Evaluating Chat GPT-4o's Comparative Performance Over GPT-4 in Japanese Medical Licensing Examination and Its Clinical Partnership Potential

Masatoshi Miyamura , Goro Fujiki , Yumiko Kanzaki , Kosuke Tsuda , Hironaka Asano , Hideaki Morita * , Masaaki Hoshiga

Posted Date: 7 December 2025

doi: 10.20944/preprints202512.0565.v1

Keywords: artificial intelligence; multimodal large language model; ChatGPT; japan national medical licensing examinations

Article

Evaluating Chat GPT-4o's Comparative Performance Over GPT-4 in Japanese Medical Licensing Examination and Its Clinical Partnership Potential

Masatoshi Miyamura [1], Goro Fujiki [2], Yumiko Kanzaki [1], Kosuke Tsuda [1], Hironaka Asano [3], Hideaki Morita [1,*] and Masaaki Hoshiga [4]

[1] Department of Cardiology, Osaka Medical and Pharmaceutical University, Daigakumachi 2-7, Takatsuki City, Osaka Prefecture, Japan
[2] Department of Cardiology, Nippon Life Hospital, Enokojima 2-1-54, Nishi-ku, Osaka City, Osaka Prefecture, Japan
[3] Department of Cardiology, Itami City Hospital, Koyaike 1-100, Itami City, Hyougo Prefecture, Japan
[4] Department of Cardiology, Osaka Medical and Pharmaceutical University Mishima-minami Hospital, Tamagawashinmachi 8-1, Takatsuki City, Osaka Prefecture, Japan
* Correspondence: hideaki.morita@ompu.ac.jp; Tel.: +81-72-683-1221

Abstract

Backgrounds: Recent advances in artificial intelligence (AI) have produced ChatGPT-4o, a multimodal large language model (LLM) capable of processing both text and image inputs. Although ChatGPT has demonstrated usefulness in medical examinations, few studies have evaluated its image analysis performance. Methods: This study compared GPT-4o and GPT-4 using public questions from the 116th–118th Japan National Medical Licensing Examinations (JNMLE), each consisting of 400 questions. Both models answered in Japanese using simple prompts, including screenshots for image-based questions. Accuracy was analyzed across essential, general, and clinical questions, with statistical comparisons by chi-square tests. Results: GPT-4o consistently outperformed GPT-4, achieving passing scores in all three examinations. In the 118th JNMLE, GPT-4o scored 457 points versus 425 for GPT-4. GPT-4o demonstrated higher accuracy for image-based questions in the 117th and 116th exams, though the difference in the 118th was not significant. For text-based questions, GPT-4o showed superior medical knowledge, clinical reasoning, and ethical response behavior, notably avoiding prohibited options. Conclusion: Overall, GPT-4o exceeded GPT-4 in both text and image domains, suggesting strong potential as a diagnostic aid and educational resource. Its balanced performance across modalities highlights its promise for integration into future medical education and clinical decision support.

Keywords: artificial intelligence; multimodal large language model; ChatGPT; japan national medical licensing examinations

1. Introduction

Recent advances in artificial intelligence (AI) have led to the development of sophisticated large-scale language models (LLM) that can process and generate human-like text. Among them, ChatGPT-4 (GPT-4) from OpenAI (OpenAI, San Francisco, CA) has shown great progress in generating and understanding text, making it a notable milestone [1]. ChatGPT can be accessed using standard computers with an internet connection. ChatGPT and its underlying model, generative pre-trained transformers (GPT), were not developed specifically for medical purposes.

GPT-4 focuses primarily on text-based interactions and has limited processing power for image inputs. Although it can process basic image data and provide basic analysis, its performance in

 doi:10.20944/preprints202512.0565.v1

complex image processing tasks is limited. GPT-4 has reportedly achieved a passing score for the medical licensing exam in non-English-speaking countries such as Japan [2–4].

However, previous studies have only considered text-based questions, not problems involving images, figures, tables, or graphs.

In contrast, ChatGPT-4o (GPT-4o) takes a big leap forward with its multimodal capabilities, allowing it to generate responses based on text, image, audio, and video inputs. One of GPT-4o's key advancements is its ability to seamlessly integrate multiple data types to provide richer, more contextually relevant interactions. This multimodal approach enables more nuanced and accurate image analysis, facilitating a variety of applications ranging from medical diagnosis to advanced visual content creation [5].

The Japanese National Medical Licensing Examinations (JNMLE) is a rigorous and comprehensive test required to obtain a medical license in Japan. This ensures only qualified individuals are licensed.

The JNMLE typically consists of multiple-choice questions covering a wide range of medical knowledge, including basic medicine, clinical medicine, and public health, divided into several sections. The purpose of this exam is to assess candidates' breadth and depth of medical knowledge and their ability to apply this knowledge in real clinical scenarios.

One of the features of the exam is its focus on clinical reasoning and decision-making skills. Candidates are tested not only on theoretical knowledge but also on their ability to interpret clinical data, perform differential diagnoses, and propose appropriate treatment plans. This reflects the importance of practical clinical skills.

