BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Analyzing evaluation methods for large language models in the medical field: a scoping review

Junbok Lee[1,2], Sungkyung Park[3], Jaeyong Shin[4,5*] and Belong Cho[2,6,7*]

Abstract

Background Owing to the rapid growth in the popularity of Large Language Models (LLMs), various performance evaluation studies have been conducted to confirm their applicability in the medical field. However, there is still no clear framework for evaluating LLMs.

Objective This study reviews studies on LLM evaluations in the medical field and analyzes the research methods used in these studies. It aims to provide a reference for future researchers designing LLM studies.

Methods & materials We conducted a scoping review of three databases (PubMed, Embase, and MEDLINE) to identify LLM-related articles published between January 1, 2023, and September 30, 2023. We analyzed the types of methods, number of questions (queries), evaluators, repeat measurements, additional analysis methods, use of prompt engineering, and metrics other than accuracy.

Results A total of 142 articles met the inclusion criteria. LLM evaluation was primarily categorized as either providing test examinations ($n=53$, 37.3%) or being evaluated by a medical professional ($n=80$, 56.3%), with some hybrid cases ($n=5$, 3.5%) or a combination of the two ($n=4$, 2.8%). Most studies had 100 or fewer questions ($n=18$, 29.0%), 15 (24.2%) performed repeated measurements, 18 (29.0%) performed additional analyses, and 8 (12.9%) used prompt engineering. For medical assessment, most studies used 50 or fewer queries ($n=54$, 64.3%), had two evaluators ($n=43$, 48.3%), and 14 (14.7%) used prompt engineering.

Conclusions More research is required regarding the application of LLMs in healthcare. Although previous studies have evaluated performance, future studies will likely focus on improving performance. A well-structured methodology is required for these studies to be conducted systematically.

Keywords Large language model, LLM, Evaluation methods

*Correspondence:
Jaeyong Shin
drshin@yuhs.ac
Belong Cho
belong@snu.ac.kr
[1]Institute for Innovation in Digital Healthcare, Yonsei University, Seoul, Republic of Korea
[2]Department of Human Systems Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea
[3]Department of Bigdata AI Management Information, Seoul National University of Science and Technology, Seoul, Republic of Korea
[4]Department of Preventive Medicine and Public Health, Yonsei University College of Medicine, 50-1, Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea
[5]Institute of Health Services Research, Yonsei University College of Medicine, Seoul, Korea
[6]Department of Family Medicine, Seoul National University Hospital, Seoul, Republic of Korea
[7]Department of Family Medicine, Seoul National University College of Medicine, 101 Daehak-ro, Jongno-gu, Seoul 03080, Republic of Korea

Introduction

A Large Language Model (LLM) is a type of artificial intelligence (AI) designed to mimic human language processing using deep learning techniques trained on large amounts of textual data from various sources [1]. The rapid increase in the popularity of LLMs has led to numerous attempts to utilize them across different fields, with many demonstrating a significant level of competence [2]. LLMs are designed to respond to a wide range of topics, making them helpful tools for customer service, chatbots, and many other applications, hence the keen interest in their use in the medical field [3–6].

Several LLMs are currently accessible to researchers, each with unique features. OpenAI's ChatGPT is widely used for its strong language understanding and generation capabilities [7]. Google's Bard leverages vast search data to provide factual and accurate information [8]. Microsoft's Bing Chat integrates chat with search for real-time information access [9]. In contrast, open-source LLMs like Meta's LLaMA and Stanford's Alpaca allow for customization and experimentation. While commercial models offer ease of use and technical support, open-source models provide flexibility and cost-effectiveness.

Various studies have been conducted in the medical field to verify the performance of LLMs. The following topics are being studied for the application of LLMs: (1) diagnostic and clinical decision support, (2) automation of medical records, (3) patient education and support, and (4) medical research and data analytics. LLMs can be utilized in diagnostic and clinical decision support to suggest possible diagnoses or treatment options based on a patient's symptoms, medical history, and test results [10]. For medical record automation, LLMs have been studied for their potential to automatically organize and document patient encounters or generate explanatory materials to provide patients with information from their medical records [11]. Additionally, LLMs can enhance patient health literacy by explaining diseases, treatment options, and medication instructions in easy-to-understand language [12].

Only when LLMs perform at a human-like level in medical knowledge and reasoning assessments can users have sufficient confidence in their responses, and LLMs are useful in medical settings [13–15]. A framework has been proposed for LLM evaluation [16]. However, no clear methodology exists for evaluating LLMs in the medical field. In this study, we review the evaluation of LLMs in medicine. Based on these findings, we discuss essential points to consider when evaluating LLMs in medical applications.

Methods & materials

Study design

A scoping review aims to systematically synthesize knowledge within a defined area and explore and map key concepts, available evidence, and shortcomings of the existing research; it was determined to be the most appropriate method for this study [17, 18]. We considered several methodological approaches, including a systematic and narrative review. Systematic reviews focus narrowly on specific questions and similar methodologies, making them unsuitable for emerging technical areas with diverse evidence [19]. Narrative reviews offer the flexibility to synthesize diverse literature but lack the systematic rigor required to comprehensively map the research landscape and identify gaps in the literature [20]. The study followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) [21].

Search strategy

We conducted a preliminary review to establish the search strategies. Initially, we searched the PubMed database using the keyword "Large Language Model*." PubMed was utilized in the preliminary review of this study because it is a commonly suggested database to use when conducting a systematic review. A total of 498 articles were retrieved, and the review found that 73 articles that evaluated LLMs were identified.

We found that papers assessing LLMs tended to use keywords such as "evaluation, assessment, performance, and comparison." In addition, we included commercially available programs such as ChatGPT, Google's Bard, and Microsoft's Bing Chat in our search strategy. Given that the terminology for LLM gained prominence after 2023, we focused our search on the literature published between January 1, 2023, and September 30, 2023. We opted not to utilize MeSH terms, as the term LLM has only been in full use since early 2023; therefore, using MeSH terms may not reflect the latest research trends. After establishing the search strategy, we systematically searched MEDLINE, PubMed, and EMBASE (Fig. 1). The final search results were sent to EndNote to remove any duplicates. The search strategy used in this study is presented in Appendix Table 2.

Selecting and screening studies

The screening process comprised two stages. Initially, articles were screened for relevance based on the information presented in the title and abstract, followed by a thorough assessment of inclusion based on the full text. Two authors (JB and SK) independently reviewed the articles. In cases of disagreement between the reviewers, a third independent reviewer (JY) was consulted to reach a consensus. For studies to be eligible for inclusion,

