



GPT-4 Turbo with Vision fails to outperform text-only GPT-4 Turbo in the Japan Diagnostic Radiology Board Examination

Yuichiro Hirano^{1,5} · Shouhei Hanaoka⁵ · Takahiro Nakao² · Soichiro Miki² · Tomohiro Kikuchi^{2,3} · Yuta Nakamura² · Yukihiko Nomura^{2,4} · Takeharu Yoshikawa² · Osamu Abe⁵

Received: 27 December 2023 / Accepted: 16 March 2024
© The Author(s) 2024

Abstract

Purpose To assess the performance of GPT-4 Turbo with Vision (GPT-4TV), OpenAI's latest multimodal large language model, by comparing its ability to process both text and image inputs with that of the text-only GPT-4 Turbo (GPT-4 T) in the context of the Japan Diagnostic Radiology Board Examination (JDRBE).

Materials and methods The dataset comprised questions from JDRBE 2021 and 2023. A total of six board-certified diagnostic radiologists discussed the questions and provided ground-truth answers by consulting relevant literature as necessary. The following questions were excluded: those lacking associated images, those with no unanimous agreement on answers, and those including images rejected by the OpenAI application programming interface. The inputs for GPT-4TV included both text and images, whereas those for GPT-4 T were entirely text. Both models were deployed on the dataset, and their performance was compared using McNemar's exact test. The radiological credibility of the responses was assessed by two diagnostic radiologists through the assignment of legitimacy scores on a five-point Likert scale. These scores were subsequently used to compare model performance using Wilcoxon's signed-rank test.

Results The dataset comprised 139 questions. GPT-4TV correctly answered 62 questions (45%), whereas GPT-4 T correctly answered 57 questions (41%). A statistical analysis found no significant performance difference between the two models ($P=0.44$). The GPT-4TV responses received significantly lower legitimacy scores from both radiologists than the GPT-4 T responses.

Conclusion No significant enhancement in accuracy was observed when using GPT-4TV with image input compared with that of using text-only GPT-4 T for JDRBE questions.

Keywords Artificial intelligence (AI) · Large language model (LLM) · ChatGPT · GPT-4 Turbo · GPT-4 Turbo with Vision · Japan Diagnostic Radiology Board Examination (JDRBE)

Introduction

Recent advancements in large language models (LLMs) have marked a significant evolution in the field of artificial intelligence (AI). Among the numerous LLM-based applications, ChatGPT, which is based on the generative pre-trained transformer (GPT) architecture, has gained widespread recognition for its extensive capabilities [1–3]. Although not specifically designed for medical applications, ChatGPT possesses a substantial repository of medical knowledge, enabling it to handle healthcare-related queries. Kung et al. reported that ChatGPT, powered by the GPT-3.5 model, attained scores above or close to passing thresholds in the United States Medical Licensing Examination (USMLE) [4]. More recent studies have indicated that the latest GPT-4 model

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

¹ Department of Radiology, The International University of Health and Welfare Narita Hospital, 852 Hatakeda, Narita, Chiba, 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
⁵ Department of Radiology, the University of Tokyo Hospital, 7-3-1 Hongo, Bunkyo-Ku, Tokyo, Japan

successfully attained passing scores in medical licensing examinations in several countries including Japan, China, Poland, and Peru [5–10]. Several studies have evaluated the performance of ChatGPT in radiology. Bhayana et al. reported that GPT-3.5 nearly passed a radiology-board-style examination that resembles the Canadian Royal College and American Board of Radiology examinations [11]. Toyama et al. reported that the GPT-4 scored slightly above the provisional passing limit when applied to questions from the Japan Radiology Board Examination [12]. However, because ChatGPT was originally unable to accept image inputs, questions necessitating the interpretation of input images were not included in these studies.