One unique characteristic of the exam is the inclusion of so-called "prohibited choices" (kinshi-shi). These are answer options that represent actions or decisions that could seriously endanger a patient's life. Candidates who select a certain number of prohibited choices—typically two or three—are disqualified from passing the exam, regardless of their overall score. This mechanism is intended to ensure a baseline of ethical and clinical safety in medical practice.

The examination is conducted annually, typically taken by medical graduates who have completed six years of medical education in Japan. The pass rate is relatively high, reflecting the rigorous training provided by Japanese medical schools. However, the exam remains challenging, requiring thorough preparation for success.

In summary the JNMLE is a critical element in ensuring that candidates possess the knowledge and skills necessary to provide high-quality healthcare. Its comprehensive scope and emphasis on clinical competence form the foundation of Japan's healthcare system.

The JNMLE frequently includes image-based questions. However, the accuracy rate for these image-related questions was low in previous versions of ChatGPT and other LLM [6,7].

We evaluated the performance of GPT4o and GPT4 by having them solve questions from the 118th, 117th, and 116th JNMLE, confirming superior capabilities in image analysis.

2. Materials and Methods

We downloaded the questions and answers for the 118th, the 117th, and the 116th JNMLE from the Japanese Ministry of Health, Labor and Welfare's website [8–10].

We assumed that the AI would be used by general medical professionals, so we did not train the AI to increase its rate of correct answers, nor did we input complex prompts to increase its rate of correct answers. Instead, we had the AI solve the problems using prompts as simple as possible.

For each question, we started by entering the initial prompt in Japanese, "Please solve the Japanese National Medical Licensing Examinations' questions." In the subsequent prompt, for textual questions, we copied and pasted the text to obtain the answers from ChatGPT (Figure 1).

 doi:10.20944/preprints202512.0565.v1

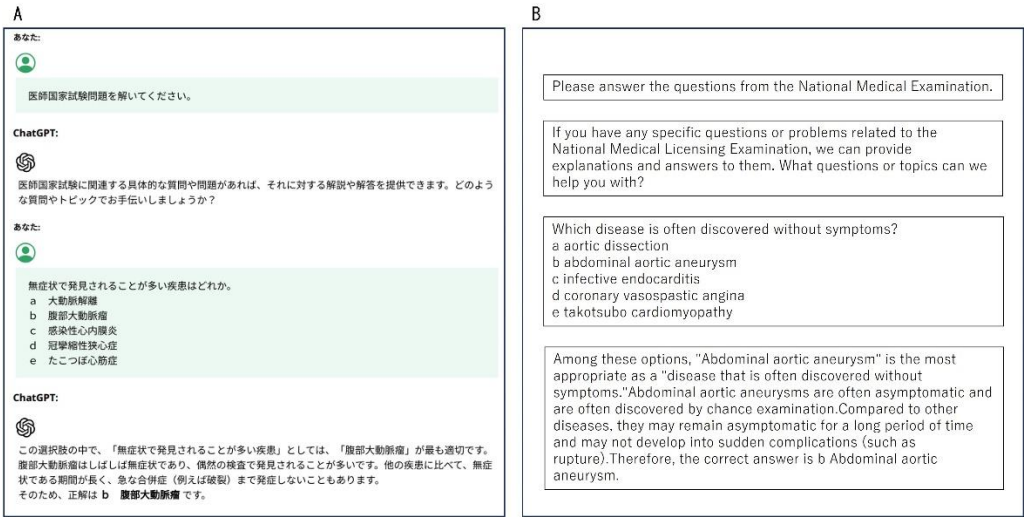

Figure 1. Example of a multiple-choice question, in which select one answer from five options, in Japanese with Chat-GPT response in a single chat box. A is an example of actually entering prompt in Japanese and getting the GPT answer. B is an English translation of the contents of A by the author.

For image-based questions, in the following prompt, we not only copied and pasted the text but also captured the image portion via a screenshot and attached it to obtain the answers (Figure 2).