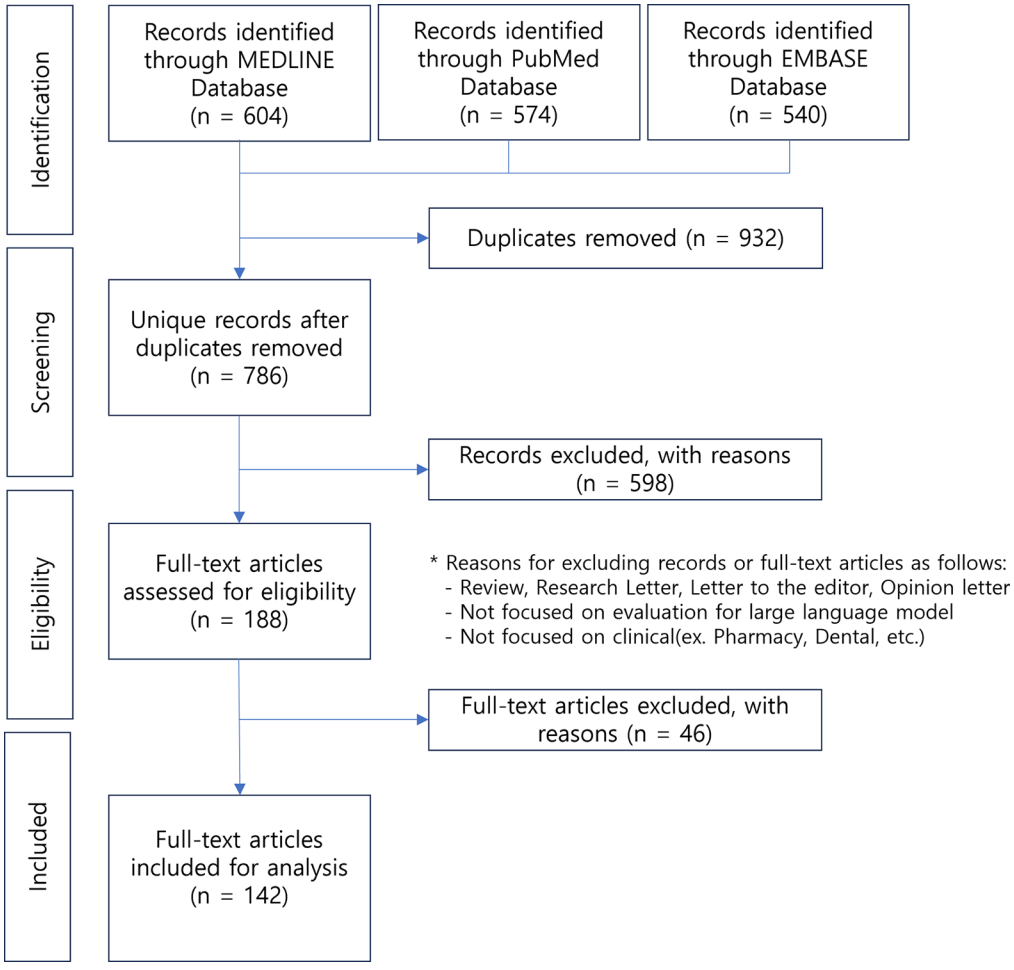

Fig. 1 Analysis of evaluation by medical professionals

they had to meet specific criteria, including being written in English and addressing the evaluation of LLMs in healthcare settings. We excluded publications such as conference abstracts, editorials, reviews, research letters, letters to the editor, and opinion letters. Articles in the pharmacy and dentistry fields were excluded from the screening process. A comprehensive list of inclusion and exclusion criteria can be found in the appendix.

Extracting and analyzing the data

We summarized information regarding the evaluation method, type of LLMs, and medical specialty for the studies included in the review.

For test-based evaluation, we analyzed the number of questions and repeated measurements, use of prompt engineering (e.g., few-shot learning, role-based prompting), additional analysis, whether the questions were analyzed for difficulty, and the primary outcomes. The number of repeated measurements refers to administering the same prompt more than once to evaluate the consistency of the responses generated by the LLMs. In prompt engineering, few-shot learning refers to a methodology where a small number of examples are provided within a prompt to guide the model in learning patterns and generating correct answers, while role-based prompting means designing prompts so that the model adopts a specific role when answering. Additional analysis refers to whether further assessments were conducted beyond accuracy, including the overall adequacy of logical reasoning and evidence provided, the frequency of hallucinations, and the error types made in incorrect responses. For the LLM evaluation by medical professionals, we analyzed the number of queries, number of repeated measurements, number of evaluators, use of prompt engineering, evaluation tools and sources, evaluation items, and scales.

In contrast to studies that conducted both test-based evaluations and assessments by healthcare professionals, some studies removed the selection of test questions and had medical professionals evaluate the LLM responses directly, a method referred to as a hybrid approach. Research that used both methods or a hybrid approach

Table 1 LLMs evaluation methods

Methods	(*n* = 142)
	N (%)
Test questions	53 (37.3)
Expert evaluation	80 (56.3)
Hybrid approach	5 (3.5)
Both	4 (2.8)

was analyzed by including them in test-based evaluations and evaluations by medical professionals.

Statistical analysis

We conducted a Kruskal-Wallis test to compare the performance differences among the four LLMs (GPT-3.5, GPT-4, Bing Chat, and Bard). This nonparametric method was chosen due to the small sample size for each model evaluation. Following this, post hoc pairwise comparisons were conducted using the Mann-Whitney U test with Bonferroni correction applied to account for multiple comparisons, identifying which specific pairs of models show significant differences. For all statistical tests, a significance level of 0.05 was used. The analysis was performed using STATA 16 (StataCorp LLC, College Station, TX, USA).

Results

Overview

Following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flowchart (Fig. 1), four review steps were performed: identification, screening, eligibility assessment, and final consensus. The initial search retrieved 1,718 unique articles. Automatic deduplication using ENDNOTE removed 894 articles, 38 of which were identified as manual duplicates. After reviewing all the abstracts, 598 (34.8%) were excluded based on the exclusion criteria. A total of 188 (10.9%) articles underwent a full-text review, of which 46 (2.7%) did not meet the inclusion criteria, leaving 142 (8.3%) for final inclusion and analysis. The Appendix Tables 8 and 9 provide data excerpts from the papers.

Characteristics of published literature

The effectiveness of the LLMs was assessed in two ways: evaluation based on test examination (n=53, 37.3%) [22–74] and evaluation by medical professionals (n=80, 56.3%) [75–154]. Others used a combination of both (n=4, 2.8%) [155–158] or a hybrid approach to evaluate the LLMs' responses to the test examination (n=5, 3.5%) [159–163] (Table 1).

Articles evaluating LLM often used several models instead of only one (n=88, 54%). A total of 218 LLMs were used in 142 studies, including this study (Table 2). The most common LLM used was the Open AI's GPT-3.5 (n=114, 52.3%), followed by GPT-4 (n=65, 29.8%). Google's Bard (n=15, 6.9%) and Microsoft's Bing Chat (n=12, 5.5%) were the third and fourth most common. A few models were developed by fine-tuning the models (n=3, 1.4%).

In terms of medical specialties, internal medicine (n=23, 16.2%) was the most common medical specialty to which the LLMs were applied (Table 3), followed by radiology (n=16, 11.3%) and ophthalmology (n=15, 10.6%). Regarding Internal Medicine, Cardiovascular Disease and Gastroenterology had the highest number

Table 3 LLMs used in medical specialties

Medical Specialty	(*n* = 142)	
	N	(%)
Anesthesiology	1	(0.7)
Dermatology	4	(2.8)
Emergency Medicine	3	(2.1)
Family Medicine	1	(0.7)
Internal Medicine	23	(16.2)
Neurological Surgery	5	(3.5)
Obstetrics and Gynecology	6	(4.2)
Ophthalmology	15	(10.6)
Orthopaedic Surgery	9	(6.3)
Otolaryngology – Head and Neck Surgery	7	(4.9)
Pathology	4	(2.8)
Pediatrics	1	(0.7)
Plastic Surgery	4	(2.8)
Psychiatry and Neurology	6	(4.2)
Radiology	16	(11.3)
Surgery	4	(2.8)
Thoracic Surgery	3	(2.1)
Urology	8	(5.6)
General Practice	19	(13.4)
ETC (clinical informatics, nursing)	3	(2.1)

Table 2 LLMs used in the evaluation

Language Model	Expert evaluation	Test questions	Hybrid approach	Both	Total
GPT-3.5	61	45	4	4	114 (52.1)
GPT-4	30	34	1	1	66 (30.1)
Bard	8	6	-	1	15 (6.8)
Bing Chat	7	4	-	1	12 (5.5)
ETC (GPT-3, GPT-2)	2	7	-	-	9 (4.1)
Fine tuning	3	-	-	-	3 (1.4)
Total	**111**	**96**	**5**	**7**	**219 (100.0)**

of LLM evaluation papers (six each). In addition, some general practices did not belong to a specific medical department ($n=19$, 13.4%) and were mainly validated by examination. We derived suggestions for systematically designing studies evaluating LLMs in healthcare based on our findings.

