



Assessing accuracy and legitimacy of multimodal large language models on Japan Diagnostic Radiology Board Examination

Yuichiro Hirano¹ · Soichiro Miki² · Yosuke Yamagishi¹ · Shouhei Hanaoka² · Takahiro Nakao² · Tomohiro Kikuchi^{2,3} · Yuta Nakamura² · Yukihiko Nomura^{2,4} · Takeharu Yoshikawa² · Osamu Abe¹

Received: 12 June 2025 / Accepted: 14 August 2025 / Published online: 12 September 2025
© The Author(s) 2025

Abstract

Purpose To assess and compare the accuracy and legitimacy of multimodal large language models (LLMs) on the Japan Diagnostic Radiology Board Examination (JDRBE).

Materials and methods The dataset comprised questions from JDRBE 2021, 2023, and 2024, with ground-truth answers established through consensus among multiple board-certified diagnostic radiologists. Questions without associated images and those lacking unanimous agreement on answers were excluded. Eight LLMs were evaluated: GPT-4 Turbo, GPT-4o, GPT-4.5, GPT-4.1, o3, o4-mini, Claude 3.7 Sonnet, and Gemini 2.5 Pro. Each model was evaluated under two conditions: with inputting images (vision) and without (text-only). Performance differences between the conditions were assessed using McNemar's exact test. Two diagnostic radiologists (with 2 and 18 years of experience) independently rated the legitimacy of responses from four models (GPT-4 Turbo, Claude 3.7 Sonnet, o3, and Gemini 2.5 Pro) using a five-point Likert scale, blinded to model identity. Legitimacy scores were analyzed using Friedman's test, followed by pairwise Wilcoxon signed-rank tests with Holm correction.

Results The dataset included 233 questions. Under the vision condition, o3 achieved the highest accuracy at 72%, followed by o4-mini (70%) and Gemini 2.5 Pro (70%). Under the text-only condition, o3 topped the list with an accuracy of 67%. Addition of image input significantly improved the accuracy of two models (Gemini 2.5 Pro and GPT-4.5), but not the others. Both o3 and Gemini 2.5 Pro received significantly higher legitimacy scores than GPT-4 Turbo and Claude 3.7 Sonnet from both raters.

Conclusion Recent multimodal LLMs, particularly o3 and Gemini 2.5 Pro, have demonstrated remarkable progress on JDRBE questions, reflecting their rapid evolution in diagnostic radiology.

Secondary abstract Eight multimodal large language models were evaluated on the Japan Diagnostic Radiology Board Examination. OpenAI's o3 and Google DeepMind's Gemini 2.5 Pro achieved high accuracy rates (72% and 70%) and received good legitimacy scores from human raters, demonstrating steady progress.

Keywords Artificial Intelligence (AI) · Large Language Model (LLM) · ChatGPT · Multimodal · Vision-language Model (VLM) · Reasoning Model

✉ Yuichiro Hirano
yhirano-ty@umin.ac.jp

- ¹ Department of Radiology, the University of Tokyo Hospital, 7-3-1 Hongo, Bunkyo-Ku, Tokyo, Japan
- ² Department of Computational Diagnostic Radiology and Preventive Medicine, The University of Tokyo Hospital, 7-3-1 Hongo, Bunkyo-Ku, Tokyo, Japan
- ³ Department of Radiology, School of Medicine, Jichi Medical University, 3311-1 Yakushiji, Shimotsuke, Tochigi, Japan
- ⁴ Center for Frontier Medical Engineering, Chiba University, 1-33 Yayoicho, Inage-Ku, Chiba, Japan

Introduction

Large language models (LLMs) possess rich medical knowledge. Studies have shown that recent LLMs can pass medical licensing examinations in multiple countries with scores far exceeding those of average human examinees [1–3]. LLMs have acquired multimodal capabilities, accepting images, video, and audio as input. Researchers have investigated the vision capabilities of multimodal LLMs, also known as vision-language models or VLMs, in various medical domains.

So far, the image interpretation capabilities of publicly available multimodal LLMs have not proven very impressive in the field of diagnostic radiology. Our previous research showed that GPT-4 Turbo with Vision, one of the earliest multimodal LLMs, did not outperform its text-only counterpart on the Japan Diagnostic Radiology Board Examination (JDRBE) [4]. Furthermore, due to the model's poor image interpretation capability, the addition of images resulted in significantly worse scores in subjective rating by radiologists.

