Evaluation of a digital ophthalmologist app built by GPT4-V(ision)

Pusheng Xu, MD[1#], Xiaolan Chen, MD[1#], Ziwei Zhao, MD[1#], Danli Shi, MD, PhD[1,2,3*]

Affiliations

1. School of Optometry, The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China.
2. Research Centre for SHARP Vision, The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China
3. Centre for Eye and Vision Research (CEVR), 17W Hong Kong Science Park, Hong Kong

# Contributed equally

Correspondence

***Dr Danli Shi,** MD, PhD., Experimental Ophthalmology, School of Optometry, The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China.

Email: danli.shi@polyu.edu.hk

Funding: This study was supported by the Start-up Fund for RAPs under the Strategic Hiring Scheme (P0048623) from HKSAR.

Disclosures: None of the authors has a financial or proprietary interest in any material or method mentioned.

Word count: Abstract: 247; **Whole paper**: 2851.

Tables: 3; Figures: 2; **Supplemental Tables**: **6. Supplemental Figures: 1**

Synopsis

Only 30.6%, 21.5%, and 55.6% responses about ocular multimodal images generated by GPT-4V(ision) were considered accurate, highly usable, no harm, respectively. Currently, GPT-4V lacks the reliability required for clinical decision-making and patient consultation in ophthalmology.

Abstract

Purpose: To evaluate the capabilities and incapabilities of a GPT-4V(ision)-based chatbot in interpreting ocular multimodal images.

Methods: We developed a digital ophthalmologist app using GPT-4V and evaluated its performance with a dataset (60 images, 60 ophthalmic conditions, 6 modalities) that included slit-lamp, scanning laser ophthalmoscopy (SLO), fundus photography of the posterior pole (FPP), optical coherence tomography (OCT), fundus fluorescein angiography (FFA), and ocular ultrasound (OUS) images. The chatbot was tested with ten open-ended questions per image, covering examination identification, lesion detection, diagnosis, and decision support. The responses were manually assessed for accuracy, usability, safety, and diagnosis repeatablity. Auto-evaluation was performed using sentence similarity and GPT-4-based auto-evaluation.

Results: Out of 600 responses, 30.6% were accurate, 21.5% were highly usable, and 55.6% were deemed as no harm. GPT-4V performed best with slit-lamp images, with 42.0%, 38.5%, and 68.5% of the responses being accurate, highly usable, and no harm, respectively. However, its performance was weaker in FPP images, with only 13.7%, 3.7%, and 38.5% in the same categories. GPT-4V correctly identified 95.6% of the imaging modalities and showed varying accuracy in lesion identification (25.6%), diagnosis (16.1%), and decision support (24.0%). The overall repeatability of GPT-4V in

diagnosing ocular images was 63% (38/60). The overall sentence similarity between responses generated by GPT-4V and human answers is 55.5%, with Spearman correlations of 0.569 for accuracy and 0.576 for usability.

Conclusion: GPT-4V currently lacks the reliability needed for clinical decision-making in ophthalmology. Our study serve as a benchmark for enhancing ophthalmic multimodal models.

Keywords: GPT-4V(ision); LLM; medical images; evaluation; ophthalmology

- **What is already known on this topic**： First, GPT-4V(ision) exhibited significant advantages in fine-grained world-knowledge-intensive visual question answering. Second, the performance of GPT-4V in the multimodal medical diagnosis domain had been evaluated through case analysis, involving 17 medical systems and 8 modalities used in clinical practice. However, ophthalmic-related images were not included in the study.
- **What this study adds**：Being the first known evaluation of GPT-4V's capabilities in processing ophthalmic multimodal images, our study adds valuable insights to the existing body of knowledge. Our study highlight the incapabilities of GPT-4V, demonstrating that it currently lacks the essential reliability required for clinical decision-making and patient consultation in ophthalmology.
- **How this study might affect research, practice or policy**： The evidence gathered in this study show that continued refinement and testing remain crucial for enhancing the effectiveness of large language models in medical applications. This work provides a benchmark for further investigation in building large language models for processing ophthalmic multimodal images.

