Research Letter

The Performance of ChatGPT-4V in Interpreting Images and Tables in the Japanese Medical Licensing Exam

Soshi Takagi[1*], BA; Masahide Koda[2*], MD, PhD; Takashi Watari[3,4*], MHQS, MD, PhD

[1]Faculty of Medicine, Shimane University, Izumo, Japan

[2]Co-learning Community Healthcare Re-innovation Office, Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama University, Okayama, Japan

[3]General Medicine Center, Shimane University Hospital, Izumo, Japan

[4]Integrated Clinical Education Center, Kyoto University Hospital, Kyoto, Japan

*all authors contributed equally

Corresponding Author:
Takashi Watari, MHQS, MD, PhD
General Medicine Center
Shimane University Hospital
89-1, Enya
Izumo, 693-8501
Japan
Phone: +81853202005
Email: wataritari@gmail.com

JMIR Med Educ 2024;10:e54283; doi: 10.2196/54283

Keywords: ChatGPT; medical licensing examination; generative artificial intelligence; medical education; large language model; images; tables; artificial intelligence; AI; Japanese; reliability; medical application; medical applications; diagnostic; diagnostics; online data; web-based data

Introduction

OpenAI's ChatGPT, a leading large language model (LLM), has shown promise for medical purposes. The program can pass the United States Medical Licensing Examination (USMLE) and the Japanese Medical Licensing Exam (JMLE) [1-3]. However, previous studies regarding this software have focused on its text-based capabilities. ChatGPT-4 Vision (ChatGPT-4V), announced on September 25, 2023, includes image input features, potentially expanding the medical applications of the program [4]. To assess the multimodal performance of ChatGPT-4V in medicine, its performance on JMLE questions involving clinical images and tables was tested.

Methods

Overview

ChatGPT-4V was used to complete the 117th JMLE in the Japanese language (Figure S1 in Multimedia Appendix 1). Its responses were compared to the passing criteria and mean human examinee score of the JMLE. This study, conducted from October 12 to 14, 2023, used the September 25, 2023, version of the LLM (ChatGPT-4V) with a knowledge cutoff date of January 2022 (Multimedia Appendix 2 [5]). Human examinees' correct response rates were obtained from statistics based on reports from actual JMLE examinees, calculated by medu4, a preparatory school for the JMLE [5,6].

Statistical Analysis

The mean and 95% CIs of the test scores are provided. A one-sample proportion test was used to compare the correct response rate of the human examinees with that of ChatGPT-4V. Statistical significance was set at *P*<.05 for all 2-tailed tests. All statistical analyses were conducted using Stata statistical software (version 17; StataCor).

Ethical Considerations

This study used previously available web-based data and did not include human participants. Therefore, Shimane University's Institutional Review Board did not mandate ethics approval.

Results

Evaluation Outcomes

The responses to 386 questions from the 117th JMLE were used in this study. Using the Ministry of Health, Labor, and Welfare criteria, GPT-4V scored 85.1% on the essential knowledge section and 76.5% on the other sections of the JMLE, meeting the passing criteria [6]. For text-only questions, ChatGPT-4V achieved a correct response rate of 84.5%, similar to the mean human examinee score (Table 1). The correct response rate for questions with images was 71.9% for ChatGPT-4V, 13.1 points below the mean human examinee score (*P*<.001). The correct response rate for questions with tables (including figures) was 35.0% for ChatGPT-4V, which was significantly lower than the mean human examinee score (83.9%; *P*<.001).

Table 1. Correct response rates of ChatGPT-4 Vision (ChatGPT-4V) and human examinees on the Japanese Medical Licensing Examination (JMLE).

Characteristics	Total, n (%)	Examinees[a], mean	GPT-4V, mean	95% CI	Difference	*P* value
All questions	386 (100)	84.9	78.2	74.1-82.4	–6.7	.003
Question category						
Essential knowledge	96 (24.9)	89.6	83.3	75.9-90.8	–6.3	.04
General clinical knowledge	144 (37.3)	83.1	70.8	63.4-78.3	–12.3	<.001
Specific diseases	146 (37.8)	83.5	82.2	76.0-88.4	–1.3	.67
Type						
General	190 (49.2)	84.6	78.9	73.2-84.7	–5.7	.03
Clinical	149 (38.6)	84.1	77.2	70.4-83.0	–6.9	.02
Clinical sentence	47 (12.2)	88.5	78.7	67.0-90.4	–9.8	.04
Imaging and table questions						
Text only	252 (65.3)	84.9	84.5	80.1-89.0	–0.4	.87
With images	114 (29.5)	85.0	71.9	63.7-80.2	–13.1	<.001
With tables	20 (5.2)	83.9	35.0	14.1-55.9	–48.9	<.001

[a] The correct response rates of human examinees are based on a survey of actual human examinees, reported by medu4, a preparatory school for the JMLE [5].

