RESEARCH

Open Access

Evaluating GPT-4-based ChatGPT's clinical potential on the NEJM quiz

Daiju Ueda[1*], Shannon L. Walston[2], Toshimasa Matsumoto[1], Ryo Deguchi[2], Hiroyuki Tatekawa[2] and Yukio Miki[2]

Abstract

Background GPT-4-based ChatGPT demonstrates significant potential in various industries; however, its potential clinical applications remain largely unexplored.

Methods We employed the New England Journal of Medicine (NEJM) quiz "Image Challenge" from October 2021 to March 2023 to assess ChatGPT's clinical capabilities. The quiz, designed for healthcare professionals, tests the ability to analyze clinical scenarios and make appropriate decisions. We evaluated ChatGPT's performance on the NEJM quiz, analyzing its accuracy rate by questioning type and specialty after excluding quizzes which were impossible to answer without images. ChatGPT was first asked to answer without the five multiple-choice options, and then after being given the options.

Results ChatGPT achieved an 87% (54/62) accuracy without choices and a 97% (60/62) accuracy with choices, after excluding 16 image-based quizzes. Upon analyzing performance by quiz type, ChatGPT excelled in the Diagnosis category, attaining 89% (49/55) accuracy without choices and 98% (54/55) with choices. Although other categories featured fewer cases, ChatGPT's performance remained consistent. It demonstrated strong performance across the majority of medical specialties; however, Genetics had the lowest accuracy at 67% (2/3).

Conclusion ChatGPT demonstrates potential for diagnostic applications, suggesting its usefulness in supporting healthcare professionals in making differential diagnoses and enhancing AI-driven healthcare.

Keywords Artificial intelligence, Large language model, Natural language processing

Introduction

In recent years, the field of artificial intelligence (AI) has witnessed rapid advancements, particularly in the domain of natural language processing (NLP) [1]. The development of advanced NLP models has revolutionized the way humans interact with computers, enabling machines to better understand and respond to complex linguistic inputs. As AI systems become increasingly intuitive and capable, they present the potential to transform a multitude of industries and improve the quality of life for millions of people worldwide [1].

The advent of ChatGPT, and specifically the GPT-4 architecture, has resulted in a multitude of applications and research opportunities [2, 3]. GPT-4 has demonstrated superior capabilities in language processing and generation, significantly outperforming its predecessors in terms of performance and versatility [4, 5]. Its ability

*Correspondence:
Daiju Ueda
ai.labo.ocu@gmail.com
[1] Smart Life Science Laboratory, Center for Health Science Innovation, Osaka Metropolitan University, 1-4-3, Asahi-Machi, Abeno-Ku, Osaka 545-8585, Japan
[2] Department of Diagnostic and Interventional Radiology, Graduate School of Medicine, Osaka Metropolitan University, 1-4-3 Asahi-Machi, Abeno-Ku, Osaka 545-8585, Japan

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

to process context, generate coherent and contextually relevant responses, and adapt to a wide range of tasks has made it an effective tool in numerous domains. As researchers and industries continue to explore the potential of GPT-4, its role in shaping the future of human–computer interaction becomes increasingly apparent.

Despite the attention that ChatGPT based on GPT-4 has received due to its superior performance compared to GPT-3 or GPT-3.5 [6], there is a significant gap in publications exploring its potential clinical applications which others have claimed will revolutionize healthcare and improve patient outcomes [7–9]. This lack of knowledge underscores the need for more in-depth investigations into the clinical capabilities of ChatGPT, including as a diagnostic support tool or second opinion.

To assess the clinical applicability of ChatGPT, we employed the New England Journal of Medicine (NEJM) quiz as a benchmark. This rigorous quiz, designed for healthcare professionals, tests the ability to analyze clinical scenarios, synthesize information, and make appropriate decisions. By analyzing ChatGPT's performance on the NEJM quiz, we sought to determine its potential to assist clinicians in their daily practice, contribute to the ever-growing field of AI-driven healthcare, and help transform the way healthcare professionals approach decision-making and patient care. This study is a preliminary examination of the usefulness of ChatGPT for differential diagnosis. This preliminary evaluation aims to determine the model's performance at healthcare question answering in well-defined question formats like those of the NEJM quiz.