Introduction

With the rapid development of artificial intelligence, large language models (LLM) have brought immense potential and opportunities to various fields, particularly in the medical domain.[1] Imaging examinations are a fundamental aspect of medical practice, particularly in ophthalmology. Ophthalmology is a highly multimodal specialty that relies on clinical records and a wide range of medical imaging modalities to make accurate diagnoses and decisions regarding treatment.[2] Currently, the applications of LLMs in ophthalmology are mainly text-based. These include answering questions in ophthalmology specialty exams such as OKAP and FRCOphth,[3 4] addressing surgical treatment-related queries in retinal diseases,[5] and providing insights on myopia-related issues.[6] However, most existing LLMs still have limitations in handling medical fields involving image content.

The recent introduction of GPT-4V(ision) has brought new hope to the medical field.[7] GPT-4V is a multimodal generalist LLM that can process both images and text, enabling various downstream tasks, including visual question answering (VQA). This means that GPT-4V can understand and answer image-related questions, providing more accurate and comprehensive information both for doctors and patients. Li et al.[8] indicated that GPT-4V exhibited significant advantages in fine-grained world-knowledge-intensive VQA. Another study evaluated the performance of GPT-4V in the multimodal

medical diagnosis domain through case analysis, involving 17 medical systems and 8 modalities used in clinical practice.[9] However, ophthalmology was not included in the study, leaving the practical application capabilities of GPT-4V in addressing image-related concerns in ophthalmology uncertain. Meanwhile, previous work on ophthalmic VQA focuses on a specific modality,[10] leaving multimodal ocular VQA unexplored.

In this study, we aim to evaluate the capabilities of a chatbot based on GPT-4V in handling queries related to ocular multimodal images.

Methods

Data

To avoid overlap between the evaluation set and the training set of GPT-4V, a private dataset of ocular multimodal images was collected from multiple tertiary centers in China (Figure 1A). All images and data were existing and deidentified. This study was approved by the institutional review board of The Hong Kong Polytechnic University.

A subset of multimodal images were randomly selected, including 1000 slit-lamp images, 500 scanning laser ophthalmoscopy (SLO) images, 519 fundus photography of the posterior pole (FPP) images, 500 optical coherence tomography (OCT) images, 200 fundus fluorescein angiography (FFA) images,

and 500 ocular ultrasound (OUS) B-scan images. All patient information underwent anonymization and de-identification processes.

The images were manually screened by three experienced ophthalmologists (PX, XC and ZZ), each with more than five years of clinical experience. 10 images representing distinct diagnoses were carefully chosen from each modality as the evaluation dataset (**Figure 1B**). The diagnosis of the selected images were listed in **Supplemental Table 1**. These images showed typical manifestations of the diseases that ophthalmologists would use to make a probable diagnosis.

App construction using GPT-4V

Since GPT-4V often refused to provide potential diagnoses when presented with only ocular examination images during our early testing, we utilized its built-in customized function to build a digital ophthalmologist app with the prompt provided in **Supplemental Figure 1**.

Generation of responses

Due to the lack of universally recognized standard questions in the field of multimodal image-based question answering in ophthalmology, ten questions were constructed to evaluate the performance of GPT-4V based on previous studies[11 12] and clinical experience of ophthalmologists. These questions cover examination identification, lesion identification, diagnosis capacity, and

decision support (**Supplemental Table 2)**. The first three dimensions each consist of one question. The decision support includes seven frequently asked questions by patients to clinical doctors, which cover the following areas: next examination, treatment, severity, complication, etiology, prognosis, and prevention. The GPT-4V was prompted with multimodal ophthalmic images from November 14th, 2023 to November 26th, 2023. An example of question-answer on ocular images using GPT-4V was shown in **Supplemental Table 3.**

Human Evaluation

The evaluation was conducted by three experienced ophthalmologists (PX, XC and ZZ), each with more than five years of clinical experience. The responses were evaluated in accuracy, usability, and safety (**Figure 1B**). Intra-grader agreement was assessed by Fleiss' Kappa.

The accuracy of the response generated by LLMs is of utmost importance. This encompasses the accuracy of factual information, as well as the accuracy of deductions and resolutions. A three-point scale was used to assess the accuracy of responses generated by GPT-4V: 1) “Correct” for factually accurate and reliable content; 2) “Partially Correct” for a mix of accurate and inaccurate content; 3) “Incorrect” for mostly inaccurate or misleading content.[13]

Evaluating the usability of a response could be quite difficult since there are no absolute measures of usability.[14] In this study, usability was assessed as follows: 1) “Highly Usable” for highly relevant, near-expert level understanding and response; 2) “Moderately Usable” for generic response covers a broader range of relevant diseases; 3) “Limited Usability” for irrelevant information or lack of domain knowledge.

Safety is a complex concept that can involve several aspects, such as physical, mental, moral, and financial harm. When evaluating the safety of the responses, raters were specifically instructed to consider only physical or mental health-related harms. We evaluated the severity and likelihood of such harm, assuming that a patient or ophthalmologist, based on the content of the answer, might take appropriate actions.[15] Safety was also evaluated on a three-point scale: 1) “No Harm” for answers that had no negative consequences or risks; 2) “Mild Harm” for answers that could cause slight vision loss or discomfort; 3) “Severe Harm” for answers that could cause significant vision loss or discomfort, potentially leading to blindness.

Question 1 (“What kind of examination is this image?”) was only evaluated in accuracy, while the other nine questions were assessed in accuracy, usability, and safety.

The evaluation of repeatability took place one week after the initial

evaluation rounds, with the history being cleared. We evaluated the repeatability of the responses only for diagnosis, as they determined the responses for lesion identification and decision support. A two-point scale was used to rate the repeatability of the responses: consistent or inconsistent. A response was consistent if it had the same meaning as the previous response, and inconsistent otherwise.

Auto Evaluation

Human answers were generated by three ophthalmologist (PX, XC and ZZ) based on textbooks and clinical consensus. Sentence similarity was compared between the responses of GPT-4V and human answers using Sentence-BERT (sentence-transformers model: paraphrase-MiniLM-L6-v2).[16] GPT-based auto evaluation[17 18] was performed with GPT-4-0613 using the same scoring criteria as human ratings as the prompt at January 26th, 2024. Spearman and Kendall-Tau correlations were calculated to compare the semantic similarity between the responses of GPT-4V and human answer as a continuous variable and categorical level, respectively.

Statistical analysis

Statistical analysis was performed with Stata software (version 17.0, StataCorp, College Station, Texas, USA). Stacked bars were graphed using

GraphPad Prism (version 9.5.0, GraphPad Software Inc. San Diego, CA, USA). The chi-square test was used to compare the difference in performance levels (in propotion) between common diseases and rare diseases with $P < 0.05$ considered statistically significant.

Results

Overall performance

Table 1 summarizes the performance of GPT-4V in terms of accuracy, usability, and safety based on the proportion of responses rated as good (score of 1) across 600 questions related to ocular multimodal images. Fleiss' Kappa values indicated good consistency among the three raters, with 0.758 for accuracy, 0.751 for usability, and 0.767 for safety. Only 30.6% (183.3/600) of responses considered correct by ophthalmologists and 21.5% (123.3/540) deemed highly usable, while approximately half of the answers (55.6% (300.3/540)) considered no harm (**Figure 2**).

Multimodal performance

GPT-4V performed best in slit-lamp images, with average proportions of correct, highly usable, and no harm answers being 42.0% (42.0/100), 38.5% (34.7/90), and 68.5% (61.7/90), respectively (**Figure 2**). However, for FPP, only 13.7% (13.7/100), 3.7% (3.3/90), and 38.5% (34.7/90) of answers were

rated as correct, highly usable, and no harm, respectively. In contrast, GPT-4V received relatively better evaluations (36.0% accuracy, 28.5% high usefulness, and 64.4% harmless) in SLO images. For detailed evaluations of the six modalities, please refer to **Supplemental Table 4**.

Image interpretation performance

We assessed the ability of GPT-4V to interpret ophthalmic images using various VQA tasks, including examination identification, lesion recognition, diagnostic capacity, and decision support (see **Table 2**). Overall, GPT-4V was able to correctly identify the imaging modality of most ophthalmic images (average 95.6% (57.3/60)).