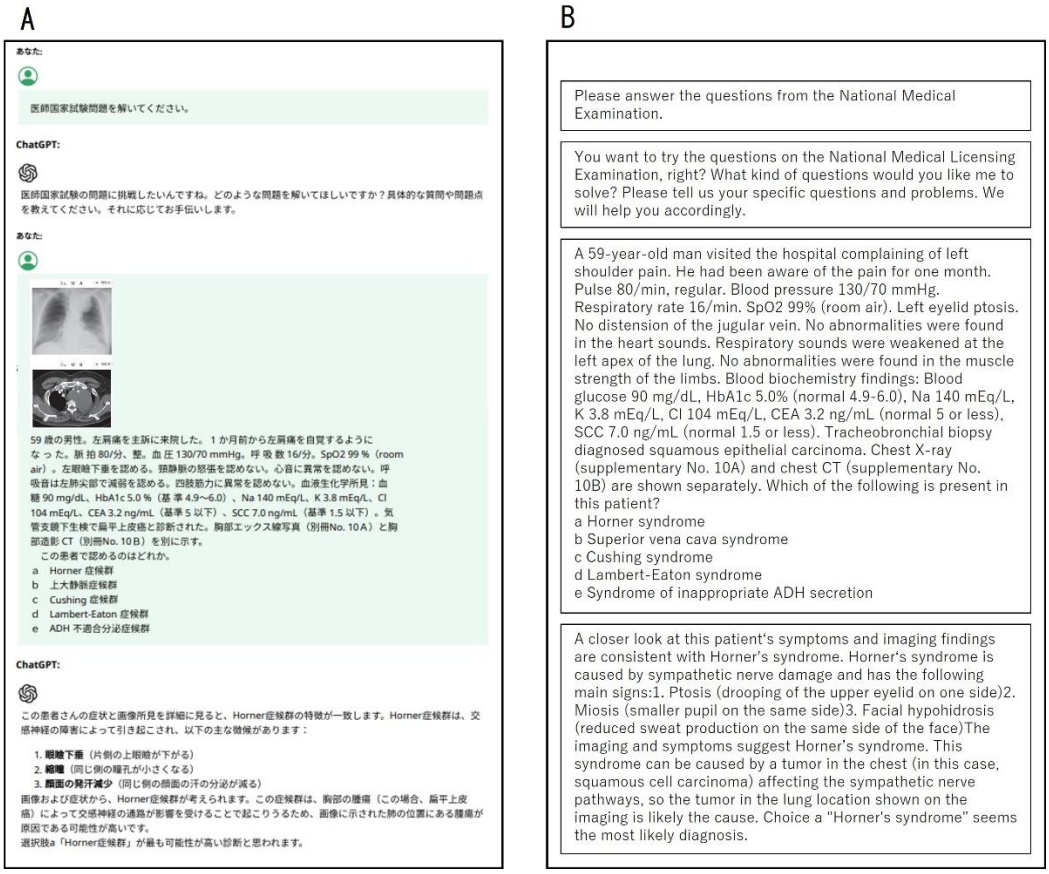

Figure 2. Example of prompt input and GPT response when there is an attached image. A is an example of actually entering prompt in Japanese and getting the GPT answer. B is an English translation of the contents of A by the author.

 doi:10.20944/preprints202512.0565.v1

Similarly, for questions involving the interpretation of tables, in the next prompt, we copied and pasted the text along with capturing the table portion via a screenshot and attached it to obtain the answers. For questions where the options were presented in a table format, we also captured the table via a screenshot and attached it to obtain the answers. In order to avoid hallucinations, we created a new Chat for each question and let ChatGPT solve the problem.

Between August and December 2024, responses to each question were obtained using ChatGPT-4 and ChatGPT-4o.

The JNMLE consists of 400 questions, with a total of 500 points, of which 100 questions are essential questions and 300 are non-essential questions.

The pass criteria for the 118th JNMLE of Japan are as follows:

Essential Questions: A score of at least 160 out of 200 points is required. This represents an 80% passing rate, which is an absolute standard. Regardless of other examinees' performance, this score is mandatory for passing. General questions will be scored as 1 point each, and clinical practical questions will be scored as 3 points each.

Non-essential questions: Each question is worth 1point, and a total score of at least 230 out of 300 points is required. This is a relative standard, meaning the passing rate can vary each year depending on the performance of all examinees. For this examination, a passing rate of 76.7% was required. Candidates receive one mark for each correct answer.

Prohibited Choices: Up to three prohibited choices are allowed. If an examinee selects two or three, they will automatically fail.

All these criteria must be met to pass the examination.

We used similar criteria to examine whether GPT-4 and GPT-4o could pass 118th JNMLE, as well as the rate of correct answers and trends in incorrect answers [11].

The 117th and 116th JNMLE also have passing criteria for three elements: compulsory questions, general questions, and prohibited questions [12,13].

Although the Japanese Ministry of Health, Labour and Welfare has not announced any information regarding prohibited choices, Medic Media Company limited (Tokyo, Japan) have published speculations, and this information was used in our investigation [14–16].

The 117th JNMLE contained two questions for which the images were not made public because they contained photographs of genitalia, and these two questions were excluded from this survey.

The passing criteria for each round are shown in the table (Table 1).

Table 1. The Passing Criteria for Each Round of Japan National Medical Licensing. Exami nation.