Materials and methods

Study design

In this study, our primary hypothesis was that ChatGPT, based on the GPT-4 architecture, could accurately evaluate and respond to the clinical scenarios presented in the NEJM quiz. The potential clinical applications of ChatGPT were demonstrated by using it as a tool for evaluating clinical scenarios and making appropriate diagnostic decisions. As ChatGPT is currently unable to handle images, they were not used as input. The requirement for informed consent of quizzes was waived by the Ethical Committee of Osaka Metropolitan University Graduate School of Medicine because this study only utilized published papers. All authors, including participating physicians, consented to the study. The study design was based on the Standards for Reporting Diagnostic Accuracy (STARD) guidelines, where applicable [10].

Data collection

The NEJM offers a weekly quiz called "Image Challenge" (https://www.nejm.org/image-challenge). Although the training data is not publicly available, ChatGPT was developed using data available up to September 2021. Taking into account the possibility that earlier NEJM quizzes may have been used for training purposes, we collected the quizzes from October 2021 to March 2023. This quiz consists of images and clinical information, with readers selecting their answers from five candidate choices. While images are undoubtedly important, many questions can be answered based on clinical information alone. Two author physicians read all the quizzes and commentaries and excluded questions from the NEJM quiz that were impossible to answer without images. A third author physician was consulted when consensus could not be reached. We categorized the quiz types as Diagnosis, Finding, Treatment, Cause, and Other based on what the reader was asked to find. Case commentaries for each quiz are featured on the "Images in Clinical Medicine"

Table 1 Accuracy summary

	Accuracy without choice	Accuracy with choice	*P*-values
Total	87%(54/62)	97%(60/62)	0.01
Types of quiz			
Diagnosis	89%(49/55)	98%(54/55)	0.11
Finding	0%(0/1)	100%(1/1)	> 0.99
Treatment	100%(2/2)	100%(2/2)	> 0.99
Cause	50%(1/2)	50%(1/2)	> 0.99
Other	100%(2/2)	100%(2/2)	> 0.99
Specialty of quiz			
Dermatology	83%(24/29)	93%(27/29)	0.02
Emergency medicine	92%(11/12)	92%(11/12)	0.08
Infectious disease	92%(12/13)	100%(13/13)	> 0.99
Radiology	88%(7/8)	100%(8/8)	> 0.99
Ophthalmology	80%(8/10)	100%(10/10)	> 0.99
Pediatrics	100%(6/6)	100%(6/6)	> 0.99
Hematology/Oncology	80%(8/10)	90%(9/10)	0.22
Gastroenterology	100%(7/7)	100%(7/7)	> 0.99
Neurology/Neurosurgery	100%(7/7)	100%(7/7)	> 0.99
Pulmonary/Critical Care	100%(3/3)	100%(3/3)	> 0.99
Surgery	100%(13/13)	100%(13/13)	> 0.99
Obstetrics/Gynecology	80%(4/5)	100%(5/5)	> 0.99
Otolaryngology	50%(1/2)	100%(2/2)	> 0.99
Nephrology	100%(4/4)	100%(4/4)	> 0.99
Genetics	67%(2/3)	67%(2/3)	0.33
Cardiology	100%(2/2)	100%(2/2)	> 0.99
Allergy/Immunology	50%(1/2)	100%(2/2)	> 0.99
Rheumatology	67%(2/3)	100%(3/3)	> 0.99
Urology/Prostate disease	100%(3/3)	100%(3/3)	> 0.99
Endocrinology	100%(3/3)	100%(3/3)	> 0.99
Toxicology	100%(2/2)	100%(2/2)	> 0.99
Orthopedics	100%(2/2)	100%(2/2)	> 0.99

website, and tags related to the specialty for the case are displayed. These specialty tags were also extracted for our analysis.

