Performance and Errors of ChatGPT-4o on the Japanese Medical Licensing Examination: Solving All Questions Including Images with Over 90% Accuracy

Yuki Miyazaki, Masahiro Hata, Hisaki Omori, Atsuya Hirashima, Yuta Nakagawa, Mitsuhiro Eto, Shun Takahashi, Manabu Ikeda

Submitted to: JMIR Medical Education
on: June 13, 2024

Table of Contents

Original Manuscript..... 5

Performance and Errors of ChatGPT-4o on the Japanese Medical Licensing Examination: Solving All Questions Including Images with Over 90% Accuracy

Yuki Miyazaki[1] MD; Masahiro Hata[1] MD, PhD; Hisaki Omori[1, 2] MD; Atsuya Hirashima[1, 3] MD; Yuta Nakagawa[1, 4] DR; Mitsuhiro Eto[1, 4] MD; Shun Takahashi[1, 5, 6, 7] MD, PhD; Manabu Ikeda[1] MD, PhD

[1]Department of Psychiatry Osaka University Graduate School of Medicine Suita JP
[2]Department of Psychiatry Shichiyama Hospital Sennan District JP
[3]Department of Psychiatry Osaka Psychiatric Medical Center Hirakata JP
[4]Department of Psychiatry Asakayama General Hospital Sakai JP
[5]Clinical Research and Education Center Asakayama General Hospital Sakai JP
[6]Graduate School of Rehabilitation Science Osaka Metropolitan University Habikino JP
[7]Department of Neuropsychiatry Wakayama Medical University Wakayama JP

Corresponding Author:
Yuki Miyazaki MD
Department of Psychiatry
Osaka University Graduate School of Medicine
2-2 D3
Yamadaoka
Suita
JP

Abstract

Background: Recent advancements in AI technology have begun to play a crucial role in medical education. AI models, such as ChatGPT, have shown promise in various applications, including answering medical questions and assisting in clinical decision-making. However, there is limited research on the performance of these models on comprehensive medical licensing exams.

Objective: This study aims to evaluate the performance of ChatGPT-4o on the 118th Japanese Medical Licensing Examination (JMLE), specifically assessing its ability to handle both text and image-based questions, and to analyze the types of errors it makes.

Methods: ChatGPT-4o was utilized to complete all 400 questions of the 118th JMLE held in February 2024. The model, updated with data up to May 13, 2023, was assessed on its ability to answer both text-only and image-based questions. Questions were directly input into the chat interface without the use of prompt engineering or memory functions. Due to the daily response limit of ChatGPT-4o, the study was conducted from May 13 to May 19, 2024. An independent samples t-test compared the correct response rates between image-based and text-only questions. Statistical significance was set at ????<.05 for all two-tailed tests.

Results: ChatGPT-4o achieved an overall correct response rate of 93.25%, with 93.48% for image-based and 93.18% for text-only questions. The difference in correct response rates between text-only and image-based questions was not statistically significant (t-value: -0.074, p-value: 0.941). The errors were classified into four categories: diagnostic errors, logical errors, medical knowledge errors, and reading comprehension errors.

Discussion
ChatGPT-4o demonstrated high proficiency in both text-centric and image-based questions, marking a significant improvement over previous iterations of GPT models. This performance meets the passing criteria set by the Ministry of Health, Labor, and Welfare for the JNMLE, which requires a total score of at least 160/200 points for compulsory questions, at least 230/300 points for non-compulsory questions, and no more than 3 incorrect choices in critical exclusion questions. Although ChatGPT-4o met the overall passing criteria, some responses indicated potentially problematic clinical judgments, such as incorrect triage decisions and prioritization errors in clinical scenarios. These findings underscore the need for improved clinical judgment capabilities in AI models.

Conclusions: ChatGPT-4o successfully met the passing criteria for the 118th JNMLE, demonstrating high proficiency in handling both text and image-based questions. This marks a significant improvement over previous iterations of GPT models, particularly in managing multimodal tasks. The model excelled in answering specific medical knowledge questions, indicating a strong grasp of medical facts and concepts. However, it struggled with clinical judgment and prioritization, as evidenced by errors in triage decisions and the selection of appropriate diagnostic procedures. These findings highlight the need for continued enhancement of AI models to ensure their reliability and accuracy in clinical decision-making. While generative AI like ChatGPT-4o shows great potential, understanding and addressing its limitations will be critical for its effective integration into medical education and practice.