The introduction of GPT-4 V(ision), an advanced iteration of GPT-4 featuring image processing capabilities, marks a significant leap beyond the initial text-only functionality of ChatGPT [3]. This version processes and interprets images in conjunction with textual data, broadening its applicability to fields that require image analysis. Yang et al. reported that the accuracy of GPT-4 improved from 83.6 to 90.7% when images were provided as input along with text for the USMLE [13]. Notably, the images utilized in the study were primarily non-radiological visuals, such as photographic images, pathological slides, electrocardiograms, and diagrams. Consequently, the diagnostic capabilities of GPT-4 V on radiological images, especially in challenging tasks, remain unexplored. Enhancing GPT-4 V to achieve a high diagnostic accuracy in interpreting radiological images may offer significant benefits to both diagnostic radiologists and physicians in clinical practice. This study was conducted to evaluate the diagnostic accuracy of GPT-4 Turbo with Vision (GPT-4TV)—the latest iteration of GPT-4 V—and compare it with that of its text-only counterpart, GPT-4 Turbo (GPT-4 T), in the Japan Diagnostic Radiology Board Examination (JDRBE) that assesses extensive expertise in diagnostic radiology. Our larger objective was to determine the impact of integrating visual data on the performance of AI models in diagnostic radiology.

Materials and methods

Study design

This retrospective 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 Japan Radiology Society (JRS) members. Furthermore, all data have been input through the OpenAI application programming interface (API), as explained in subsequent sections. OpenAI ensures that data submitted via the

API are encrypted, securely retained with strict access controls, deleted from the systems after 30 days, and not used for model training [14]. Therefore, approval from the Institutional Review Board was waived.

Questions dataset

All questions used in our experiments were sourced from the JDRBE, which assesses in-depth knowledge of diagnostic radiology. To be eligible for the JDRBE, candidates must initially acquire the Japan Radiology Specialist certification, which involves completing a minimum of a 3-year training program and passing the Japan Radiology Board Examination. Furthermore, an additional 2-year training period in diagnostic radiology is mandatory for JDRBE eligibility.

The examination papers are exclusively accessible to JRS members via the website. Each paper was originally provided in the Portable Document Format (PDF). To extract texts and images, we converted the PDF files into the eXtensive Markup Language (XML) format using Adobe Acrobat (Adobe, San Jose, California, US). All extracted images retained their original resolutions, and were in PNG or JPEG format. Heights ranged from 134 to 1708 pixels (mean: 447), and widths ranged from 143 to 950 pixels (mean: 474). For the extracted texts, we only used the main text from each question; any other texts, including the captions of input images, were discarded. Note that as the problem statement details each input image, an understanding of what each image represents is conveyed even in the absence of captions. Figure 1 exemplifies a question extracted in this way.

Our dataset consisted of questions from the JDRBE 2021 and 2023. Questions from JDRBE 2022 were not included because we failed to extract the input data due to inconsistency in the provided PDF file. Questions without any accompanying images were excluded, and each of the remaining questions was accompanied by one to four images. Each question had five possible choices, and approximately 90% of the questions were of the single-answer type, requiring the selection of one correct option out of five. The remaining 10% were two-answer questions, wherein participants had to choose two correct options from the five available, and a response was deemed correct only if both options were correct. The required number of correct options was specified in the text.

Because the answers were not officially published, six board-certified diagnostic radiologists (Y.N., T.K., T.N., S.M., S.H., and T.Y.; with 6, 7, 8, 16, 21, and 28 years of experience in diagnostic radiology, respectively) determined the ground-truth answers, consulting relevant literature as necessary. Three or more radiologists were assigned to each question, and consensus was reached through discussions.

Fig. 1 Example of text and image extraction from a question. The main text and input images were extracted and fed to the models. The question number (“23”) and image captions (“inhalation” and “exhalation”) were omitted from the input. In this question, the main text states that “axial CT images during inhalation and exhalation are shown”