Processes for input and output into the ChatGPT interface

We used the GPT-4-based ChatGPT (Mar 23 Version; OpenAI; https://chat.openai.com/). One case at a time, the quizzes were entered and answers were obtained from ChatGPT. For each case, we obtained the output from ChatGPT (Step 1: Generate answer without choices). Then we input the answer choices and asked ChatGPT to choose one of them (Step 2: Generate answer with choices). Two author physicians confirmed whether the answer generated by ChatGPT matched the ground truth. If there was a discrepancy, a third author physician made the decision. We introduced this process of confirmation in case the difference was purely linguistic.

Statistical analysis

The percentage of correct responses generated by ChatGPT with and without candidate choices was evaluated by quiz type and specialty. We verified the reproducibility by obtaining the responses again using the same prompt, and comparing the results using Fisher's exact test for paired data and the chi-square test. We extracted the percentage of correct choices for each case from the NEJM Image Challenge website and compared this to ChatGPT's accuracy using Spearman's correlation analysis. Cases with a lower percentage of correct choices were considered more difficult questions for medical professionals, while those with a higher percentage were considered easier questions. A p-value of 0.05 was used to determine statistical significance. All analyses were performed using R (version 4.0.0, 2020; R Foundation for Statistical Computing; https://R-project.org).

Results

Evaluation

In our study, we assessed ChatGPT's performance on the NEJM quiz questions which span different types and medical specialties. The results demonstrated varying levels of accuracy depending on the specific context, summarized in Table 1. Eligibility is shown in Fig. 1. Overall, ChatGPT correctly answered 87% (54/62) of the questions without candidate choices, and this accuracy increased to 97% (60/62) with the choices after excluding 16 quizzes which required images. The results show that the best performing category was Diagnosis, although the number of cases was too small for accuracy in the other categories. This is shown in Fig. 2.

Overall, ChatGPT performed well on the NEJM quiz across a range of medical specialties. In most cases, the model's accuracy improved when given choices. Several specialties showcased a remarkable 100% accuracy rate in both scenarios while Genetics had the lowest accuracy at 67% (2/3) both with and without choices. Accuracy for a few specialties, including Otolaryngology, Allergy/Immunology, and Rheumatology, improved when choices were provided. This is shown in Fig. 2. In assessing ChatGPT's reproducibility, the initial test yielded accuracies of 97% (60/62) and 87% (54/62) for tasks with and without choices, respectively, while the retest produced accuracies of 94% (58/62) and 84% (52/62). Chi-square tests showed no statistically significant differences between the two tests, with p-values of 0.5 and 0.69 for tasks with and without choices, respectively. No significant differences were found between

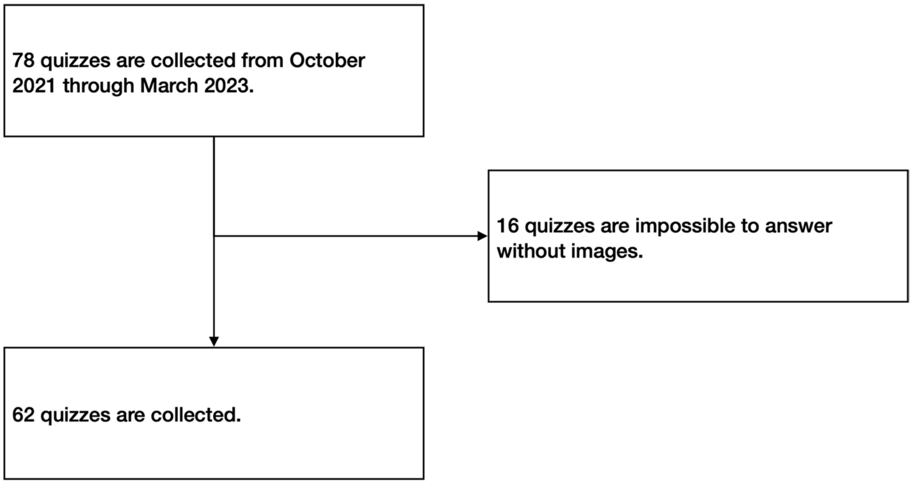
78 quizzes are collected from October 2021 through March 2023.
16 quizzes are impossible to answer without images.
62 quizzes are collected.