Evaluation based on test examination

Regarding the number of questions used for evaluation, less than 100 were the most common ($n=18$, 29.0%), followed by 200–300 ($n=14$, 22.6%), then 100–200 ($n=11$, 17.7%), and 500 or more ($n=11$, 17.7%) (Table 4). Regarding repeated measures, about three-quarters of the studies did not perform any repeated measures ($n=47$, 75.8%). Five papers (8.1%) did this twice, six papers did it three times (9.7%), and four papers did it four or more times (6.5%). Eight (12.9%) studies applied prompt engineering to improve the LLM performance. Seven studies employed role-based prompting, while one used the few-shot learning method with examples. Eighteen papers (29.0%) conducted additional analyses beyond simply measuring correct responses to the questions, and fourteen papers (14%) conducted analyses based on question difficulty.

Table 4 Analysis of evaluation based on test examinations

	(*n*=62)	
	N	(%)
Number of questions		
1–100	18	(29.0)
101–200	11	(17.7)
201–300	14	(22.6)
301–400	5	(8.1)
401–500	3	(4.8)
501-	11	(17.7)
Number of repeated measurements		
0	47	(75.8)
2	5	(8.1)
3	6	(9.7)
above 4	4	(6.5)
Prompt engineering		
Yes	8	(12.9)
No	54	(87.1)
Additional analysis		
Yes	18	(29.0)
No	44	(71.0)
Difficulty		
Yes	14	(22.6)
No	48	(77.4)

The performance of LLMs is illustrated in Fig. 2. Among the models evaluated, GPT-4 exhibited the highest accuracy (mean: 76.47, median: 79.65, SD: 12.57, IQR: 12.30), while GPT-3.5 (mean: 57.62, median: 57.00, SD: 13.26, IQR: 16.25) and Bing Chat (mean: 57.61, median: 68.33, SD: 21.10, IQR: 18.95) demonstrated lower accuracy scores. Bard had the lowest accuracy (mean: 49.63, median: 46.67, SD: 14.94, IQR: 21.82. We conducted a Kruskal-Wallis test to assess the differences in performance across the models statistically. The analysis revealed significant model differences ($H=35.51$, $p<0.001$). Post-hoc analysis using the Mann-Whitney U test indicated significant differences between GPT-4 and GPT-3.5 ($z = -5.50$, $p<0.001$), GPT-4 and Bard ($z = -3.52$, $p<0.001$), and GPT-4 and Bing Chat ($z = -2.00$, $p=0.045$). However, no significant differences were found between GPT-3.5 and Bard, GPT-3.5 and Bing Chat, or Bard and Bing Chat ($p>0.05$). It is important to note that the number of studies involving Bard and Bing Chat is limited, and results should be interpreted cautiously.

Evaluation by medical professionals

Regarding the number of queries, 50 or fewer was the most common ($n=54$, 64.3%), followed by 50–100 ($n=14$, 16.7%), then 151–200 ($n=7$, 8.3%), 101–150 ($n=6$, 7.1%), and 150–200 ($n=7$, 8.3%) (Table 5). Regarding repeated measures, about 70% of the studies did not perform any repeated measures ($n=63$, 70.8%). Eleven papers (12.4%) did it twice, ten papers did it three times (11.2%), and five papers did it five times (5.6%). Among the experts who evaluated the LLMs, two were the most common ($n=43$,

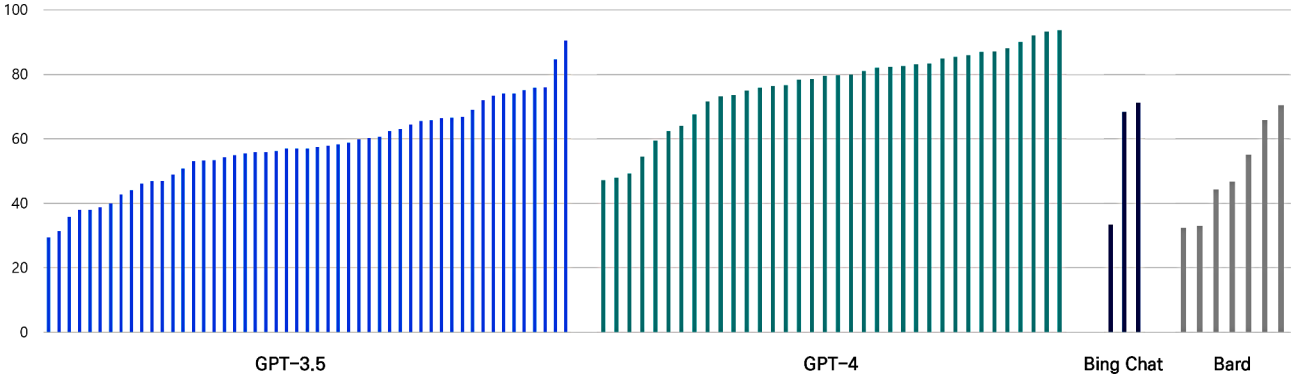

Fig. 2 Performance of GPT-3.5, GPT-4, Bing Chat and Bard

Table 5 Analysis of evaluation by medical professionals

	(*n* = 89)	
Number of queries	**N**	**(%)**
1–50	54	(64.3)
51–100	14	(16.7)
101–150	6	(7.1)
151–200	7	(8.3)
201 above	3	(3.6)
Number of repeat measurements	**N**	**(%)**
0	63	(70.8)
2	11	(12.4)
3	10	(11.2)
5	5	(5.6)
Number of evaluators	**N**	**(%)**
1	5	(5.6)
2	43	(48.3)
3	13	(14.6)
4	3	(3.4)
5 above	17	(19.1)
(not indicated)	8	(9.0)
Prompt Tuning	**N**	**(%)**
None	76	(85.4)
Role-based prompting	6	(6.7)
Few shots learning	2	(2.3)
Explain context	2	(2.3)
Template	3	(3.4)

48.3%), followed by 3 (n=13, 14.6) and by 1 (n=5). Thirteen (14.7%) studies applied prompt engineering to improve the LLM performance.

In addition to accuracy, we identified several metrics used for LLM evaluation (Table 6). The papers we reviewed evaluated whether the LLM's responses were in concord with guidelines or expert opinions (n=12) or whether the responses were appropriate (n=9), complete (n=8), or of high quality (n=8). A few studies also assessed the safety (n=5) or readability (n=3) of the responses, as well as their clarity (n=3).