(JMIR Preprints 13/06/2024:63129)
DOI: https://doi.org/10.2196/preprints.63129

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Research Letter
[Title]

Performance and Errors of ChatGPT-4o on the Japanese Medical Licensing Examination: Solving All Questions Including Images with Over 90% Accuracy

[Authors]
Yuki Miyazaki[1*], Masahiro Hata[1], Hisaki Omori[1,2], Atsuya Hirashima[1,3], Yuta Nakagawa[1,4], Mitsuhiro Eto[1,4], Shun Takahashi[1,5,6,7], Manabu Ikeda[1].

1. Department of Psychiatry, Osaka University Graduate School of Medicine, Suita, Osaka, Japan
2. Shichiyama Hospital, Sennan District, Osaka, Japan
3. Osaka Psychiatric Medical Center, Hirakata, Osaka, Japan
4. Department of Psychiatry, Asakayama General Hospital, Osaka, Japan
5. Clinical Research and Education Center, Asakayama General Hospital, Sakai, Osaka, Japan
6. Graduate School of Rehabilitation Science, Osaka Metropolitan University, Habikino, Osaka, Japan
7. Department of Neuropsychiatry, Wakayama Medical University, Kimiidera, Wakayama, Japan

*Corresponding author: Yuki Miyazaki
Department of Psychiatry, Osaka University Graduate School of Medicine. 2-2 D3 Yamadaoka, Suita, Osaka 565-0871, Japan
Tel: +81-6-6879-3051, Fax: +81-6-6879-3054
E-mail: miyazaki@psy.med.osaka-u.ac.jp

Keywords: Medical Education; Artificial Intelligence; Clinical Decision-Making

Introduction

Artificial Intelligence (AI) models, such as ChatGPT [1], have shown promise in various applications, including answering medical questions and assisting in clinical decision-making. Previous studies have evaluated AI models on medical exams such as the USMLE, where ChatGPT-3 achieved correct response rates of 42-64% on Step 1 and 2 exams [2]. In Japan, studies on the Japanese Medical Licensing Examination (JMLE) reported that GPT-4 achieved a 77.7% correct response rate on 292 questions in 2022 (116th JMLE) [3] and 79.9% on 254 questions in 2023 (117th JMLE) [4]. A study with GPT-4, using prompt tuning, achieved 82.7% on essential questions and 77.2% on basics and clinical questions out of 336 questions [5]. For image-inclusive exams, GPT-4 Vision scored 78.2% on 386 questions, with a significantly lower performance on image (71.9%) and table questions (35.0%) [6]. No prior studies have evaluated an AI model on all 400 questions of the JMLE. The recently released ChatGPT-4o (GPT-4 omni) on May 13, 2024, represents a significant step towards more natural human-computer interaction, capable of accepting inputs in text, audio, image, and video formats, and generating outputs in text, audio, and image formats [7], promising improved performance on image-based questions. Recent comparisons of ChatGPT-4 with other AI models in psychiatric licensing exams have also shown its superior performance, emphasizing its potential in various medical fields [8]. As generative AI is increasingly applied in medical education, understanding its limitations will be essential for effectively integrating it into learning and practice. This study aims to examine the current strengths and weaknesses of generative AI using the JMLE.

Methods

Overview

ChatGPT-4o was utilized to complete all 400 questions of the 118th Japanese Medical Licensing Examination (JMLE) in February 2024 [9]. The model, updated with data up to May 2023, was assessed on its ability to answer both text-only and image-based questions. The questions were directly input into the chat interface without the use of prompt engineering or memory functions. Due to the daily response limit of ChatGPT-4o, the study was conducted over a period from May 13 to May 19, 2024.

Statistical Analysis

To compare the correct response rates between the image-based and text-only questions, an independent samples t-test was used. Statistical significance was set at $P<.05$ for all two-tailed tests. All statistical analyses were conducted using Python's SciPy library (v1.13.1) for statistical computations.

Ethical Considerations

This study used previously available web-based data and did not include human participants. Therefore, ethics approval was not mandated.