Fig. 1 Eligibility flowchart

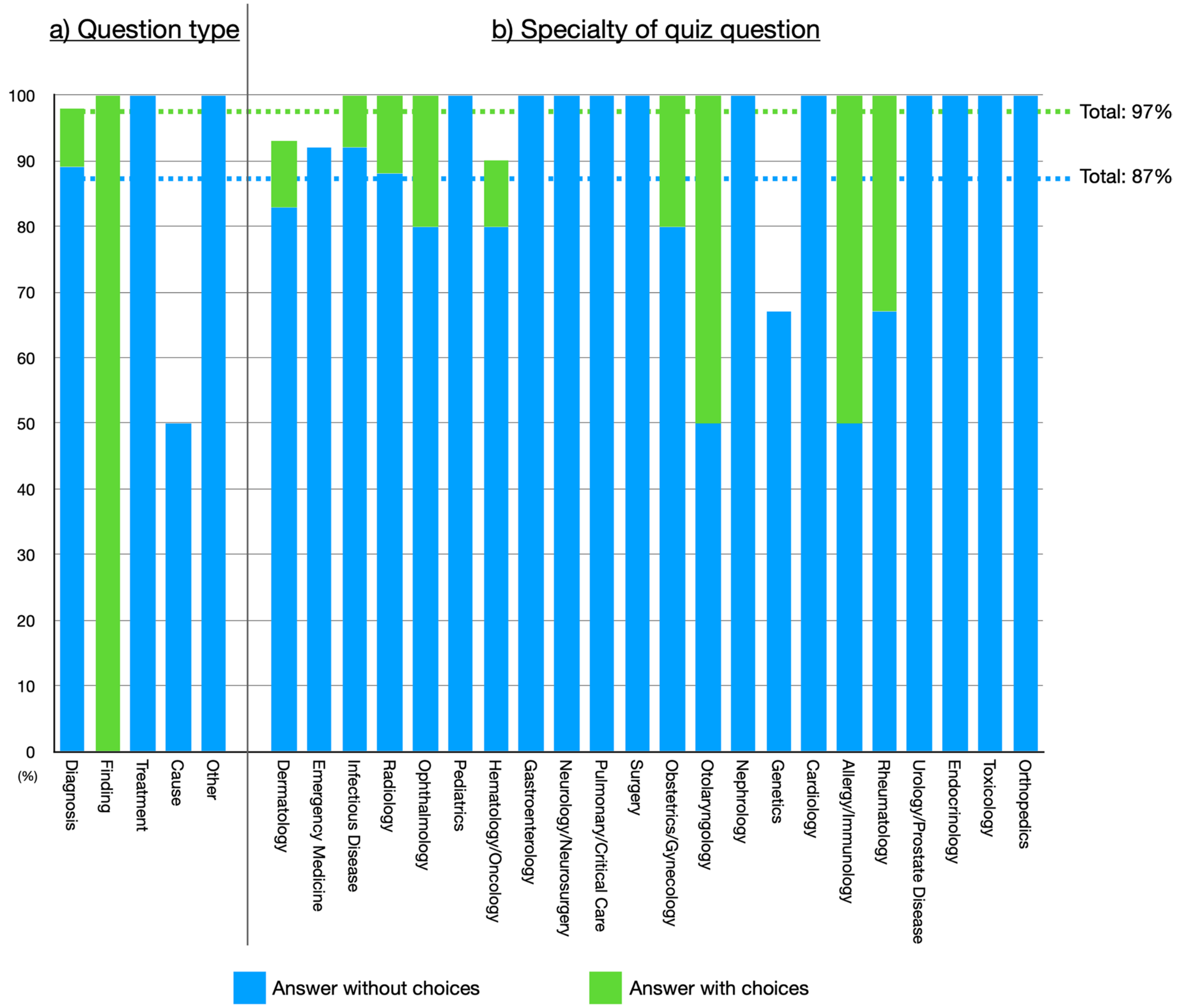

Fig. 2 Results by answer type and specialty. These are the accuracy rates for various types and specialties of quizzes from the New England Journal of Medicine. The blue bar is the accuracy without choices and the green bar is the accuracy with choices. Dotted lines show total accuracy with and without choices

the percentage of correct choices and ChatGPT's accuracy both with choices ($r = -0.0034$, p-value $= 0.98$) as well as without choices ($r = 0.075$, p-value $= 0.52$). The percentage of correct choices by those who attempted the image challenge did not significantly correlate with ChatGPT's accuracy, as shown in Fig. 3. ChatGPT maintained consistent performance regardless of the perceived difficulty.