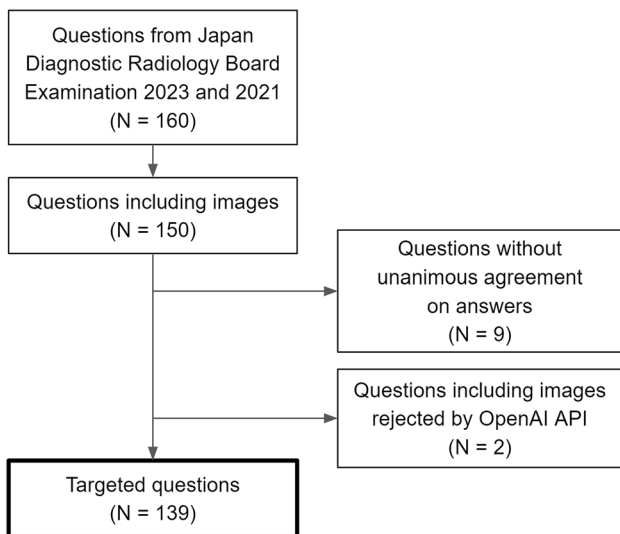
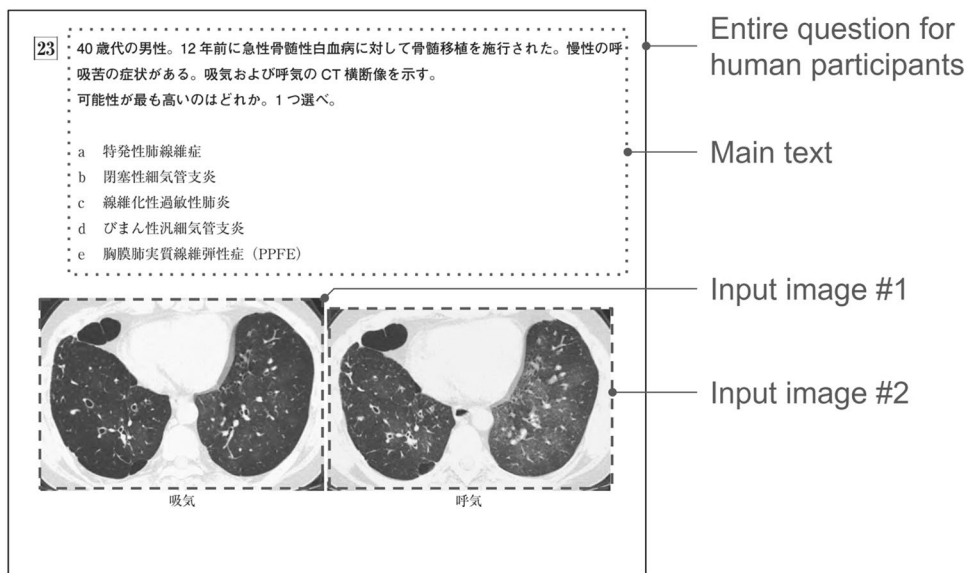


Fig. 2 Summary of questions included in this study

Questions without unanimous agreement on answers were excluded. All questions were classified into the following 11 subspecialties: breast, cardiovascular, gastrointestinal, genitourinary, head and neck, musculoskeletal, neuroradiology, pediatric, thoracic, interventional radiology, and nuclear medicine.

During the experiments, we further excluded two questions from the dataset due to the rejection of their corresponding images by the OpenAI API, which flagged them for potentially containing unsafe content. Figure 2 illustrates a flow chart detailing the inclusion and exclusion processes for questions.

Experimental details

We evaluated the performance of GPT-4TV and GPT-4 T on the prepared dataset. Because GPT-4TV is designed to accept visual input, it was passed both text and image data from the questions, whereas GPT-4 T solely received textual input without any accompanying images. Questions were submitted through the OpenAI API. The internal version of GPT-4TV was *gpt-4-vision-preview*, whereas that of GPT-4 T was *gpt-4-1106-preview*. Both models were released on November 2023 and trained on data dating up to April 2023 [15]. We set the *max_tokens* parameter of the API to the maximum of 4,096, and used the default settings for all other parameters. Because all questions from the examinations were in Japanese, the textual data were passed to the models without translation. We provided a shared prompt in Japanese along with the questions, as shown in Table 1. The prompt included the instruction, “Even if you are not confident, you will always be forced to select and provide an answer,” ensuring that the models answer as many questions as possible. In cases where the models initially refused to provide an answer, or if an error occurred during the experiment, we re-entered the same prompt and question until we successfully received an answer. All experiments were conducted between January 16–17, 2024.

To assess the radiological credibility of responses generated by GPT-4TV and GPT-4 T, two diagnostic radiologists, Y.H. and S.M., with 2 and 16 years of diagnostic radiology experience, respectively (S.M. is board-certified), independently assigned legitimacy scores using a five-point Likert scale (5 = Excellent, 4 = Good, 3 = Fair, 2 = Poor, and 1 = Very poor). The radiologists were blinded to each other’s assessments, and all responses were presented in a random order. Both radiologists were informed of the model

Table 1 Prompts used in the experiments

	Prompt in Japanese	English translation
GPT-4TV	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。	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
GPT-4 T	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。	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

(GPT-4TV or GPT-4 T) associated with each response. Legitimacy scores were rated subjectively based on how reasonable the response was according to the information provided to each model (i.e., for GPT-4 T, a response was considered excellent if it made a reasonable guess from what could be determined solely from the textual information). The quadratic weighted kappa coefficient [16] was calculated to measure the degree of mutual agreement between the two raters.

Statistical analysis

Differences in performance between GPT-4TV and GPT-4 T were analyzed using McNemar's exact test, with subgroup analyses conducted for single- and two-answer questions. Differences in legitimacy scores between GPT-4TV and GPT-4 T were analyzed using Wilcoxon's signed-rank test. Statistical significance was set at $P < 0.05$. All analyses were conducted using the Python software (version 3.11.4) along with SciPy (version 1.12.0) and statsmodels (version 0.14.1) libraries.

Results

The dataset encompassed 139 questions. Table 2 displays the frequency of modalities and planes across all questions in the dataset. Figure 3 illustrates an example question along with corresponding responses (translated into English by us) from GPT-4TV and GPT-4 T. Table 3 lists the performance metrics achieved by GPT-4TV and GPT-4 T for the dataset. GPT-4TV achieved a correct answer rate of 45% (62 out of 139 questions), whereas GPT-4 T achieved a correct answer

Table 2 Details of the questions used

	No. of questions	No. of images
Total	139	290
Average number of images		2.1
Modalities		
CT	65	115
Axial	62	95
Coronal	12	12
Sagittal	4	4
Other	3	4
MRI	51	96
Axial	33	54
Coronal	9	13
Sagittal	11	17
Other	8	12
Nuclear medicine	30	56
Other	20	23

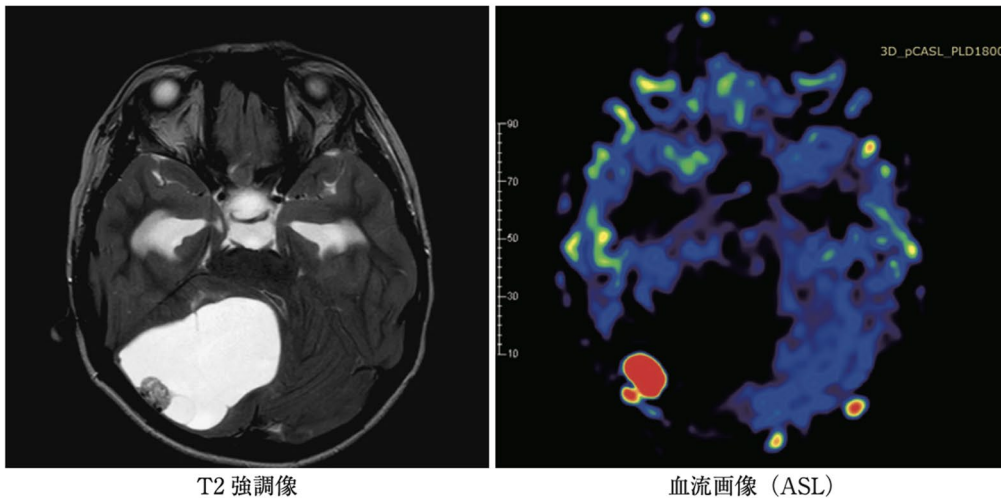
For modalities, each value of "No. of questions" represents the number of questions with at least one associated image of that modality (and plane, if specified)

rate of 41% (57 out of 139 questions). The difference in accuracy between the two conditions was not statistically significant ($P = 0.44$). The two models selected the same option(s) for 86 questions (62%). In the subgroup analyses, no significant difference in accuracy between GPT-4TV and GPT-4 T was observed for the single- and two-answer cohorts. Table 4 presents a contingency table describing the numbers of correct and incorrect answers from GPT-4TV and GPT-4 T.

A

36 12歳の女児。嘔気と頭痛を訴える。MRIのT2強調像とMRI血流画像（ASL：arterial spin labeling）を示す。
適切な診断はどれか。1つ選べ。

- a 結節性硬化症
- b von Hippel-Lindau 病
- c Sturge-Weber 症候群
- d neurofibromatosis type1 (NF-1)
- e neurofibromatosis type2 (NF-2)



B

GPT-4TV's response

The diagnosis is c) Sturge-Weber syndrome. Here's the reasoning behind this decision:

On the T2-weighted image, we observe abnormal hyperintensity in the left side of the brain, particularly prominent along the cortical surface. Additionally, arterial spin labeling (ASL) MRI reveals areas of increased blood flow on the left side. This lesion is caused by a cerebral hemangioma along the cortical surface, which is consistent with findings associated with Sturge-Weber syndrome.

These findings are not typically seen in the other options of tuberous sclerosis, von Hippel-Lindau disease, neurofibromatosis type 1, and neurofibromatosis type 2. Therefore, options a, b, d, and e are excluded.

Based on these findings, the diagnosis of Sturge-Weber syndrome is ultimately selected.

Fig. 3 a Question 36 from the Japan Diagnostic Radiology Board Examination 2021, representing a clinical scenario of a 12-year-old girl with complaints of nausea and headaches. The question asks to identify the most probable diagnosis from the following options: (a) tuberous sclerosis, (b) von Hippel-Lindau disease (vHL), (c) Sturge-Weber syndrome, (d) neurofibromatosis type 1, and (e) neurofibromatosis type 2. Axial MRI scans of T2-weighted image and arterial spin labeling image are included. The correct answer is (b) vHL. **b**

GPT-4TV's response to the question presented in Fig. 3a, translated into English. Text highlighted in red indicates inaccurate image interpretation. **c** GPT-4 T's response to the question presented in Fig. 3a, translated into English. Text highlighted in green indicates medically accurate descriptions of the provided options. Text highlighted in yellow indicates terminology that is not strictly accurate

C

GPT-4T's response
<p>The information directly inferred from the problem statement is limited, and without images, we cannot draw specific conclusions. However, let's examine each diagnostic option:</p> <p>a) Tuberous sclerosis - Typically characterized by heart tumors, skin lesions, and facial keratoses, as well as subependymal nodules and subependymal giant cell astrocytomas in the brain. Given that headaches and nausea can result from increased intracranial pressure, this option cannot be ruled out.</p> <p>b) von Hippel-Lindau disease - Characterized by hemangiomas in the pancreas, kidneys, spinal cord, and particularly in the cerebellum and retina. Although symptoms related to the central nervous system may suggest von Hippel-Lindau disease, onset at age 12 may be slightly early.</p> <p>c) Sturge-Weber syndrome - Characterized by skin lesions (unilateral facial hemangioma), ophthalmologic diseases (such as glaucoma), and cerebral vascular anomalies. Since headaches may be associated with cerebral vascular anomalies, this option cannot be excluded.</p> <p>d) Neurofibromatosis type 1 (NF-1) - Features include café-au-lait spots, neurofibromas of the skin and eyes, Lisch nodules, short stature, and learning disabilities. While symptoms related to the brain may include optic pathway gliomas, diagnosis based solely on symptoms of headaches and nausea is difficult.</p> <p>e) Neurofibromatosis type 2 (NF-2) - Mainly characterized by tumors in the central nervous system, particularly bilateral acoustic neuromas. However, diagnosis based solely on symptoms of headaches and nausea is difficult.</p> <p>Considering the options, confirmation cannot be obtained without image information. However, based on the characteristics of each condition and considering the age, it may be more reasonable to consider a) tuberous sclerosis or c) Sturge-Weber syndrome, as they are more commonly associated with symptoms at that age. Nonetheless, since these syndromes rely on skin symptoms and other physical features for diagnosis, confirmation cannot be made solely based on information about headaches and nausea.</p> <p>In situations where definitive images are not available during an exam and no additional information is provided, one must rely on experience and general knowledge to select the most likely answer. Given the information provided, a) tuberous sclerosis may theoretically be considered more likely, as it encompasses many phenotypes of the central nervous system and may be associated with headaches and nausea.</p> <p>In conclusion, the most appropriate diagnosis may be a) tuberous sclerosis, but without images, confirmation cannot be made. Clinical information and results of physical examinations, in addition to symptoms of nausea and headaches, will play an important role.</p>

Fig. 3 (continued)

Table 5 shows the distribution of legitimacy scores for GPT-4TV and GPT-4 T responses. The quadratic weighted kappa coefficient between the two raters was 0.517, indicating moderate agreement [17]. Both raters provided significantly lower legitimacy scores for GPT-4TV responses than those for GPT-4 T responses.

Discussion

In our investigation, we compared the performance of GPT-4TV with that of GPT-4 T on questions from the JDRBE. We found no statistically significant difference in accuracy between the two models (Table 3). Moreover, the two models selected the same option(s) for a substantial portion of the questions (62%). These results suggest that GPT-4TV primarily depends on linguistic cues for decision-making, with images playing a supplementary role.

As shown in Table 5, GPT-4 T received exceptionally high legitimacy scores (medians of 5 and 4) in subjective analysis, partly because it accurately recalled diseases (e.g. autosomal dominant polycystic kidney disease, multiple endocrine neoplasia, Birt-Hogg-Dubé syndrome) from patient characteristics and options, even without images. Despite the absence of image data, GPT-4 T often selected the most plausible option based on epidemiological knowledge, contributing to its high subjective scores. In contrast, GPT-4TV received significantly lower subjective scores, primarily stemming from numerous image interpretation errors, including ones that are considered basic by radiologists (e.g. mislabeling hyperintensity as hypointensity, misidentifying lesion locations including laterality). Because the evaluators were specialized radiologists, they were potentially more likely to note image interpretation errors, leading to a more critical evaluation.

Table 3 Performance comparison between GPT-4TV and GPT-4 T

	No. of questions	GPT-4TV	GPT-4 T	P-value
All questions	139	62 (45%)	57 (41%)	0.44
Single-answer questions	123	55 (45%)	50 (41%)	0.44
Two-answer questions	16	7 (44%)	7 (44%)	1.0
Subspecialty				
Breast	6	2 (33%)	3 (50%)	
Cardiovascular	9	2 (22%)	2 (22%)	
Gastrointestinal	22	8 (36%)	4 (18%)	
Genitourinary	11	3 (27%)	4 (36%)	
Head and Neck	5	3 (60%)	2 (40%)	
Musculoskeletal	11	4 (36%)	2 (18%)	
Neuroradiology	15	8 (53%)	7 (47%)	
Pediatric	6	2 (33%)	2 (33%)	
Thoracic	20	12 (60%)	11 (55%)	
Interventional radiology	4	3 (75%)	3 (75%)	
Nuclear medicine	30	15 (50%)	17 (57%)	

Table 4 Numbers of correct and incorrect responses from GPT-4TV and GPT-4 T

No. of questions		GPT-4TV		
		Correct	Incorrect	Total
GPT-4 T	Correct	46 (33%)	11 (8%)	57 (41%)
	Incorrect	16 (12%)	66 (47%)	82 (59%)
	Total	62 (45%)	77 (55%)	139 (100%)

Table 5 Distribution of legitimacy scores for responses generated by GPT-4TV and GPT-4 T

Score	Rater #1 (Y.H.)		Rater #2 (S.M.)	
	GPT-4TV	GPT-4 T	GPT-4TV	GPT-4 T
1 (Very poor)	32	0	4	1
2 (Poor)	24	3	37	11
3 (Fair)	38	6	29	29
4 (Good)	19	19	26	53
5 (Excellent)	26	111	43	45
Median	3	5	3	4
P value	<0.001		<0.001	

In Fig. 3, both GPT-4TV and GPT-4 T selected an incorrect option for the given question. GPT-4 T carefully assessed the likelihood of each option solely based on the clinical information provided in the problem statement and identified tuberous sclerosis as the most probable diagnosis. Though this conclusion was incorrect, it received legitimacy scores of 5 (Excellent) and 3 (Fair) from the

raters. Conversely, the image interpretation by GPT-4TV was highly inaccurate, resulting in a legitimacy score of 1 (Very poor) from both raters. This case underscores the limited proficiency of GPT-4TV in image interpretation and its negative impact on the subjective impressions of radiologists, despite both models having arrived at an incorrect conclusion.

Although GPT-4TV received significantly lower legitimacy scores with addition of input images, the final accuracy rate did not decrease. This may be because, as mentioned earlier, ChatGPT does not place much emphasis on image information. The nature of prioritizing linguistic information over image information does not align well with the responsibilities of radiologists. If this is a common characteristic of LLMs, it could pose a barrier when constructing general-purpose image diagnostic AI systems to assist radiologists.

The ability of multimodal GPT models to interpret medical images is an active research area. Yang et al. [13] reported a notable improvement in GPT-4's performance on USMLE questions with the addition of images to supplement text inputs. In contrast, our previous study [18] demonstrated that GPT-4 V did not significantly outperform the text-only GPT-4 in answering questions from the Japanese National Medical Licensing Examination. Our present findings align with this, as we observed no significant enhanced accuracy in GPT-4TV compared to GPT-4 T. This variability in performance improvement suggests that the efficacy of integrating both text and image inputs, as opposed to relying solely on text, may vary depending on the nature of the input question. Notably, the images used in previous studies largely represented non-radiological visuals including photographic images, pathological slides, electrocardiograms, and diagrams. In contrast, the present study predominantly focused on radiological images, particularly CT and MRI, as well as nuclear medicine imagery. Furthermore, the JDRBE targets board-certified radiologists with a minimum of five years of radiological experience, making it more challenging than the examinations utilized in prior studies. These differences in the data may account for the observed variations in accuracy improvement. Another key difference is the language of the input texts: Yang et al. used English, whereas our studies used Japanese. Although the input language has been noted to affect GPT model performance [3], the extent of this impact in our studies remains unclear. Future research should explore how input languages influence performance, perhaps by comparing the outcomes between Japanese texts provided as is and those translated into English.

This study has several limitations. First, the inherent generative nature of ChatGPT can result in different outputs for identical prompts and questions, which may have affected study outcomes. Notably, we recorded only a single

Table 6 Questions where (A) only GPT-4TV was correct and (B) only GPT-4 T was correct

No. of questions	16	No. of total input images	33
(A) Correct responses from GPT-4TV and incorrect responses from GPT-4 T			
Single-answer questions	16	CT	13
Two-answer questions	0	Axial	12
		Coronal	1
Subspecialty		Sagittal	0
Gastrointestinal	4	Other	0
Neuroradiology	3	MRI	13
Thoracic	3	Axial	8
Head and Neck	2	Coronal	2
Musculoskeletal	2	Sagittal	2
Nuclear medicine	2	Other	1
		Nuclear medicine	3
		Other	4
No. of questions	11	No. of total input images	28
(B) Incorrect responses from GPT-4TV and correct responses from GPT-4 T			
Single-answer questions	11	CT	10
Two-answer questions	0	Axial	10
		Coronal	0
Subspecialty		Sagittal	0
Nuclear medicine	4	Other	0
Neuroradiology	2	MRI	11
Thoracic	2	Axial	5
Breast	1	Coronal	2
Genitourinary	1	Sagittal	4
Head and Neck	1	Other	0
		Nuclear medicine	7
		Other	0

response from each model per question without examining the potential variability in responses. Ideally, a more extensive analysis must be conducted to investigate the extent of this variability; however, this aspect was not explored in our study. Second, the training data for GPT-4TV and GPT-4 T were dated up to April 2023, which was more recent than the online disclosure of examination papers. Although access to examination data is restricted to JRS members, some of the questions may have been included in ChatGPT's training dataset through various scenarios, such as if a JRS member had input the questions into ChatGPT's web interface or inadvertently published them online before April 2023. Third, our experiments employed only one prompt per model, potentially overlooking more effective prompts. Lastly, the raters were aware of the model (GPT-4TV or GPT-4 T) linked to each response, which may have introduced cognitive bias.