Lesion identification

An average of 25.6% (15.3/60) of answers were considered correct, 47.2% (28.3/60) partially correct, and 27.2% (16.3/60) incorrect. In terms of usability, an average of 23.3% (14.0/60) of answers provided useful recognition of the image's lesions, 43.3% (26.0/60) were moderately usable, and 33.3% (20.0/60) had limited usability. Regarding potential harm caused by the responses, less than half (46.1% (27.7/60)) were considered unsafe, moreover,15.6% (9.3/60) were deemed to have serious potential harm.

Diagnosis capacity

Disease staging and classification were not necessary for this evaluation.

If the response achieved top-1 accuracy, meaning that the prediction with the highest probability matched the expected answer exactly,[19] it was considered "correct" during evaluation. If the response achieved top-5 accuracy, which means that the expected answer matched any of the top-5 highest probability predictions, it was assessed as "partially correct". Otherwise, it was recorded as "incorrect". The results showed that, in terms of accuracy, usability, and safety, only 16.1% (9.7/60), 15.0% (9.0/60), and 31.1% (18.7/60) met the standards of clinical doctors, respectively.

Decision support

This involves a range of topics including further examinations, treatment options, visual impact, complications, etiology, prognosis, and prevention measures, requiring the model to have the ability to integrate image information and professional knowledge. We established a relevant series of questions (**Supplemental Table 2**) to test this. The results showed that GPT-4V only met clinical standards (score of 1) in terms of accuracy and usability in 24.0% (101.0/420) and 22.1% (93.0/420) of responses, respectively. In terms of safety, 59.4% (249.3/420) were considered no harm.

Evaluation of Repeatability in Diagnosis

As shown in **Supplemental Table 5**, the responses of GPT-4V had the highest repeatability in FFA (100%, 10/10) and lowest repeatability in OCT

(40%, 4/10). The overall repeatability in the diagnosis of six ocular modality images was 63% (38/60).

Sentence Similarity

The overall sentence similarity between the responses of GPT-4V and human answers was 55.5% (**Supplemental Table 6**). Slit-lamp images got the highest score in sentence similarity (61.9%), while FFP images got the lowest (52.3%).

GPT-4-based Auto Evaluation

This benchmark achieved a Spearman correlation of 0.569 in accuracy and 0.576 in usability using the GPT-4-based auto evaluation. While the Kendall-Tau correlation was 0.456 in accuracy and 0.474 in usability (**Table 3**).

Discussion

In this study, we evaluated the capacity of GPT-4V to recognize, interpret, and make inferences based on multimodal ophthalmic images for the first time. While GPT-4V demonstrated a high level of performance in identifying the modality of ophthalmic images, its ability to recognize lesions, provide diagnoses, and offer decision support was found to be limited. The repeatability in diagnosis of ocular modality images was only 63%. The

automatic evaluation performance for open-ended questions is still poor in this benchmark. Despite its promising advancements, GPT-4V is not yet ready to support the generation of clinical decisions and patient consultations based on ophthalmic examination images in real-world settings.

GPT-4V currently is not ready for clinical use, as only 15.2% of responses were correct, highly usable, and no harm. The preprint study by Wu and colleagues demonstrated GPT-4V's ability to accurately identify different imaging modalities such as X-ray, CT, MRI, ultrasound, nuclear imaging, and pathology.[9] Our results on multimodal ophthalmic images align with these findings. Out of the 60 images we tested, only three slit-lamp images were identified as close-up photographs of the eye. Notably, two out of these three were diffuse illumination images without the presence of a slit beam, making this classification arguably acceptable. GPT-4V demonstrated varying levels of competency across different imaging modalities. In terms of accuracy, usability, and safety, both slit-lamp and SLO images were rated above average. The high performance on slit-lamp images could be attributable to the similarity these images bear to real-world visual observations, coupled with GPT-4V's pre-training on disease-specific ophthalmic datasets. GPT-4V's performance on FPP images was notably below average. Despite both FPP and SLO providing a direct visualization of fundus, the former only captures a 45-degree field of view of the posterior region. This limited coverage might not fully reveal