Discussion

Our study assessed ChatGPT's performance on the NEJM quiz, encompassing various medical specialties and question types. The sample size was relatively small, limiting the generalizability of the findings. However, it provides a preliminary assessment of the potential clinical applications of GPT-4-based ChatGPT. Overall, ChatGPT achieved an 87% accuracy without choices and a 97% accuracy with choices, after excluding image-dependent questions. When examining performance by quiz type, ChatGPT excelled in the Diagnosis category, securing an 89% accuracy without choices and a 98% accuracy with choices. Although other categories contained fewer cases, ChatGPT's performance remained consistent across the spectrum. ChatGPT exhibited high accuracy in most specialties, however Genetics registered the lowest at 67%. This could be due to the amount of available Genetics-related data for training, or due to the complexity and specificity of the language used in this field.

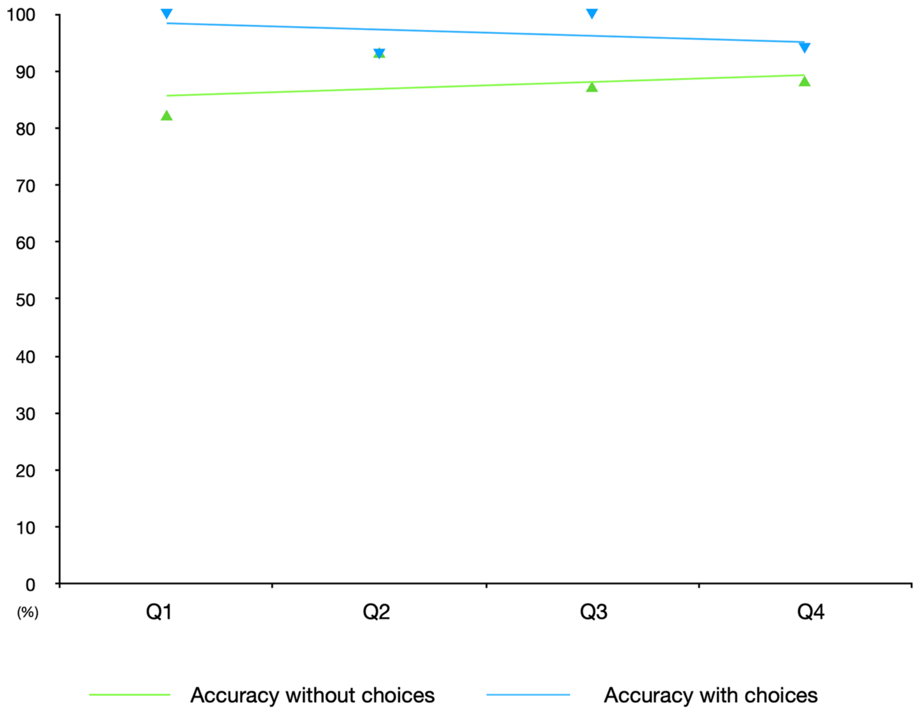

Fig. 3 Relationship between perceived difficulty of quiz questions and ChatGPT's accuracy. The x-axis represents the percentages of correct choices made by the participants, grouped in quartiles, from most difficult (Q1) to easiest (Q4). The number of correct and total choices published on the New England Journal of Medicine Image Challenge website was used as a proxy for perceived difficulty. The y-axis represents the accuracy of ChatGPT's responses, both with choices (blue line) and without choices (green line). ChatGPT's accuracy remained consistent across the perceived difficulty quartiles of the quiz questions

While this analysis highlighted the potential for clinical applications of ChatGPT, it also revealed some strengths and weaknesses, emphasizing the importance of understanding and leveraging these performance insights to optimize its use.