Discussion

This study aimed to analyze methods for evaluating LLMs in medicine. For LLM performance evaluation, two main methods were used: evaluation based on test examinations and evaluation by medical professionals. In addition, there is a method that uses both methods together and a hybrid method. Evaluation based on a test examination was used to evaluate LLM performance according to the medical specialty, and evaluation by medical professionals was mainly used when LLMs were utilized for particular purposes, such as clinical decision support or answering questions. Based on our findings, we derived suggestions for systematically designing studies evaluating LLMs in healthcare (Fig. 3).

Table 6 Metrics used for LLMs evaluation (except accuracy)

Metric	No.
Concordance / Agreement with guidelines or experts	12
Appropriateness	9
Completeness	8
Quality	8
Reproducibility	7
Safety, Extent of harm	6
Readability	5
Relevance	5
Clarity	3
Acceptability	2
Comprehensiveness	2
Currency of information	2
Efficacy	2
Empathetic	2
Helpfulness	2
Reliability	2
Understandability	2
Usefulness	2
Preference	1
Satisfaction	1
Specificity	1
Validity of references	1
Over-conclusiveness	1
Supplemental information	1
Objectivity	1
Bias	1
Adaptiveness	1

Better evaluations based on test examination

For evaluations based on a test examination, beyond simply presenting the percentage of correct responses, studies should also be reviewed to ensure the evidence is presented. Some studies have performed additional evaluations, such as checking for concordance with the proposed correct answer or reviewing responses to ensure that they are well-founded with an appropriate rationale; however, the number of such studies remains relatively limited [25, 26, 34, 37, 39, 41, 46]. Given the high number of reports on hallucinations in the LLMs, additional reviews are needed to ensure that hallucinations, poorly supported answers, and reasoning are accurately reported [164–166].

In a test examination-based evaluation, the difficulty of the questions is a critical factor to consider. Approximately a quarter of the papers included in this study have evaluated performance based on question difficulty. By analyzing performance about difficulty, researchers can determine the level of complexity at which the LLMs perform optimally or begin to degrade. It is also essential to assess how the LLMs respond to varying difficulty levels, as clinical settings often present questions of differing complexities. Finally, such analysis can guide future research to improve model performance through prompt

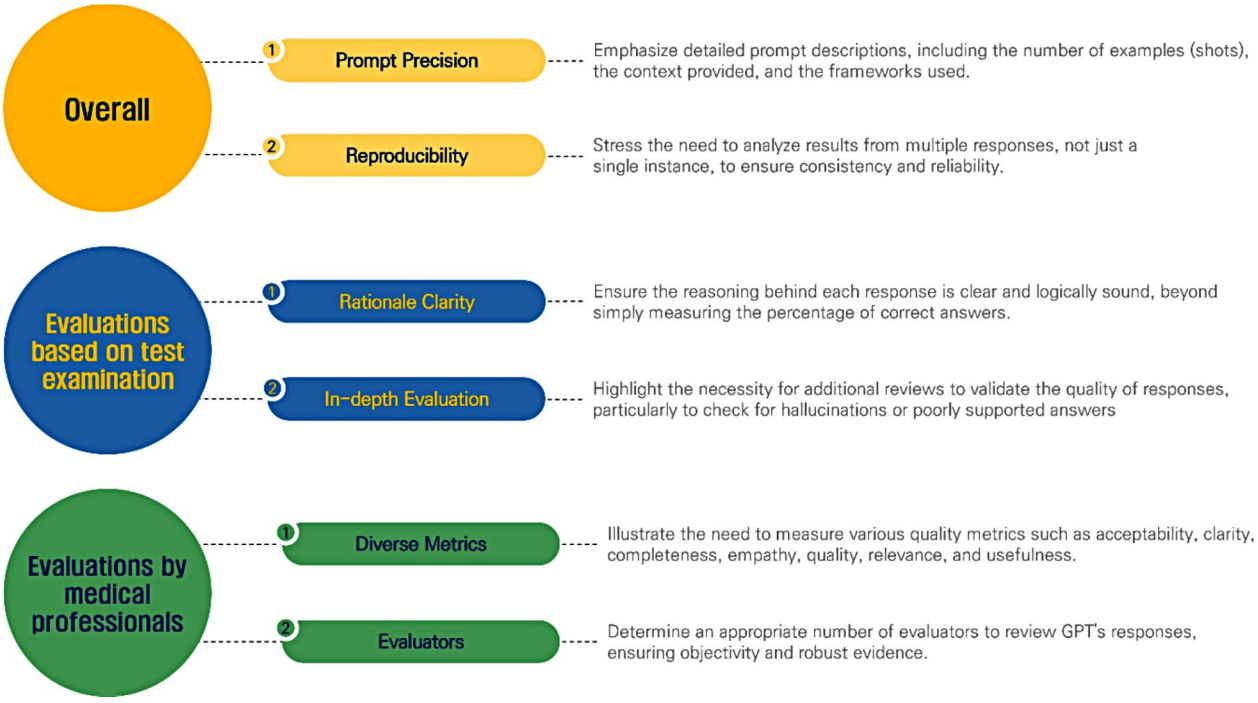

Fig. 3 Improving the evaluation of LLMs

engineering or fine-tuning approaches. Developing an evaluation framework to verify the reasoning behind LLMs is required. Some studies analyzed incorrect answers by categorizing them as logical, informational, or statistical errors, and one study proposed a CVSA (Concordance, Validity, Safety, and Accuracy) model [49].

Better evaluations by medical professionals

Most studies on LLMs have measured the accuracy. However, it is also necessary to measure various other metrics. In addition to accuracy, the reviewed studies measured concordance with guidelines or expert opinion and the responses' appropriateness, completeness, quality, safety, readability, and clarity. Although it may vary from one medical field to another or depending on the purpose of the study, an evaluation frame or guideline for the evaluation of LLMs is also needed. Because different people may have different ideas about a term, researchers must precisely describe what they measure and present the scale.

In addition, most studies used two people to evaluate the LLM responses, but two people should not be considered an appropriate number for evaluating the LLM performance. When future guidelines for LLM evaluation are developed, an appropriate number of evaluators should be considered to ensure representativeness.

Need for considering reproducibility

A study design that considers reproducibility is required. Some studies have performed two or three repeated measurements to ensure reproducibility. Because the LLMs do not always provide the same response, we believe it is better to draw results and analyze them for multiple responses rather than just one. Studies that have validated reproducibility have reported reproducibility rates of 90–100% [88, 89, 113, 142]. While 5–10% may not seem like a lot, given the specificity of the medical field, we believe that reproducibility should be considered.

Need for accurate prompt descriptions

Lastly, an accurate description of the prompts is necessary. The LLMs can produce very different results depending on how the prompts are written. Various engineering methods have been proposed to improve the LLM performance. Therefore, researchers must be precise regarding prompting. For example, it is necessary to be precise about the number of examples for few-shot learning, whether roles-based prompting is given, and the use of frameworks. Some studies did not provide supplementary materials or figures for the prompts, so checking how the prompts were written was impossible. Therefore, it would be helpful for follow-up studies to provide supplementary materials about the writing of the prompts. Additionally, it would be helpful for researchers to maintain a version of the prompts when recording or revising a study.

Limitations

The limitations of this study were as follows. First, some of the LLM evaluation studies may not have been included due to the scope of our search strategy. Specifically, we did not include open-source models such as LLAMA or ALPACA, which may have led to the omission of relevant studies. Additionally, while we utilized representative databases for our search, similar research may exist in other databases not included in our review. However, we believe the large number of reviewed papers (142) mitigates this limitation. In future research, including more databases and expanding the search strategy to cover additional models could help address these limitations. Additionally, the scoping review required a qualitative evaluation of the studies, but this was not performed because there is no established evaluation methodology for LLMs. We hope this study will contribute to developing criteria for the qualitative evaluation of LLM research.

