Original Paper

Capability of GPT-4V(ision) in the Japanese National Medical Licensing Examination: Evaluation Study

Takahiro Nakao[1], MD, PhD; Soichiro Miki[1], MD, PhD; Yuta Nakamura[1], MD, PhD; Tomohiro Kikuchi[1,2], MD, PhD; Yukihiro Nomura[1,3], PhD; Shouhei Hanaoka[4], MD, PhD; Takeharu Yoshikawa[1], MD, PhD; Osamu Abe[4], MD, PhD

[1]Department of Computational Diagnostic Radiology and Preventive Medicine, The University of Tokyo Hospital, Bunkyo-ku, Tokyo, Japan

[2]Department of Radiology, School of Medicine, Jichi Medical University, Shimotsuke, Tochigi, Japan

[3]Center for Frontier Medical Engineering, Chiba University, Inage-ku, Chiba, Japan

[4]Department of Radiology, The University of Tokyo Hospital, Bunkyo-ku, Tokyo, Japan

Corresponding Author:
Takahiro Nakao, MD, PhD
Department of Computational Diagnostic Radiology and Preventive Medicine
The University of Tokyo Hospital
7-3-1 Hongo
Bunkyo-ku, Tokyo, 113-8655
Japan
Phone: 81 358008666
Email: tanakao-tky@umin.ac.jp

Abstract

Background: Previous research applying large language models (LLMs) to medicine was focused on text-based information. Recently, multimodal variants of LLMs acquired the capability of recognizing images.

Objective: We aim to evaluate the image recognition capability of generative pretrained transformer (GPT)-4V, a recent multimodal LLM developed by OpenAI, in the medical field by testing how visual information affects its performance to answer questions in the 117th Japanese National Medical Licensing Examination.

Methods: We focused on 108 questions that had 1 or more images as part of a question and presented GPT-4V with the same questions under two conditions: (1) with both the question text and associated images and (2) with the question text only. We then compared the difference in accuracy between the 2 conditions using the exact McNemar test.

Results: Among the 108 questions with images, GPT-4V's accuracy was 68% (73/108) when presented with images and 72% (78/108) when presented without images (P=.36). For the 2 question categories, clinical and general, the accuracies with and those without images were 71% (70/98) versus 78% (76/98; P=.21) and 30% (3/10) versus 20% (2/10; P≥.99), respectively.

Conclusions: The additional information from the images did not significantly improve the performance of GPT-4V in the Japanese National Medical Licensing Examination.

(JMIR Med Educ 2024;10:e54393) doi: 10.2196/54393

KEYWORDS

AI; artificial intelligence; LLM; large language model; language model; language models; ChatGPT; GPT-4; GPT-4V; generative pretrained transformer; image; images; imaging; response; responses; exam; examination; exams; examinations; answer; answers; NLP; natural language processing; chatbot; chatbots; conversational agent; conversational agents; medical education

Introduction

The field of natural language processing is rapidly developing with the advent of large language models (LLMs). LLMs are models trained with massive text data sets and achieve the capability to understand and generate text in natural languages. With the introduction of ChatGPT (OpenAI) [1] and other LLM-based chatbot services, many people have started to benefit from the use of LLMs. Although ChatGPT and its underlying model, generative pretrained transformer (GPT) [2,3], were not specifically developed for medical purposes, they possess a considerable amount of medical knowledge. They have achieved good scores in the United States Medical Licensing Examination [4] and are being explored for various applications for clinical and educational purposes [5-7]. GPT can also understand

languages other than English. The latest model, GPT-4, has been reported to achieve passing scores in medical licensing examinations in non–English speaking countries such as Japan, China, Poland, and Peru [8-13].

Despite these successes, there is still a significant challenge in applying LLMs to real-world problems with non–text-based information. Radiological, pathological, and many other types of visual information play a crucial role in determining a patient's management. Very recently, researchers have proposed multimodal variants of LLMs that can handle not only text but various types of input including images [14]. Providing medical images to multimodal LLMs may realize an even higher accuracy in solving medical-related problems. However, in previous studies on the accuracy rate of medical licensing examinations, questions with images were either not mentioned at all or explicitly excluded from the studies. To the best of our knowledge, no study directly evaluated the performance in solving questions with images. Therefore, in this study, we investigated the image recognition capabilities and limitations of GPT-4V [3,15], one of the most potent publicly available multimodal (vision and language) models, in solving medical questions. We focused on the Japanese National Medical Licensing Examination to examine how the visual information affects GPT-4V's performance.

