大阪公立大学
Osaka Metropolitan University
学術情報リポジトリ

Diagnostic accuracy of vision-language models on Japanese diagnostic radiology, nuclear medicine, and interventional radiology specialty board examinations

メタデータ	言語: English 出版者: Springer Nature 公開日: 2025-01-09 キーワード (Ja): キーワード (En): Large language models, Vision-language models, Certifcation tests, Diagnostic radiology, Nuclear medicine, Interventional radiology 作成者: Oura, Tatsushi, Tatekawa, Hiroyuki, Horiuchi, Daisuke, Matsushita, Shu, Takita, Hirotaka, Atsukawa, Natsuko, Mitsuyama, Yasuhito, Yoshida, Atsushi, Murai, Kazuki, Tanaka, Rikako, Shimono, Taro, Yamamoto, Akira, Miki, Yukio, Ueda, Daiju メールアドレス: 所属:
URL	http://hdl.handle.net/10466/0002001515

This work is licensed under a Creative Commons
Attribution 4.0 International License.

Japanese Journal of Radiology (2024) 42:1392–1398
https://doi.org/10.1007/s11604-024-01633-0

ORIGINAL ARTICLE

Diagnostic accuracy of vision-language models on Japanese diagnostic radiology, nuclear medicine, and interventional radiology specialty board examinations

Tatsushi Oura[1] · Hiroyuki Tatekawa[1] · Daisuke Horiuchi[1] · Shu Matsushita[1] · Hirotaka Takita[1] · Natsuko Atsukawa[1] · Yasuhito Mitsuyama[1] · Atsushi Yoshida[2] · Kazuki Murai[1] · Rikako Tanaka[1] · Taro Shimono[1] · Akira Yamamoto[1] · Yukio Miki[1] · Daiju Ueda[1,3]

Received: 2 June 2024 / Accepted: 16 July 2024 / Published online: 20 July 2024
© The Author(s) 2024

Abstract

Purpose The performance of vision-language models (VLMs) with image interpretation capabilities, such as GPT-4 omni (GPT-4o), GPT-4 vision (GPT-4V), and Claude-3, has not been compared and remains unexplored in specialized radiological fields, including nuclear medicine and interventional radiology. This study aimed to evaluate and compare the diagnostic accuracy of various VLMs, including GPT-4 + GPT-4V, GPT-4o, Claude-3 Sonnet, and Claude-3 Opus, using Japanese diagnostic radiology, nuclear medicine, and interventional radiology (JDR, JNM, and JIR, respectively) board certification tests.
Materials and methods In total, 383 questions from the JDR test (358 images), 300 from the JNM test (92 images), and 322 from the JIR test (96 images) from 2019 to 2023 were consecutively collected. The accuracy rates of the GPT-4 + GPT-4V, GPT-4o, Claude-3 Sonnet, and Claude-3 Opus were calculated for all questions or questions with images. The accuracy rates of the VLMs were compared using McNemar's test.
Results GPT-4o demonstrated the highest accuracy rates across all evaluations with the JDR (all questions, 49%; questions with images, 48%), JNM (all questions, 64%; questions with images, 59%), and JIR tests (all questions, 43%; questions with images, 34%), followed by Claude-3 Opus with the JDR (all questions, 40%; questions with images, 38%), JNM (all questions, 42%; questions with images, 43%), and JIR tests (all questions, 40%; questions with images, 30%). For all questions, McNemar's test showed that GPT-4o significantly outperformed the other VLMs (all $P < 0.007$), except for Claude-3 Opus in the JIR test. For questions with images, GPT-4o outperformed the other VLMs in the JDR and JNM tests (all $P < 0.001$), except Claude-3 Opus in the JNM test.
Conclusion The GPT-4o had the highest success rates for questions with images and all questions from the JDR, JNM, and JIR board certification tests.