Conclusion

LLMs are being applied in various ways, and we expect them to become more advanced. They have several potential applications in medicine. However, according to the medical field's characteristics, accuracy is critical, and incorrect information should not be provided to patients. It is necessary to increase reliability through various evaluations before LLMs can be used in the medical field. Future studies should conduct additional analyses to examine factors such as reasoning ability, hallucinations, and the difficulty of test questions. Moreover, these studies should consider applying metrics beyond accuracy and ensure reproducibility.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-024-02709-7.

Supplementary Material 1

Acknowledgements
Not applicable.

Author contributions
Concept and design: JB, BL. Acquisition, analysis, or interpretation of data: All authors. Drafting of the manuscript: JB, SK. Critical revision of the manuscript for important intellectual content: All authors. Administrative, technical, or material support: BL, JY. Supervision: BL, JY.

Funding
This research was supported by the Technology Innovation Program (20018246), funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea).

Data availability
The data from this study is available in the supplementary file. Additional data can be requested from the corresponding author if needed.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 19 January 2024 / Accepted: 3 October 2024
Published online: 29 November 2024

References

1. Thirunavukarasu AJ, et al. Large language models in medicine. Nat Med. 2023;29:1930–40.
2. Lund BD, Wang T. Chatting about ChatGPT: how may AI and GPT impact academia and libraries? Libr Hi Tech News. 2023;40:26–9.
3. Abd-Alrazaq A, et al. Large Language models in Medical Education: opportunities, challenges, and future directions. JMIR Med Educ. 2023;9:e48291.
4. Iannantuono GM, et al. Applications of large language models in cancer care: current evidence and future perspectives. Front Oncol. 2023;13:1268915.
5. Qiu J et al. Large ai models in health informatics: Applications, challenges, and the future. *IEEE Journal of Biomedical and Health Informatics* (2023). (2023).
6. Temsah M-H et al. MDPI,. Chatgpt and the future of digital health: a study on healthcare workers' perceptions and expectations. In *Healthcare* 1812 (2023).
7. Wu T, et al. A brief overview of ChatGPT: the history, status quo and potential future development. IEEE/CAA J Automatica Sinica. 2023;10:1122–36.
8. Rahaman MS et al. The AI race is on! Google's Bard and OpenAI's ChatGPT head to head: an opinion article. *Mizanur and Rahman, Md Nafizur, The AI Race is on* (2023). (2023).
9. Hill JE, Harris C, Clegg A. Methods for using Bing's AI-powered search engine for data extraction for a systematic review. Res Synthesis Methods. 2024;15:347–53.
10. Liu S, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc. 2023;30:1237–45.
11. Guo E, et al. Automated paper screening for clinical reviews using large language models: data analysis study. J Med Internet Res. 2024;26:e48996.
12. Subramanian CR, Yang DA, Khanna R. Enhancing health care communication with large language models—the role, challenges, and future directions. JAMA Netw Open. 2024;7:e240347–240347.
13. Karabacak M, Margetis K. Embracing large Language models for Medical Applications: opportunities and challenges. Cureus 15 (2023).
14. Choudhury A, Shamszare H. Investigating the impact of user trust on the adoption and use of ChatGPT: Survey Analysis. J Med Internet Res. 2023;25:e47184.
15. Shahsavar Y, Choudhury A. User intentions to Use ChatGPT for self-diagnosis and health-related Purposes: cross-sectional survey study. JMIR Hum Factors. 2023;10:e47564.
16. Reddy S. Evaluating large language models for use in healthcare: A framework for translational value assessment. *Informatics in Medicine Unlocked* (2023). 101304 (2023).
17. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol. 2005;8:19–32.
18. Colquhoun HL, et al. Scoping reviews: time for clarity in definition, methods, and reporting. J Clin Epidemiol. 2014;67:1291–4.
19. Munn Z, et al. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol. 2018;18:1–7.
20. Pavlenko A. Narrative analysis. Blackwell Guide Res Methods Biling Multiling. 2008;311–325:2008.
21. Tricco AC, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. 2018;169:467–73.
22. Ali R et al. Performance of ChatGPT, GPT-4, and Google bard on a neurosurgery oral boards preparation question bank. *Neurosurgery* (2022). 10.1227 (2022).
23. Ali R et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. *medRxiv* (2023). 2023.2003. 2025.23287743 (2023).