Likewise, GPT-4o, Gemini Flash 1.5, Gemini Pro 1.5, and Claude 3 Opus failed to benefit from the addition of images [5, 6]. Similar findings have been reported for the Japanese Nuclear Medicine Board Examination [7]. Meanwhile, Kurokawa et al. found that diagnostic performance of Claude 3 Opus and Claude 3.5 Sonnet significantly improved when images were provided, using cases from *Radiology's* "Diagnosis Please" [8]. Others have investigated the performance of multimodal LLMs on Japanese [9], European [10], and Australian [11] board certification examinations for diagnostic radiology, although they did not directly compare performance with and without images.

For now, the strength of "multimodal" LLMs appears to lie largely in their robust text-based knowledge rather than their ability to directly interpret medical images. Even if their scores are high in an image-rich examination like JDRBE, it cannot be definitively concluded that they have exceeded the image interpretation abilities of human radiologists. Furthermore, the ability of multimodal LLMs to provide diagnostic reasoning based on image interpretation has been insufficiently evaluated and remains undemonstrated in prior studies.

In early 2025, multiple new multimodal LLMs were released by major vendors. Some of these are reasoning models designed to solve complex tasks by decomposing them into smaller steps and applying logical reasoning. In April, OpenAI released o3 and o4-mini, describing o3 as their most powerful reasoning model and o4-mini as a smaller, cost-efficient alternative. They also released non-reasoning models: GPT-4.5 in February and GPT-4.1 in April. Anthropic released Claude 3.7 Sonnet in February, and Google DeepMind introduced Gemini 2.5 Pro in May, both of which are reasoning models. To date, their performance in diagnostic radiology, especially in image interpretation, remains largely unexplored.

This study aimed to evaluate the performance of recent vision-enabled LLMs on the JDRBE, an exam requiring comprehensive expertise in diagnostic radiology. In addition to measuring accuracy, we also assessed the legitimacy of responses based on subjective ratings by radiologists.

Materials and methods

Study design

This retrospective study did not directly involve human subjects. All data are devoid of any information that can identify individuals, and are available online to members of Japan Radiological Society (JRS). Model inputs were submitted via application programming interfaces (APIs) of OpenAI, Anthropic, Google Cloud, or Azure AI Foundry, as detailed in later sections. Their privacy policies guarantee that data submitted via their APIs are securely handled and not used for model training. Therefore, Institutional Review Board approval was waived.

Question dataset

The questions included in our study were entirely derived from the JDRBE, which evaluates comprehensive knowledge of diagnostic radiology. Candidates must complete at least five years of training of radiology to be eligible for the JDRBE.

For the JDRBE 2021 and 2023, we used the same dataset as in our previous report [9]. We additionally prepared the dataset for the 2024 examination following the same method. Briefly, we downloaded the examination papers in the Portable Document Format (PDF) from the member-only section of the JRS website, and extracted text and images using Adobe Acrobat (Adobe, San Jose, CA), preserving the original image resolutions and file formats. For most questions, the extracted images were used as-is. For some questions with inappropriate extracted images (such as those with multiple images superimposed), we used screenshots captured from the PDF files in PNG format instead. The images were either in PNG or JPEG format, with heights ranging from 134 to 1,708 pixels (mean, 456) and widths from 143 to 1,255 pixels (mean, 482). Figure 1 shows an example of a question processed using this method. Questions from the 2022 examination were excluded because we failed to extract relevant data from the PDF file.

Questions without images were excluded; the remaining questions contained one to four images each. All questions had five answer choices. Approximately 90% were single-answer questions, while the remaining 10% were two-answer questions that required choosing both correct answers. The required number of choices was specified in each question statement.

Since there were no officially published answers to the examinations, ground-truth answers were determined

Fig. 1 Example of text and image extraction from a question. The main text and input images were extracted and provided to the model, while the question number (“18”) and image captions (“reconstructed sagittal slice” and “axial slice at the level of the first lumbar spine”) were omitted. In this example, the main text states, “A reconstructed sagittal slice and an axial slice at the level of the first lumbar spine are shown”

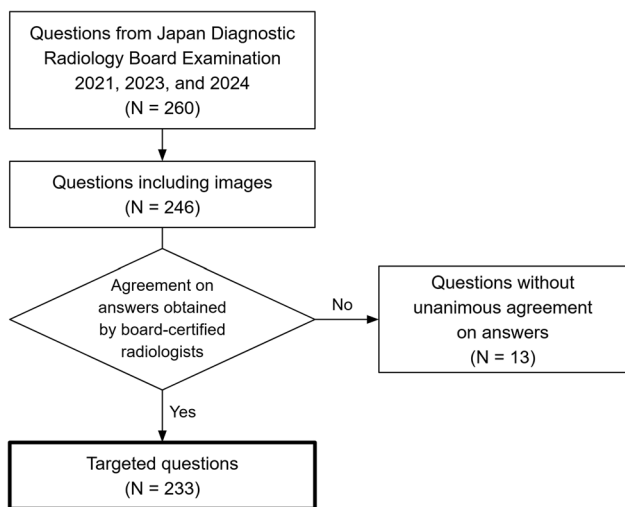
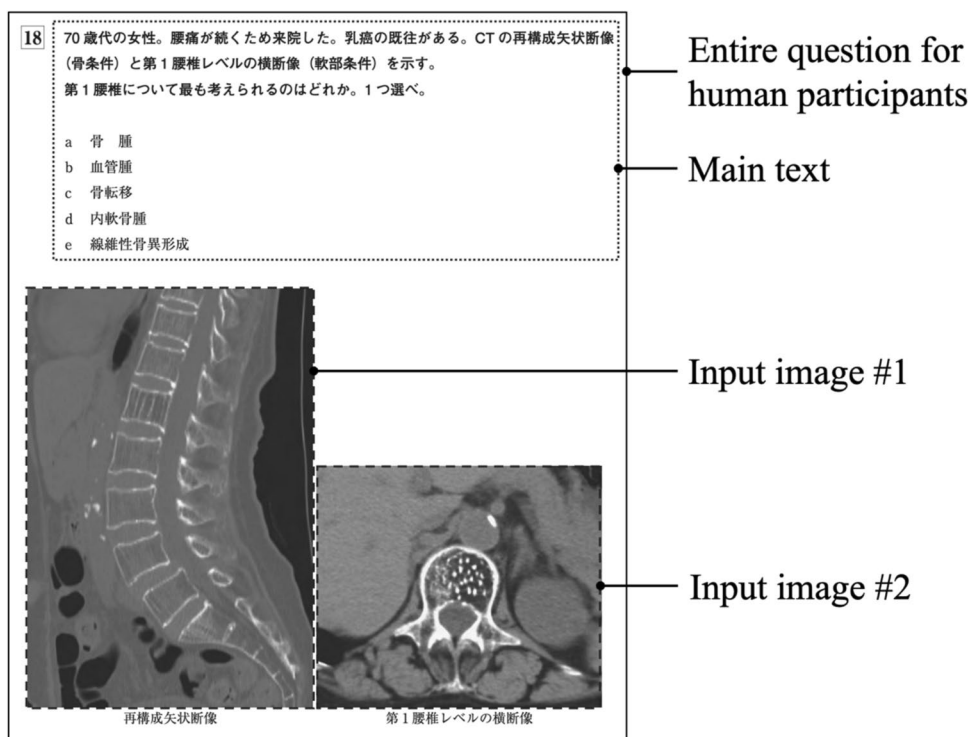


Fig. 2 Summary of questions included in this study

through consensus by three or more board-certified diagnostic radiologists. The answers for the 2024 examination were determined by S.M., S.H., and T.Y., with 18, 23, and 30 years of experience in diagnostic radiology, respectively. Questions without unanimous agreement on answers were excluded from the study. Figure 2 illustrates a flow chart detailing the inclusion and exclusion processes for questions.

Model evaluations

The following eight models were evaluated: GPT-4 Turbo, GPT-4o, GPT-4.5, GPT-4.1, o3, and o4-mini (all developed by OpenAI, San Francisco, CA); Claude 3.7 Sonnet (Anthropic, San Francisco, CA); and Gemini 2.5 Pro (Google DeepMind, London, UK). Details of the models are described in Table 1. Some of these models are reasoning models designed to solve complex tasks by employing logical reasoning [12–14]. GPT-4 Turbo and GPT-4o were chosen as baselines for comparison with our previous reports; the others are recent models released between February and April 2025. Since the GPT-4 Turbo model used in our previous study was a preview version and has since become unavailable, we used the closest available version in the same model family. All models were tested under two conditions: one using both text and image inputs (hereafter, “vision”) and one using only text inputs (“text-only”).

We used the official Anthropic API for Claude 3.7 Sonnet, the Google Cloud API for Gemini 2.5 Pro, and either the OpenAI API or the Azure AI Foundry (Microsoft, Redmond, WA) API for OpenAI models. For Claude 3.7 Sonnet, the required *max_tokens* parameter was set to 4,096; all other parameters for this model and all parameters for the other models, including the temperature parameter (if applicable), were left at their default values. All questions from the examinations were in Japanese, and the textual data were passed to the models without translation. We provided a system prompt similar to the one used in our previous

