scientific reports

Check for updates

OPEN

Large language models' capabilities in responding to tuberculosis medical questions: testing ChatGPT, Gemini, and Copilot

Meisam Dastani[1], Jalal Mardaneh[2] & Morteza Rostamian[3]✉

This study aims to evaluate the capability of Large Language Models (LLMs) in responding to questions related to tuberculosis. Three large language models (ChatGPT, Gemini, and Copilot) were selected based on public accessibility criteria and their ability to respond to medical questions. Questions were designed across four main domains (diagnosis, treatment, prevention and control, and disease management). The responses were subsequently evaluated using DISCERN-AI and NLAT-AI assessment tools. ChatGPT achieved higher scores (4 out of 5) across all domains, while Gemini demonstrated superior performance in specific areas such as prevention and control with a score of 4.4. Copilot showed the weakest performance in disease management with a score of 3.6. In the diagnosis domain, all three models demonstrated equivalent performance (4 out of 5). According to the DISCERN-AI criteria, ChatGPT excelled in information relevance but showed deficiencies in providing sources and information production dates. All three models exhibited similar performance in balance and objectivity indicators. While all three models demonstrate acceptable capabilities in responding to medical questions related to tuberculosis, they share common limitations such as insufficient source citation and failure to acknowledge response uncertainties. Enhancement of these models could strengthen their role in providing medical information.

Keywords Large language models (LLMs), ChatGPT, Gemini, Copilot, Tuberculosis, Medical questions

Large Language Models (LLMs) have made remarkable progress in artificial intelligence and natural language processing in recent years. These models, utilizing deep neural networks and learning from extensive textual data, have acquired the ability to comprehend and generate human-like text. Key capabilities of LLMs include question answering, translation, summarization, and text generation[1]. Consequently, the use of LLMs is rapidly increasing, with tools like Bard, Bing, and ChatGPT (OpenAI) providing users access to extensive services[2,3]. Some experts believe these models could soon replace search engines and play crucial roles in various software domains[4]. Initial evaluations indicate that LLMs possess strong semantic and syntactic understanding across many natural languages[5,6] and can effectively perform natural language processing operations. These models also demonstrate proficiency in answering questions related to mathematics, science, programming, logical reasoning, and humanities[7,8]. LLMs have captured public attention due to their potential to improve traditional approaches across various fields[9].

In healthcare, ChatGPT stands out as a notable example, showing promising features in generating human-like textual communications[10]. These capabilities have led to exploratory applications of ChatGPT in tasks such as answering medical questions and creating accurate medical content. Additionally, ChatGPT has demonstrated successful potential in various aspects including diagnosis[11], treatment recommendations[12], patient education[13], and medical image interpretation[1]. LLMs like GPT-3.5 and GPT-4 can process and synthesize vast amounts of medical texts and patient data, potentially reducing the information burden on healthcare professionals[14]. This capability becomes particularly relevant during periods of medical specialist shortages and increasing wait times, potentially leading to a preference for LLM-based chatbots like ChatGPT over consultation with trained specialists[15,16]. However, significant challenges remain regarding the use of LLMs in sensitive healthcare domains,

[1]Infectious Diseases Research Center, Gonabad University of Medical Sciences, Gonabad, Iran. [2]Department of Microbiology, Infectious Diseases Research Center, School of Medicine, Gonabad University of Medical Sciences, Gonabad, Iran. [3]English Department, School of Medicine, Gonabad University of Medical Sciences, Gonabad, Iran. ✉email: m.rostamian.edu@gmail.com

including ethical concerns, patient privacy, and information security, as well as worries about perpetuating existing biases or causing unintended harm through these models[17–19].