This is our initial investigation exploring the potential clinical applications of GPT-4-based ChatGPT to clinical decision-making quizzes, marking an important milestone. Our study highlights the novelty of assessing GPT-4-based ChatGPT's potential for clinical applications, specifically its ability to handle well-defined problems in the medical field, setting it apart from earlier research on GPT-3-based ChatGPT. Compared to GPT-3, GPT-4 demonstrates improved performance in processing complex linguistic inputs and generating contextually relevant responses, making it more suitable for specialized domains such as healthcare [2, 3]. A previous study applied GPT-3-based ChatGPT to the United States Medical Licensing Examination and found that it achieved 60% accuracy [11]. This outcome hinted at its potential for medical education and future incorporation into clinical decision-making. Another study evaluated the diagnostic accuracy of GPT-3-based ChatGPT in generating differential diagnosis lists for common clinical vignettes [12]. Results showed that it can generate diagnosis lists with good accuracy, but physicians still outperformed the AI chatbot (98.3% vs. 83.3%, $p = 0.03$).

The results of this study reveal that ChatGPT, based on the GPT-4 architecture, demonstrates promising potential in various aspects of healthcare. With an accuracy rate of 97% for answers with choices and 87% for answers without choices, ChatGPT has shown its capability in analyzing clinical scenarios and making appropriate diagnostic decisions. There is no evident correlation between the proportion of respondents choosing the correct answer, which is believed to reflect the difficulty of the quiz, and the accuracy of ChatGPT. This suggests that ChatGPT might be able to provide correct answers regardless of the question's difficulty. One key implication is the potential use of ChatGPT as a diagnostic support tool. Healthcare professionals may utilize ChatGPT to help them with differential diagnosis after taking into consideration its strengths and weaknesses as demonstrated in this study. By streamlining workflows and reducing cognitive burden, ChatGPT could enable more efficient and accurate decision-making [13, 14]. In addition to supporting diagnostic decisions, ChatGPT's performance on the NEJM quiz suggests that it could be a valuable resource for medical education [15–20]. By providing students, professionals, and patients with a dynamic and interactive learning tool, ChatGPT could enhance understanding and retention of medical knowledge.

This study has several limitations that should be considered when interpreting the results. Firstly, it focused solely on text-based clinical information, which might have affected ChatGPT's performance due to the absence of crucial visual data. The sample size was relatively small and limited to the NEJM quizzes, which may not fully represent the vast array of clinical scenarios encountered in real-world medical practice, limiting the generalizability of the findings. Additionally, the study did not evaluate the impact of ChatGPT's use on actual clinical outcomes, patient satisfaction, or healthcare provider workload, leaving the real-world implications of using ChatGPT in clinical practice uncertain. Another limitation is the absence of a comparative analysis with human performance on the same quiz. Lastly, potential biases in GPT-4's training data, as well as potential clinician biases for or against AI-provided results, may lead to disparities in the quality and accuracy of AI-driven recommendations for specific clinical scenarios or populations [21].

In conclusion, this study demonstrates the potential of GPT-4-based ChatGPT for diagnosis by evaluating its performance on the NEJM quiz. While the results show promising accuracy rates, several limitations highlight the need for further research. Future studies should focus on expanding the range of clinical scenarios, assessing the impact of ChatGPT on actual clinical outcomes and healthcare provider workload, and exploring the performance of ChatGPT in diverse language settings and healthcare environments. Additionally, the importance of incorporating image analysis in future models should not be overlooked. By addressing these limitations and integrating image analysis, the potential of ChatGPT to revolutionize healthcare and improve patient outcomes can be more accurately understood and harnessed.

Acknowledgements
We have used ChatGPT to generate a portion of the manuscript, but the output was confirmed by the authors.

Authors' contributions
D.U. performed the analysis and wrote the main manuscript text. D.U. and H.T. designed the study. Model test was performed by D.U., S.W., and T.M. Consensus included D.U., R.D., and H.T. S.W. proofread the manuscript and Y.M. supervised the study. All authors reviewed the manuscript.

Funding
This is a collaborative research project with Iida Group Holdings Co.,Ltd.

Availability of data and materials
The data used to support the findings of the study can be obtained from the corresponding author upon request.