Table 1 Details of tested large language models

Model name	Developer	First release date	Version	Knowledge cutoff	Reasoning model
GPT-4 Turbo*	OpenAI	6 Nov 2023	<i>gpt-4-turbo-2024-04-09</i>	Dec 2023	No
GPT-4o	OpenAI	13 May 2024	<i>gpt-4o-2024-11-20</i>	Oct 2023	No
GPT-4.5	OpenAI	27 Feb 2025	<i>gpt-4.5-preview-2025-02-27</i>	Oct 2023	No
GPT-4.1	OpenAI	14 Apr 2025	<i>gpt-4.1-2025-04-14</i>	Jun 2024	No
o3*	OpenAI	16 Apr 2025	<i>o3-2025-04-16</i>	Jun 2024	Yes
o4-mini	OpenAI	16 Apr 2025	<i>o4-mini-2025-01-31</i>	Jun 2024	Yes
Claude 3.7 Sonnet*	Anthropic	24 Feb 2025	<i>claude-3-7-sonnet-20250219</i>	Nov 2024	Yes
Gemini 2.5 Pro*	Google DeepMind	28 Mar 2025	<i>gemini-2.5-pro-preview-03-25</i>	Jan 2025	Yes

* Included in legitimacy evaluation

Table 2 Prompts used in the experiments

	Prompt in Japanese	English translation
Vision	あなたはこれから放射線科診断専門医試験を受験する放射線科医です。以下に専門医試験の問題を提示しますので、解答をお願いします。また、解答に至った思考過程も簡潔に説明してください。自信がない場合でも必ず解答を選択して答えることが強制されます。最終行に"ANSWER: x"または"ANSWER: x, y"の形で選択した選択肢を明示してください。	You are a radiologist who is about to take the Japan Diagnostic Radiology Board Examination. Below we present a question for the examination and ask you to answer it. Please also briefly explain the thought process that led you to your answer. Even if you are not confident, you will always be forced to select and provide an answer. Please clearly indicate your answer in the format of "ANSWER: x" or "ANSWER: x, y" at the last line
Text-only	あなたはこれから放射線科診断専門医試験を受験する放射線科医です。以下に専門医試験の問題を提示しますので、解答をお願いします。また、解答に至った思考過程も簡潔に説明してください。ただし、問題文のみで、画像は与えられません。自信がない場合でも必ず解答を選択して答えることが強制されます。最終行に"ANSWER: x"または"ANSWER: x, y"の形で選択した選択肢を明示してください。	You are a radiologist who is about to take the Japan Diagnostic Radiology Board Examination. Below we present a question for the examination and ask you to answer it. Please also briefly explain the thought process that led you to your answer. Note that you will be given only the text of the questions, without any images. Even if you are not confident, you will always be forced to select and provide an answer. Please clearly indicate your answer in the format of "ANSWER: x" or "ANSWER: x, y" at the last line

report, as shown in Table 2. All experiments were conducted between April 18 and May 1, 2025.

To examine which image modalities led to greater performance gains with input images, we categorized the images into five modalities (CT, MRI, X-ray, nuclear medicine, and others) and counted the number of correct responses for each modality.

Legitimacy assessment

To assess the legitimacy of model responses, two diagnostic radiologists with different levels of experience (Y.Y., 2 years; S.M., 18 years, board-certified) independently

rated the responses to all 92 questions from the 2024 examination. For this evaluation, we included one representative model from each LLM vendor—Claude 3.7 Sonnet from Anthropic, Gemini 2.5 Pro from Google DeepMind, and the best-performing model from OpenAI. GPT-4 Turbo was also included as a baseline model. A five-point Likert scale (1 = very poor to 5 = excellent) was used to rate each of the 368 responses based on a comprehensive assessment of response quality, including image interpretation and explanation. The responses were presented in randomized order, and the raters were blinded to the model identities.

Statistical analysis

Differences in performance between the vision and text-only results were analyzed using McNemar's exact test. For the legitimacy scores, we first applied Friedman's test, followed by Wilcoxon's signed-rank test with Holm's correction for post-hoc pairwise comparisons. In addition, the quadratic weighted kappa was calculated to assess agreement between the two raters. Statistical significance was set at $P < 0.05$. All analyses were conducted using Python (version 3.12.6) with the scipy (version 1.15.2) and statsmodels (version 0.14.4) libraries.

Results

The dataset comprised 233 questions with 477 images (including 184 CT, 159 MRI, 15 X-ray, and 90 nuclear medicine images). Of these, 210 were single-answer and 23 were two-answer questions. Figure 3 provides an example question, along with summarized responses from the four models rated by the radiologists. The original full responses for this question (translated into English) are available in Online Resource 1.

Table 3 shows the number of correct answers for each model under each condition. The best-performing model was o3 under the vision condition, correctly answering 168 questions (72%). For GPT-4.5 and Gemini 2.5 Pro, significantly higher accuracy was observed in the vision condition compared to the text-only condition. For the other models, no significant difference in accuracy was observed between the two conditions.

Table 4 presents the number of correct responses from all models for each imaging modality. We observed a trend where the inclusion of input images improved the accuracy of questions involving CT or MRI, particularly for high-performing models such as o3, o4-mini, and Gemini 2.5 Pro.