Methods

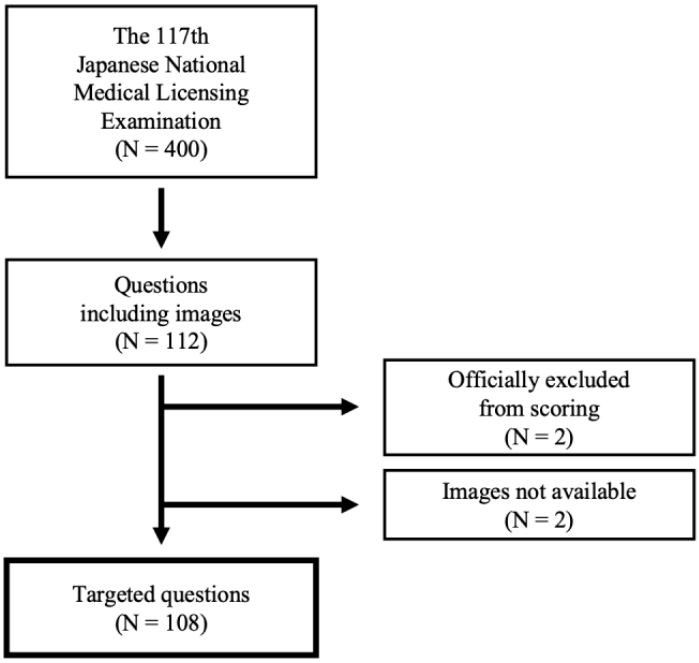
Overview

From the questions of the 117th Japanese National Medical Licensing Examination, held in February 2023, we focused on those that included images as part of a question. Since some of these questions can be answered correctly without interpreting images, we measured the benefit of adding image information by comparing the accuracy rates of ChatGPT under two different conditions: (1) with both the question text and associated images and (2) with the question text only.

Data Set Details

Figure 1 shows the summary of our data set. The questions and correct answers of the 117th Japanese National Medical Licensing Examination are publicly available for download on the official website of the Ministry of Health, Labour and Welfare [16]. All the questions are in a format in which a specified number of choices, typically 1, are to be selected from 5 options. Of the questions that had images, 2 were officially excluded from scoring because they were either too difficult or inappropriate. Additionally, for 2 questions, images of female genitals were not made public on the aforementioned website. These 4 questions were excluded from our study.

Figure 1. Summary of the questions included in this study.



The questions in the Japanese National Medical Licensing Examination were divided into 2 categories: clinical questions and general questions. In clinical questions, clinical information about a specific case is first presented, such as medical history and test results, and answers to questions about the case are required. General questions are about basic medical knowledge, and one is required to choose the correct answer among options for a short question text (typically of 1 or 2 sentences) with an image.

Some clinical questions consisted of multiple subquestions, in which case the background common to all the subquestions was first described, followed by the subquestions. In such cases, each subquestion was individually included in the following analysis if either the subquestion itself or the background part contained an image.

As a result, counting subquestions individually, out of 400 questions, we collected 108 questions that had images, such as photographs of lesions, radiographic images, histopathological images, electrocardiograms, and graphs representing statistical data. Among them, 98 were clinical questions and 10 were general questions.

Experimental Details

We used ChatGPT (September 25, 2023, version) enabled with GPT-4V, which is a multimodal model capable of processing both text and images. This version of ChatGPT asserts it was trained with information up to January 2022, meaning that it had no direct prior knowledge about our target examination. All the question statements and images were manually entered through ChatGPT's web interface. One of the authors, TN, who has 10 years of experience as a medical doctor, reviewed the outputs to interpret the response output by ChatGPT.

A new chat session was created for each question and each condition (ie, with or without images). For questions that comprised multiple subquestions, the background information part and each subquestion were entered into ChatGPT in this order within the same chat session. Subquestions without images were also input to provide ChatGPT with enough context, but they were excluded from the accuracy calculations and the subsequent statistical analysis described below.

The questions were presented to ChatGPT without any preceding or custom instructions. Sometimes, ChatGPT did not respond with the specified number of choices, in which case an additional instruction, such as "select only one option" or "select two options," was provided in Japanese. This additional instruction produced the correct number of options for all the questions.

Statistical Analysis

The difference in ChatGPT's performance between the 2 conditions (ie, with or without images) was analyzed using the exact McNemar test. A *P* value of less than .05 was considered statistically significant. The analysis was conducted using R (version 4.3.1; R Foundation for Statistical Computing).