Keywords Large language models · Vision-language models · Certification tests · Diagnostic radiology · Nuclear medicine · Interventional radiology

Introduction

In recent years, the field of artificial intelligence (AI) has witnessed remarkable advancements, particularly in the development of large language models (LLMs) [1–4]. LLMs such as ChatGPT and Claude have demonstrated the ability to understand and generate human-like text across a wide range of domains, showing excellent performance in various medical fields [5, 6]. Several studies have investigated the performance of LLMs in the field of radiology [7–10]. These studies revealed that LLMs exhibit high diagnostic accuracy not only in general radiological knowledge but also

✉ Hiroyuki Tatekawa
htatekawa@omu.ac.jp

[1] Department of Diagnostic and Interventional Radiology, Graduate School of Medicine, Osaka Metropolitan University, 1-4-3, Asahi-machi, Abeno-ku, Osaka 545-8585, Japan

[2] Department of Nuclear Medicine, Graduate School of Medicine, Osaka Metropolitan University, Osaka, Japan

[3] Department of Artificial Intelligence, Graduate School of Medicine, Osaka Metropolitan University, Osaka, Japan

in specialized areas such as thoracic radiology, neuroradiology, and musculoskeletal radiology [11–14]. The recent emergence of LLMs with image interpretation capabilities such as GPT-4 with vision (GPT-4V) and Claude, which are often referred to as vision-language models (VLMs), has opened new possibilities for AI-assisted medical support. VLMs are designed to process and understand both visual and textual information, which enables them to analyze medical images and provide diagnostic insights. Among these models, the GPT-4 omni (GPT-4o), released by OpenAI in May 2024, has recently gained attention as a VLM that demonstrates high performance in multilingual support and image understanding.

Despite the growing interest in VLMs, only a few studies have evaluated their diagnostic accuracy in radiology quiz cases and specialty board examinations [15, 16]. In particular, regarding GPT-4o, no reports have assessed the diagnostic accuracy in the field of radiology or compared its diagnostic performance among different VLMs. Furthermore, the diagnostic accuracy of VLMs in radiology subspecialties, including nuclear medicine (NM) and interventional radiology (IR), remains unknown. Diagnosing in these specialized fields requires the interpretation of various images and modalities related to diseases, making it crucial to evaluate the performance of VLMs. As the application of VLMs continues to expand in healthcare, it is essential to investigate how well these advanced models can handle the complexities and nuances of NM and IR as well as diagnostic radiology (DR) fields.

This study aimed to evaluate various VLMs, including GPT-4V, GPT-4o, Claude-3 Sonnet, and Claude-3 Opus, and compare their diagnostic accuracy in the Japanese diagnostic radiology (JDR), Japanese nuclear medicine (JNM), and Japanese interventional radiology (JIR) board certification tests. By examining the diagnostic accuracy of these questions, the extent to which VLMs can be performed in highly specialized fields, such as advanced DR, NM, and IR, can be assessed. Understanding the areas in which VLMs are proficient and those in which they are less effective can be useful when considering their future use in supporting medical care with VLMs.

Materials and methods

The protocol of this study was approved by the institutional review board (IRB approval number: 2023–015). As this study only utilized the publicly available data, the requirement for informed consent was waived.

Data collection

The case vignettes were collected consecutively over a 5-year period as following sources: the 28th (August 23, 2019), 29th (August 28, 2020), 30th (August 20, 2021), 31st (August 26, 2022), and 32nd (August 25, 2023) JDR tests; the 16th (July 7, 2019), 17th (October 25, 2020), 18th (June 20, 2021), 19th (June 26, 2022), and 20th (July 2, 2023) JNM tests; and the 18th (November 17, 2019), 19th (November 8, 2020), 20th (November 14, 2021), 21st (November 13, 2022), and 22nd (November 12, 2023) JIR tests. The tests were downloaded from the official websites of the respective societies. Duplicate vignettes during the data collection period were excluded. A selection flowchart of the questions is shown in Fig. 1.