Previous studies have extensively evaluated the comprehensiveness of LLMs responses. For instance, Shao et al. (2023) demonstrated that ChatGPT performed successfully in providing comprehensive and appropriate responses to chest surgery-related questions, with over 92% of responses being validated. This study also noted the model's significant role in improving patient satisfaction and reducing anxiety[20]. Lahat et al. (2023) examined ChatGPT's application in answering gastrointestinal health questions, finding that response accuracy varied by question type, with treatment-related questions showing higher accuracy than diagnostic ones[21]. Furthermore, Yeo et al. (2023) evaluated ChatGPT's performance in responding to questions about liver cirrhosis and hepatocellular carcinoma. Results indicated that while the model provides appropriate information, weaknesses were observed in diagnostic and preventive domains[11]. Moreover, Sarangi et al. (2023) evaluated the performance of four large language models in providing clinical decision support for imaging in cases suspected of pulmonary embolism. The results demonstrated variable accuracy in their responses: Perplexity excelled in open-ended questions, while Bing showed superior performance in multiple-choice questions[22].

Additionally, a systematic review by Omar et al. (2023) examining language models' performance in managing infectious diseases showed that while these tools succeed in diagnosing certain diseases, they still require improvement[23]. Response quality heavily depends on the quality of available online data, which can lead to errors[11,21]. The benefits of chatbots include improved patient access to information, facilitated therapeutic communications, and increased efficiency in disease diagnosis. For example, Zhang and Song (2023) achieved 97.50% accuracy in chronic disease diagnosis using a GPT-2-based system[24]. Likewise, Mondal et al. (2023) demonstrated that ChatGPT could provide accurate information about lifestyle-related diseases and serve as a preliminary tool for patient consultation[25].

Despite significant advances in large language models, accurately evaluating their performance in specialized fields like medicine remains challenging. Specifically, these models' ability to answer medical questions related to particular diseases like tuberculosis requires careful examination. Tuberculosis is one of the world's most significant infectious diseases, affecting millions annually and requiring accurate, up-to-date information for diagnosis and treatment. TB represents a major global health concern requiring serious attention, where early diagnosis and appropriate treatment can aid in disease control[26]. Given these considerations, this study aims to evaluate the performance of LLMs like ChatGPT in answering medical questions related to tuberculosis. Considering the importance of access to accurate and comprehensive information, increasing drug resistance, and high economic costs for health ministries and families, this study examines the accuracy and efficiency of responses generated by LLMs in various tuberculosis-related areas. Additionally, this research aims to identify these models' limitations and strengths in providing medical information and examine potential risks that might arise from incorrect or incomplete information. The study's results could contribute to improving the use of LLMs in healthcare and provide strategies for optimizing their performance in dealing with infectious diseases like tuberculosis.

Methods

In this research, the study population comprised publicly available LLMs. The inclusion criteria were defined as public accessibility to these models and their capability to respond to research questions. Based on these criteria, three large language models—ChatGPT, Gemini, and Copilot—were selected as samples. These models were chosen due to their easy accessibility and ability to provide textual responses.

Question design and categorization

The medical questions were designed based on study objectives across four main domains in the field of tuberculosis:

- Disease Diagnosis
- Disease Treatment
- Disease Prevention and Control
- Disease Management

Additionally, a segment of questions was dedicated to topics related to children and vulnerable populations affected by tuberculosis. These questions were developed in consultation with infectious disease specialists to ensure comprehensive coverage of tuberculosis-related information needs. The resulting checklist comprised 23 questions, distributed as follows:

- Five questions related to diagnosis
- Five questions related to treatment
- Five questions related to prevention and control
- Five questions related to disease management
- Three questions related to children and vulnerable people

Data collection

The designed questions were posed separately to each language model (ChatGPT, Gemini, and Copilot) in English on December 23, 2024, following the Meskó 2023 prompt engineering guidelines[27]. All responses were systematically recorded and stored for subsequent analysis.

	ChatGPT	Copilot	Gemini
Diagnostic	4	4	3.6
Disease Management	4	3.6	4
Prevention and control	4	3.6	4.4
Treatment	4	4	3.8

Table 1. Mean scores across different categories for each chatbot based on NLAT-AI criteria.