Ethical Considerations

This study was conducted solely using publicly available resources, therefore, approval from the institutional review board of our institution was not required.

Results

Table 1 shows the results of our experiment. ChatGPT correctly answered 68% (73/108) of image-based questions when provided with both the question text and images, whereas it correctly answered 72% (78/108) of image-based questions when only the question text was provided. There was no significant difference in accuracy between these 2 conditions (*P*=.36). For the clinical questions, the accuracies when presented with and without images were 71% (70/98) and 78% (76/98), respectively. For the general questions, the accuracies were 30% (3/10) when presented with images and 20% (2/10) without images. We have included examples of the input and output along with their English translations in Multimedia Appendix 1, and we have also provided a summary of image interpretation for each question where the results differed depending on the presence of image input (N=7+12) in Multimedia Appendix 2.

Table 1. Performance of ChatGPT in answering questions from the 117th Japanese National Medical Licensing Examination, when presented with or without associated images for each question.

	With images		
	Correct	Incorrect	Total
Overall (*P*=.36)			
Without images, n (%)			
Correct	66 (61)	12 (11)	78 (72)
Incorrect	7 (6)	23 (21)	30 (28)
Total	73 (68)	35 (32)	108 (100)
Clinical (*P*=.21)			
Without images , n (%)			
Correct	65 (66)	11 (11)	76 (78)
Incorrect	5 (5)	17 (17)	22 (22)
Total	70 (71)	28 (29)	98 (100)
General (*P*≥.99)			
Without images, n (%)			
Correct	1 (10)	1 (10)	2 (20)
Incorrect	2 (20)	6 (60)	8 (80)
Total	3 (30)	7 (70)	10 (100)

Discussion

Principal Results

In this study, we examined the image recognition capabilities of GPT-4V using questions associated with images from the Japanese National Medical Licensing Examination. To the best of our knowledge, this is the first study in which the capability of multimodal LLM for the Japanese National Medical Licensing Examination was investigated. Contrary to our initial expectations, the inclusion of image information did not result in any improvement in accuracy. Instead, we even observed a slight decrease, albeit not significant. This indicates that, at the moment, GPT-4V cannot effectively interpret images related to medicine. The passing score rate for the 117th Japanese National Medical Licensing Examination is approximately 75% (and 80% for some questions marked as "essential") [16]. In this study, GPT-4V failed to reach this passing score rate for the questions it was tested on. Considering that 92% of human candidates passed, it implies that the image interpretation skills of GPT-4V will fall short of those possessed by many medical students.

For the clinical questions, in which sufficient clinical information including patient history was available in the text form, GPT-4V was able to choose the correct answers solely from the textual information in the majority (76/98, 78%) of questions, but the addition of images did not improve the accuracy. On the other hand, for the general questions, there was little information in the question text, and GPT-4V had to determine the correct answer by interpreting the images. For these, GPT-4V yielded an accuracy rate that was hardly any better than random guessing even when presented with images. Our results suggest that, for both categories of questions, GPT-4V failed to use visual information to improve its accuracy. We observed that GPT-4V often either explicitly stated that it was unable to interpret the images or failed to provide information beyond what was evident from the question text. In our retrospective review, even in questions where GPT-4V gave correct answers only when presented with images, there were only 2 out of 7 questions where it provided a correct interpretation of the image and used that as a critical clue. Conversely, in questions where GPT-4V provided incorrect answers only when presented with images, it sometimes made incorrect or insufficient interpretations of the images, leading to incorrect answers (4 out of 12).

ChatGPT may serve as a valuable teaching assistant in medical education; however, the inaccuracies in its responses are a significant concern [5,7]. Our current findings suggest that, especially with medical-related images, GPT-4V should not be relied upon as a primary source of information for medical education or practice. If used, extreme caution should be exercised regarding the accuracy of its responses. OpenAI officially states [15] that they "do not consider the current version of GPT-4V to be fit for performing any medical function or substituting professional medical advice, diagnosis, or treatment, or judgment" due to its imperfect performance in the medical domain. Yang et al [17] have comprehensively examined the capabilities of GPT-4V in various tasks including medical image understanding and radiology report generation, and they stated that GPT-4V could correctly diagnose some medical images. However, as they acknowledge, their results contained a considerable number of errors, such as overlooking obvious lesions and errors in laterality. According to the case studies by Wu et al [18], GPT-4V could recognize the modality and anatomy of medical images, but it could hardly make accurate diagnoses and its prediction relied heavily on the patient's medical history. The results of our experiment supported these previous reports.