24. Antaki F et al. Evaluating the performance of chatgpt in ophthalmology: An analysis of its successes and shortcomings. *Ophthalmology Science* (2023). 100324 (2023).
25. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: Insights into current strengths and limitations. *Radiology* (2023). 230582 (2023).
26. Cai LZ et al. Performance of Generative Large Language Models on Ophthalmology Board Style Questions. *American Journal of Ophthalmology* (2023). (2023).
27. Chen TC et al. Chat GPT as a Neuro-Score Calculator: Analysis of a Large Language Model's Performance on Various Neurological Exam Grading Scales. *World Neurosurgery* (2023). (2023).
28. Cohen A et al. Performance of ChatGPT in Israeli Hebrew OBGYN national residency examinations. *Archives of Gynecology and Obstetrics* (2023). 1–6 (2023).
29. Cuthbert R, Simpson AI. Artificial intelligence in orthopaedics: can chat generative pre-trained transformer (ChatGPT) pass Sect. 1 of the Fellowship of the Royal College of Surgeons (Trauma & Orthopaedics) examination? Postgrad Med J. 2023;99:1110–4.
30. Deebel NA, Terlecki R. ChatGPT performance on the American Urological Association (AUA) Self-Assessment Study Program and the potential influence of artificial intelligence (AI) in urologic training. *Urology* (2023). (2023).
31. Friederichs H, Friederichs WJ, März M. ChatGPT in medical school: how successful is AI in progress testing? Med Educ Online. 2023;28:2220920.
32. Gencer A, Aydin S. Can ChatGPT pass the thoracic surgery exam? Am J Med Sci. 2023;366:291–5.
33. Giannos P. Evaluating the limits of AI in medical specialisation: ChatGPT's performance on the UK Neurology Specialty Certificate Examination. BMJ Neurol Open 5 (2023).
34. Gilson A, et al. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. 2023;9:e45312.
35. Guerra GA et al. GPT-4 artificial intelligence model outperforms ChatGPT, medical students, and neurosurgery residents on neurosurgery written board-like questions. *World Neurosurgery* (2023). (2023).
36. Guigue PA et al. Performance of ChatGPT in French language Parcours d'Accès Spécifique Santé test and in OBGYN. *International Journal of Gynecology & Obstetrics* (2023). (2023).
37. Gupta R, et al. Performance of ChatGPT on the plastic surgery inservice training examination. Aesthetic Surg J. 2023;sjad128:2023.
38. Hoch CC et al. ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification preparation questions. *European Archives of Oto-Rhino-Laryngology* (2023). 1–8 (2023).
39. Holmes J et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. *arXiv preprint arXiv:2304.01938* (2023). (2023).
40. Hopkins BS, et al. ChatGPT versus the neurosurgical written boards: a comparative analysis of artificial intelligence/machine learning performance on neurosurgical board–style questions. J Neurosurg. 2023;139:904–11.
41. Huang RS, et al. Assessment of resident and AI chatbot performance on the University of Toronto family medicine residency progress test: comparative study. JMIR Med Educ. 2023;9:e50514.
42. Humar P, Asaad M, Bengur FB, Nguyen V. ChatGPT is equivalent to first year plastic surgery residents: evaluation of ChatGPT on the plastic surgery inservice exam. *Aesthetic Surgery Journal* (2023). sjad130 (2023).
43. Hurley NC, Schroeder KM, Hess AS. Would doctors dream of electric blood bankers? Large language model-based artificial intelligence performs well in many aspects of transfusion medicine. Transfusion. 2023;63:1833–40.
44. Kaneda Y et al. Assessing the performance of GPT-3.5 and GPT-4 on the 2023 Japanese nursing examination. Cureus 15 (2023).
45. Kumah-Crystal Y, Mankowitz S, Embi P, Lehmann CU. ChatGPT and the clinical informatics board examination: the end of unproctored maintenance of certification? *Journal of the American Medical Informatics Association* (2023). ocad104 (2023).
46. Kung JE, et al. Evaluating ChatGPT performance on the Orthopaedic In-Training examination. JBJS Open Access. 2023;8:e23.
47. Lewandowski M, Łukowicz P, Świetlik D, Barańska-Rybak W. An original study of ChatGPT-3.5 and ChatGPT-4 dermatological knowledge level based on the dermatology specialty certificate examinations. *Clinical and Experimental Dermatology* (2023). lad255 (2023).
48. Li Q, Min X. Unleashing the Power of Language Models in Clinical Settings: A Trailblazing Evaluation Unveiling Novel Test Design. *medRxiv* (2023). 2023.2007. 2011.23292512 (2023).
49. Long C et al. Evaluating ChatGPT-4 in Otolaryngology–Head and Neck Surgery Board Examination using the CVSA Model. *medRxiv*https://doi.org/10.1101/2023.05.30.23290758 (2023). 2023.2005.2030.23290758 (2023).
50. Lum ZC. Can artificial intelligence pass the American Board of Orthopaedic Surgery examination? Orthopaedic residents versus ChatGPT. *Clinical Orthopaedics and Related Research®*. 2022:10.1097.
51. Madrid-García A, et al. Harnessing ChatGPT and GPT-4 for evaluating the rheumatology questions of the Spanish Access exam to Specialized Medical Training. medRxiv. 2023;20232007. 2021.23292821. (2023).
52. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT–3.5, ChatGPT-4, and Orthopaedic Resident Performance on Orthopaedic Assessment Examinations. *JAAOS-Journal of the American Academy of Orthopaedic Surgeons* (2022). 10.5435 (2022).
53. Meo SA et al. ChatGPT knowledge evaluation in basic and clinical medical sciences: multiple choice question examination-based performance. In *Healthcare* 2046 (MDPI, 2023).
54. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. *JAMA ophthalmology* (2023). (2023).
55. Moshirfar M et al. Artificial Intelligence in Ophthalmology: a comparative analysis of GPT-3.5, GPT-4, and human expertise in answering StatPearls questions. Cureus 15 (2023).
56. Noda R et al. Performance of ChatGPT and Bard in Self-Assessment Questions for Nephrology Board Renewal. *medRxiv* (2023). 2023.2006. 2006.23291070 (2023).
57. Oh N, Choi G-S, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. Annals Surg Treat Res. 2023;104:269.
58. Oztermeli AD, Oztermeli A. ChatGPT performance in the medical specialty exam: an observational study. Medicine. 2023;102:e34673.
59. Passby L, Jenko N, Wernham A. Performance of ChatGPT on dermatology Specialty Certificate Examination multiple choice questions. *Clin Exp Dermatol*https://doi.org/10.1093/ced/llad197 (2023). (2023).
60. Patil NS, Huang RS, van der Pol CB, Larocque N. Comparative performance of ChatGPT and bard in a text-based radiology knowledge assessment. *Canadian Association of Radiologists Journal* (2023). 08465371231193716 (2023).
61. 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. 2023;9:e46482.
62. Rosoł M et al. Evaluation of the performance of GPT-3.5 and GPT-4 on the Medical Final Examination. *medRxiv* (2023). 2023.2006. 2004.23290939 (2023).
63. Saad A, Iyengar KP, Kurisunkal V, Botchu R. Assessing ChatGPT's ability to pass the FRCS orthopaedic part a exam: a critical analysis. Surgeon. 2023;21:263–6.
64. Schubert MC, Wick W, Venkataramani V. Evaluating the Performance of Large Language Models on a Neurology Board-Style Examination. *medRxiv* (2023). 2023.2007. 2013.23292598 (2023).
65. Shetty M, Ettlinger M, Lynch M. GPT-4, an artificial intelligence large language model, exhibits high levels of accuracy on dermatology specialty certificate exam questions. *medRxiv* (2023). 2023.2007. 2013.23292418 (2023).
66. Smith J, Choi PM, Buntine P. Will code one day run a code? Performance of language models on ACEM primary examinations and implications. *Emergency Medicine Australasia* (2023). (2023).
67. Suchman K, Garg S, Trindade AJ. Chat Generative Pretrained Transformer fails the Multiple-Choice American College of Gastroenterology Self-Assessment Test. Official J Am Coll Gastroenterology| ACG. 2022. 10.14309. (2022).
68. Taira K, Itaya T, Hanada A. Performance of the large language model ChatGPT on the national nurse examinations in Japan: evaluation study. JMIR Nurs. 2023;6:e47305.
69. 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. 2023;9:e48002.
70. Tanaka Y et al. Performance of Generative Pretrained Transformer on the National Medical Licensing Examination in Japan. *medRxiv* (2023). 2023.2004. 2017.23288603 (2023).
71. Teebagy S, et al. Improved performance of chatgpt-4 on the OKAP examination: a comparative study with chatgpt-3.5. J Acad Ophthalmol. 2023;15:e184–7.
72. Thirunavukarasu AJ, et al. Trialling a large language model (ChatGPT) in general practice with the Applied Knowledge Test: observational study