Results

Evaluation Outcomes

The overall accuracy was 93.25%, with 93.48% for image-based and 93.18% for text-only questions. The following table summarizes the correct response rates for each type of question:

[Table 1]

Table 1. Performance Comparison of ChatGPT-4o Across Different Sections in 118th Japanese Medical Licensing Examination (JMLE).

Characteristics	Total Correct [%]	Total Questions	Text-only Correct [%]	Text-only Questions	Image-based Correct [%]	Image-based Questions
Total	93.25	400	93.18	308	93.48	92
A	94.67	75	97.67	43	90.63	32
B	92.00	50	90.70	43	100.00	7
C	90.67	75	89.71	68	100.00	7
D	94.67	75	95.56	45	93.33	30
E	96.00	50	95.83	48	100.00	2
F	92.00	75	91.80	61	92.86	14

The difference in correct response rates between text-only and image-based questions was not statistically significant (t-value: -1.190, p-value: 0.257).

Error Classification

The errors made by ChatGPT-4o were further analyzed and classified into four categories: diagnostic errors, logical errors, medical knowledge errors, and clinical judgment errors. The detailed error classification is provided below:

[Table 2]

Table 2. Classification and Details of All Errors of ChatGPT-4o in 118th JMLE.

Problem No.	Classification	Error Details
A021	Diagnostic Error	Incorrect diagnosis.
A039	Logical Error	Incorrect logic regarding the risk reduction of PT sheet ingestion.
A059	Medical Knowledge Error	Incorrect use of medical knowledge regarding the urea breath test.
A061	Logical Error	Incorrect final answer despite correct assessment of individual questions.
B021	Medical Knowledge Error	Incorrect medical knowledge regarding the risk relationship of latex allergy after banana ingestion.
B038	Medical Knowledge Error	Incorrect medical knowledge for classifying activity restriction.
B047	Medical Knowledge Error	Incorrect medical knowledge about social support systems.
B049	Medical Knowledge Error	Incorrect medical knowledge for describing Trousseau sign.
C012	Logical Error	Correct medical knowledge but incorrect final answer (confusion between right and left).
C020	Medical Knowledge Error	Incorrect medical knowledge regarding occupational cataract risk.
C040	Clinical Judgment Error	Incorrect triage decision, suggesting black tag for a critically ill patient.
C043	Clinical Judgment Error	Incorrect clinical judgment prioritizing ultrasound over Cardiotocogram (CTG).
C055	Medical Knowledge Error	Incorrect medical knowledge regarding fetal rotation.
C056	Logical Error	Incorrect interpretation of the problem statement.
C074	Medical Knowledge	Incorrect medical knowledge for selecting the appropriate type of infusion.

	Error	
D012	Medical Knowledge Error	Incorrect medical knowledge regarding CKD severity classification.
D017	Diagnostic Error	Incorrect diagnosis.
D035	Medical Knowledge Error	Incorrect medical knowledge for selecting the appropriate type of infusion.
D047	Diagnostic Error	Incorrect diagnosis.
E034	Medical Knowledge Error	Incorrect medical knowledge regarding postprandial blood glucose targets in gestational diabetes management.
E041	Medical Knowledge Error	Incorrect medical knowledge for GCS motor response.
F001	Medical Knowledge Error	Incorrect medical knowledge regarding the design principles of tactile paving.
F010	Medical Knowledge Error	Incorrect medical knowledge regarding the peak population year in Japan.
F018	Medical Knowledge Error	Correct image interpretation but incorrect medical knowledge regarding sagittal suture alignment.
F054	Clinical Judgment Error	Incorrect decision between referring to a specialized hospital and a community support hospital.
F066	Logical Error	Incorrect interpretation and judgment regarding wheelchair options.
F068	Logical Error	Incorrect interpretation of the problem statement regarding creatinine clearance calculation.

Discussion

Principal Results

ChatGPT-4o achieved an overall correct response rate of 93.25% on the 2024 (118th) JMLE without the use of prompt engineering or memory functions, surpassing the results of prior studies involving earlier versions of GPT models. The performance of ChatGPT-4o did not decline on image-based or table questions, marking a significant improvement in multimodal question handling. The improved performance of ChatGPT-4o suggests that the integration of multimodal capabilities may have significantly enhanced its clinical decision-making skills. Additionally, ChatGPT-4o's real-time speech capabilities could revolutionize both exam performance and clinical interactions, enhancing diagnostic and consultative processes in real-world settings.