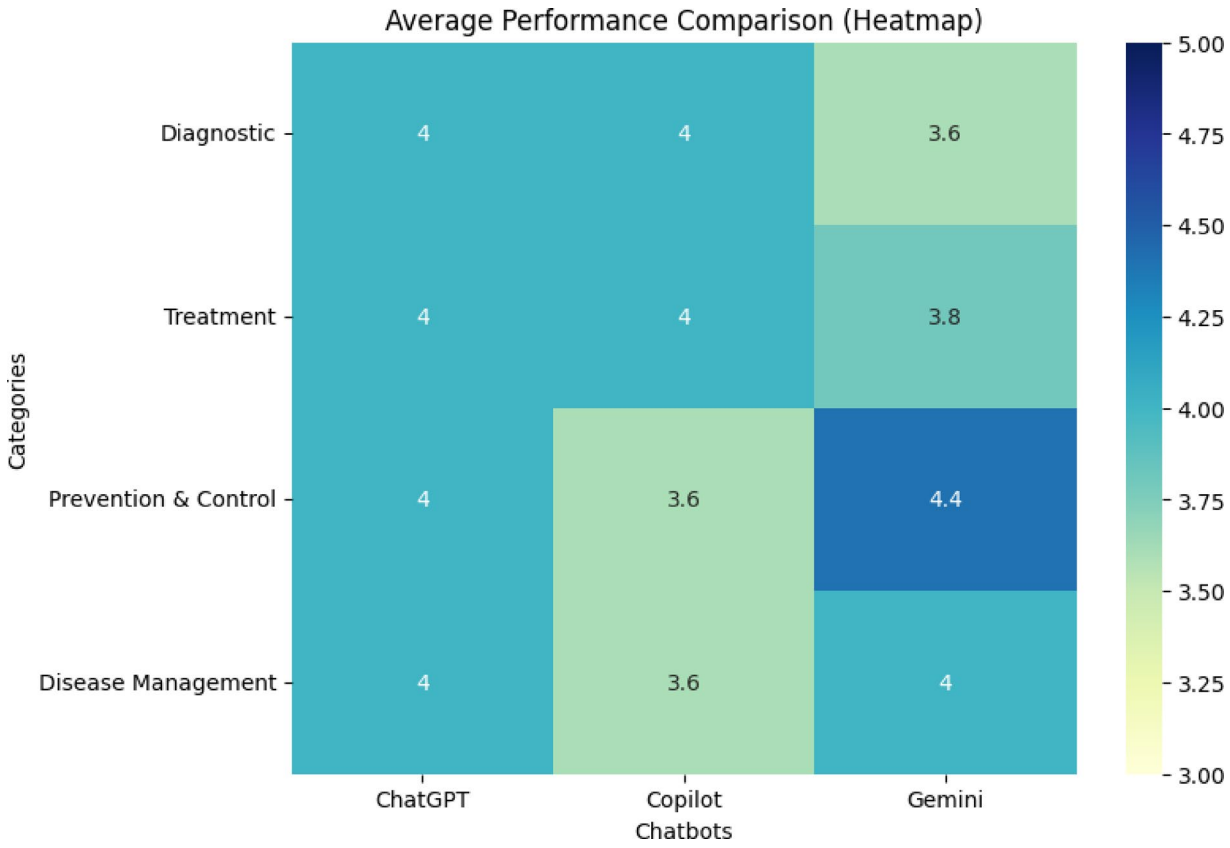

Fig. 1. Heatmap of mean scores across different categories for each chatbot based on NLAT-AI criteria.

Evaluation process

The DISCERN-AI Tool and the Natural Language Assessment Tool for AI [NLAT-AI] were used to assess the data. These tools, previously validated in language model evaluation studies, demonstrate acceptable validity and reliability. The assessment checklists evaluate various criteria including accuracy, transparency, and comprehensiveness of the models' responses[28]. The DISCERN-AI tool is a modified version of the validated DISCERN instrument, which is used to assess the quality of health care treatment information. In this study, seven questions from the original DISCERN were selected and adapted to a 3-point scale for evaluating the responses generated by language models. Based on the total scores, the quality of the generated content was categorized into five levels: very poor, poor, moderate, good, and excellent[29]. The NLAT-AI tool comprises five key components—accuracy, safety, appropriateness, actionability, and effectiveness—each of which is assessed using a 5-point Likert scale[28]. The questions and answers generated by each chatbot under study (ChatGPT, Gemini, and Copilot), along with the evaluation tools DISCERN-AI and NLAT-AI and explanatory information regarding each chatbot and tool, were provided to a subject-matter expert. The expert selected to evaluate the responses in this study met the following criteria: (a) not a member of the research team; (b) a qualified physician, and (c) possessing a scientific and research background with experience in the field of infectious diseases. After thorough review and completion of the questionnaires, the collected data were analyzed and reported using descriptive statistics.