demonstrating opportunities and limitations in primary care. JMIR Med Educ. 2023;9:e46599.
73. Valdez D et al. Performance of progressive generations of GPT on an exam designed for certifying physicians as Certified Clinical Densitometrists. *medRxiv* (2023). 2023.2007. 2025.23293171 (2023).
74. Wang H, et al. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: pave the way for medical AI. Int J Med Informatics. 2023;177:105173.
75. Abi-Rafeh J et al. Complications Following Facelift and Neck Lift: Implementation and Assessment of Large Language Model and Artificial Intelligence (ChatGPT) Performance Across 16 Simulated Patient Presentations. *Aesthetic Plastic Surgery* (2023). 1–8 (2023).
76. Ali MJ. ChatGPT and lacrimal drainage disorders: performance and scope of improvement. Ophthal Plast Reconstr Surg. 2023;39:221.
77. Allahqoli L et al. The Diagnostic and Management Performance of the ChatGPT in Obstetrics and Gynecology. *Gynecologic and Obstetric Investigation* (2023). (2023).
78. Athavale A, Baier J, Ross E, Fukaya E, THE POTENTIAL OF CHATBOTS IN CHRONIC VENOUS DISEASE PATIENT MANAGEMENT. *JVS-Vascular Insights* (2023). 100019 (2023).
79. Ayers JW et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine* (2023). (2023).
80. Ayoub NF, Lee YJ, Grimm D, Divi V. Head-to-Head Comparison of ChatGPT Versus Google Search for Medical Knowledge Acquisition. *Otolaryngology–Head and Neck Surgery* (2023). (2023).
81. Barash Y, Klang E, Konen E, Sorin V. ChatGPT-4 assistance in optimizing emergency department radiology referrals and imaging selection. J Am Coll Radiol. 2023;20:998–1003.
82. Bellinger JR et al. BPPV Information on Google Versus AI (ChatGPT). *Otolaryngology–Head and Neck Surgery* (2023). (2023).
83. Benirschke RC et al. Assessment of a large language model's utility in helping pathology professionals answer general knowledge pathology questions. *American Journal of Clinical Pathology* (2023). aqad106 (2023).
84. Bernstein IA, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. JAMA Netw Open. 2023;6:e2330320–2330320.
85. Birkun AA, Gautam A. Large language model-based chatbot as a source of advice on first aid in heart attack. *Current Problems in Cardiology* (2023). 102048 (2023).
86. Biswas S et al. Assessing the utility of ChatGPT as an artificial intelligence-based large language model for information to answer questions on myopia. *Ophthalmic and Physiological Optics* (2023). (2023).
87. Cadamuro J, et al. Potentials and pitfalls of ChatGPT and natural-language artificial intelligence models for the understanding of laboratory medicine test results. An assessment by the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Working Group on Artificial Intelligence (WG-AI). Clin Chem Lab Med (CCLM). 2023;61:1158–66.
88. Caglar U et al. Evaluating the performance of ChatGPT in answering questions related to pediatric urology. *Journal of pediatric urology* (2023). (2023).
89. Cakir H et al. Evaluating the performance of ChatGPT in answering questions related to urolithiasis. *International Urology and Nephrology* (2023). 1–5 (2023).
90. Chen S et al. The utility of ChatGPT for cancer treatment information. *medRxiv* (2023). 2023.2003. 2016.23287316 (2023).
91. Chiesa-Estomba CM et al. Exploring the potential of Chat-GPT as a supportive tool for sialendoscopy clinical decision making and patient information support. *European Archives of Oto-Rhino-Laryngology* (2023). 1–6 (2023).
92. Clough RA et al. Transforming healthcare documentation: Harnessing the potential of AI to generate discharge summaries. *BJGP open* (2023). (2023).
93. Cocci A et al. Quality of information and appropriateness of ChatGPT outputs for urology patients. *Prostate cancer and prostatic diseases* (2023). 1–6 (2023).
94. Coskun B, Ocakoglu G, Yetemen M, Kaygisiz O. Can ChatGPT, an artificial intelligence language model, provide accurate and high-quality patient information on prostate cancer? *Urology* 180, 35–58 (2023).
95. Coskun BN et al. Assessing the accuracy and completeness of artificial intelligence language models in providing information on methotrexate use. *Rheumatology International* (2023). 1–7 (2023).
96. Davis R, et al. Evaluating the effectiveness of artificial intelligence–powered large language models application in disseminating appropriate and readable health information in urology. J Urol. 2023;210:688–94.
97. Delsoz M et al. Performance of chatgpt in diagnosis of corneal eye diseases. *medRxiv* (2023). (2023).
98. Delsoz M et al. The Use of ChatGPT to Assist in Diagnosing Glaucoma Based on Clinical Case Reports. *Ophthalmology and Therapy* (2023). 1–12 (2023).
99. Duey AH, et al. Thromboembolic prophylaxis in spine surgery: an analysis of ChatGPT recommendations. Spine J. 2023;23:1684–91.
100. Fink MA, et al. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. Radiology. 2023;308:e231362.
101. Gorelik Y, Ghersin I, Maza I, Klein A. Harnessing language models for streamlined postcolonoscopy patient management: a novel approach. Gastrointest Endosc. 2023;98:639–41. e634.
102. Haemmerli J et al. ChatGPT in glioma adjuvant therapy decision making: ready to assume the role of a doctor in the tumour board? BMJ Health Care Inf 30 (2023).
103. Henson JB et al. Evaluation of the Potential Utility of an Artificial Intelligence Chatbot in Gastroesophageal Reflux Disease Management. *Official journal of the American College of Gastroenterology| ACG* (2022). 10.14309 (2022).
104. Hirosawa T, et al. 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. 2023;20:3378.
105. Hirosawa T, Mizuta K, Harada Y, Shimizu T. Comparative Evaluation of Diagnostic Accuracy between Google Bard and Physicians. Am J Med. 2023;136:1119–23. e1118.
106. Hristidis V, et al. ChatGPT vs Google for queries related to dementia and other Cognitive decline: comparison of results. J Med Internet Res. 2023;25:e48966.
107. Hu X et al. What can GPT-4 do for Diagnosing Rare Eye Diseases? A Pilot Study. *Ophthalmology and Therapy* (2023). 1–8 (2023).
108. Hung Y-C, et al. Comparison of Patient Education materials generated by Chat Generative Pre-trained Transformer Versus experts: an innovative way to increase readability of Patient Education materials. Ann Plast Surg. 2023;91:409–12.
109. Johnson D et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. *Research square* (2023). (2023).
110. Kaarre J, et al. Exploring the potential of ChatGPT as a supplementary tool for providing orthopaedic information. Knee Surg Sports Traumatol Arthrosc. 2023;31:5190–8.
111. Kao H-J et al. Assessing ChatGPT's capacity for clinical decision support in pediatrics: a comparative study with pediatricians using KIDMAP of Rasch analysis. Medicine 102 (2023).
112. King RC et al. A multidisciplinary assessment of ChatGPTs knowledge of amyloidosis. *medRxiv* (2023). 2023.2007. 2017.23292780 (2023).
113. King RC et al. Appropriateness of ChatGPT in answering heart failure related questions. *medRxiv* (2023). 2023.2007. 2007.23292385 (2023).
114. Kiyohara Y et al. Large language models to differentiate vasospastic angina using patient information. *medRxiv* (2023). 2023.2006. 2026.23291913 (2023).
115. Krusche M, Callhoff J, Knitza J, Ruffer N. Diagnostic accuracy of a large language model in rheumatology: comparison of physician and ChatGPT-4. *Rheumatology International* (2023). 1–4 (2023).
116. Kuckelman IJ et al. Assessing ai-powered patient education: a case study in radiology. *Academic Radiology* (2023). (2023).
117. Kumari A et al. Large language models in hematology case solving: a comparative study of ChatGPT-3.5, Google Bard, and Microsoft Bing. Cureus 15 (2023).
118. Kuroiwa T, et al. The potential of ChatGPT as a Self-Diagnostic Tool in Common Orthopedic diseases: exploratory study. J Med Internet Res. 2023;25:e47621.
119. Kusunose K, Kashima S, Sata M. Evaluation of the accuracy of ChatGPT in answering clinical questions on the Japanese Society of Hypertension guidelines. Circ J. 2023;87:1030–3.
120. Lahat A et al. Evaluating the Utility of a Large Language Model in Answering Common Patients' Gastrointestinal Health-Related Questions: Are We There Yet? *Diagnostics* 13, 1950 (2023).
121. Lim ZW et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *EBioMedicine* 95 (2023).
122. Liu S et al. Assessing the value of ChatGPT for clinical decision support optimization. *MedRxiv* (2023). 2023.2002. 2021.23286254 (2023).
123. Lukac S et al. Evaluating ChatGPT as an Adjunct for the Multidisciplinary Tumor Board Decision-Making in Primary Breast Cancer Cases. (2023). (2023).
124. Lyons RJ et al. Artificial intelligence chatbot performance in triage of ophthalmic conditions. *Canadian Journal of Ophthalmology* (2023). (2023).