Considering the well-known high performance of GPT-4V in more generic image recognition tasks [3,17], the most probable reason for its limited image recognition performance in the medical field is that it may simply not have been trained with a sufficient number of medical-related images. LLMs are trained with a vast data set available on the internet, but medical images are not as readily accessible, partly due to privacy concerns. Some researchers are now working on developing multimodal LLMs specialized for medicine based on open-source LLMs [19,20]. These models use publicly available data sets that combine medical images and text, including MIMIC-CXR [21], which contains chest x-ray images with their associated reports, and PMC-OA [22], a compilation of the figures and captions from open-access medical journal papers. The rise of multimodal LLMs is expected to stimulate the publication of more such data sets, thereby advancing the development of multimodal LLMs in the medical field. Moreover, although there are limited medical-related images publicly available on the internet, hospitals have a vast amount of image data. A large part of this is accompanied by textual interpretations in the form of reports or medical records, which may serve as an ideal data set for training multimodal LLMs. In highly specialized domains such as medicine, there remains a significant value in developing domain-specific models using such medical data sets.

Limitations

This study had several limitations. First, ChatGPT was not provided any prior instructions and was directly presented with only the questions themselves. This might have negatively affected its capability to interpret images as the capabilities of LLMs are known to be affected by such "prompt engineering." This will be a subject for future investigation. Second, this study specifically targeted the Japanese National Medical Licensing Examination, and thus, further analysis is necessary to determine whether its conclusions can be generalized to questions in other languages or of different types. However, as mentioned earlier, the limited capability of GPT-4V to interpret medical images has also been demonstrated in other studies focusing on English [17,18], and our results are consistent with those findings. Since ChatGPT's proficiency in non-English interpretation is known to be inferior to that in English interpretation, translating the question text into English before inputting it to ChatGPT might have improved the model's image interpretation capability. However, in a previous study by Yanagita et al [10], in which nonimage questions from the Japanese National Medical Licensing Examination were the target, satisfactory results were achieved even when the questions were input in Japanese. Thus, we adopted the same approach in our study. Third, although our results were based on the same version of ChatGPT and the

same question was evaluated with and without images on the same day, we cannot exclude the possibility that different models were used internally. Lastly, only a single evaluation was conducted for each condition and question. ChatGPT's outputs have some randomness, and responses may differ across multiple evaluations. With ChatGPT's application programming interface, users can programmatically control the degree of randomness by specifying a parameter called *temperature* and obtain mostly deterministic responses. However, during the time of this study, the application programming interface for GPT-4V was not available.

Conclusions

At present, GPT-4V's capability to interpret medical images may be insufficient. In highly specialized fields such as medicine, it is considered meaningful to develop field-specific multimodal models.

Acknowledgments

The Department of Computational Diagnostic Radiology and Preventive Medicine, The University of Tokyo Hospital, is sponsored by HIMEDIC Inc and Siemens Healthcare K.K.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Examples of inputs and outputs from GPT-4V.
[PDF File (Adobe PDF File), 997 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Summary of image interpretation by GPT-4V.
[DOC File , 50 KB-Multimedia Appendix 2]