Based on the quantitative results, we included the following four models in the legitimacy assessment: GPT-4 Turbo (baseline), Claude 3.7 Sonnet, o3, and Gemini 2.5 Pro. Figure 4 illustrates the distribution of the legitimacy scores for each rater. The quadratic weighted kappa between the two raters was 0.855, indicating a high level of agreement [15]. Gemini 2.5 Pro received the best scores from both raters (medians, 4 and 5), followed by o3 (4 and 4.5), Claude 3.7 Sonnet (3 and 3), and GPT 4 Turbo (2 and 2). Friedman's test indicated significant differences among the four models for both raters ($P < 0.001$). For Rater #1, all six pairwise comparisons showed significant differences after Holm correction: GPT-4 Turbo vs. all others ($P < 0.001$), Claude 3.7 Sonnet vs. o3 ($P = 0.027$), Claude 3.7 Sonnet vs. Gemini

2.5 Pro ($P < 0.001$), and o3 vs. Gemini 2.5 Pro ($P = 0.027$). For Rater #2, five of the six comparisons were significant: GPT-4 Turbo vs. all others ($P < 0.001$), Claude 3.7 Sonnet vs. o3 ($P = 0.006$), and Claude 3.7 Sonnet vs. Gemini 2.5 Pro ($P = 0.004$). The difference was not significant between o3 and Gemini 2.5 Pro ($P = 0.196$) for Rater #2.

Discussion

Since our last report in June 2024, LLMs have significantly improved their accuracy and legitimacy on JDRBE. Both GPT-4 Turbo and GPT-4o were outperformed by all the other models. Particularly, OpenAI's o3 and Google DeepMind's Gemini Pro 2.5 achieved a substantial leap in performance, although these models were released no more than 18 months after the debut of GPT-4 Turbo with Vision. To our knowledge, this is the first study that showed that the addition of images achieved a statistically significant accuracy improvement in the JDRBE.

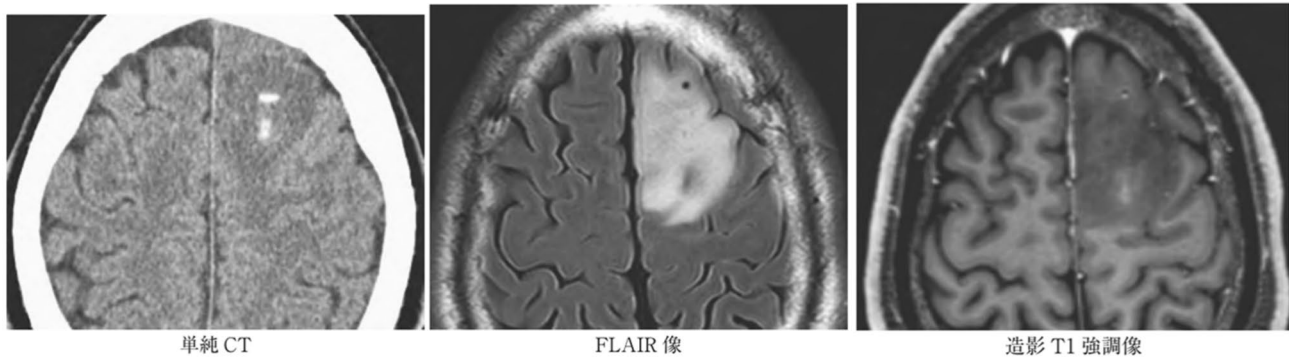
OpenAI's o3 was the best-performing model under both conditions, achieving an accuracy of 72% with image input. Notably, it correctly answered 67% of the questions even without image input, suggesting its strong medical knowledge and ability to reason the most likely answer based solely on the question text and the provided options. The addition of images did not significantly improve its accuracy, which may indicate that o3 heavily relies on its text-based reasoning capability when choosing an answer. OpenAI explains that reasoning models "think before they answer," generating a long internal chain of thought before responding to the user [12]. Our results showed that models with reasoning capabilities generally performed better, suggesting that the reasoning ability may be one of the contributing factors in the field of diagnostic radiology.

Gemini 2.5 Pro showed the greatest improvement in accuracy with the addition of images (from 59 to 70%). Furthermore, its legitimacy scores tended to be higher than those of o3, with a statistically significant difference from one of the raters. Although o3 was better in terms of accuracy, these findings suggest that Gemini 2.5 Pro can perform more reasonable medical image interpretation than o3. The sub-analysis revealed that Gemini 2.5 Pro showed performance gains across all image modalities with the addition of input images, with the most substantial improvement observed for MRI (from 49 to 67%). GPT-4.5 was another model that showed significantly better accuracy with image input.