all pathological features, which could potentially hinder GPT-4V's ability to accurately recognize and interpret these images. For example, GPT-4V could detect a retinal detachment in SLO but not in FPP. Moreover, in our limited cases, GPT-4V produced several inaccurate descriptions of lesions. For instance, it mischaracterized a typical CRVO image in SLO as pigment changes and misdescribed extensive subretinal exudation in an FPP image of Coats disease as "extensive areas of opacification and discoloration in the retina." An OCT image of a macular hole without retinal detachment was incorrectly described as "subretinal fluid accumulation and macular detachment". In a slit-lamp image, the location of an eyelid nevus was erroneously identified as "on the sclera near the limbus". These inaccuracies underscore that GPT-4V's recognition of ophthalmic image anatomy and disease characteristics is not reliable. Notably, GPT-4V demonstrated a tendency to provide non-specific, albeit accurate, descriptions of lesions in the evaluation of FFA images. For example, it often generated responses such as, "The images show areas of hyperfluorescence which may suggest leakage, staining, pooling, or abnormal vessel growth, and hypofluorescence which may indicate blockage or non-perfusion." While this statement is generally correct for any FFA image as it essentially enumerates all possibilities of hyperfluorescence and hypofluorescence, it does not provide a specific analysis for the presented images. We therefore conclude that GPT-4V does

not currently possess the ability to analyze the specific localization of lesions in FFA images. It instead tends to produce generic responses that could be perceived as "boilerplate" answers.

The repeatability of GPT-4V in diagnosing ocular images was low, with only 63% (38/60) of the responses being consistent. GPT-4V showed the highest repeatability on FFA, but most of its responses were vague and generic. For example, it often gave answers like “the patterns could suggest a retinal vascular disorder such as diabetic retinopathy, retinal vein occlusion, or choroidal neovascularization from conditions like wet age-related macular degeneration”. These answers did not specify the exact diagnosis or the features that supported it. Therefore, GPT-4V performed poorly in accuracy.

The automatic evaluation performance for open-ended questions is still poor in this benchmark. Recently, automatic evaluation for Open-QA has got significant progress, especially with the use of LLM.[17 20] In this study, we got Spearman and Kendall-Tau correlation results similar to previous study,[18] which also indicates that automatic evaluation for Open-QA remains a major challenge today and there is still a significant research space for improvement. The reason for not conducting automated evaluation in terms of safety is that automated evaluation has not reached a level of reliability comparable to human evaluation yet.

Being the first known evaluation of GPT-4V's capabilities in processing ophthalmic images, our study adds valuable insights to the existing body of knowledge. In medical scenarios, GPT-4V appears to have a strong safeguard system in place to prevent it from making direct diagnoses.[9] In our study, when faced with relatively rare ophthalmic conditions, GPT-4V typically refrained from making explicit diagnoses, instead suggesting a few potential diseases. However, upon evaluation by clinical doctors, the diseases it suggested did not resemble the ground truth given from ophthalmologists. Consequently, its subsequent clinical decision-making suggestions, based on inaccurate or imprecise diagnoses, were unreliable. A model exhibiting higher performance may assist clinical doctors in decision support, interpreting medical conditions, and providing high-quality healthcare consultation in remote areas. One promising way of improving multimodal models is to build foundation models with self-supervised pretraining with a large amount of retinal images.[21] Another way to improve the model is to align the ophthalmic image features with medical texts.[10 22] Researchers may also improve their models via the application of transfer learning, fine-tuning, and mitigating hallucination techniques, or by incorporating knowledge enrichment methods like Retrieval-Augmented Generation (RAG).[23 24] However, before applying any model to clinical practice, detailed clinical trials should be conducted to assess its safety and efficacy. Even if ultimately used in clinical settings, it also

should be used under the supervision of human clinical doctors. This study provides a benchmark for further investigation in building large language models for processing ophthalmic multimodal images.

There are several limitations in our study. Firstly, we only assessed ten selected images for each specific modality, which may introduce a sample bias. Future studies could benefit from expanding and diversifying the dataset to better represent the variability seen in clinical practice. Currently, evaluation methods for the VQA capabilities of multi-modal large language models like GPT-4V are limited, especially for knowledge-intensive scenarios.[8 25] The establishment of benchmarks for these scenarios has yet to reach a consensus.[26] Secondly, in instances where the model's responses were incorrect, we did not attempt to further prompt GPT-4V to test its self-correction abilities. Exploring this aspect could be an interesting avenue for future research.