	118th JNMLE Passing Criteria	117th JNMLE Passing Criteria	116th JNMLE Passing Criteria
Essential Questions	Each general question is worth 1 point, and each clinical practical question is worth 3 points, total score is 160 points or more / 200 points.	Each general question is worth 1 point, and each clinical practical question is worth 3 points, total score is 160 points or more / 200 points.	Each general question is worth 1 point, and each clinical practical question is worth 3 points, total score is 158 points or more / 197 points. Excluded questions: B6, B43, E16 E16Candidates who answer correctly will be included in the marks, and candidates who answer incorrectly will not be included in the marks.
Non-essential General and Clinical Questions	Each question is worth 1 point, total score is 230 points or more / 300 points.	Each question is worth 1 point, total score is 220 points or more / 295 points. Excluded questions: C15, C60, D38, D53, F42	Each question is worth 1 point, total score is 214 points or more / 297 points. Excluded questions: A34、A71、C36、D64
Prohibited Choices (Critical Questions)	3 questions or less	2 questions or less	3 questions or less

The overall correct answer rate for each round, as well as the correct answer rates for required questions, general/clinical questions, and each text- and image-based question, were compared between GPT-4 and GPT-4o. The number of prohibited options selected was also compared. Since

 doi:10.20944/preprints202512.0565.v1

the number of image questions was small (approximately 100 questions), the correct answer rates were compared not only for each round but also for the total of the three rounds. In addition, we divided the test questions into general questions that ask simple knowledge and clinical questions that are in the form of case studies, and compared the overall correct answer rate and each section.

The correct answer rates were statistically analyzed using the chi-square test, and a P value of less than 0.05 was considered statistically significant.

The JNMLE and Chat GPT used in this study are publicly accessible online. Therefore, ethical approval was not required.

3. Results

3.1. Performance in the 118th JNMLE

3.1.1. Overall Results

The 118th JNMLE included 101 image-based questions, consisting of 10 essential image questions and 91 non-essential image questions. GPT-4o passed the examination with a total score of 457 points, including 190 in essential questions (general 49, clinical 141), 267 in non-essential questions, and zero prohibited choices. GPT-4 also passed, scoring 425 points (essential 181, non-essential 244), but selected one prohibited option.

3.1.2. Accuracy Comparison

GPT-4o demonstrated a significantly higher overall accuracy rate than GPT-4. No significant difference was observed in image-based questions. GPT-4o showed significantly higher accuracy in text-based questions. In the essential section, GPT-4o showed superior accuracy for text-based items, while image-based accuracy was comparable between models. In both general and clinical domains, GPT-4o demonstrated significantly higher overall accuracy and higher text-based accuracy, with no significant difference in image-based accuracy.

3.2. Performance in the 117th JNMLE

3.2.1. Overall Results

The 117th examination included 127 image-based questions (16 essential and 111 non-essential). GPT-4o passed with 446 total points (essential 190, non-essential 256, prohibited choices 0). GPT-4 narrowly passed with 392 points (essential 161, non-essential 231), selecting two prohibited choices.

3.2.2. Accuracy Comparison

GPT-4o again demonstrated significantly superior overall performance.

GPT-4o achieved significantly higher accuracy in image-based questions, with no difference in text-based accuracy. In essential questions, GPT-4o surpassed GPT-4 overall and in text-based accuracy. In non-essential questions, GPT-4o excelled overall and in image-based accuracy. In general questions, no significant differences were observed. In clinical questions, GPT-4o demonstrated significantly higher accuracy overall and in image-based questions.

3.3. Performance in the 116th JNMLE

3.3.1. Overall Results

The 116th JNMLE included 94 image-based questions (13 essential, 81 non-essential). GPT-4o passed with 190 essential points (general 46, clinical 144), 272 non-essential points, and zero prohibited choices. GPT-4 passed with 173 essential points, 247 non-essential points, and two prohibited choices.

 doi:10.20944/preprints202512.0565.v1

3.3.2. Accuracy Comparison

GPT-4o demonstrated significantly higher overall accuracy compared with GPT-4.

GPT-4o achieved significantly higher image-based accuracy, while text-based accuracy was similar between models. In essential questions, GPT-4o showed significantly higher overall and text-based accuracy. In non-essential questions, GPT-4o outperformed GPT-4 across overall, image-based, and text-based categories. In general questions, GPT-4o showed significantly higher overall and text-based accuracy. In clinical questions, GPT-4o demonstrated superior overall and image-based accuracy, with comparable text-based accuracy.

These findings are summarized in Table 2.

Table 2. Number of Questions, Number of Correct Answers, and Correct Answer Rate for Each Rounds and Question Types for Chat GPT-4o and GPT-4.

Question Section	Question Type	GPT-4o Number of Correct Answers	GPT-4o Number of Questions	GPT-4o Correct Answer Rate	GPT-4 Number of Correct Answers	GPT-4 Number of Questions	GPT-4 Correct Answer Rate	P Value
118 All questions	All Questions	363	400	90.8%	333	400	83.3%	0.0016
118 All questions	Picture-based Questions	85	101	84.2%	76	101	75.2%	0.1153
118 All questions	Text-based Questions	278	299	93.0%	257	299	86.0%	0.0050
118 Essential Section	All Questions	96	100	96.0%	89	100	89.0%	0.0602
118 Essential Section	Picture-based Questions	8	10	80.0%	6	10	60.0%	0.3291
118 Essential Section	Text-based Questions	88	90	97.8%	83	90	92.2%	0.0005
118 Non-essential Sections	All Questions	267	300	89.0%	244	300	81.3%	0.0084
118 Non-essential Sections	Picture-based Questions	77	91	84.6%	70	91	76.9%	0.1879
118 Non-essential Sections	Text-based Questions	190	209	90.9%	174	209	83.3%	0.0192
118 General Section	All Questions	133	150	88.7%	123	150	82.0%	0.1026
118 General Section	Picture-based Questions	5	10	50.0%	7	10	70.0%	0.3613
118 General Section	Text-based Questions	128	140	91.4%	116	140	82.9%	0.0321
118 Clinical Section	All Questions	230	250	92.0%	210	250	84.0%	0.0059