Figure 3 highlights the ability of o3 and Gemini 2.5 Pro to accurately recognize and describe key image findings. GPT-4 Turbo incorrectly identified the lesion as being in the "right frontal lobe," reflecting the same left-right

- 4 30 歳代の男性。一過性発語障害を主訴に来院した。頭部単純 CT、頭部 MRI の FLAIR 像および造影 T1 強調像を示す。
可能性が最も高いのはどれか。1 つ選べ。

- a 膠芽腫
- b 血管芽腫
- c 転移性脳腫瘍
- d 乏突起神経膠腫
- e 中枢神経原発悪性リンパ腫



Summary of responses from four models

Model name	GPT-4 Turbo	Claude 3.7 Sonnet	o3	Gemini 2.5 Pro
Image interpretation				
Lesion location	Right frontal lobe <i>(incorrect)</i>	Left frontal lobe <i>(correct)</i>	Frontal lobe <i>(correct)</i>	Left frontal lobe <i>(correct)</i>
Contrast enhancement	Strong <i>(incorrect)</i>	Homogeneous <i>(incorrect)</i>	Minimal <i>(correct)</i>	Minimal <i>(correct)</i>
Detected calcification	No	No	Yes	Yes
Final answer	(e) PCNSL <i>(incorrect)</i>	(e) PCNSL <i>(incorrect)</i>	(d) Oligodendroglioma <i>(correct)</i>	(d) Oligodendroglioma <i>(correct)</i>
Legitimacy scores (Rater #1 / #2)	1 / 1	1 / 1	5 / 5	5 / 5

Fig. 3 Question 4 from the Japan Diagnostic Radiology Board Examination 2024, representing a clinical scenario of a man in his 30s presented with transient dysphasia. The question asks to identify the most probable diagnosis from the following options: **a** glioblastoma, **b** hemangioblastoma, **c** metastatic brain tumor, **d** oligodendroglioma,

and **e** primary central nervous system lymphoma (PCNSL). The correct answer is **d** oligodendroglioma. The figure also includes a summary of responses from four large language models, along with their legitimacy scores rated by diagnostic radiologists

confusion reported in our previous study [4]. It selected an incorrect option (malignant lymphoma) and received legitimacy scores of 1 (very poor) from both raters. Claude 3.7 Sonnet successfully identified the lesion location but

inaccurately described the lesion as having “homogeneous contrast enhancement,” selecting the same incorrect option (malignant lymphoma) and receiving scores of 1 from both raters. In contrast, both o3 and Gemini 2.5 Pro accurately

Table 3 Number of correct responses out of 233 questions, with (vision) and without (text-only) input images

Model name	Vision	Text-only	P-value (vision vs. text-only)
GPT-4 Turbo	97 (42%)	102 (44%)	0.583
GPT-4o	121 (52%)	113 (48%)	0.302
GPT-4.5	155 (67%)	138 (59%)	0.030*
GPT-4.1	141 (61%)	129 (55%)	0.134
o3	168 (72%)	156 (67%)	0.126
o4-mini	163 (70%)	151 (65%)	0.134
Claude 3.7 Sonnet	127 (55%)	132 (57%)	0.590
Gemini 2.5 Pro	162 (70%)	137 (59%)	0.001*

*P < 0.05

described the contrast enhancement as minimal and identified small calcification, leading to the correct diagnosis of oligodendroglioma. Both models received legitimacy scores of 5 (excellent) from both raters.

This study has several limitations. First, we did not assess the variability of responses across multiple API calls for the same question. Some reasoning models,

namely o3 and o4-mini, do not provide a configurable “temperature” parameter, making deterministic outputs unattainable. In addition, while it has been reported that different versions of the same model may exhibit substantial performance variation [16], we did not investigate such differences in this study. Second, although reasoning models tended to perform better, we could not conclusively determine whether this was truly owing to their reasoning ability or simply due to increased knowledge. Third, some of the examinations analyzed in this study took place before the knowledge cutoffs of certain LLMs. While access to the exam data is restricted to JRS members, we could not fully exclude the possibility of data leak. Fourth, we used only a single prompt throughout the experiments; alternative prompt strategies might have yielded better performance. Lastly, due to time constraints, we could not conduct legitimacy analysis for all eight models.