Results

Table 1 presents the mean scores for responses to medical questions related to tuberculosis based on the NLAT-AI criteria across different categories for each chatbot: ChatGPT, Copilot, and Gemini. Additionally, Fig. 1 displays

	ChatGPT	Copilot	Gemini
Accuracy	4	4	4
Safety	4	4	4
Appropriateness	4	4	3
Actionability	4	4	4
Effectiveness	4	4	3

Table 2. Chatbot scores on diagnostic domain indices based on the NLAT-AI criteria.

	ChatGPT	Copilot	Gemini
Accuracy	5	4	4
Safety	4	4	4
Appropriateness	3	4	3
Actionability	4	4	4
Effectiveness	4	4	3

Table 3. Chatbot scores on treatment domain indices based on the NLAT-AI criteria.

	ChatGPT	Copilot	Gemini
Accuracy	4	3	4
Safety	4	4	5
Appropriateness	3	4	4
Actionability	4	4	5
Effectiveness	4	3	4

Table 4. Chatbot scores in prevention and control based on NLAT-AI criteria.

a heatmap showing the average performance of the three chatbot models (ChatGPT, Copilot, and Gemini) across the four main categories (*Diagnostic*, *Treatment*, *Prevention & Control*, and *Disease Management*).

The data from Table 1 and Fig. 1 demonstrate that in the *Diagnostic* category, all three chatbots achieved an identical score of 4.0, indicating comparable and effective performance. This suggests equivalent competency in diagnostic capabilities across all three models, with no significant variations observed. In the *Treatment* category, Gemini showed slightly lower performance (3.8) compared to ChatGPT and Copilot (both 4.0). This discrepancy may be attributed to Gemini's relative limitations in certain domain indicators, warranting further investigation. In the *Prevention & Control* category, Copilot demonstrated the lowest performance with a score of 3.6, while Gemini excelled with 4.4, suggesting superior capabilities in prevention and control-related queries. Regarding *Disease Management*, ChatGPT and Gemini showed equivalent performance (4.0), while Copilot scored lower (3.6), indicating relatively weaker performance in disease management compared to the other models.

Table 2 presents mean scores for diagnostic indicators across all three platforms.

The data presented in Table 2 illustrates the performance of chatbots in responding to queries related to the diagnosis of brucellosis. The chatbots Gemini, Copilot, and ChatGPT exhibited comparable performance, scoring 4 across most indices. However, the Gemini chatbot demonstrated slightly weaker performance in the indices of *Appropriateness and Effectiveness*, with a score of 3.

Table 3 presents the scores for the chatbots in the Treatment domain, as evaluated by the NLAT-AI criteria.

Data from Table 3, assessing chatbot performance in answering questions pertaining to brucellosis treatment, reveals that ChatGPT achieved a higher score in the Accuracy metric (5 out of 5) compared to both Copilot and Gemini. However, in the Appropriateness metric, the scores for Gemini and ChatGPT were lower than that of Copilot.

Table 4 presents the scores for the chatbots in the domains of Prevention and Control, as evaluated by the NLAT-AI criteria.

The data presented in Table 4 indicates that Gemini demonstrated superior performance in the domains of *Prevention & Control*. Specifically, Gemini achieved the highest scores in the *Safety* (5 out of 5) and *Actionability* (5 out of 5) metrics.

Table 5 presents the scores for the chatbots in the Disease Management domain, as evaluated by the NLAT-AI criteria.

The evaluation results of the data presented in Table 5 indicate a similar performance among ChatGPT, Copilot, and Gemini. All three chatbots achieved high scores (4 out of 5) on most accuracy metrics. However, Copilot's performance was lower (3 out of 5) than ChatGPT and Gemini specifically in the Accuracy and Effectiveness indices.

	ChatGPT	Copilot	Gemini
Accuracy	4	3	4
Safety	4	4	4
Appropriateness	4	4	4
Actionability	4	4	4
Effectiveness	4	3	4

Table 5. Chatbot scores in disease management based on NLAT-AI criteria.