 doi:10.20944/preprints202512.0565.v1

118 Clinical Section	Picture-based Questions	80	91	87.9%	69	91	75.8%	0.0343
118 Clinical Section	Text-based Questions	150	159	94.3%	141	159	88.7%	0.0701
117 All questions	All Questions	350	393	89.1%	312	393	79.4%	0.0016
117 All questions	Picture-based Questions	107	127	84.3%	83	127	65.4%	0.0005
117 All questions	Text-based Questions	243	266	91.4%	229	266	86.1%	0.0550
117 Essential Section	All Questions	94	100	94.0%	81	100	81.0%	0.0054
117 Essential Section	Picture-based Questions	13	16	81.3%	10	16	62.5%	0.2381
117 Essential Section	Text-based Questions	81	84	96.4%	71	84	84.5%	0.0085
117 Non-essential Sections	All Questions	256	293	87.4%	231	293	78.8%	0.0058
117 Non-essential Sections	Picture-based Questions	94	111	84.7%	73	111	65.8%	0.0010
117 Non-essential Sections	Text-based Questions	162	182	89.0%	158	182	86.8%	0.5201
117 General Section	All Questions	128	148	86.6%	119	148	80.4%	0.0042
117 General Section	Picture-based Questions	9	14	64.3%	7	14	50.0%	0.4450
117 General Section	Text-based Questions	117	132	88.6%	110	132	83.3%	0.2145
117 Clinical Section	All Questions	222	246	90.2%	193	246	78.5%	0.0003
117 Clinical Section	Picture-based Questions	96	111	86.5%	74	111	66.7%	0.0004
117 Clinical Section	Text-based Questions	126	134	94.0%	119	134	88.8%	0.1268
116 All questions	All Questions	366	395	92.7%	330	394	83.8%	0.0016

 doi:10.20944/preprints202512.0565.v1

116 All questions	Picture-based Questions	84	94	89.4%	71	94	75.5%	0.0126
116 All questions	Text-based Questions	282	301	93.7%	259	300	86.3%	0.0027
116 Essential Section	All Questions	94	98	95.9%	83	97	85.6%	0.0130
116 Essential Section	Picture-based Questions	12	13	92.3%	11	13	85.6%	0.5393
116 Essential Section	Text-based Questions	82	85	96.5%	72	84	85.7%	0.0145
116 Non-essential Sections	All Questions	272	297	91.6%	247	297	83.2%	0.0020
116 Non-essential Sections	Picture-based Questions	72	81	88.9%	60	81	74.1%	0.0152
116 Non-essential Sections	Text-based Questions	200	216	92.6%	187	216	86.6%	0.0406
116 General Section	All Questions	139	147	94.6%	123	146	84.2%	0.0042
116 General Section	Picture-based Questions	4	7	57.1%	3	7	42.9%	0.5929
116 General Section	Text-based Questions	135	140	96.4%	120	139	86.3%	0.0027
116 Clinical Section	All Questions	227	248	91.5%	207	248	83.5%	0.0066
116 Clinical Section	Picture-based Questions	80	87	92.0%	68	87	78.2%	0.0107
116 Clinical Section	Text-based Questions	147	161	91.3%	139	161	86.3%	0.1571

3.4. Combined Analysis of All Three Examinations

3.4.1. Integrated Overall Accuracy

When combining all 1,200 questions across the three examinations, GPT-4o demonstrated significantly higher overall accuracy than GPT-4. GPT-4o also achieved superior performance in both image-based and text-based questions (Figure 3).

3.4.2. Section-Based Performance

Essential questions: GPT-4o showed significantly higher overall and text-based accuracy; image-based accuracy did not differ significantly. Non-essential questions: GPT-4o significantly outperformed GPT-4 overall, in image-based accuracy, and in text-based accuracy.

 doi:10.20944/preprints202512.0565.v1

3.4.3. Question-Type Performance

General questions: GPT-4o showed significantly higher overall and text-based accuracy; image-based accuracy was similar.

Clinical questions: GPT-4o demonstrated significantly higher accuracy overall and in both image-based and text-based categories.

