, Samaan JS, Yeo YH, Peng Y, Kunkel DC, Habib AA, et al. A multidisciplinary assessment of ChatGPT's knowledge of amyloidosis: observational study. JMIR Cardio. Apr 19, 2024;8:e53421. [FREE Full text] [doi: 10.2196/53421] [Medline: 38640472]
147. Pinto VBP, de Azevedo MF, Wroclawski ML, Gentile G, Jesus VLM, de Bessa Junior J, et al. Conformity of ChatGPT recommendations with the AUA/SUFU guideline on postprostatectomy urinary incontinence. Neurourol Urodyn. Apr 07, 2024;43(4):935-941. [doi: 10.1002/nau.25442] [Medline: 38451040]
148. Momenaei B, Wakabayashi T, Shahlaee A, Durrani AF, Pandit SA, Wang K, et al. Assessing ChatGPT-3.5 versus ChatGPT-4 performance in surgical treatment of retinal diseases: a comparative study. Ophthalmic Surg Lasers Imaging Retina. Aug 2024;55(8):481-482. [FREE Full text] [doi: 10.3928/23258160-20240227-02] [Medline: 38531015]
149. Stevenson E, Walsh C, Hibberd L. Can artificial intelligence replace biochemists? A study comparing interpretation of thyroid function test results by ChatGPT and Google Bard to practising biochemists. Ann Clin Biochem. Mar 20, 2024;61(2):143-149. [doi: 10.1177/00045632231203473] [Medline: 37699796]
150. Dronkers EAC, Geneid A, Al Yaghchi C, Lechien JR. Evaluating the potential of AI chatbots in treatment decision-making for acquired bilateral vocal fold paralysis in adults. J Voice. Apr 06, 2024:1. [FREE Full text] [doi: 10.1016/j.jvoice.2024.02.020] [Medline: 38584026]
151. Rahimli Ocakoglu S, Coskun B. The emerging role of AI in patient education: a comparative analysis of LLM accuracy for pelvic organ prolapse. Med Princ Pract. Mar 25, 2024;33(4):330-337. [FREE Full text] [doi: 10.1159/000538538] [Medline: 38527444]
152. Gandhi AP, Joesph FK, Rajagopal V, Aparnavi P, Katkuri S, Dayama S, et al. Performance of ChatGPT on the India undergraduate community medicine examination: cross-sectional study. JMIR Form Res. Mar 25, 2024;8:e49964. [FREE Full text] [doi: 10.2196/49964] [Medline: 38526538]
153. Tariq R, Malik S, Khanna S. Evolving landscape of large language models: an evaluation of ChatGPT and Bard in answering patient queries on colonoscopy. Gastroenterology. Jan 2024;166(1):220-221. [doi: 10.1053/j.gastro.2023.08.033] [Medline: 37634736]
154. Li P, Zhang X, Zhu E, Yu S, Sheng B, Tham YC, et al. Potential multidisciplinary use of large language models for addressing queries in cardio-oncology. J Am Heart Assoc. Mar 19, 2024;13(6):e033584. [FREE Full text] [doi: 10.1161/JAHA.123.033584] [Medline: 38497458]
155. Sosa BR, Cung M, Suhardi VJ, Morse K, Thomson A, Yang HS, et al. Capacity for large language model chatbots to aid in orthopedic management, research, and patient queries. J Orthop Res. Jun 21, 2024;42(6):1276-1282. [doi: 10.1002/jor.25782] [Medline: 38245845]
156. 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]
157. Warrier A, Singh R, Haleem A, Zaki H, Eloy JA. The comparative diagnostic capability of large language models in otolaryngology. Laryngoscope. Sep 02, 2024;134(9):3997-4002. [doi: 10.1002/lary.31434] [Medline: 38563415]