Discussion

Principal Results

Although ChatGPT-4V demonstrated proficiency in text-centric questions, the correct response rates were significantly lower for image and table-oriented questions. ChatGPT-4V may have poorer text comprehension skills compared to ChatGPT-4, even when image processing is not required [7]. Additionally, a language bias may obscure the image context when interpreting images and texts simultaneously, potentially leading to an overreliance on prior text information, even when it contradicts the image context, a phenomenon called “hallucination” [8]. These factors may have led to ChatGPT-4V’s lower rate of correct responses to questions involving images.

Furthermore, responding to questions with tables requires interpreting the Japanese characters within the tables. OpenAI has verified that its GPT-4V model misrecognizes symbols, including image characters [4]. Previous studies have noted that GPT-4V relies on text-based information rather than an analysis of tables when answering questions [8]. In addition, the program’s performance diminishes when interpreting characters in non-Latin languages [9]. These factors may explain the observed decline in performance when interpreting tables containing Japanese characters.

The multimodal LLM GPT-4V is unreliable in interpreting information presented in image or tables, especially for medical purposes [4]. Further development of the program is required for diagnostic applications.

Limitations

This study has several limitations. First, different results may be obtained even when using the same methods owing to the inherent randomness of ChatGPT or version changes in ChatGPT. A report indicates that test results can vary with repeated responses from ChatGPT [10]. Furthermore, when providing images to ChatGPT, we did not remove blank spaces, indicating that the quality of images sent to ChatGPT could also affect the outcomes. Second, the JMLE includes options that, if selected twice or more, will result in failure. However, these options are not publicly disclosed, making them unaccounted for in this study [5]. Finally, although this study focused on ChatGPT, ongoing advancements in other multimodal LLMsshould also be considered.

Conclusions

ChatGPT-4V successfully passed the 117th JMLE, demonstrating proficiency in handling including image- and table-based questions. However, more developments are needed to improve its ability to interpret tables. Further research should assess the safety and efficacy of ChatGPT-4V as a multimodal LLM in supporting medical practice, facilitating learning in clinical environments and advancing medical education.

Acknowledgments

We would like to thank Dr Kota Sakaguchi, Shimane University Hospital, for his careful support throughout this study. We would also like to thank Dr Sanjay Saint, a professor at the University of Michigan, for his numerous contributions and support in this work.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Additional statistics.
[DOCX File (Microsoft Word File), 2285 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Detailed methodology.
[DOCX File (Microsoft Word File), 17 KB-Multimedia Appendix 2]

References

1. Introducing ChatGPT. OpenAI. 2022. URL: https://openai.com/blog/chatgpt/ [Accessed 2023-11-30]
2. Gilson A, Safranek CW, Huang T, 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. Feb 8, 2023;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
3. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: comparison study. JMIR Med Educ. Jun 29, 2023;9:e48002. [doi: 10.2196/48002] [Medline: 37384388]
4. GPT-4V(Ision) system card. OpenAI. 2023. URL: https://cdn.openai.com/papers/GPTV_System_Card.pdf [Accessed 2023-10-26]
5. Announcement of successful passage of the 117th National Medical Examination (Japanese) [Article in Japanese]. Ministry of Health. 2023. URL: https://www.mhlw.go.jp/general/sikaku/successlist/2023/siken01/about.html [Accessed 2023-10-26]
6. Searching questions [Article in Japanese]. Medu4. 2023. URL: https://medu4.com/quizzes/search [Accessed 2023-10-26]
7. Wu Y, Wang S, Yang H, et al. An early evaluation of GPT-4V(ision). arXiv. Preprint posted online on Oct 25, 2023. URL: https://arxiv.org/abs/2310.16534 [Accessed 2024-05-14]
8. Liu F, Lin K, Li L, Wang J, Yacoob Y, Wang L. Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning. arXiv. Preprint posted online on Jun 26, 2023. URL: https://arxiv.org/abs/2306.14565 [Accessed 2024-05-14]
9. Shi Y, Peng D, Liao W, Lin Z, Chen X, Liu C, et al. Exploring OCR capabilities of GPT-4V(ision): a quantitative and in-depth evaluation. arXiv. Preprint posted online on Oct 25, 2023. URL: https://arxiv.org/abs/2310.16809 [Accessed 2024-05-14]
10. Zhu L, Mou W, Yang T, Chen R. ChatGPT can pass the AHA exams: open-ended questions outperform multiple-choice format. Resuscitation. Jul 2023;188:109783. [doi: 10.1016/j.resuscitation.2023.109783] [Medline: 37349064]

Abbreviations

ChatGPT-4V: ChatGPT 4 Vision
JMLE: Japanese Medical Licensing Examination
LLM: large language model
USMLE: United States Medical Licensing Examination

Edited by Gunther Eysenbach, Taiane de Azevedo Cardoso; peer-reviewed by Fuxiao Liu, Lingxuan Zhu, Tianyu Ma; submitted 06.11.2023; final revised version received 09.04.2024; accepted 22.04.2024; published 23.05.2024

Please cite as:
Takagi S, Koda M, Watari T
The Performance of ChatGPT-4V in Interpreting Images and Tables in the Japanese Medical Licensing Exam
JMIR Med Educ 2024;10:e54283
URL: https://mededu.jmir.org/2024/1/e54283
doi: 10.2196/54283