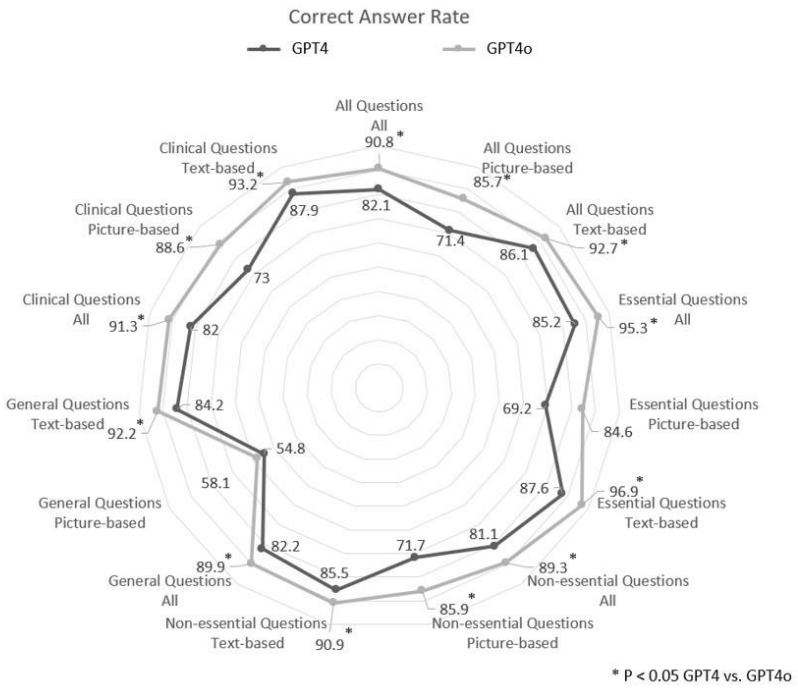

Figure 3. Radar chart of all answered questions and correct answer rate for each section for GPT-4 and GPT-4o.

3.4.4. Prohibited Choices

GPT-4 selected several prohibited choices across the three examinations, whereas GPT-4o selected none. This difference further supports the relative safety of GPT-4o's clinical decision-making tendencies (Table 3).

Table 3. Number of Prohibited Answers, Number of Prohibited Questions, and Prohibit Answer Rate for Each Round of Chat GPT-4o and GPT-4.

	GPT-4o Number of Prohibit Answers	GPT-4o Number of Questions	GPT-4o Prohibit Answer Rate	GPT-4 Number of Prohibit Answers	GPT-4 Number of Questions	GPT-4 Prohibit Answer Rate
118th JNMLE	0	9	0.0%	1	9	1.1%
117th JNMLE	0	11	0.0%	2	11	8.2%
116th JNMLE	0	9	0.0%	2	9	22.2%

4. Discussion

Both GPT-4o and GPT-4 passed each round of the JNMLE. GPT-4o not only outperformed GPT-4 in overall scores, but also performed well in both text-based and graphic-based questions. A comprehensive analysis of the three JNMLE showed that GPT-4o outperformed GPT-4 in almost all

 doi:10.20944/preprints202512.0565.v1

aspects, including general questions that ask basic medical knowledge, clinical questions that require text comprehension, and analysis of test values and images. In the mandatory image-based questions and general image-based questions, GPT-4o's correct answer rate was higher than GPT-4's, but the difference was not significant. This is likely due to the small number of questions, with 39 essential image-based questions and 31 general image-based questions in a total of three tests. In addition, GPT-4o did not choose prohibited options, unlike GPT-4. GPT-4o has been reported to improve the accuracy rate of image questions to the same level as that of text questions, but comparisons with previous models and consideration of prohibited options have not yet been reported. (17)

Apart from mental stress and time management, the skills required to pass the exam are considered to be as follows.

1. Depth and breadth of medical knowledge: A wide range of knowledge is required from basic medicine to clinical medicine.

2. Clinical reasoning ability: The ability to analyze cases and develop appropriate diagnoses and treatment plans.

3. Image reading ability: The ability to accurately read images such as X-rays, CT scans, and MRIs.

4. Communication ability: Effective communication with patients and medical teams.

5. Ethical judgment: The ability to make decisions based on medical ethics.

Based on these aspects, we considered the advantages of GPT-4o over GPT-4.

1. Depth and breadth of medical knowledge: GPT-4o showed a higher accuracy rate than GPT-4 throughout the exam, especially in text-based questions. This indicates that GPT-4o has access to deeper and wider medical information.

2. Clinical reasoning ability: GPT-4o also showed a high accuracy rate in clinical questions, indicating its superior ability to analyze and diagnose complex cases. This indicates its usefulness as a diagnostic support tool for doctors, .

3. Image interpretation skills: Although there was no significant difference between the models in image-based questions, the high overall accuracy rate of GPT-4o suggests that it can be a reliable partner in image diagnosis. However, there are also reports that there was no significant difference in the accuracy rate when GPT-4 was used with and without images attached, suggesting that image interpretation may not necessarily be necessary in the national medical examination (18).