158. 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]
159. Hirosawa T, Harada Y, Mizuta K, Sakamoto T, Tokumasu K, Shimizu T. Diagnostic performance of generative artificial intelligences for a series of complex case reports. Digit Health. Jul 21, 2024;10:20552076241265215. [FREE Full text] [doi: 10.1177/20552076241265215] [Medline: 39229463]
160. Mandalos A, Tsouris D. Artificial versus human intelligence in the diagnostic approach of ophthalmic case scenarios: a qualitative evaluation of performance and consistency. Cureus. Jun 2024;16(6):e62471. [doi: 10.7759/cureus.62471] [Medline: 39015855]
161. Krusche M, Callhoff J, Knitza J, Ruffer N. Diagnostic accuracy of a large language model in rheumatology: comparison of physician and ChatGPT-4. Rheumatol Int. Feb 24, 2024;44(2):303-306. [FREE Full text] [doi: 10.1007/s00296-023-05464-6] [Medline: 37742280]
162. 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]
163. 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]
164. 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]
165. Albaladejo A, Lorleac'h A, Allain J. [The spring of artificial intelligence: AI vs. expert for internal medicine cases]. Rev Med Interne. Jul 2024;45(7):409-414. [FREE Full text] [doi: 10.1016/j.revmed.2024.01.012] [Medline: 38331591]
166. 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]
167. 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]
168. 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. [FREE Full text] [doi: 10.3390/ijerph20043378] [Medline: 36834073]
169. 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 03, 2023;11:e49995. [FREE Full text] [doi: 10.2196/49995] [Medline: 37788063]
170. Rojas-Carabali W, Cifuentes-González C, Wei X, Putera I, Sen A, Thng ZX, et al. Evaluating the diagnostic accuracy and management recommendations of ChatGPT in uveitis. Ocul Immunol Inflamm. Oct 18, 2024;32(8):1526-1531. [doi: 10.1080/09273948.2023.2253471] [Medline: 37722842]
171. Gräf M, Knitza J, Leipe J, Krusche M, Welcker M, Kuhn S, et al. Comparison of physician and artificial intelligence-based symptom checker diagnostic accuracy. Rheumatol Int. Dec 2022;42(12):2167-2176. [FREE Full text] [doi: 10.1007/s00296-022-05202-4] [Medline: 36087130]
172. 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]
173. 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.e18. [doi: 10.1016/j.amjmed.2023.08.003] [Medline: 37643659]
174. 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]
175. 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 16, 2024;281(5):2717-2721. [doi: 10.1007/s00405-024-08509-z] [Medline: 38365990]
176. Shemer A, Cohen M, Altarescu A, Atar-Vardi M, Hecht I, Dubinsky-Pertzov B, et al. Diagnostic capabilities of ChatGPT in ophthalmology. Graefes Arch Clin Exp Ophthalmol. Jul 06, 2024;262(7):2345-2352. [doi: 10.1007/s00417-023-06363-z] [Medline: 38183467]
177. Gunes Y, Cesur T. The diagnostic performance of large language models and general radiologists in thoracic radiology cases: a comparative study. J Thorac Imaging. Sep 13, 2024:2024. [doi: 10.1097/RTI.0000000000000805] [Medline: 39269227]
178. Sarangi PK, Irodi A, Panda S, Nayak DSK, Mondal H. Radiological differential diagnoses based on cardiovascular and thoracic imaging patterns: perspectives of four large language models. Indian J Radiol Imaging. Apr 28, 2024;34(2):269-275. [FREE Full text] [doi: 10.1055/s-0043-1777289] [Medline: 38549881]

179. Berg HT, van Bakel B, van de Wouw L, Jie KE, Schipper A, Jansen H, et al. ChatGPT and generating a differential diagnosis early in an emergency department presentation. Ann Emerg Med. Jan 2024;83(1):83-86. [doi: 10.1016/j.annemergmed.2023.08.003] [Medline: 37690022]
180. Haider SA, Pressman SM, Borna S, Gomez-Cabello CA, Sehgal A, Leibovich BC, et al. Evaluating large language model (LLM) performance on established breast classification systems. Diagnostics (Basel). Jul 11, 2024;14(14):1491. [FREE Full text] [doi: 10.3390/diagnostics14141491] [Medline: 39061628]
181. Gan RK, Ogbodo JC, Wee YZ, Gan AZ, González PA. Performance of Google bard and ChatGPT in mass casualty incidents triage. Am J Emerg Med. Jan 2024;75:72-78. [FREE Full text] [doi: 10.1016/j.ajem.2023.10.034] [Medline: 37967485]
182. Aiumtrakul N, Thongprayoon C, Arayangkool C, Vo KB, Wannaphut C, Suppadungsuk S, et al. Personalized medicine in urolithiasis: AI chatbot-assisted dietary management of oxalate for kidney stone prevention. J Pers Med. Jan 18, 2024;14(1):107. [FREE Full text] [doi: 10.3390/jpm14010107] [Medline: 38248809]
183. Wang H, Gao C, Dantona C, Hull B, Sun J. DRG-LLaMA: tuning LLaMA model to predict diagnosis-related group for hospitalized patients. NPJ Digit Med. Jan 22, 2024;7(1):16. [FREE Full text] [doi: 10.1038/s41746-023-00989-3] [Medline: 38253711]
184. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. Aug 2023;620(7972):172-180. [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
185. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. Aug 17, 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
186. Researcher Access Program application. OpenAI. URL: https://platform.openai.com/docs/model-index-for-researchers [accessed 2025-04-21]
187. Zhou J, He X, Sun L, Xu J, Chen X, Chu Y, et al. Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. Nat Commun. Jul 05, 2024;15(1):5649. [FREE Full text] [doi: 10.1038/s41467-024-50043-3] [Medline: 38969632]
188. 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]
189. Bagde H, Dhopte A, Alam MK, Basri R. A systematic review and meta-analysis on ChatGPT and its utilization in medical and dental research. Heliyon. Dec 2023;9(12):e23050. [FREE Full text] [doi: 10.1016/j.heliyon.2023.e23050] [Medline: 38144348]
190. Levin G, Horesh N, Brezinov Y, Meyer R. Performance of ChatGPT in medical examinations: a systematic review and a meta-analysis. BJOG. Feb 2024;131(3):378-380. [doi: 10.1111/1471-0528.17641] [Medline: 37604703]

Abbreviations

API: application programming interface
CINEMA: Confidence in Network Meta-Analysis
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
LLM: large language model
NMA: network meta-analysis
OR: odds ratio
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SUCRA: surface under the cumulative ranking curve

Edited by X Ma; submitted 18.07.24; peer-reviewed by M Lotfinia, X-M Zhang; comments to author 14.10.24; revised version received 04.02.25; accepted 03.04.25; published 30.04.25

Please cite as:
Wang L, Li J, Zhuang B, Huang S, Fang M, Wang C, Li W, Zhang M, Gong S
Accuracy of Large Language Models When Answering Clinical Research Questions: Systematic Review and Network Meta-Analysis
J Med Internet Res 2025;27:e64486
URL: https://www.jmir.org/2025/1/e64486
doi: 10.2196/64486
PMID: 40305085