References

1. Introducing ChatGPT. OpenAI. URL: https://openai.com/blog/chatgpt [accessed 2023-10-23]
2. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. ArXiv. [FREE Full text] [doi: 10.48550/arXiv.2005.14165]
3. OpenAI. GPT-4 technical report. ArXiv. Preprint posted online December 19, 2023. [FREE Full text] [doi: 10.48550/arXiv.2303.08774]
4. Brin D, Sorin V, Konen E, Nadkarni G, Glicksberg BS, Klang E. How large language models perform on the United States medical licensing examination: a systematic review. medRxiv. [FREE Full text] [doi: 10.1101/2023.09.03.23294842]
5. Lee H. The rise of ChatGPT: exploring its potential in medical education. Anat Sci Educ. 2023. [FREE Full text] [doi: 10.1002/ase.2270] [Medline: 36916887]
6. Cooper A, Rodman A. AI and medical education - a 21st-century Pandora's Box. N Engl J Med. 2023;389(5):385-387. [doi: 10.1056/NEJMp2304993] [Medline: 37522417]
7. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel). 2023;11(6):887. [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
8. Tanaka Y, Nakata T, Aiga K, Etani T, Muramatsu R, Katagiri S, et al. Performance of Generative Pretrained Transformer on the National Medical Licensing Examination in Japan. PLOS Digit Health. Jan 2024;3(1):e0000433. [FREE Full text] [doi: 10.1371/journal.pdig.0000433] [Medline: 38261580]
9. 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. [FREE Full text] [doi: 10.2196/48002] [Medline: 37384388]
10. Yanagita Y, Yokokawa D, Uchida S, Tawara J, Ikusaka M. Accuracy of ChatGPT on medical questions in the national medical licensing examination in Japan: evaluation study. JMIR Form Res. 2023;7:e48023. [FREE Full text] [doi: 10.2196/48023] [Medline: 37831496]
11. Fang C, Wu Y, Fu W, Ling J, Wang Y, Liu X, et al. How does ChatGPT-4 preform on non-english national medical licensing examination? An evaluation in Chinese language. PLOS Digit Health. 2023;2(12):e0000397. [FREE Full text] [doi: 10.1371/journal.pdig.0000397] [Medline: 38039286]
12. Flores-Cohaila JA, García-Vicente A, Vizcarra-Jiménez SF, De la Cruz-Galán JP, Gutiérrez-Arratia JD, Torres BGQ, et al. Performance of ChatGPT on the Peruvian national licensing medical examination: cross-sectional study. JMIR Med Educ. 2023;9:e48039. [FREE Full text] [doi: 10.2196/48039] [Medline: 37768724]

13. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish medical final examination. Sci Rep. 2023;13(1):20512. [FREE Full text] [doi: 10.1038/s41598-023-46995-z] [Medline: 37993519]
14. Zhang J, Huang J, Jin S, Lu S. Vision-language models for vision tasks: a survey. ArXiv. [FREE Full text] [doi: 10.48550/arXiv.2304.00685]
15. GPT-4V(ision) System Card. OpenAI. Sep 25, 2023. URL: https://cdn.openai.com/papers/GPTV_System_Card.pdf [accessed 2023-10-23]
16. 第１１７回医師国家試験問題および正答について [The 117th national medical licensing examination questions and correct answers]. Ministry of Health, Labour and Welfare. URL: https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp230502-01.html [accessed 2023-10-24]
17. Yang Z, Li L, Lin K, Wang J, Lin CC, Liu Z, et al. The dawn of LMMs: preliminary explorations with GPT-4V(ision). ArXiv. [FREE Full text] [doi: 10.48550/arXiv.2309.17421]
18. Wu C, Lei J, Zheng Q, Zhao W, Lin W, Zhang X, et al. Can GPT-4V(ision) serve medical applications? Case studies on GPT-4V for multimodal medical diagnosis. ArXiv. [FREE Full text] [doi: 10.48550/arXiv.2310.09909]
19. Moor M, Huang Q, Wu S, Yasunaga M, Zakka C, Dalmia Y, et al. Med-flamingo: a multimodal medical few-shot learner. ArXiv. [FREE Full text] [doi: 10.48550/arXiv.2307.15189]
20. Xu S, Yang L, Kelly C, Sieniek M, Kohlberger T, Ma M, et al. ELIXR: Towards a general purpose X-Ray artificial intelligence system through alignment of large language models and radiology vision encoders. ArXiv. [FREE Full text] [doi: 10.48550/arXiv.2308.01317]
21. Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng CY, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data. 2019;6(1):317. [FREE Full text] [doi: 10.1038/s41597-019-0322-0] [Medline: 31831740]
22. Lin W, Zhao Z, Zhang X, Wu C, Zhang Y, Wang Y, et al. PMC-CLIP: Contrastive language-image pre-training using biomedical documents. ArXiv. [FREE Full text] [doi: 10.48550/arXiv.2303.07240]

Abbreviations

GPT: generative pretrained transformer
LLM: large language model

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 08.11.23; peer-reviewed by D Hu, M Chatzimina; comments to author 07.12.23; revised version received 26.12.23; accepted 16.02.24; published 12.03.24

Please cite as:
Nakao T, Miki S, Nakamura Y, Kikuchi T, Nomura Y, Hanaoka S, Yoshikawa T, Abe O
Capability of GPT-4V(ision) in the Japanese National Medical Licensing Examination: Evaluation Study
JMIR Med Educ 2024;10:e54393
URL: https://mededu.jmir.org/2024/1/e54393
doi: 10.2196/54393
PMID: 38470459