In conclusion, this study assessed the performance of multiple vision-enabled LLMs on the Japan Diagnostic Radiology Board Examination. Recent LLMs, particularly o3 and Gemini 2.5 Pro, demonstrated improved accuracy

Table 4 Number of correct responses from all models by modality. For each modality, we counted questions that included at least one image of the corresponding modality. Note that the total does not match the overall number of questions, as some involved multiple modalities

	CT	MRI	X-ray	Nuclear medicine	Others
No. of questions	107	79	13	51	23
GPT-4 Turbo					
Vision	41 (38%)	32 (41%)	10 (77%)	25 (49%)	9 (39%)
Text-only	46 (43%)	30 (38%)	8 (62%)	28 (55%)	11 (48%)
GPT-4o					
Vision	48 (45%)	39 (49%)	10 (77%)	32 (63%)	13 (57%)
Text-only	46 (43%)	36 (46%)	10 (77%)	30 (59%)	11 (48%)
GPT-4.5					
Vision	69 (64%)	45 (57%)	10 (77%)	33 (65%)	19 (83%)
Text-only	60 (56%)	42 (53%)	8 (62%)	38 (75%)	14 (61%)
GPT-4.1					
Vision	58 (54%)	49 (62%)	11 (85%)	32 (63%)	16 (70%)
Text-only	52 (49%)	44 (56%)	8 (62%)	32 (63%)	13 (57%)
o3					
Vision	73 (68%)	58 (73%)	11 (85%)	37 (73%)	17 (74%)
Text-only	67 (63%)	51 (65%)	10 (77%)	39 (76%)	15 (65%)
o4-mini					
Vision	65 (61%)	59 (75%)	10 (77%)	38 (75%)	15 (65%)
Text-only	64 (60%)	47 (59%)	11 (85%)	37 (73%)	15 (65%)
Claude 3.7 Sonnet					
Vision	48 (45%)	47 (59%)	8 (62%)	31 (61%)	11 (48%)
Text-only	56 (52%)	41 (52%)	9 (69%)	35 (69%)	14 (61%)
Gemini 2.5 Pro					
Vision	70 (65%)	53 (67%)	11 (85%)	37 (73%)	18 (78%)
Text-only	63 (59%)	39 (49%)	9 (69%)	34 (67%)	14 (61%)

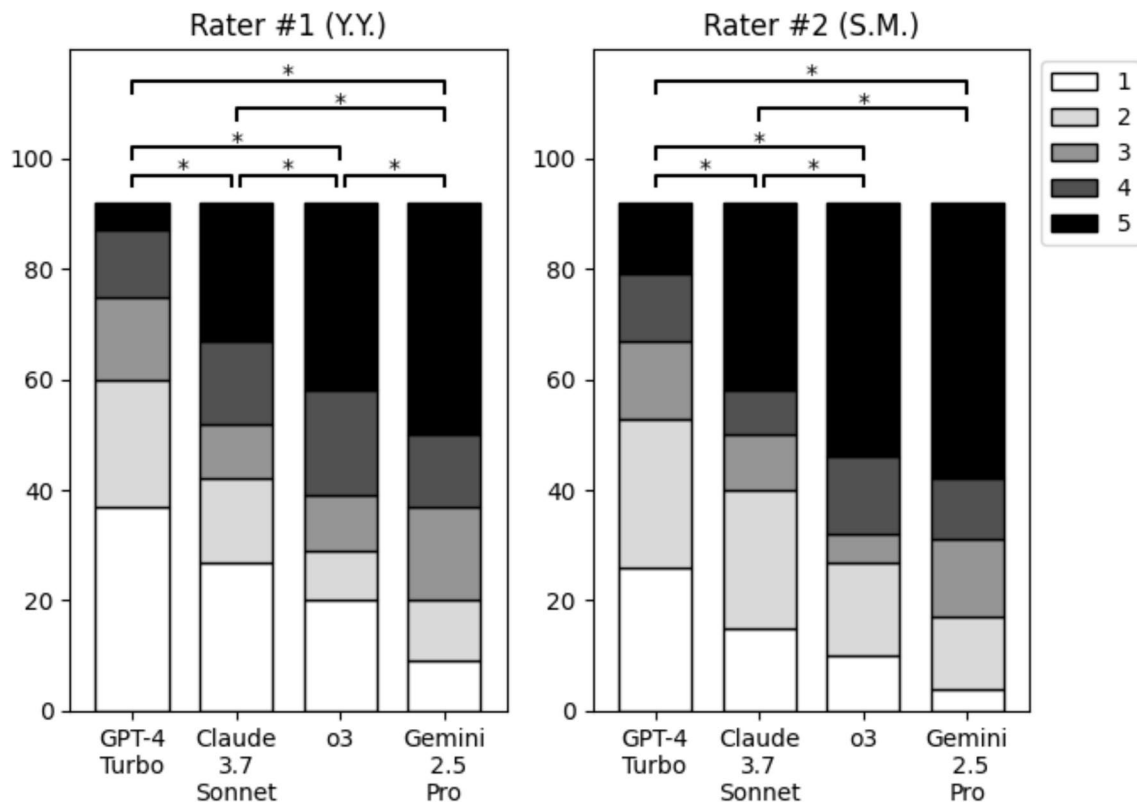


Fig. 4 Distribution of legitimacy scores from two raters. * $P < 0.05$ (after Holm correction)

and legitimacy, reflecting notable advancements in their abilities in diagnostic radiology.