4. Communication skills: GPT-4o had a high accuracy rate for text-based questions. Although Although GPT lacks genuine communication skills, it is suggested that it refers not only to medical information about the disease, but also to information about the patient and the medical team.

5. Ethical judgment: Ethical judgment cannot be considered in this study. It is impossible to evaluate whether the failure to choose an option that would threaten the patient's life is due to ethical judgment. In a study using GPT-4 to solve national examination questions, the accuracy rate was lower when images were attached than when they were not.

The accuracy rate of general questions based on images and required questions based on images was lower. According to Liu et al., GPT-4o performed better than GPT-4, as well as Gemini 1.5 Pro and Claude 3 Opus. This report suggests that academic publications may be one of the sources of training data for LLM, and it is speculated that LLM performance may be poor in areas with insufficient publications. The low accuracy rates of image-based general and essential questions may be due to the fact that these questions utilize basic textbook figures and conceptual graphs, limiting the effectiveness of reasoning or referencing statistics (7).

We also analyzed four prohibited answers selected by GPT-4. In three of the four questions, GPT recognized that an image was attached and which part and modality it was, but was unable to read the findings on the image. As a result, the image findings could not be taken into account, leading to incorrect choices that put the patient's life at risk. On the other hand, GPT-4o not only recognized which part and which modality the image was, but also pointed out whether there was an abnormality and what kind of abnormality it was, and was able to take the findings into account, thus avoiding prohibited choices and selecting correctly. In one of the five questions, by giving a lot

 doi:10.20944/preprints202512.0565.v1

of text information as the question text, a drug that would endanger the patient's life was selected as the treatment. GPT-4o also selected the correct answer for this question. From these results, it can be said that while the questions in the exam require test takers to have image diagnosis skills, GPT-4o has the minimum image diagnosis skills required in actual clinical practice to pass the exam without any special prompts or training. In addition, these results were obtained with simple prompts, without complex prompts or deep knowledge of AI, which is important. Thus, even non-experts can achieve good results. From these points, it is suggested that GPT-4o has superior performance compared to GPT-4 and can be useful as a partner to doctors in clinical medicine if there is basic medical knowledge to determine the authenticity of information obtained by AI.

Notably, both ChatGPT-4 and GPT-4o often provide explanations in a confident and coherent tone, even when their answers are incorrect. As a result, users may find it difficult to discern whether the response is factually accurate based solely on the language and reasoning provided. Therefore, physicians and medical students utilizing ChatGPT must possess sufficient domain-specific knowledge to critically evaluate the content generated by the model. ChatGPT should be regarded as a supportive tool in clinical practice and not as a substitute for medical expertise. Relying solely on AI-generated responses in unfamiliar clinical domains poses a potential risk and should be avoided.

5. Limitation

This study ChatGPT was asked to answer questions with the simplest possible prompts, so it is possible that ChatGPT's maximum functionality was not utilized. It is impossible to determine whether the literature or information on the web on which the AI based its answers is correct. It is also not possible to determine whether the reference papers are old and do not reflect new concepts. It is also not possible to determine whether the AI is being misled by the large amount of non-medical incorrect information from ordinary citizens on the web, which has overpowered the correct medical information. Since we cannot expect ethics from AI, human judgment is required to determine whether the answers it arrives at are correct or not. Therefore, although it can serve as a partner in supporting diagnosis, it is not advisable to accept AI's diagnostic results without doubting them.

The Ministry of Health, Labor and Welfare's web page publishes the pass criteria and pass rate, but not the average pass score, so direct comparison with test takers is not possible. However, based on the pass rates of 91.7% for the 116th test, 91.7% for the 117th test, and 92.4% for the 118th test, it can be inferred that ChatGPT's correct answer rate was at least in the top 90% of examinees of the exam.

This study the exam questions were entered in Japanese as prompts, and further study is required to determine whether the results can be applied to medical examination questions in multiple languages, such as English.

5. Conclusions

GPT-4o outperformed GPT-4 on the exam questions. It performed well not only on text-based questions but also on image-based questions. GPT-4o's high accuracy rate, extensive knowledge, advanced image analysis capabilities, and multifaceted considerations are likely to expand its potential as a diagnostic support and educational tool in the medical field for doctors with correct basic medical knowledge.

Author Contributions: Conceptualization, Masatoshi Miyamura and Goro Fujiki; methodology, Masatoshi Miyamura; software, Masatoshi Miyamura; validation, Masatoshi Miyamura; formal analysis, Masatoshi Miyamura; investigation, Masatoshi Miyamura, Hironaka Asano, Khosuke Tsuda, Yumiko Kanzaki, Hideaki Morita and Masaaki Hoshiga; resources, Masatoshi Miyamura; data curation, Masatoshi Miyamura; writing—original draft preparation, Masatoshi Miyamura; writing—review and editing, Masatoshi Miyamura, Hideaki Morita and Masaaki Hoshiga; visualization, Masatoshi Miyamura; supervision, Masaaki Hoshiga, Hideaki

 doi:10.20944/preprints202512.0565.v1

Morita; project administration, Masaaki Hoshiga, Hideaki Morita; funding acquisition, Masatoshi Miyamura. All authors have read and agreed to the published version of the manuscript.