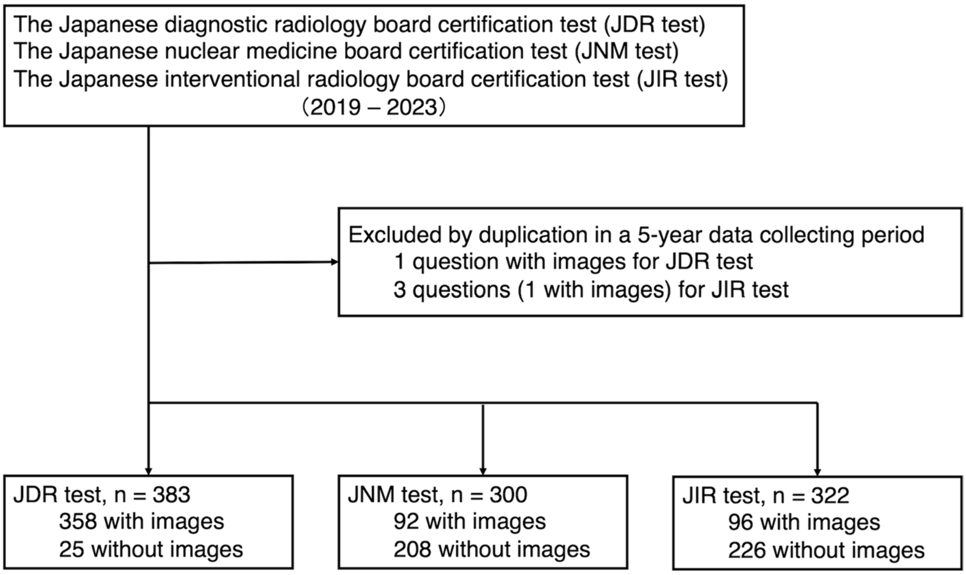

The Japanese diagnostic radiology board certification test (JDR test)
The Japanese nuclear medicine board certification test (JNM test)
The Japanese interventional radiology board certification test (JIR test)
（2019 – 2023）
Excluded by duplication in a 5-year data collecting period
1 question with images for JDR test
3 questions (1 with images) for JIR test
JDR test, n = 383
358 with images
25 without images
JNM test, n = 300
92 with images
208 without images
JIR test, n = 322
96 with images
226 without images

Fig. 1 Selection flowchart of the questions. *JDR* Japanese diagnostic radiology, *JNM* Japanese nuclear medicine, *JIR* Japanese interventional radiology

Data assessment

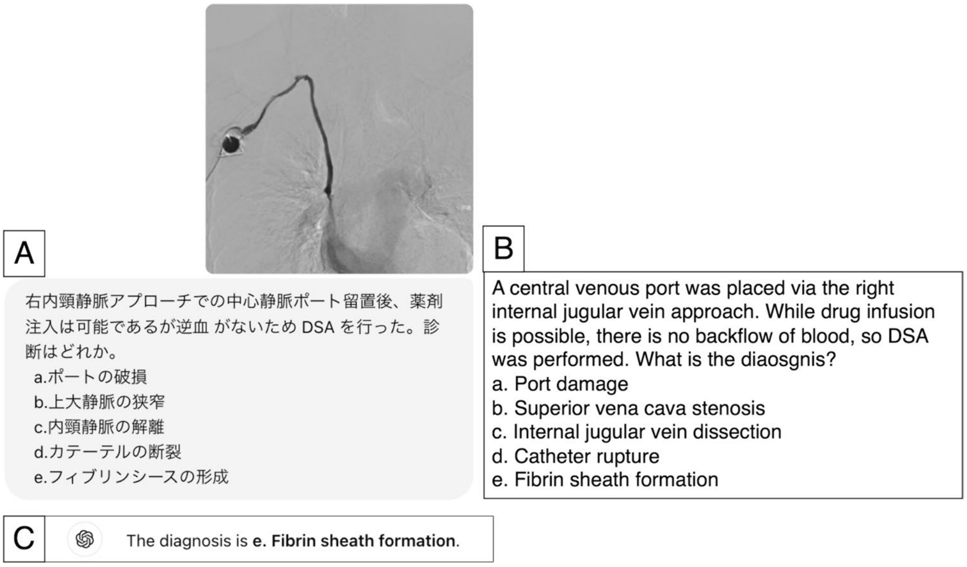
Question prompts, patient histories, and images (if available) from each case vignette were provided as inputs to the language models. In this study, the same prompt was used for each VLM. We initiated the input prompt for each case as follows: "As a highly experienced professor of radiology with 30 years of expertise, you assist in radiology cases. Your role is to analyze questions, patient histories, and imaging findings to determine correct answer(s)," in English in accordance with a previous study [12]. Subsequently, the text of the questions and options was input into the VLMs' input field in Japanese, as they were originally presented, to generate responses. In addition, when images were available, they were input simultaneously. An example of the input and output of the VLMs is shown in Fig. 2.