Acknowledgements The Department of Computational Diagnostic Radiology and Preventive Medicine, The University of Tokyo Hospital, is sponsored by HIMEDIC Inc. and Siemens Healthcare K.K. We thank the Japan Radiological Society for granting permission to cite questions from the Japan Diagnostic Radiology Board Examination.

Funding Open Access funding provided by The University of Tokyo. No funding was received for conducting this study.

Declarations

Conflict of interest The authors declare no conflicts of interest.

Ethical approval This study did not directly involve human subjects. All data used in this study have been anonymized and are devoid of any information that could identify individuals, and these are available online to all JRS members. Therefore, approval from the Institutional Review Board was waived.

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/>.

References

- Chen Y, Huang X, Yang F, Lin H, Lin H, Zheng Z, et al. Performance of ChatGPT and Bard on the medical licensing examinations varies across different cultures: a comparison study. *BMC Med Educ.* 2024;24(1):1372.
- Miyazaki Y, Hata M, Omori H, Hirashima A, Nakagawa Y, Eto M, et al. Performance of ChatGPT-4o on the Japanese Medical Licensing Examination: Evaluation of accuracy in text-only and image-based questions. *JMIR Med Educ.* 2024;10(1): e63129.
- Kipp M. From GPT-3.5 to GPT-4.O: A leap in AI's medical exam performance. *Information (Basel).* 2024;15(9):543.
- Hirano Y, Hanaoka S, Nakao T, Miki S, Kikuchi T, Nakamura Y, et al. Gpt-4 turbo with vision fails to outperform text-only GPT-4 turbo in the Japan Diagnostic Radiology Board Examination. *Jpn J Radiol.* 2024;42(8):918–26.
- Hirano Y, Hanaoka S, Nakao T, Miki S, Kikuchi T, Nakamura Y, et al. No improvement found with GPT-4o: results of additional experiments in the Japan Diagnostic Radiology Board Examination. *Jpn J Radiol.* 2024;42(11):1352–3.
- Nakaura T, Yoshida N, Kobayashi N, Nagayama Y, Uetani H, Kidoh M, et al. Performance of multimodal large language models

- in Japanese Diagnostic Radiology Board Examinations (2021–2023). *Acad Radiol.* 2025;32(5):2394–401.
7. Watanabe T, Baba A, Fukuda T, Watanabe K, Woo J, Ojiri H. Role of visual information in multimodal large language model performance: an evaluation using the Japanese nuclear medicine board examination. *Ann Nucl Med.* 2025;39(2):217–24.
 8. Kurokawa R, Ohizumi Y, Kanzawa J, Kurokawa M, Sonoda Y, Nakamura Y, et al. Diagnostic performances of Claude 3 Opus and Claude 3.5 Sonnet from patient history and key images in Radiology’s “Diagnosis Please” cases. *Jpn J Radiol.* 2024;42(12):1399–402.
 9. Oura T, Tatekawa H, Horiuchi D, Matsushita S, Takita H, Atsukawa N, et al. Diagnostic accuracy of vision-language models on Japanese diagnostic radiology, nuclear medicine, and interventional radiology specialty board examinations. *Jpn J Radiol.* 2024;42(12):1392–8.
 10. Beşler MS, Oleaga L, Junquero V, Merino C. Evaluating GPT-4o’s performance in the official European Board of radiology exam: A comprehensive assessment. *Acad Radiol.* 2024;31(11):4365–71.
 11. Beşler MS. The accuracy of large language models in RAN-ZCR’s clinical radiology exam sample questions. *Jpn J Radiol.* 2024;42(9):1080.
 12. Reasoning models [Internet]. [cited 2025 May 23]. Available from: <https://platform.openai.com/docs/guides/reasoning?api-mode=responses>
 13. Gemini 2.5 Pro [Internet]. Google DeepMind. [cited 2025 May 23]. Available from: <https://deepmind.google/models/gemini/pro/>
 14. Claude 3.7 Sonnet and Claude Code [Internet]. [cited 2025 May 21]. Available from: <https://www.anthropic.com/news/claude-3-7-sonnet>
 15. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33(1):159–74.
 16. Li D, Gupta K, Bhaduri M, Sathiadoss P, Bhatnagar S, Chong J. Comparing GPT-3.5 and GPT-4 accuracy and drift in radiology diagnosis please cases. *Radiology.* 2024;310(1): e232411.

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