In conclusion, this study found no notable benefit in employing GPT-4TV with image inputs to respond to JDRBE questions compared with that of using GPT-4 T solely with text. The outcomes of this study underscore the

need for future research to explore more sophisticated methodologies for multimodal models, particularly in challenging domains such as those exemplified by the JDRBE.

Appendix

Questions where only one of the two models was correct

Table 6 lists detailed information pertaining to the 27 questions for which only one of the models was correct. Among these, four gastrointestinal questions received correct responses from GPT-4TV and incorrect responses from GPT-4 T, whereas no gastrointestinal questions received incorrect responses from GPT-4TV and correct responses from GPT-4 T. This led to an 18% discrepancy in accuracy on gastrointestinal questions between GPT-4TV and GPT-4 T, as shown in Table 3. It is possible that GPT-4TV is slightly more accurate in answering gastrointestinal questions. Apart from this, there was no observable trend

regarding the difference in modalities or subspecialties between the two groups.

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 Radiology Society for granting permission to cite questions from the Japan Diagnostic Radiology Board Examination. We thank Editage (www.editage.jp) for English language editing.

Funding Open Access funding provided by The University of Tokyo.

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

1. OpenAI. Introducing ChatGPT [Internet]. [cited 2023 Nov 14]. Available from: <https://openai.com/blog/chatgpt>
2. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language Models are Few-Shot Learners [Internet]. arXiv [cs.CL]. 2020. Available from: <http://arxiv.org/abs/2005.14165>
3. OpenAI. GPT-4 Technical Report [Internet]. arXiv [cs.CL]. 2023. Available from: <http://arxiv.org/abs/2303.08774>
4. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198.
5. 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. bioRxiv. 2023. <https://doi.org/10.1101/2023.04.17.23288603.abstract>.
6. 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.
7. 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;13(7):e48023.
8. Fang C, Ling J, Zhou J, Wang Y, Liu X, Jiang Y, et al. How does ChatGPT4 perform on Non-English National Medical Licensing Examination? An evaluation in Chinese Language. bioRxiv. 2023. <https://doi.org/10.1101/2023.05.03.23289443.abstract>.
9. Flores-Cohaila JA, García-Vicente A, Vizcarra-Jiménez SF, De la Cruz-Galán JP, Gutiérrez-Arratia JD, Quiroga Torres BG, et al. Performance of ChatGPT on the Peruvian national licensing medical examination: cross-sectional study. *JMIR Med Educ*. 2023;28(9):e48039.
10. Rosoł M, Gašior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Medical Final Examination. bioRxiv. 2023. <https://doi.org/10.1101/2023.06.04.23290939.abstract>.
11. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. *Radiology*. 2023;307(5):e230582.
12. Toyama Y, Harigai A, Abe M, Nagano M, Kawabata M, Seki Y, et al. Performance evaluation of ChatGPT, GPT-4, and Bard on the official board examination of the Japan Radiology Society. *Jpn J Radiol*. 2023. <https://doi.org/10.1007/s11604-023-01491-2>.
13. Yang Z, Yao Z, Tasmin M, Vashisht P, Jang WS, Ouyang F, et al. Performance of multimodal GPT-4V on USMLE with image: potential for imaging diagnostic support with explanations. medRxiv. 2023. <https://doi.org/10.1101/2023.10.26.23297629v3.abstract>.
14. Enterprise privacy at OpenAI [Internet]. [cited 2024 Jan 21]. Available from: <https://openai.com/enterprise-privacy>
15. Models - OpenAI API [Internet]. [cited 2024 Jan 21]. Available from: <https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo>
16. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ Psychol Meas*. 1973;33(3):613–9.
17. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–74.
18. Nakao T, Miki S, Nakamura Y, Kikuchi T, Nomura Y, Hanaoka S, et al. Capability of GPT-4V(ision) in Japanese national medical licensing examination. bioRxiv. 2023. <https://doi.org/10.1101/2023.11.07.23298133v1.abstract>.

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