Responses from all VLMs were collected between March 31, 2024, and May 22, 2024. The data collection approach differed for each language model, where GPT-4-based ChatGPT (version gpt-4-0125-preview) provided answers only for questions without images, and GPT-4V-based ChatGPT (version gpt-4-vision-preview) provided answers only for questions with images. In contrast, GPT-4o-based ChatGPT (version gpt-4o-2024-05-13), Claude-3 Sonnet, and Claude-3 Opus (accessed on 15th Mar 2024, https://claude.ai/) provided answers to all questions, regardless of image presence. With regard to the characteristics of the collected questions, the JNM test followed a format in which one correct answer was chosen from five options. In contrast, the JRD and JIR tests included questions where one correct answer was chosen from five options as well as questions where multiple correct answers (mainly two) were chosen from five options. For questions with multiple correct answers, only responses that perfectly matched all the correct answers were considered accurate. For all language models, each question was answered three times with the system being reset between each attempt, and the most frequent answer was considered the final answer because of the limited repeatability and robustness of the current language models [13]. If the VLMs generated three different answers to a question, the answer was considered incorrect. Additionally, if the image of a question was determined to be in violation of the terms of service of the VLMs and no response was generated, the question was considered to be answered incorrectly.

As official answers were not available, two DR specialists (HT, a radiologist with 14 years of experience, and DH, a radiologist with 7 years of experience), two NM specialists (AY, a radiologist with 15 years of experience, and HT, a radiologist with 14 years of experience), and two IR specialists (HT, a radiologist with 14 years of experience, and KM, a radiologist with 9 years of experience) independently evaluated each case in their respective fields and provided their answers. If needed, they referred to textbooks and the literature to determine the correct answers. If the answers were in agreement, they were considered correct. In case of disagreement, a consensus was reached to determine the correct answer.

Statistical analyses

The accuracy rates of GPT-4 + GPT-4V, GPT-4o, Claude-3 Sonnet, and Claude-3 Opus were calculated for all questions, questions with images, questions without images, single-answer questions, and multi-answer questions. Diagnostic accuracy rates were compared among the language models using Cochran's Q test and post hoc McNemar's test [17]. P values < 0.05 were considered statistically significant.



右内頸静脈アプローチでの中心静脈ポート留置後、薬剤注入は可能であるが逆血 がないため DSA を行った。診断はどれか。

a.ポートの破損
b.上大静脈の狭窄
c.内頸静脈の解離
d.カテーテルの断裂
e.フィブリンシースの形成

A central venous port was placed via the right internal jugular vein approach. While drug infusion is possible, there is no backflow of blood, so DSA was performed. What is the diaosgnis?

a. Port damage
b. Superior vena cava stenosis
c. Internal jugular vein dissection
d. Catheter rupture
e. Fibrin sheath formation

The diagnosis is **e. Fibrin sheath formation**.

Fig. 2 An example of the input and output of the VLMs. Response samples of correct answers generated by GPT4o-based ChatGPT on a Japanese interventional radiology question. **A** The question and accompanying image were input manually. **B** The English version of the question. **C** The response generated by GPT4o-based ChatGPT. The question and figure are cited from the Japan Society of Interventional Radiology website (https://www.jsir.or.jp/docter_test/oldtest/), and permission was obtained from the Japanese Society of Interventional Radiology to use this question and figure

Table 1 Correct answer rates in the Japanese diagnostic radiology board certification test

	All questions	Question with images	Question without images	Single-answer question	Multi-answer question
No. of questions	383	358	25	323	60
GPT-4 + GPT-4 V	145 (38%)	128 (36%)	17 (68%)	125 (39%)	20 (33%)
GPT-4o	188 (49%)	172 (48%)	16 (64%)	169 (52%)	19 (32%)
Claude-3 Sonnet	120 (31%)	110 (31%)	10 (40%)	105 (33%)	15 (25%)
Claude-3 Opus	152 (40%)	135 (38%)	17 (68%)	126 (39%)	26 (43%)
Cochran's *Q*	40.37	37.31	8.16	45.34	5.9
P value	< 0.001*	< 0.001*	0.043*	0.041*	0.085

*Statistically significant

Table 2 Correct answer rates in the Japanese nuclear medicine board certification test

	All questions	Question with images	Question without images
No. of questions	300	92	208
GPT-4 + GPT-4 V	149 (50%)	35 (38%)	114 (55%)
GPT-4o	191 (64%)	54 (59%)	137 (66%)
Claude-3 Sonnet	83 (27%)	27 (29%)	56 (27%)
Claude-3 Opus	125 (42%)	40 (43%)	85 (41%)
Cochran's *Q*	87.85	18.68	75.71
P value	< 0.001*	< 0.001*	< 0.001*

*Statistically significant

Statistical analyses were performed using Python version 3.11.8 (Python Software Foundation, Wilmington, DE, USA).

Use of large language models

This manuscript was proofread with the assistance of ChatGPT (GPT-4o architecture; OpenAI, https://chat.openai.com/), and all outputs were confirmed by the authors.

Results

A total of 383 questions from the JDR test (358 with images and 60 with multiple answers), 300 questions from the JNM test (92 with images), and 322 questions from the JIR test (96 with images and 177 with multiple answers) were included. One question from the JDR tests and three questions from the JIR tests were excluded because they were duplicated in the 5-year data collecting period.

The number of correct answers and accuracy rates for the JDR, JNM, and JIR tests are presented in Tables 1, 2, and 3, respectively. The GPT-4o-based ChatGPT demonstrated the highest accuracy across all tests (JDR: all questions, 49%; questions with images, 48%; JNM: all questions, 64%; questions with images, 59%; JIR: all questions, 43%; questions with images, 34%). Claude-3 Opus showed mostly consistent performance across all evaluation categories (JDR test: all questions, 40%; questions with images, 38%; JNM: all questions, 42%; questions with images, 43%; JIR test: all questions, 40%; questions with images, 30%), ranking mostly second in accuracy rate after GPT-4o. All performances, except for Claude-3 Sonnet and Opus in the JNM test, exhibited higher accuracy rates for non-image-based questions than for image-based questions. No responses were

Table 3 Correct answer rates in the Japanese interventional radiology board certification test

	All questions	Question with images	Question without images	Single-answer question	Multi-answer question
No. of questions	322	96	226	145	177
GPT-4 + GPT-4 V	111 (34%)	29 (30%)	82 (36%)	58 (40%)	53 (30%)
GPT-4o	138 (43%)	33 (34%)	105 (46%)	78 (54%)	60 (34%)
Claude-3 Sonnet	98 (30%)	27 (28%)	71 (31%)	52 (36%)	46 (26%)
Claude-3 Opus	130 (40%)	29 (30%)	101 (45%)	66 (46%)	64 (36%)
Cochran's *Q*	17.29	0.88	18.97	14.57	6.04
P value	< 0.001*	0.83	< 0.001*	0.004	0.11

*Statistically significant

generated for one question in the JDR test and two questions in the JNM test for either GPT-4V or GPT-4o, while responses were generated from all questions for Claude-3 Sonnet and Claude-3 Opus. The accuracy rates by year for each test are presented in Supplemental Tables 1, 2, and 3.

The results of the pairwise McNemar's tests are shown in Table 4, where Cochran's Q test showed statistically significant differences. For all questions, McNemar's tests showed that GPT-4o-based ChatGPT significantly outperformed the other VLMs (all $P < 0.007$), except for Claude-3 Opus in the JIR test. For questions with images, the GPT-4o-based ChatGPT outperformed the other VLMs in the JDR and JNM tests (all $P < 0.001$), except for Claude-3 Opus in the NM test.

Discussion

This study evaluated the diagnostic accuracy of various VLMs, including GPT-4V, GPT-4o, Claude-3 Sonnet, and Claude-3 Opus, in the Japanese radiological board certification tests, including the JDR, JNM, and JIR tests. The results demonstrated that GPT-4o exhibited the highest accuracy rates across all tests, whereas Claude-3 Opus mostly ranked second. These findings highlight the potential of VLMs in supporting medical care in highly specialized fields, such as advanced DR, NM, and IR.

To the best of our knowledge, this is the first study to evaluate and compare the diagnostic accuracy of multiple VLMs, including GPT-4V, GPT-4o, Claude-3 Sonnet, and Claude-3 Opus, using the JDR, JNM, and JIR tests, and the results showed that the GPT-4o-based ChatGPT had the highest accuracy rates for questions with images as well as all questions. The superior performance of GPT-4o-based ChatGPT can be attributed to the broader and more recent dataset on which it was trained. Unlike Claude-3, which was trained only on data up to August 2023, GPT-4o-based ChatGPT incorporated data available up to December 2023. This extended training period likely provides more up-to-date information and advancements, contributing to higher diagnostic accuracy. In addition, GPT-4o not only has improved image recognition capabilities compared with GPT-4 but also shows remarkable enhancements in non-English languages, including Japanese. This could explain the better performance of the model using the GPT-4o in the evaluation of Japanese tests. By contrast, LLM drift, which refers to the deterioration in the performance of LLMs, may have influenced the performance of relatively old VLMs because of the impact of model updates and weight changes on maintaining the reliability of language models [13].

Given the varied difficulty levels of the questions, a simple comparison is not feasible. However, when comparing questions with images to questions without images, all VLMs demonstrated higher accuracy rates in questions without images, with the exception of Claude-3 Sonnet and Opus in the NM test. These results suggest that current VLMs have an inadequate capability to process radiological images and extract imaging findings. OpenAI, which developed ChatGPT, officially commented that the current GPT-4V is unsuitable for medical image interpretation and cannot replace professional medical diagnoses [18]. Furthermore, a previous study indicated that GPT-4V primarily depends on linguistic cues for decision-making with images supplementary [15]. Thus, future update and weight changes of VLMs in decision-making may vary the results. Techniques, such as retrieval-augmented generation, fine-tuning with reinforcement learning from appropriate feedback, and training vision models on a wide range of medical images, may also improve the performances [19]. Nonetheless, GPT-4o, the latest model of ChatGPT, showed the best performance in answering the test; therefore, VLMs are gradually improving their ability to recognize medical images.

In the JIR tests, although GPT-4o exhibited the highest accuracy rates among the other VLMs, the differences in performance between the models were less pronounced, and the overall accuracy was lower in the JIR test than in the other tests. This could be attributed to the trends in the JIR

Table 4 *P* values of the pairwise McNemar's test for each board certification test

	All question			Questions with images			Questions without images		
	JDR test	JNM test	JIR test	JDR test	JNM test	JIR test	JDR test	JNM test	JIR test
4+4 V vs. Sonnet	0.02*	<0.001*	0.21	0.09	0.23	–	0.07	<0.001*	0.21
4+4 V vs. Opus	0.63	0.06	0.094	0.63	0.55	–	0.99	0.006*	0.09
4+4 V vs. 4o	<0.001*	<0.001*	0.007*	<0.001*	<0.001*	–	0.99	0.005*	0.007*
Sonnet vs. Opus	<0.001*	<0.001*	0.003*	0.02*	0.059	–	0.07	0.003*	0.003*
Sonnet vs. 4o	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	–	0.11	<0.001*	<0.001*
Opus vs. 4o	<0.001*	<0.001*	0.40	<0.001*	0.059	–	0.99	<0.001*	0.40

JDR Japanese diagnostic radiology, *JNM* Japanese nuclear medicine, *JIR* Japanese interventional radiology

*Statistically significant

tests, which predominantly featured multiple-choice questions (JDR test, 60/383, 16%; JNM test, 0/300, 0%; JIR test, 177/322, 55%). Although a simple comparison between single- and multiple-answer questions might be inappropriate because of the varied difficulty levels of the questions, most VLMs demonstrated higher accuracy rates for single-answer questions. This trend suggests that language models may be better at handling questions in which only one correct answer needs to be identified, potentially owing to less complexity in the decision-making processes [20]. Additionally, JIR tests often require decisions not only for diagnosis but also for treatment options, which can vary based on the clinical scenario. This variability, along with questions demanding detailed anatomical knowledge, might have influenced the lower performance rates observed among the VLMs in this specialized field.

In recent years, advancements in technologies such as CT and MRI have led to an increase in the number of these examinations, subsequently raising demands on radiologists [21, 22]. Using VLMs as diagnostic support tools can help mitigate these demands, improve diagnostic efficiency, reduce errors, and enhance patient outcomes. Therefore, it would be essential for radiologists and other healthcare professionals to understand the latest performance, strengths, and limitations of VLMs to use them effectively. Future research should broaden the evaluation to include clinical problems and other medical specialties, enabling a more comprehensive assessment of VLMs' utility in healthcare. This approach would highlight the strengths and weaknesses of each model across various domains.

This study had several limitations. First, the questions used in this study may have been included in the training data of VLMs, which could introduce potential bias [23]. This bias may lead to overestimation of the diagnostic accuracy of VLMs [24]. Second, answering each question three times and using the most frequent response as the final answer may not be sufficient, as there could be variability with only three responses. This could lead to underestimation or overestimation of the performance of the VLMs. Third, this study evaluated the performance of Japanese questions. There may be differences in the performance of VLMs when using other languages, such as English. Fourth, because the official correct answers were not publicly available, there is a possibility that the answers provided by the specialists were incorrect, which might have prevented a fair evaluation of the performance of the LLMs.

In conclusion, this study evaluated the diagnostic accuracy of various VLMs in the JDR, JNR, and JIR board certification tests. The results demonstrated that GPT-4o exhibited the highest accuracy rates across all tests, whereas Claude-3 Opus consistently ranked second. The superior performance of GPT-4o can be attributed to its more recent and broader training dataset as well as its improved image recognition capabilities and enhancements in non-English languages. However, current VLMs have limitations in processing radiological images and extracting imaging findings. Despite these limitations, this study highlights the potential of VLMs to support medical care in highly specialized fields.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11604-024-01633-0.

Acknowledgements This study was supported by Guerbet and Iida Group Holdings.

Author contributions Conceptualization: Tatsushi Oura and Hiroyuki Tatekawa; Methodology: Tatsushi Oura, Hiroyuki Tatekawa, Daisuke Horiuchi, and Daiju Ueda; Formal analysis and investigation: Tatsushi Oura and Hiroyuki Tatekawa; Writing—original draft preparation: Tatsushi Oura and Hiroyuki Tatekawa; Writing—review and editing: Daisuke Horiuchi, Shu Matsushita, Hirotaka Takita, Natsuko Atsukawa, Yasuhito Mitsuyama, Atsushi Yoshida, Kazuki Murai, Rikako Tanaka, Taro Shimono, Akira Yamamoto, Yukio Miki, and Daiju Ueda; Supervision: Yukio Miki and Daiju Ueda.

Funding This study was supported by Guerbet and Iida Group Holdings.

Data, material and/or code availability The data that support the findings of this study are available on request to the corresponding authors.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This study was conducted in accordance with the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards. The protocol of this study was approved by the institutional review board. As this study only utilized the publicly available data, the requirement for informed consent was waived.

Consent to participate Not applicable.

Consent to publish Not applicable.

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. 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.
2. OpenAI. GPT-4 Technical Report [Internet]. arXiv [cs.CL]. 2023. Available from: http://arxiv.org/abs/2303.08774.
3. Eloundou T, Manning S, Mishkin P, Rock D. GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models [Internet]. arXiv [econ.GN]. 2023. Available from: http://arxiv.org/abs/2303.10130.
4. Anthropic AI. The claude 3 model family: Opus, sonnet, haiku. Claude-3 Model Card. 2024.
5. Takita H, Kabata D, Walston SL, Tatekawa H, Saito K, Tsujimoto Y, et al. Diagnostic performance comparison between generative AI and physicians: a systematic review and meta-analysis [Internet]. bioRxiv. 2024. Available from: https://doi.org/10.1101/2024.01.20.24301563v2.
6. Ueda D, Walston SL, Matsumoto T, Deguchi R, Tatekawa H, Miki Y. Evaluating GPT-4-based ChatGPT's clinical potential on the NEJM quiz. BMC Digit Health. 2024;2:1–7. https://doi.org/10.1186/s44247-023-00058-5.
7. Ueda D, Mitsuyama Y, Takita H, Horiuchi D, Walston SL, Tatekawa H, et al. ChatGPT's diagnostic performance from patient history and imaging findings on the diagnosis please quizzes. Radiology. 2023;308: e231040. https://doi.org/10.1148/radiol.231040.
8. Sasaki F, Tatekawa H, Mitsuyama Y, Kageyama K, Jogo A, Yamamoto A, et al. Bridging language and stylistic barriers in ir standardized reporting: enhancing translation and structure using ChatGPT-4. J Vasc Interv Radiol. 2023. https://doi.org/10.1016/j.jvir.2023.11.014.
9. Nakaura T, Yoshida N, Kobayashi N, Shiraishi K, Nagayama Y, Uetani H, et al. Preliminary assessment of automated radiology report generation with generative pre-trained transformers: comparing results to radiologist-generated reports. Jpn J Radiol. 2024;42:190–200. https://doi.org/10.1007/s11604-023-01487-y.
10. Nakaura T, Naganawa S. Writing medical papers using large-scale language models: a perspective from the Japanese Journal of Radiology. Jpn J Radiol. 2023;41:457–8. https://doi.org/10.1007/s11604-023-01408-z.
11. 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. 2024;42:201–7. https://doi.org/10.1007/s11604-023-01491-2.
12. Güneş YC, Cesur T. Diagnostic accuracy of large language models in the European Board of Interventional Radiology Examination (EBIR) sample questions. Cardiovasc Intervent Radiol. 2024. https://doi.org/10.1007/s00270-024-03674-4.
13. Krishna S, Bhambra N, Bleakney R, Bhayana R. Evaluation of reliability, repeatability, robustness, and confidence of GPT-3.5 and GPT-4 on a radiology board-style examination. Radiology. 2024;311: e232715. https://doi.org/10.1148/radiol.232715.
14. Horiuchi D, Tatekawa H, Shimono T, Walston SL, Takita H, Matsushita S, et al. Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. Neuroradiology. 2023. https://doi.org/10.1007/s00234-023-03252-4.
15. 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. https://doi.org/10.1007/s11604-024-01561-z.
16. Horiuchi D, Tatekawa H, Oura T, Oue S, Walston SL, Takita H, et al. Comparing the diagnostic performance of GPT-4-based ChatGPT, GPT-4V-based ChatGPT, and radiologists in challenging neuroradiology cases. Clin Neuroradiol. 2024. https://doi.org/10.1007/s00062-024-01426-y.
17. McNEMAR Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947;12:153–7. https://doi.org/10.1007/BF02295996.
18. GPT-4V(ision) System Card [Internet]. Open AI; 2023 Sep. Available from: https://cdn.openai.com/papers/GPTV_System_Card.pdf.
19. Bhayana R. Chatbots and large language models in radiology: a practical primer for clinical and research applications. Radiology. 2024;310: e232756. https://doi.org/10.1148/radiol.232756.
20. Zheng C, Zhou H, Meng F, Zhou J, Huang M. Large Language Models Are Not Robust Multiple Choice Selectors [Internet]. arXiv [cs.CL]. 2023. Available from: http://arxiv.org/abs/2309.03882.
21. Hendee WR, Becker GJ, Borgstede JP, Bosma J, Casarella WJ, Erickson BA, et al. Addressing overutilization in medical imaging. Radiology. 2010;257:240–5. https://doi.org/10.1148/radiol.10100063.
22. Kumamaru KK, Machitori A, Koba R, Ijichi S, Nakajima Y, Aoki S. Global and Japanese regional variations in radiologist potential workload for computed tomography and magnetic resonance imaging examinations. Jpn J Radiol. 2018;36:273–81. https://doi.org/10.1007/s11604-018-0724-5.
23. Ueda D, Kakinuma T, Fujita S, Kamagata K, Fushimi Y, Ito R, et al. Fairness of artificial intelligence in healthcare: review and recommendations. Jpn J Radiol. 2023;42:3–15. https://doi.org/10.1007/s11604-023-01474-3.
24. Walston SL, Seki H, Takita H, Mitsuyama Y, Sato S, Hagiwara A, et al. Data Set Terminology of Artificial Intelligence in Medicine: A Historical Review and Recommendation [Internet]. arXiv [cs.AI]. 2024. Available from: http://arxiv.org/abs/2404.19303.

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