125. Lyu Q, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Visual Comput Ind Biomed Art. 2023;6:9.
126. Mika AP et al. Assessing ChatGPT Responses to Common Patient Questions Regarding Total Hip Arthroplasty. *JBJS* 105, 1519–1526 (2023).
127. Mishra A et al. Exploring the intersection of artificial intelligence and neurosurgery: Let us be cautious with ChatGPT. *Neurosurgery* (2022). 10.1227 (2022).
128. Momenaei B et al. Appropriateness and Readability of ChatGPT-4 generated Responses for Surgical Treatment of Retinal Diseases. *Ophthalmology Retina* (2023). (2023).
129. Nakaura T et al. Preliminary assessment of automated radiology report generation with generative pre-trained transformers: comparing results to radiologist-generated reports. *Japanese Journal of Radiology* (2023). 1–11 (2023).
130. O'Hagan R et al. Trends in accuracy and appropriateness of alopecia areata information obtained from a popular online large language model, ChatGPT. *Dermatology (Basel, Switzerland)* (2023). (2023).
131. Qu RW, Qureshi U, Petersen G, Lee SC. Diagnostic and management applications of ChatGPT in structured otolaryngology clinical scenarios. OTO open. 2023;7:e67.
132. Rahsepar AA, et al. How AI responds to common lung Cancer questions: ChatGPT vs Google Bard. Radiology. 2023;307:e230922.
133. Rao A et al. Evaluating GPT as an Adjunct for Radiologic Decision Making: GPT-4 Versus GPT-3.5 in a Breast Imaging Pilot. *Journal of the American College of Radiology* (2023). (2023).
134. Rao A, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. J Med Internet Res. 2023;25:e48659.
135. Rau A et al. A context-based chatbot surpasses trained radiologists and generic ChatGPT in following the ACR appropriateness guidelines. *medRxiv* (2023). 2023.2004. 2010.23288354 (2023).
136. Reese JT et al. On the limitations of large language models in clinical diagnosis. *medRxiv* (2023). (2023).
137. Rogasch JM et al. ChatGPT: Can You Prepare My Patients for [18F] FDG PET/CT and Explain My Reports? *Journal of Nuclear Medicine* (2023). (2023).
138. Rojas-Carabali W et al. Evaluating the Diagnostic Accuracy and Management Recommendations of ChatGPT in Uveitis. *Ocular Immunology and Inflammation* (2023). 1–6 (2023).
139. Russe MF, et al. Performance of ChatGPT, human radiologists, and context-aware ChatGPT in identifying AO codes from radiology reports. Sci Rep. 2023;13:14215.
140. Salazar GZ et al. Efficacy of AI chats to determine an emergency: a comparison between OpenAI's ChatGPT, Google Bard, and Microsoft Bing AI Chat. Cureus 15 (2023).
141. Samaan JS, et al. ChatGPT's ability to comprehend and answer cirrhosis related questions in Arabic. Arab J Gastroenterol. 2023;24:145–8.
142. Samaan JS et al. Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. *Obesity surgery* (2023). 1–7 (2023).
143. Sarbay İ, Berikol GB, Özturan İU. Performance of emergency triage prediction of an open access natural language processing based chatbot application (ChatGPT): a preliminary, scenario-based cross-sectional study. Turkish J Emerg Med. 2023;23:156.
144. Shao C-y, et al. Appropriateness and comprehensiveness of using ChatGPT for perioperative patient education in thoracic surgery in different language contexts: survey study. Interact J Med Res. 2023;12:e46900.
145. 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. *Annals of Clinical Biochemistry* (2023). 00045632231203473 (2023).
146. Sütcüoğlu BM, Güler M. Appropriateness of premature ovarian insufficiency recommendations provided by ChatGPT. Menopause. 2023;30:1033–7.
147. Suthar PP et al. Artificial Intelligence (AI) in Radiology: A Deep Dive Into ChatGPT 4.0's Accuracy with the American Journal of Neuroradiology's (AJNR) Case of the Month. Cureus 15 (2023).
148. Ueda D, et al. Diagnostic performance of ChatGPT from Patient History and Imaging findings on the diagnosis please quizzes. Radiology. 2023;308:e231040.
149. Uz C, Umay E. Dr ChatGPT: Is it a reliable and useful source for common rheumatic diseases? *International Journal of Rheumatic Diseases* (2023). (2023).
150. Vaira LA et al. Accuracy of ChatGPT-Generated Information on Head and Neck and Oromaxillofacial Surgery: A Multicenter Collaborative Analysis. *Otolaryngology–Head and Neck Surgery* (2023). (2023).
151. Wagner MW, Ertl-Wagner BB. Accuracy of information and references using ChatGPT-3 for retrieval of clinical radiological information. *Canadian Association of Radiologists Journal* (2023). 08465371231171125 (2023).
152. Wang C, Liu S, Li A, Liu J. Text dialogue analysis based ChatGPT for primary screening of mild cognitive impairment. *medRxiv* (2023). 2023.2006. 2027.23291884 (2023).
153. Whiles BB, et al. Caution! AI bot has entered the patient chat: ChatGPT has limitations in providing accurate urologic healthcare advice. Urology. 2023;180:278–84.
154. Yeo YH et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *medRxiv* (2023). 2023.2002. 2006.23285449 (2023).
155. Angel M, Rinehart J, Canneson M, Baldi PF. Clinical Knowledge and Reasoning Abilities of AI Large Language Models in Anesthesiology: A Comparative Study on the ABA Exam. *medRxiv* (2023). 2023.2005. 2010.23289805 (2023).
156. Chervenak J, Lieman H, Blanco-Breindel M, Jindal S. The promise and peril of using a large language model to obtain clinical information: ChatGPT performs strongly as a fertility counseling tool with limitations. *Fertility and Sterility* (2023). (2023).
157. Copeland-Halperin LR, O'Brien L, Copeland M. Evaluation of Artificial intelligence–generated responses to common plastic surgery questions. Plast Reconstr Surgery–Global Open. 2023;11:e5226.
158. Harskamp RE, De Clercq L. Performance of ChatGPT as an AI-assisted decision support tool in medicine: a proof-of-concept study for interpreting symptoms and management of common cardiac conditions (AMSTELHEART-2). *medRxiv* (2023). 2023.2003. 2025.23285475 (2023).
159. Beaulieu-Jones BR et al. Evaluating Capabilities of Large Language Models: Performance of GPT4 on Surgical Knowledge Assessments. *medRxiv* (2023). (2023).
160. Fang C et al. How does ChatGPT4 preform on Non-English National Medical Licensing Examination? An Evaluation in Chinese Language. *medRxiv* (2023). 2023.2005. 2003.23289443 (2023).
161. Huynh LM et al. New Artificial Intelligence ChatGPT Performs Poorly on the 2022 Self-assessment Study Program for Urology. Urology Practice. 2023. https://doi.org/10.1097/UPJ.0000000000000406.
162. Kung TH, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLoS Digit Health. 2023;2:e0000198.
163. Strong E et al. Performance of ChatGPT on free-response, clinical reasoning exams. *medRxiv* (2023). 2023.2003. 2024.23287731 (2023).
164. Athaluri SA et al. Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. Cureus 15 (2023).
165. Gilbert S et al. Large language model AI chatbots require approval as medical devices. *Nature Medicine* (2023). 1–3 (2023).
166. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. Sci Rep. 2023;13:14045.

Publisher's note