Criteria	ChatGPT	Copilot	Gemini
Relevance of information	Relevant	Partially relevant	Partially relevant
Information sources	Not mentioned	Partially specified	Partially specified
Date of information production	Not provided	Not provided	Not provided
Balance and impartiality	Completely neutral and balanced	Completely neutral and balanced	Completely neutral and balanced
Additional sources	Not provided	Limited details on additional sources	Incomplete details on additional sources
Indication of uncertainty	Not indicated	Not indicated	Not indicated
Overall quality	Average	Average	Average

Table 6. Evaluation of chatbots in responding to medical questions related to brucellosis based on the DISCERN-AI criteria.

The results of evaluating the three chatbots, ChatGPT, Copilot, and Gemini, in response to medical inquiries related to brucellosis, based on the DISCERN-AI criteria, are shown in Table 6.

The data presented in Table 6 indicate that ChatGPT performed better in the information relevance index, providing more relevant responses, while Copilot and Gemini provided responses that were only partially relevant. Regarding information sources, ChatGPT did not provide any sources, whereas Copilot and Gemini partially cited information along with sources. This could positively impact the transparency of information from these two models. In the information production date index, none of the three models provided a date, which could limit the assessment of the responses' timeliness. Additionally, all three models performed similarly in the balance and impartiality index, providing completely neutral and balanced information, indicating their accuracy in presenting information without bias. In the additional sources index, ChatGPT did not provide any additional resources, while Copilot provided limited details of additional resources, and Gemini included incomplete details of additional resources or related references. This indicates that all three models have shortcomings in providing additional sources. In the uncertainty indication index, none of the models referred to uncertainty in the responses, which can be considered a common weakness. Finally, the overall quality of all three models was evaluated as average, indicating that although the responses provided are acceptable in some cases, there is still a need for improvement in certain aspects. Therefore, while ChatGPT excelled in information relevance, Copilot and Gemini compensated through partial source provision. However, all three models showed limitations in dating information, acknowledging uncertainty, and providing comprehensive references, indicating areas for potential improvement.

Discussion

The findings of the present study indicate that LLMs-based chatbots—ChatGPT, Copilot, and Gemini—demonstrate comparable performance across various domains in responding to medical questions related to tuberculosis. However, certain language models exhibited higher or lower performance on specific indicators. Results revealed that ChatGPT generally outperformed Gemini and Copilot. This model achieved high scores across all main categories (Diagnostic, Treatment, Prevention & Control, and Disease Management), demonstrating its consistency and capability in providing accurate and reliable responses. In contrast, Gemini and Copilot showed weaker performance in certain areas, with Copilot particularly showing the lowest performance in Prevention & Control and Disease Management categories. These results suggest that ChatGPT, as a large language model, demonstrates relative superiority in terms of comprehensiveness and accuracy in responding to medical questions.

Recent studies evaluating the performance of LLMs-based chatbots in healthcare applications have identified distinct strengths and weaknesses in these systems. For instance, Huo et al. examined the performance of chatbots including ChatGPT-4, Copilot, Google Bard, and Perplexity AI in providing recommendations for surgical management of Gastroesophageal Reflux Disease (GERD). Results indicated that Google Bard provided the most accurate recommendations for both physicians and patients, with ChatGPT-4 ranking second, while Copilot and Perplexity AI demonstrated lower accuracy[30]. Another study by Masalkhi et al. compared the capabilities of large language models, particularly Gemini AI and ChatGPT, in healthcare applications. Gemini AI showed superior performance in language comprehension and multimodal processing, while ChatGPT demonstrated stronger capabilities in medical knowledge, visual analysis, and providing personalized guidance[31].