Declarations

Ethics approval and consent to participate
The requirement for informed consent of quizzes was waived by the Ethical Committee of Osaka Metropolitan University Graduate School of Medicine because this study only utilized published papers. All authors, including participating physicians, consented to the study. All experimental protocols were approved by the Ethical Committee of Osaka Metropolitan University Graduate School of Medicine. All experiments were performed in accordance with the Declaration of Helsinki.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 7 June 2023 Accepted: 5 December 2023
Published online: 11 January 2024

References

1. Hirschberg J, Manning CD. Advances in natural language processing. Science. 2015;349:261–6.
2. OpenAI. GPT-4 Technical Report. arXiv [cs.CL]. 2023.
3. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language Models are Few-Shot Learners. arXiv [cs.CL]. 2020;:1877–901.
4. Eloundou T, Manning S, Mishkin P, Rock D. GPTs are GPTs: An early look at the labor market impact potential of large language models. arXiv [econ.GN]. 2023.
5. Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv [cs.CL]. 2023.
6. Ueda D, Mitsuyama Y, Takita H, Horiuchi D, Walston SL, Tatekawa H, et al. ChatGPT's Diagnostic Performance from Patient History and Imaging Findings on the Diagnosis Please Quizzes. Radiology. 2023;308:e231040.
7. Zheng Y, Wang L, Feng B, Zhao A, Wu Y. Innovating Healthcare: The Role of ChatGPT in Streamlining Hospital Workflow in the Future. Ann Biomed Eng. 2023. https://doi.org/10.1007/s10439-023-03323-w.
8. Lecler A, Duron L, Soyer P. Revolutionizing radiology with GPT-based models: Current applications, future possibilities and limitations of ChatGPT. Diagn Interv Imaging. 2023;104:269–74.
9. Xiao D, Meyers P, Upperman JS, Robinson JR. Revolutionizing Healthcare with ChatGPT: An Early Exploration of an AI Language Model's Impact on Medicine at Large and its Role in Pediatric Surgery. J Pediatr Surg. 2023. https://doi.org/10.1016/j.jpedsurg.2023.07.008.
10. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. Radiology. 2015;277:826–32.
11. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2:e0000198.
12. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic Accuracy of Differential-Diagnosis Lists Generated by Generative Pretrained Transformer 3 Chatbot for Clinical Vignettes with Common Chief Complaints: A Pilot Study. Int J Environ Res Public Health. 2023;20:3378.
13. Glover WJ, Li Z, Pachamanova D. The AI-enhanced future of health care administrative task management. NEJM Catal Innov Care Deliv. https://doi.org/10.1056/cat.21.0355.
14. Sandhu S, Lin AL, Brajer N, Sperling J, Ratliff W, Bedoya AD, et al. Integrating a Machine Learning System Into Clinical Workflows: Qualitative Study. J Med Internet Res. 2020;22:e22421.
15. Eysenbach G. The Role of ChatGPT, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation With ChatGPT and a Call for Papers. JMIR Med Educ. 2023;9:e46885.
16. Kundu S. How will artificial intelligence change medical training? Commun Med. 2021;1:8.

17. Rampton V, Mittelman M, Goldhahn J. Implications of artificial intelligence for medical education. Lancet Digit Health. 2020;2:e111–2.
18. Jayakumar P, Moore MG, Furlough KA, Uhler LM, Andrawis JP, Koenig KM, et al. Comparison of an Artificial Intelligence-Enabled Patient Decision Aid vs Educational Material on Decision Quality, Shared Decision-Making, Patient Experience, and Functional Outcomes in Adults With Knee Osteoarthritis: A Randomized Clinical Trial. JAMA Netw Open. 2021;4:e2037107.
19. Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of Breast Cancer Prevention and Screening Recommendations Provided by ChatGPT. Radiology. 2023;307(4):230424.
20. Shaban-Nejad A, Michalowski M, Buckeridge DL. Health intelligence: how artificial intelligence transforms population and personalized health. NPJ Digit Med. 2018;1:53.
21. Ueda D, Kakinuma T, Fujita S, Kamagata K, Fushimi Y, Ito R, et al. Fairness of artificial intelligence in healthcare: review and recommendations. Jpn J Radiol. 2023. https://doi.org/10.1007/s11604-023-01474-3.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC