Clinical Ophthalmology

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Image Recognition Performance of GPT-4V(ision) and GPT-4o in Ophthalmology: Use of Images in Clinical Questions

Kosei Tomita[1], Takashi Nishida[2], Yoshiyuki Kitaguchi[3], Koji Kitazawa[4,*], Masahiro Miyake[5,*]

[1]Department of Ophthalmology, Kawasaki Medical School, Okayama, Japan; [2]Hamilton Glaucoma Center, Shiley Eye Institute, Viterbi Family Department of Ophthalmology, University of California, San Diego, La Jolla, CA, USA; [3]Department of Ophthalmology, Osaka University Graduate School of Medicine, Osaka, Japan; [4]Department of Ophthalmology, Kyoto Prefectural University of Medicine, Kyoto, Japan; [5]Department of Ophthalmology and Visual Sciences, Kyoto University Graduate School of Medicine, Kyoto, Japan

*These authors contributed equally to this work

Correspondence: Takashi Nishida, University of California, 9415 Campus Point Drive, San Diego, La Jolla, CA, 92093-0946, USA, Email t.nishida.opt@gmail.com

Clinical Ophthalmology downloaded from https://www.dovepress.com/
For personal use only.

Purpose: To compare the diagnostic accuracy of Generative Pre-trained Transformer with Vision (GPT)-4, GPT-4 with Vision (GPT-4V), and GPT-4o for clinical questions in ophthalmology.

Patients and Methods: The questions were collected from the "Diagnosis This" section on the American Academy of Ophthalmology website. We tested 580 questions and presented ChatGPT with the same questions under two conditions: 1) multimodal model, incorporating both the question text and associated images, and 2) text-only model. We then compared the difference in accuracy using McNemar tests among multimodal (GPT-4o and GPT-4V) and text-only (GPT-4V) models. The percentage of general correct answers was also collected from the website.

Results: Multimodal GPT-4o performed the best accuracy (77.1%), followed by multimodal GPT-4V (71.0%), and then text-only GPT-4V (68.7%); (P values < 0.001, 0.012, and 0.001, respectively). All GPT-4 models showed higher accuracy than the general correct answers on the website (64.6%).

Conclusion: The addition of information from images enhances the performance of GPT-4V in diagnosing clinical questions in ophthalmology. This suggests that integrating multimodal data could be crucial in developing more effective and reliable diagnostic tools in medical fields.

Keywords: ChatGPT, large language model, GPT-4o, ophthalmology

Introduction

Artificial intelligence (AI) is advancing rapidly, with growing interest in its applications across various medical fields.[1] In particular, AI is being explored as a tool to assist in diagnostic imaging, which plays a crucial role in disease detection and monitoring. Studies have demonstrated AI's effectiveness in identifying a range of conditions, including cancer, cardiovascular diseases, neurological disorders, and musculoskeletal abnormalities.[2] In ophthalmology, where imaging is essential, AI-driven analysis of fundus photography is being used to detect diabetic retinopathy, glaucoma, age-related macular degeneration, and retinopathy of prematurity, and myopia, as well as to monitor disease progression.[3]

Generative AI models have been explored in healthcare for biomedical text processing (BioBERT),[4] medical question answering (Med-PaLM),[5] and clinical decision support, including Gemini (by Google), GPT (by OpenAI), Claude (by Anthropic) and LLaMA (by Meta AI).[1] Multimodal AI models integrating text and image analysis have shown promise in radiology and pathology, yet their role in ophthalmology remains underexplored. Meanwhile, the field of natural language processing has developed rapidly with the advent of large language models (LLMs),[6] which are being applied in medical diagnosis, clinical report generation, medical education, robotic assistance in surgery, and medical language translation.[7] The efficacy of these models has been explored in various exams, including the Basic and Clinical Science

Received: 14 October 2024
Accepted: 9 April 2025
Published: 8 May 2025

© 2025 Tomita et al. This work is published and licensed by Dove Medical Press Limited. The full terms of this license are available at https://www.dovepress.com/terms.php and incorporate the Creative Commons Attribution – Non Commercial (unported, v4.0) License (http://creativecommons.org/licenses/by-nc/4.0/). By accessing the work you hereby accept the Terms. Non-commercial uses of the work are permitted without any further permission from Dove Medical Press Limited, provided the work is properly attributed. For permission for commercial use of this work, please see paragraphs 4.2 and 5 of our Terms (https://www.dovepress.com/terms.php).

Course (BCSC) Self-Assessment Program, OphthoQuestions, and FRCOphth examinations. The BCSC Self-Assessment Program and OphthoQuestions are commonly used in the United States for board exam preparation, while the FRCOphth examinations serve as part of the ophthalmology certification process in the United Kingdom.[8–10] OpenAI recently introduced Generative Pre-trained Transformer (GPT)-4V(ision) in September 2023 aimed at delivering accurate and efficient language processing capabilities.[11] Additionally, GPT-4o was released in May 2024, an optimized version of GPT-4V, which enhances the image analysis capabilities, making it more advanced for multimodal tasks.[12] GPT-4V and GPT-4o are robust multimodal models excelling in image interpretation. These models are versatile LLMs that can process both images and text, facilitating tasks such as visual question answering (VQA),[11,12] which integrates computer vision and natural language processing.[13] This integration allows the models to analyze and understand visual data, enhancing its capability to interpret and respond to text. The introduction of multimodal LLMs capable of processing both text and images represents a major development in AI-assisted diagnostics.[14] Integrating medical imaging into these models could refine their ability to address clinical questions and enhance diagnostic decision-making.

While AI has been extensively studied in image analysis and text-based question answering, the comparative diagnostic accuracy of text-based and multimodal LLMs remains unclear. This study aims to evaluate and compare the diagnostic accuracy of text-based and multimodal LLMs in ophthalmology, thereby deepening our understanding of advanced AI models in healthcare, especially in image-centric disciplines like ophthalmology. It has the potential to improve AI-assisted diagnostics in eye care, benefiting both practitioners and patients.

Materials and Methods

In our study, we employed the latest version of the language model available at the time of study (GPT-4 Turbo and GPT-4o, OpenAI; https://chat.openai.com/), which is a multimodal model capable of processing both text and images. GPT-4V was trained with information available up to April 2023. We tested the same questions under two conditions: 1) multimodal model, incorporating both the question text and associated images, and 2) text-only model. Text and images were manually input using the ChatGPT web-based interface, which allows users to enter both textual and visual data for model interaction. To prevent the possibility of past responses affecting the output of ChatGPT, a new chat session was created for each question and each condition (ie, with and without images). We used the following standardized prompt for all cases, regardless of whether image data were available:

> I am conducting an experiment on ophthalmic clinical case discussions to compare your diagnostic conclusions with those of clinicians. Each case is derived from the American Academy of Ophthalmology (AAO) website. You are not trying to treat the patients. At the end of the case, you will be presented with four choices. Please select the most likely answer. If no background information or images are provided, please solve the question with only the available information. There's no need to elaborate on your reasoning.

This prompt was used in a zero-shot manner, without iterative refinements or modifications for all cases. The comparison between GPT-4V (text-only) and GPT-4V (text+image) was performed between December 18 and December 26, 2023. During the study, GPT-4o became available, allowing us to expand our analysis to include a comparison between GPT-4V (text+image) and GPT-4o (text+image) from June 6 to June 8, 2024.The internet browsing function was not utilized in this study, and the Memory feature was disabled to prevent any influence from past interactions on the responses.

Before using ChatGPT on a web platform, we opted out of data sharing to ensure that information is not used for training purposes. Additionally, private browsing mode was employed to prevent the storage of search history and cookies. Although the cases from the "Diagnosis This" are not personally identifiable, further anonymization was conducted, such as changing "a 58-year-old woman" to "a woman in her 50s." Permission was obtained from the AAO to use the Academy's content.

The questions were collected from the "Diagnosis This" section on the AAO website (https://www.aao.org/education/education-browse?filter=diagnose-this). Access to this content requires an AAO membership login. Since November 2009, the "Diagnosis This" program has presented one case almost every week, accumulating a total of 677 cases. Each case consists of both image and text, and one answer is chosen from four options. The order of these options is randomized on the website, and subsequently, the order of the four options entered into the prompt is also

randomized. After responding, the website displays the correct answer and explanation for the case, along with the rates of general correct answers. It is important to note that the rates for general correct answers are reference values, as the clinical level or expertise of those providing the answers is not disclosed. The purpose of this study was to compare the performance of multimodal models and text-only models; therefore, the uses of non-specific, generic images not to related to the cases—such as stock medical illustrations, anatomical diagrams, or unrelated ophthalmic images that did not correspond directly to the clinical case (n=57) were excluded. Furthermore, cases where the text-only model failed to generate a response (n=27) were excluded, as these primarily involved questions requiring image references, which the text-only model could not process. Additionally, duplicated questions (n=8), questions lacking four answer choices (n=4), and questions no longer available on the website (n=2) were also excluded. Consequently, 580 questions were included in the dataset (Figure 1). The multiple-choice answers were recorded in Excel (KT, TN) and used for analysis. To assess the outputs generated by the AI models, responses were evaluated based on their accuracy in selecting the correct answer from the four options provided. The evaluation criteria included correctness of the selected answer and consistency across different conditions (with and without images). The subspecialty in question was also recorded.

It has not been disclosed whether the AAO website is included in the training data for ChatGPT. Therefore, to avoid the influence of training on the responses, subsets of data up to April 2023 and from May 2023 onwards were created and compared. Furthermore, for the subset after May 2023, prompts were repeated five times for both the multimodal and text models, with the order of the multiple-choice options randomized, and an intraclass correlation coefficient (ICC) was calculated to assess the reproducibility of the correct answers.

The performance of different GPT models was evaluated using 2×2 contingency tables, McNemar tests, and Kappa statistics to compare correct answer rates and measure agreement beyond chance. McNemar tests were used to compare the correct answer rates among multimodal (GPT-4o and GPT-4V) and text-only (GPT-4V) models. Statistical analyses were performed using Stata version 16.0 (StataCorp, College Station, TX) and python 3.11.1 (Python Software Cooperation, Wilmington, DE, USA). No corrections for P values were made for multiple comparisons, acknowledging the potential for Type I errors. The decision not to adjust P values was based on the exploratory nature of the study and the focus on initial observations rather than definitive conclusions. Accordingly, models were not adjusted for multiple comparisons and were considered exploratory. All P values were two-sided. This research did not involve human

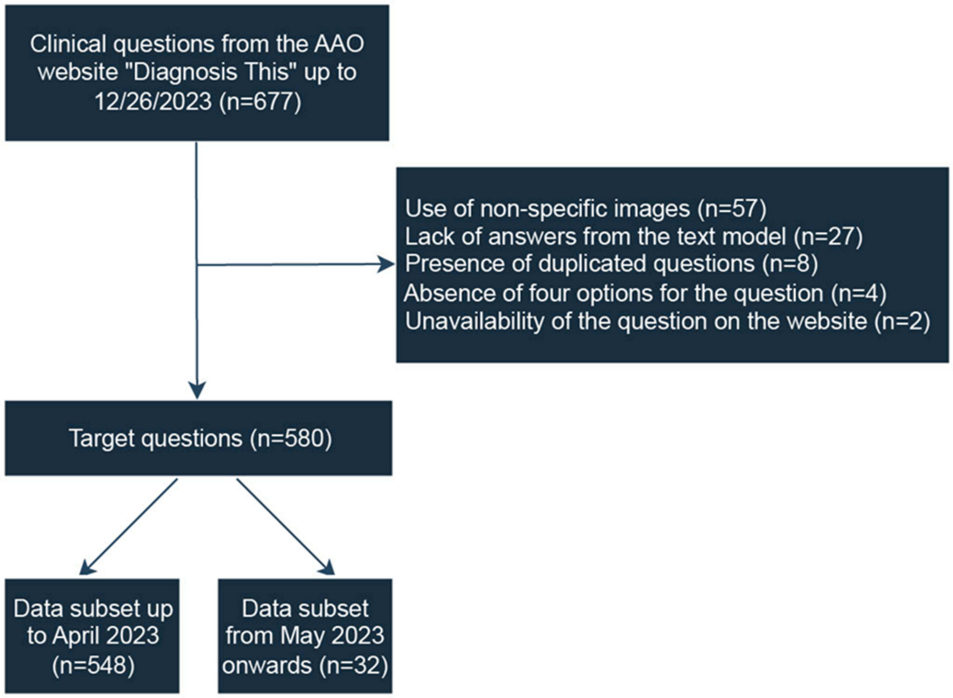

Figure 1 Diagram for the study.

https://doi.org/10.2147/OPTH.S494480

Table 1 Accuracy of Multimodal Models, Text Models, and General Correct Answers on the Website Divided Into Subsets as of the End of April 2023

Correct	Overall	Subset up to 4/2023	Subset from 5/2023
GPT-4o: Image + Text	77.1 (73.6 to 80.5)	77.0 (73.5 to 80.5)	78.1 (63.0 to 93.3)
GPT-4V: Image + Text	71.0 (67.3 to 74.7)	71.0 (67.2 to 74.8)	71.9 (55.4 to 88.3)
GPT-4V: Text only	66.7 (62.9 to 70.6)	66.6 (62.6 to 70.6)	68.8 (51.8 to 85.7)
General correct answers on the website	64.6 (62.9 to 66.3)	64.6 (62.8 to 66.4)	64.7 (56.8 to 72.6)

Notes: Values are shown in mean (95% confidence interval).

subjects, and therefore, the University of California, San Diego Institutional Review Board determined that this study was exempt for approval. The study adhered to the tenets of the Declaration of Helsinki and complied with the Health Insurance Portability and Accountability Act.

Results

Table 1 shows the accuracy of the multimodal model, text model, and rate of general correct answers on the website. "GPT-4o: Image + Text" (77.1% [95% CI, 73.6 to 80.5]) and "GPT-4V: Image + Text" (71.0% [95% CI, 67.3 to 74.7]) models had higher correct response rates compared to "GPT-4V: Text only" (66.7% [95% CI, 62.9 to 70.6]). The analysis divided the data into subsets before and after May 2023, showing similar accuracy rates in each category. For multimodal models (with images), the correct answer rates were 77.0% (95% CI, 73.5 to 80.5) before and 78.1% (95% CI, 63.0 to 93.3) after May 2023 for GPT-4V, and 71.0% (95% CI, 67.2 to 74.8) before and 71.9% (95% CI, 55.4 to 88.3) after May 2023 for GPT-4V, respectively. For the text-only model (without images), the correct answer rates were 66.6% (95% CI, 62.6 to 70.6) before and 68.8% (95% CI, 51.8 to 85.7) after May 2023. The correct answer rates on the website were 64.6% (95% CI, 62.8 to 66.4) before and 64.7% (95% CI, 56.8 to 72.6) after May 2023. ICC was 0.91 (95% CI, 0.87 to 0.94) for the data subset from May 2023 onwards.

Figure 2 illustrates the 2×2 contingency tables for each LLM model. "GPT-4o: Image + Text" model performed the best accuracy (77.1%), followed by multimodal "GPT-4V: Image + Text" (71.0%), and then "GPT-4V: Text only" (68.7%); (McNemar tests P values < 0.001, 0.012, and 0.001, respectively). "GPT-4V: Image + Text" and "GPT-4V: Text only" showed the highest agreement (Kappa=0.672, P<0.001), followed by "GPT-4o: Image + Text" and "GPT-4V: Text only" (Kappa=0.654, P <0.001), and "GPT-4o: Image + Text" and "GPT-4V: Image + Text" (Kappa=0.591, P <0.001).

Figure 3 shows the percentage of correct answers for each subspecialty. The accuracy rate was generally around 60–80% across various subspecialties. However, in the categories of retina and vitreous (n=84) with a correct rate of

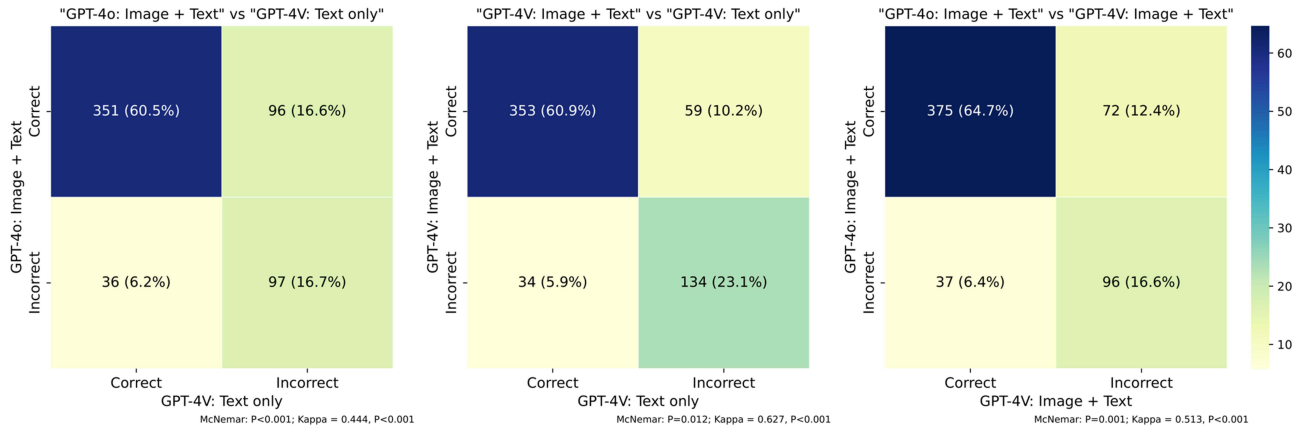

Figure 2 The 2×2 contingency tables for each LLM model.

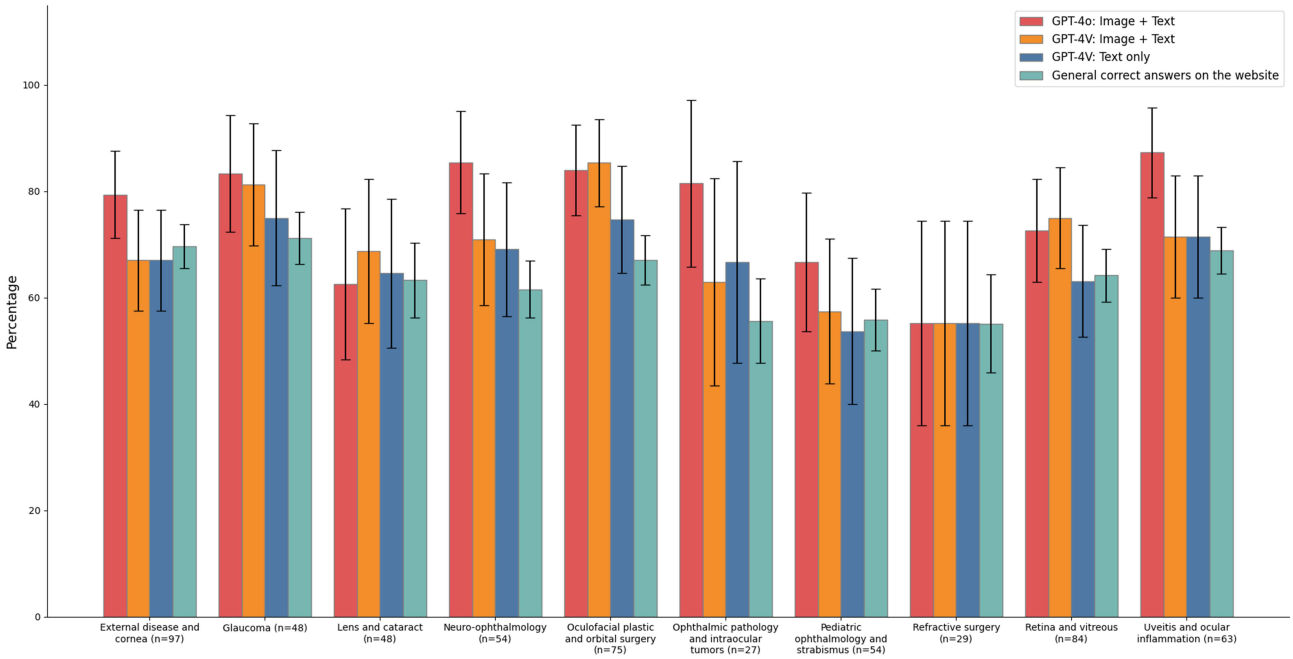

Figure 3 The percentage of correct answers for each subspecialty.

76.2%, oculofacial plastic and orbital surgery (n=75) with 87.8%, and glaucoma (n=48) with 79.2%, the multimodal model demonstrated a higher accuracy rate compared to the text-only model and the general correct answers on the website.

Discussion

In the current study, we compared the diagnostic accuracy of GPT-4-based text model and multimodal model for ophthalmic clinical questions. By incorporating information from images in addition to text, the performance of LLMs in diagnosing clinical questions improved. The improvement was notable in oculofacial plastic and orbital surgery, retina and vitreous, and glaucoma. Despite notable benefits, the overall accuracy rate was not consistently high depending on the subspecialities, indicating that there are still hurdles to its clinical use. Nevertheless, GPT-4o outperformed general respondents in clinical question accuracy, highlighting the potential for future developments in LLMs.

A Study in other medical specialties have also evaluated the performance of LLMs in complex clinical cases. One such study reported that GPT-4 correctly diagnosed 57% of cases, outperforming 99.98% of simulated human readers generated from online responses. These findings suggest that LLMs have the potential to support diagnostic decision-making across various medical fields.

Previous studies have explored the accuracy of LLMs in answering standardized text-only ophthalmology questions. One such study randomly selected 260 text-only questions from BCSC and OphthoQuestions, adjusted for the level of cognition and question difficulty to investigate the accuracy of GPT-4. The finding indicated a combined accuracy rate of 72.9%.[10] In the current study, the accuracy rate of GPT-4V on "Diagnose This", for the text model was 66.7%. This lower accuracy in our study may result from not adjusting for difficulty levels, leading to considerable variation in the difficulty of the problem statements. Even under these conditions, the increase in the accuracy rate for the multimodal models to 77.1% for GPT-4o and 71.0% for GPT-4V may suggest its capability to correctly answer more challenging questions. For instance, in a case presenting central visual loss (Supplemental Case 1), a full-thickness macular hole was diagnosed from optical coherence tomography (OCT) and fundus photograph, leading to a correct treatment plan. The text-only model provided an incorrect response to this question. In this study, multimodal models demonstrated proficiency in processing various medical images and identifying specific features, yet it is important to note that the model occasionally failed to recognize overt findings. There is a case, like the misidentification of acute corneal hydrops in keratoconus as Acanthamoeba keratitis, indicating areas for

further improvement (Supplemental Case 2). Supplemental Cases 3–6 present examples where GPT-4o accurately interpreted image information, responded correctly without a definitive diagnosis, outperformed GPT-4V, or provided more detailed reasoning leading to the correct answer. The diagnostic performance of LLMs has also been evaluated in other medical fields. A study on complex clinical cases reported that GPT-4V correctly diagnosed 57% of cases, outperforming 99.98% of simulated human readers generated from online responses.[15] Despite occasional misinterpretations, the ability of multimodal models to integrate text and image data represents a noteworthy advancement, demonstrating their potential to support clinical decision-making.

Our study found that the overall accuracy rate of the multimodal model was higher than that of text-only model, with varying degrees of improvement across different subspecialties. In specific subspecialties, such as glaucoma, oculofacial plastic and orbital surgery, and retina and vitreous, accuracy rates notably increased, likely due to image availability. In ophthalmology, imaging is crucial for diagnosis and management in clinical practice. For example, glaucoma and retinal diseases are particularly reliant on OCT. Similarly, oculofacial plastic and orbital surgery questions often involved computerized tomography (CT) and magnetic resonance imaging scans. GPT-4o demonstrated proficiency in interpreting and responding to clinical imaging, suggesting that image availability may enhance response accuracy. In contrast, for some subspecialties, the use of imaging information did not improve accuracy. There is a possibility that information related to medical imaging, such as OCT and CT scans, which contribute to our clinical practice, is widely available and publicly accessible on the web. This suggests that sub-specialties leveraging these imaging resources might particularly benefit from the use of image-assisted LLMs. Moreover, beyond the combination of images and textual queries, multimodal models have the potential to integrate other structured clinical information, such as patient history, electronic health records, or clinician-generated notes, further enhancing diagnostic accuracy and decision support. On the other hand, for subspecialties that did not readily benefit from the presence of images in this study, it may be possible to enhance overall answer accuracy in the future by incorporating specialized training datasets tailored to their specific clinical contexts.

LLMs are increasingly valuable as innovative tools for interpreting the visual world. They provide descriptions of photographs from smartphones for individuals who are blind or have low vision, detailing the surrounding environment, object locations, and character recognition. Ongoing efforts aim to employ these technologies in patient care, aiming to alleviate the workload of healthcare providers in clinical settings. Inaccuracies in responses have been reported, indicating areas of concern.[16] OpenAI explicitly states that the GPT-4V is unsuitable for any medical function, including providing professional medical advice, diagnoses, treatments, or judgments, due to its suboptimal performance in the medical field.[11] There have been cases where its use in interpreting medical images led to severe errors, such as incorrect identification of lesion laterality.[11,17] These issues highlight the need for further enhancements in model accuracy, validation, regulatory compliance, and ethical considerations before it can be safely applied in clinical contexts. Based on our findings, GPT-4o using multimodal model showed an increased accuracy compared to text-only model. However, when considering its immediate applicability in clinical settings, an accuracy rate of less than 80% might be a suboptimal result. The potential for LLMs in medicine is promising, yet it's important to acknowledge that resources like BCSC, OphthoQuestions, and Diagnose This are ultimately structured as question texts, presumably with hints provided within the text for answering. In real clinical scenarios, the challenge extends beyond interpreting physical signs to accurately gathering patient history for effective treatment. Effective communication between healthcare professionals and patients is essential for precise information collection.[10] Without such interaction, research using LLMs might remain theoretical and not practically applicable, potentially delaying their real-world clinical implementation to a distant future. A systematic review highlights the growing anticipation for patient communication-focused LLMs, particularly in extracting patient information.[18] Looking ahead, the development of medically specialized LLMs is anticipated, alongside advancements in patient communication capabilities.

Effective use of LLMs in information gathering depends on the user's understanding and the quality of prompts. Therefore, effective collaboration between users and LLMs is important. Future advancement is expected to produce more sophisticated LLMs, driven by expanded training in specific domains. However, generally available LLMs may not yet be adequately trained on medical images due to the limited availability and privacy restrictions associated with such data. While LLMs benefit from extensive web text data, accessing medical imagery is challenging. Some researchers are now focusing on creating specialized multimodal LLMs for medical applications,[19] utilizing open-source technologies and public resources.[20–22] Despite the scarcity of medical images online, those from electronic medical records, paper-based documentation, and the scientific articles could

serve as valuable training datasets for these more focused multimodal LLMs. The development of domain-specific models holds considerable promise in specialized areas like medicine.[8] LLM landscape has recently expanded with several models, not limited to ChatGPT, capable of processing images. Some of these models are noted for their superior performance. However, this study did not incorporate these recent other models. Integrating these models could facilitate more detailed and nuanced clinical assessments. Additionally, GPT-4o outperformed GPT-4V in several contexts, emphasizing the need for ongoing improvements in multimodal LLMs to achieve their full potential in medical applications. Although GPT-4V and GPT-4o were released only a little over a year apart, the differences for accuracies between them were more than 6% in our study. This highlights the rapid advancements in LLM technology and the importance of continual development and refinement. Not only for the diagnosis of individual diseases but also for comprehensive preventive medicine in systemic diseases, research utilizing AI and LLMs is expected to drive advancements in Oculomics. Oculomics is an emerging field that leverages ocular biomarkers to detect and assess systemic diseases, with applications already reported in neurodegenerative and cardiovascular risk assessment.[23] Building on these findings, further exploration of LLMs' capabilities could enhance their role in clinical decision-making, including information gathering, diagnostics, therapeutics, prophylactics, and patient care.

The limitations of this study should be acknowledged. First, the latest version of GPT-4o has been trained on data available up to October 2023, and it cannot be ruled out that some of the questions used in our dataset may have been included in its training data. The specific details of the training data have not been disclosed publicly. However, it is to be noted that on the AAO website, the "Diagnosis This" section requires users to create an ID, log in, and follow specific links to access the questions and answers. Since these questions are embedded and not part of the website's source code, it is unlikely that they were included in the training data. Second, a comparison of GPT-4's performance before and after May 2023 showed no substantial changes, further suggesting limited influence from such specific datasets. Third, systematic prompt engineering was not performed, which may influence the consistency and reproducibility of responses. Slight variations in input phrasing could potentially impact outcomes. Last, this study included various image modalities, such as OCT, fundus photographs, anterior segment images, and non-ophthalmic images related to systemic diseases. A detailed subgroup analysis by image type was not conducted. Future research using more targeted datasets could provide further insights.

Conclusion

In conclusion, this study demonstrates that incorporating image data with GPT-4o and GPT-4V improves diagnostic accuracy in ophthalmic clinical problems, indicating the potential of multimodal LLMs in medical applications. However, challenges such as AI hallucinations, errors, and limitations related to model design and policy, alongside the occasional failure in interpreting medical images, underscore the need for caution and further enhancements before these technologies can be safely implemented in clinical settings. Despite these challenges, the evolving landscape of LLMs, including the development of specialized multimodal models for medicine, holds promise for more sophisticated and nuanced clinical assessments in the future.

Abbreviations

AAO, American Academy of Ophthalmology; AI, artificial intelligence; BCSC, Basic and Clinical Science Course; CT, computerized tomography; GPT, Generative Pre-trained Transformer with Vision; GPT-4V, GPT-4 with Vision; ICC, intraclass correlation coefficient; OCT, optical coherence tomography; LLM, large language model; VQA, visual question answering.

Acknowledgments

We sincerely appreciate the American Academy of Ophthalmology for kindly granting us permission to use the foundational materials from *Diagnose This*. This research was supported by the Japan Agency for Medical Research and Development (grant number: JP22uk1024006). The funder had no role in the design or conduct of the study, collection, management, analysis, or interpretation of the data; preparation, reviews or approval of the manuscript; or the decision to submit the manuscript for approval. This paper has been uploaded to Medrxiv as a preprint: https://www.medrxiv.org/content/10.1101/2024.01.26.24301802v1

Disclosure

KT received lecture fees from Senju Pharmaceutical, Chugai Pharmaceutical, and KOWA; TN is a consultant of Topcon. MM received the grant from AMED. The authors report no other conflicts of interest in this work.

References

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930–1940. doi:10.1038/s41591-023-02448-8
2. Mello-Thoms C, Mello CAB. Clinical applications of artificial intelligence in radiology. *Br J Radiol*. 2023;96(1150):20221031. doi:10.1259/bjr.20221031
3. Ting DSW, Peng L, Varadarajan AV, et al. Deep learning in ophthalmology: the technical and clinical considerations. *Prog Retin Eye Res*. 2019;72:100759. doi:10.1016/j.preteyeres.2019.04.003
4. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234–1240. doi:10.1093/bioinformatics/btz682
5. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172–180. doi:10.1038/s41586-023-06291-2
6. OpenAI. Introducing ChatGPT. Available from: https://openai.com/index/chatgpt/.
7. Liu L, Yang X, Lei J, et al. A survey on medical large language models: technology, application, trustworthiness, and future directions. *arXiv preprint arXiv:2406 03712*. 2024. doi:10.48550/arXiv.2406.03712
8. Tan TF, Thirunavukarasu AJ, Campbell JP, et al. Generative artificial intelligence through ChatGPT and other large language models in ophthalmology: clinical applications and challenges. *Ophthalmol Sci*. 2023;3(4):100394. doi:10.1016/j.xops.2023.100394
9. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. *Ophthalmol Sci*. 2023;3(4):100324. doi:10.1016/j.xops.2023.100324
10. Antaki F, Milad D, Chia MA, et al. Capabilities of GPT-4 in ophthalmology: an analysis of model entropy and progress towards human-level medical question answering. *Br J Ophthalmol*. 2023;107:90–95. doi:10.1136/bjo-2023-324438
11. OpenAI. GPT-4V(ision) system card. 2023. Available from: https://openai.com/index/gpt-4v-system-card/.
12. OpenAI. Hello GPT-4o; 2024. Available from: https://openai.com/index/hello-gpt-4o/.
13. Jha S, Dey A, Kumar R, Kumar-Solanki V. A novel approach on visual question answering by parameter prediction using faster region based convolutional neural network. *Int J Interact Multimed Artif Intell*. 2019;5(5):30–37. doi:10.9781/ijimai.2018.08.004
14. Zhang JY, Huang J, Jin S, Lu S. Vision-language models for vision tasks: a survey. *arXiv:2304.00685*. 2023. doi:10.48550/arXiv.2304.00685
15. Eriksen AV, Möller S, Ryg J. Use of GPT-4 to diagnose complex clinical cases. *Nejm Ai*. 2024;1(1). doi:10.1056/AIp2300031
16. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *MDPI*. 2023:887.
17. Yang Z, Li L, Lin K, et al. The Dawn of lmms: preliminary explorations with gpt-4v (ision). *arXiv:2309.17421*. 2023;9(1):1. doi:10.48550/arXiv.2309.17421
18. Garg RK, Urs VL, Agarwal AA, Chaudhary SK, Paliwal V, Kar SK. Exploring the role of ChatGPT in patient care (diagnosis and treatment) and medical research: a systematic review. *Health Promot Perspect*. 2023;13(3):183–191. doi:10.34172/hpp.2023.22
19. Singhal K, Tu T, Gottweis J, et al. Towards expert-level medical question answering with large language models. *arXiv:2305.09617*. 2023. doi:10.48550/arXiv.2305.09617
20. Toma A, Lawler PR, Ba J, Krishnan RG, Rubin BB, Wang B. Clinical camel: an open-source expert-level medical language model with dialogue-based knowledge encoding. *arXiv:2305.12031*. 2023. doi:10.48550/arXiv.2305.12031
21. Singhal K, Azizi S, Tu T, et al. Publisher Correction: large language models encode clinical knowledge. *Nature*. 2023;620(7973):E19. doi:10.1038/s41586-023-06455-0
22. Tu T, Azizi S, Driess D, et al. Towards generalist biomedical ai. *arXiv.2307.14334*. 2023. doi:10.48550/arXiv.2307.14334
23. Chew EY, Burns SA, Abraham AG, et al. Standardization and clinical applications of retinal imaging biomarkers for cardiovascular disease: a Roadmap from an NHLBI workshop. *Nat Rev Cardiol*. 2025;22(1):47–63. doi:10.1038/s41569-024-01060-8

Clinical Ophthalmology

Dovepress
Taylor & Francis Group

Publish your work in this journal

Clinical Ophthalmology is an international, peer-reviewed journal covering all subspecialties within ophthalmology. Key topics include: Optometry; Visual science; Pharmacology and drug therapy in eye diseases; Basic Sciences; Primary and Secondary eye care; Patient Safety and Quality of Care Improvements. This journal is indexed on PubMed Central and CAS, and is the official journal of The Society of Clinical Ophthalmology (SCO). The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use. Visit http://www.dovepress.com/testimonials.php to read real quotes from published authors.

Submit your manuscript here: https://www.dovepress.com/clinical-ophthalmology-journal