Funding: None.

Informed Consent Statement: None.

Data Availability Statement: The questions and answers for the Japanese National Medical Licensing Examination used in this study can be viewed and downloaded without restriction from the Ministry of Health, Labour and Welfare's website (Japanese only). https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp240424-01.html [accessed 2025 Oct 21] https://perma.cc/F9DY-KCKF https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp230502-01.html [accessed 2025 Oct 21] https://perma.cc/2SSR-2WMC https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp220421-01.html [accessed 2025 Oct 21] https://perma.cc/6ZH9-MBNC https://www.mhlw.go.jp/general/sikaku/successlist/2024/siken01/about.html [accessed 2025 Oct 21] https://perma.cc/S286-GW5U https://www.mhlw.go.jp/general/sikaku/successlist/2023/siken01/about.html [accessed 2025 Oct 21] https://perma.cc/FDQ2-62TL https://www.mhlw.go.jp/general/sikaku/successlist/2022/siken01/about.html [accessed 2025 Oct 21] https://perma.cc/98MC-JUJT.

Acknowledgments: None.

Conflicts of Interest: None.

Abbreviations

The following abbreviations are used in this manuscript:

JNMLE Japan National Medical Licensing Examination

References

1. Introducing ChatGPT. Open AI. URL: https//openai.com/blog/chatgpt [accessed 2025 Oct 21]. https://perma.cc/MF36-PUJM
2. Tanaka Y, Nakata T, Aiga K, et al. Performance of Generative Pretrained Transformer on the National Medical Licensing Examination in Japan. PLOS Digit Health. Jan 2024; 3(1): e0000433
3. Takagi S, Watari T, Erabi A, et al. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. JMIR Med Educ. 2023; 9: e48002.
4. Yanagita Y, Yokokawa D, Uchida S, et al. Accuracy of ChatGPT on medical questions in the national medical Licensing examination in Japan: evaluation study. JMIR Form Res. 2023; 7: e48023.
5. Murad IA, Khaleel MI, Shakor MY. Unveiling GPT-4o: Enhanced Multimodal Capabilities and Comparative Insights with ChatGPT-4. International Journal of Electronics and Communications System Volume 4, Issue 2, 127-136. ISSN: 2798-2610 http://ejournal.radenintan.ac.id/index.php/IJECS/index DOI: 10.24042/ijecs.v4i2.25079
6. Nakao T, Miki S, Nakamura Y, et al. Capability of GPT-4V(ision) in Japanese National Medical Licensing Examination. JMIR Med Educ. 2024 Mar 12: 10: e54393. doi: 10.2196/54393.
7. Liu M, Okuhara T, Dai Z, et al. Evaluating the Effectiveness of advanced large language models in medical Knowledge: A Comparative study using Japanese national medical examination. Int J Med Inform. 2025 Jan: 193: 105673. doi: 10.1016/j.ijmedinf.2024.105673. Epub 2024 Oct 28.
8. https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp240424-01.html [accessed 2025 Oct 21] https://perma.cc/F9DY-KCKF
9. https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp230502-01.html [accessed 2025 Oct 21] https://perma.cc/2SSR-2WMC

 doi:10.20944/preprints202512.0565.v1

10. https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp220421-01.html [accessed 2025 Oct 21] https://perma.cc/6ZH9-MBNC
11. https://www.mhlw.go.jp/general/sikaku/successlist/2024/siken01/about.html [accessed 2025 Oct 21] https://perma.cc/S286-GW5U
12. https://www.mhlw.go.jp/general/sikaku/successlist/2023/siken01/about.html [accessed 2025 Oct 21] https://perma.cc/FDQ2-62TL
13. https://www.mhlw.go.jp/general/sikaku/successlist/2022/siken01/about.html [accessed 2025 Oct 21] https://perma.cc/98MC-JUJT
14. https://informa.medilink-study.com/web-informa/post41529.html/ [accessed 2025 Oct 21] https://perma.cc/RQ8V-4RP5
15. https://informa.medilink-study.com/web-informa/post39343.html/ [accessed 2025 Oct 21] https://perma.cc/NPW4-HDFT
16. https://informa.medilink-study.com/web-informa/post36171.html/ [accessed 2025 Oct 21] https://perma.cc/UF3L-8PSE
17. Miyazaki Y, Hata M, Omori H, et al. Performance of ChatGPT-4o on the Japanese Medical Licensing Examination: Evalution of Accuracy in Text-Only and Image-Based Questions. JMIR Med Educ. 2024 Dec 24: 10: e63129. doi: 10.2196/63129.
18. Nakao Y, Miki S, Nakamura Y, et al. Capability of GPT-4V(ision) in the Japanese National Medical Licensing Examination: Evaluation Study. JMIR Med Educ 2024; 10: e54393