Duran et al. evaluated the performance of various models including ChatGPT-4, Gemini, and Copilot based on readability, clarity, and accuracy of their responses to questions related to cosmetic surgeries. Results showed that ChatGPT-4 excelled in producing accurate and comprehensive medical content for patients, highlighting the distinct strengths of different models in medical communications[32]. The study by Reyhan et al. revealed that in evaluating chatbot responses to medical questions related to keratoconus, Gemini and Copilot models demonstrated better performance in terms of reliability and overall quality compared to other models, including ChatGPT. However, ChatGPT-3.5 and ChatGPT-4.0 also showed acceptable performance, though scoring lower than Gemini and Copilot in some indicators such as readability and overall quality[33]. Shanmugam & Browning's comparative study on the performance of LLMs in analyzing and managing complex ophthalmological cases showed that ChatGPT-3.5, Claude Pro, and Copilot Pro demonstrated higher performance compared to other models[34].

The evaluation of ChatGPT, Gemini, and Copilot's performance in responding to medical questions based on the Discern-AI criteria revealed that ChatGPT performed better in the Relevance of information index and provided more relevant responses. However, this model showed weaknesses in providing information sources and production dates. Conversely, Copilot and Gemini partially compensated for this weakness by providing limited details from information sources. Overall, all three models demonstrated similar performance in indicators such as balance and impartiality, providing unbiased and balanced information. This demonstrates their accuracy and impartiality in information delivery. However, a common weakness across all models was their failure to address uncertainty and provide information production dates, which could be considered a serious challenge. These weaknesses might negatively impact users' trust in the provided responses.

Accordingly, despite the generally acceptable performance of the models on many metrics, common weaknesses—such as their inability to provide the date of information generation or precise sources—can have significant implications for patient care and public health. Incomplete or incorrect information may lead to inappropriate treatment decisions[35] or alter user or patient behavior regarding disease management and treatment[36]. This is especially concerning in areas like tuberculosis control and prevention, where precise health strategies are required. Such shortcomings may reduce patients' trust in AI systems, and—if users or patients rely on inaccurate information and neglect professional medical advice—they could even result in dangerous consequences. Therefore, enhancing the ability of these models to provide reliable sources, up-to-date dates, and clarity in cases of uncertainty is a necessary step to ensure the safety and effectiveness of their use in public health.

Specific recommendations for improving these models include developing mechanisms for regularly updating medical information from trustworthy sources, strengthening their ability to accurately and transparently reference scientific sources, and designing protocols to explicitly indicate uncertainty in their responses. In particular, models such as ChatGPT and Gemini could improve their evidence-based content delivery by integrating structured data from reputable medical databases (such as UpToDate or the WHO). Adding functionalities to report the date of information and specify the confidence level in responses could further aid users in making informed decisions and increase trust in these tools. Gibson et al.'s study demonstrated that ChatGPT-4 performs well in answering common questions about prostate cancer and can serve as a useful tool in patient education. Based on quality assessment tools, this model's outputs were generally deemed reliable, safe, and appropriate, with PEMAT-AI comprehensibility scores being very good and DISCERN-AI ratings categorized as "good quality"[28]. Additionally, Hancı et al.'s study, aimed at evaluating the quality, reliability, and readability of responses from Bard, Copilot, Perplexity, ChatGPT, and Gemini chatbots to patient questions about "palliative care," showed that responses from all five chatbots had readability levels higher than the recommended level for patient educational materials, though the quality and readability of responses related to palliative care were insufficient across all chatbots[37].

Conclusion

The present study demonstrates that large language models such as ChatGPT, Gemini, and Copilot possess significant potential in providing accurate and practical medical responses, although their performance varies across different domains and evaluation metrics. ChatGPT generally exhibited superior performance in delivering comprehensive and precise responses, while Gemini and Copilot showed certain limitations in areas such as disease prevention, control, and management. Nevertheless, all three models demonstrated comparable performance in metrics such as information impartiality and balance, indicating their capability to provide unbiased and balanced information. However, common challenges persist, including the failure to acknowledge uncertainties, lack of information generation dates, and inadequate citation of credible sources, highlighting the need for further development and enhancement of these technologies.

The findings of this study also emphasize that the selection of an appropriate model for medical applications depends on specific user requirements and targeted domains. For instance, ChatGPT may be the preferred option for diagnostic and therapeutic inquiries, while Gemini demonstrates acceptable performance in areas related to disease prevention and control. However, to enhance user confidence and improve response quality, developers must focus on improving transparency, incorporating credible source citations, and providing up-to-date information. These advancements could facilitate the safe and effective expansion of these technologies in healthcare, strengthening their role as auxiliary tools in patient education and medical decision-making.