ChatGPT-4o’s performance meets the 118th JMLE passing criteria [10], which requires:

1. At least 160/200 points for compulsory questions (sections B and F).
2. At least 230/300 points for non-compulsory questions (sections A, C, D, and E).
3. No more than 3 incorrect choices in critical exclusion questions.

Error Analysis

The specific critical exclusion questions are not publicly disclosed. Although ChatGPT-4o met the overall passing criteria, some responses indicate potentially problematic clinical judgments. For instance, in question C040, the model incorrectly suggested a black tag for a critically ill patient during triage. In question C043, the model incorrectly prioritized ultrasound over Cardiotocogram (CTG) in a clinical decision. These errors highlight the potential for clinical judgment mistakes in AI models, as ChatGPT-4o struggled with questions requiring prioritization in clinical settings. This is a critical skill that will become increasingly important in medical education.

Conclusion

ChatGPT-4o successfully met the passing criteria for the 118th JMLE, demonstrating high proficiency in handling both text- and image-based questions. This marks a significant improvement over previous iterations of GPT models, particularly in managing multimodal tasks. The model excelled in answering specific medical knowledge questions, indicating a strong grasp of medical facts and concepts. However, it struggled with clinical judgment and prioritization, as evidenced by errors in triage decisions and the selection of appropriate diagnostic procedures. These findings underscore the need for continued enhancement of AI models to ensure their reliability and accuracy in clinical decision-making. While generative AI like ChatGPT-4o shows great potential, understanding and addressing its limitations will be critical for its effective integration into medical education and practice.

Limitations

This study revealed limitations of ChatGPT.

First, the JMLE includes questions related to Japan's healthcare system. For instance, ChatGPT-4o misclassified a question about social support systems (B047), which might be answered correctly if it pertained to another country's system.

Second, the error analysis was conducted based on the explanations provided by ChatGPT-4o. It remains unclear whether ChatGPT-4o actually reasons based on the provided justifications or if it merely generates plausible explanations post-hoc. Given the multiple-choice nature of the exam, there is a possibility that ChatGPT-4o could randomly select an answer and then generate a convincing rationale that might not reflect its true reasoning process.

Conflicts of Interest

None declared.

Abbreviations

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

References

1. ChatGPT. OpenAI. 2024. https://openai.com/chatgpt/ [Accessed 2024-5-31]
2. Gilson A, Safranek C, Huang T, Socrates V, Chi L, Taylor R, Chartash D. 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 2023;9:e45312.
DOI: 10.2196/45312
3. 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.
DOI: 10.2196/48023
4. 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.
DOI: 10.2196/48002
5. Tanaka Y, Nakata T, Aiga K, Etani T, Muramatsu R, et al. Performance of Generative Pretrained Transformer on the National Medical Licensing Examination in Japan. PLOS Digital Health 2024;3(1): e0000433.
DOI: https://doi.org/10.1371/journal.pdig.0000433
6. 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.
DOI: 10.2196/54283
7. Hello GPT-4o. OpenAI. 2024. https://openai.com/index/hello-gpt-4o/ [Accessed 2024-5-31]
8. Li DJ, Kao YC, Tsai SJ, Bai YM, Yeh TC, Chu CS, Hsu CW, Cheng SW, Hsu TW, Liang CS, Su KP. Comparing the performance of ChatGPT GPT-4, Bard, and Llama-2 in the Taiwan Psychiatric Licensing Examination and in differential diagnosis with multi-center psychiatrists. Psychiatry and Clinical Neurosciences. 2024;78(6):347-352. DOI: 10.1111/pcn.13656.
9. Ministry of Health, Labour and Welfare. The 118th National Medical Examination Questions and Correct Answers (Japanese). 2024 https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp240424-01.html [Accessed 2024-5-13]
10. Ministry of Health, Labour and Welfare. Announcement of Successful Passage of the 118th National Medical Examination (Japanese). 2024. https://www.mhlw.go.jp/content/10803000/001226841.pdf [Accessed 2024-5-31]