Finally, the findings of this study indicate that AI tools based on large language models can play a significant role in supporting decision-making for both health professionals and patients. In clinical settings, chatbots capable of providing accurate, balanced, and comprehensible responses can serve as valuable adjuncts for patient education, strengthening clinical decision-making, and promoting patient self-management. Therefore, choosing an appropriate model is crucial, as relying on a poorly performing model may result in incomplete, incorrect, or misleading information. This not only reduces the quality of care but may also pose risks to patient

safety. For this reason, specialized users (such as physicians) and health organizations should carefully evaluate the performance of these models in specific domains, consider their informational limitations, and take into account the intended application—such as education, consultation, or clinical decision support—when selecting and implementing such tools.

Data availability

"The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request."

Received: 14 February 2025; Accepted: 19 May 2025
Published online: 23 May 2025

References

1. Srivastav, S. et al. ChatGPT in radiology: the advantages and limitations of artificial intelligence for medical imaging diagnosis. *Cureus* **15** (2023).
2. Buhr, C. R. et al. ChatGPT versus consultants: blinded evaluation on answering otorhinolaryngology case–based questions. *JMIR Med. Educ.* **9**, e49183 (2023).
3. Tussie, C. & Starosta, A. Comparing the dental knowledge of large language models. *Br. Dent. J.* https://doi.org/10.1038/s41415-024-8015-2 (2024).
4. Grant, N. & Metz, C. A New Chat Bot Is a'Code Red'for Google's Search Business. *International New York Times*, NA-NA (2022).
5. Rathje, S. et al. GPT is an effective tool for multilingual psychological text analysis. (2023).
6. Dentella, V., Günther, F., Murphy, E., Marcus, G. & Leivada, E. Testing AI on language comprehension tasks reveals insensitivity to underlying meaning. *Sci. Rep.* **14**, 28083. https://doi.org/10.1038/s41598-024-79531-8 (2024).
7. Bongini, P., Becattini, F. & Del Bimbo, A. in *European Conference on Computer Vision*. 268–281 (Springer).
8. Shakarian, P., Koyyalamudi, A., Ngu, N. & Mareedu, L. An independent evaluation of ChatGPT on mathematical word problems (MWP). *arXiv preprint* arXiv:2302.13814 (2023).
9. Bi, K. et al. Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **619**, 533–538 (2023).
10. Ray, P. P. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet Things Cyber-Phys Syst* **3**, 121–154 (2023).
11. Yeo, Y. H. et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin. Mol. Hepatol.* **29**, 721–732 (2023).
12. Howard, A., Hope, W. & Gerada, A. ChatGPT and antimicrobial advice: the end of the consulting infection doctor?. *Lancet. Infect. Dis* **23**, 405–406 (2023).
13. Nakhleh, A., Spitzer, S. & Shehadeh, N. ChatGPT's response to the diabetes knowledge questionnaire: implications for diabetes education. *Diabetes Technol. Ther.* **25**, 571–573 (2023).
14. Brown, T. B. Language models are few-shot learners. *arXiv preprint* arXiv:2005.14165 (2020).
15. Gilson, A. 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.* **9**, e45312 (2023).
16. Kung, T. H. et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digit. Health* **2**, e0000198 (2023).
17. Huang, K., Altosaar, J. & Ranganath, R. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint* arXiv:1904.05342 (2019).
18. Lee, J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
19. Clusmann, J. et al. The future landscape of large language models in medicine. *Commun. Med.* **3**, 141. https://doi.org/10.1038/s43856-023-00370-1 (2023).
20. 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.* **12**, e46900. https://doi.org/10.2196/46900 (2023).
21. Lahat, A., Shachar, E., Avidan, B., Glicksberg, B. & Klang, E. Evaluating the utility of a large language model in answering common patients' gastrointestinal health-related questions: Are we there yet?. *Diagnostics* **13**, 1950 (2023).
22. Sarangi, P. K. et al. Radiologic decision-making for imaging in pulmonary embolism: accuracy and reliability of large language models—bing, claude, ChatGPT, and perplexity. *Indian J. Radiol. Imaging* **34**, 653–660. https://doi.org/10.1055/s-0044-1787974 (2024).
23. Omar, M., Brin, D., Glicksberg, B. & Klang, E. Utilizing natural language processing and large language models in the diagnosis and prediction of infectious diseases: A systematic review. *Am. J. Infect. Control* **52**, 992–1001. https://doi.org/10.1016/j.ajic.2024.03.016 (2024).
24. Zhang, S. & Song, J. A chatbot based question and answer system for the auxiliary diagnosis of chronic diseases based on large language model. *Sci. Rep.* **14**, 17118. https://doi.org/10.1038/s41598-024-67429-4 (2024).
25. Mondal, H., Dash, I., Mondal, S. & Behera, J. K. ChatGPT in answering queries related to lifestyle-related diseases and disorders. *Cureus* **15**, e48296. https://doi.org/10.7759/cureus.48296 (2023).
26. Natarajan, A., Beena, P., Devnikar, A. V. & Mali, S. A systemic review on tuberculosis. *Indian J. Tuberculosis* **67**, 295–311 (2020).
27. Meskó, B. Prompt engineering as an important emerging skill for medical professionals: Tutorial. *J. Med. Internet Res.* **25**, e50638. https://doi.org/10.2196/50638 (2023).
28. Gibson, D. et al. Evaluating the efficacy of ChatGPT as a patient education tool in prostate cancer: Multimetric assessment. *J. Med. Internet Res.* **26**, e55939. https://doi.org/10.2196/55939 (2024).
29. Siu, A. H. Y. et al. ChatGPT as a patient education tool in colorectal cancer—an in-depth assessment of efficacy, quality and readability. *Colorectal Dis.* **27**, e17267. https://doi.org/10.1111/codi.17267 (2025).
30. Huo, B. et al. The performance of artificial intelligence large language model-linked chatbots in surgical decision-making for gastroesophageal reflux disease. *Surg. Endosc.* **38**, 2320–2330. https://doi.org/10.1007/s00464-024-10807-w (2024).
31. Masalkhi, M., Ong, J., Waisberg, E. & Lee, A. G. Google DeepMind's gemini AI versus ChatGPT: a comparative analysis in ophthalmology. *Eye (Lond)* **38**, 1412–1417. https://doi.org/10.1038/s41433-024-02958-w (2024).
32. Duran, A., Cortuk, O. & Ok, B. Future perspective of risk prediction in aesthetic surgery: Is artificial intelligence reliable?. *Aesthet. Surg. J.* **44**, NP839–NP849. https://doi.org/10.1093/asj/sjae140 (2024).
33. Reyhan, A. H., Mutaf, Ç., Uzun, İ & Yüksekyayla, F. A performance evaluation of large language models in keratoconus: A comparative study of ChatGPT-35, ChatGPT-40, Gemini, Copilot, Chatsonic, and perplexity. *J. Clin. Med.* **13**, 6512 (2024).
34. Shanmugam, S. K. & Browning, D. J. Comparison of large language models in diagnosis and management of challenging clinical cases. *Clin. Ophthalmol.* **18**, 3239–3247. https://doi.org/10.2147/OPTH.S488232 (2024).
35. Thapa, D. K., Visentin, D. C., Kornhaber, R., West, S. & Cleary, M. The influence of online health information on health decisions: A systematic review. *Patient Educ. Couns.* **104**, 770–784. https://doi.org/10.1016/j.pec.2020.11.016 (2021).

36. Bujnowska-Fedak, M. M. & Węgierek, P. The impact of online health information on patient health behaviours and making decisions concerning health. *Int. J. Environ. Res. Public Health* **17**, 880 (2020).
37. Hanci, V. et al. Assessment of readability, reliability, and quality of ChatGPT®, BARD®, Gemini®, Copilot®, Perplexity® responses on palliative care. *Medicine (Baltimore)* **103**, e39305. https://doi.org/10.1097/md.0000000000039305 (2024).

Author contributions

M.D., J.M conceptualized, curated and analyzed data, reviewed existing literature, and drafted the first manuscript. M.R. revised, edited and provided intellectual feedback. All authors read and approved the final version of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.