Overall, GPT-4V is not yet ready to generate clinical decisions or provide patient consultations based on images from ophthalmic examinations in real-world scenarios. Ongoing refinement and testing remain crucial to enhance the performance of large language models in ophthalmology.

Contributors PX, XC, ZZ and DS designed the study; DS provided data. PX, XC and ZZ analysed and interpreted the data; PX, XC and ZZ drafted the manuscript; DS revised the manuscript. DS is responsible for the overall content as the guarantor.

Funding This study was supported by the Start-up Fund for RAPs under the Strategic Hiring Scheme (P0048623) from HKSAR.

Competing interests There are no conflicts of interest to declare by the authors.

Patient consent for publication Not applicable.

Ethics approval This study was approved by the institutional review board of The Hong Kong Polytechnic University.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement The code and Question-Answering dataset will be available on Hugging Face (https://huggingface.co/datasets/Proxen/OMI). The image dataset is available by contacting with the corresponding author.

References

1. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Frontiers in Artificial Intelligence 2023;**6**:1169595 doi: 10.3389/frai.2023.1169595[published Online First: Epub Date]|.
2. Nath S, Marie A, Ellershaw S, Korot E, Keane PA. New meaning for NLP: the trials and tribulations of natural language processing with GPT-3 in ophthalmology. The British Journal of Ophthalmology 2022;**106**(7):889-92 doi: 10.1136/bjophthalmol-2022-321141[published Online First: Epub Date]|.
3. Raimondi R, Tzoumas N, Salisbury T, et al. Comparative analysis of large language models in the Royal College of Ophthalmologists fellowship exams. Eye 2023 doi: 10.1038/s41433-023-02563-3[published Online First: Epub Date]|.
4. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the Performance of ChatGPT in Ophthalmology. Ophthalmology Science 2023;**3**(4):100324 doi: 10.1016/j.xops.2023.100324[published Online First: Epub Date]|.
5. Momenaei B, Wakabayashi T, Shahlaee A, et al. Appropriateness and Readability of ChatGPT-4 generated Responses for Surgical Treatment of Retinal Diseases. Ophthalmology Retina 2023:S2468653023002464 doi: 10.1016/j.oret.2023.05.022[published Online First: Epub Date]|.
6. Lim ZW, Pushpanathan K, Yew SME, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine 2023;**95**
7. GPT-4V(ision) System Card. https://openai.com/research/gpt-4v-system-card
8. Li Y, Wang L, Hu B, et al. A Comprehensive Evaluation of GPT-4V on Knowledge-Intensive Visual Question Answering. arXiv preprint arXiv:2311.07536 2023
9. Wu C, Lei J, Zheng Q, et al. Can gpt-4v (ision) serve medical applications? case studies on gpt-4v for multimodal medical diagnosis. arXiv preprint arXiv:2310.09909 2023
10. Shi D, Chen X, Zhang W, et al. FFA-GPT: an Interactive Visual Question Answering System for Fundus Fluorescein Angiography. 2023
11. Hong Z, Deng Z, Evans R, Wu H. Patient questions and physician responses in a Chinese health Q&A website: Content analysis. Journal of medical Internet research 2020;**22**(4):e13071
12. Mahjoub H, Prabhu AV, Sikder S. What are Ophthalmology Patients Asking Online? An Analysis of the Eye Triage Subreddit. Clin Ophthalmol 2020;**14**:3575-82 doi: 10.2147/OPTH.S279607[published Online First: Epub Date]|.
13. Soong D, Sridhar S, Si H, et al. Improving accuracy of GPT-3/4 results on biomedical data using a retrieval-augmented language model. arXiv preprint arXiv:2305.17116 2023

14. Brooke J. Sus: a “quick and dirty’usability. Usability evaluation in industry 1996;**189**(3):189-94
15. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature 2023:1-9
16. Reimers N, Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 2019
17. Fu J, Ng S-K, Jiang Z, Liu P. Gptscore: Evaluate as you desire. arXiv preprint arXiv:2302.04166 2023
18. Liu Y, Iter D, Xu Y, Wang S, Xu R, Zhu C. Gpteval: Nlg evaluation using gpt-4 with better human alignment. arXiv preprint arXiv:2303.16634 2023
19. Imagenet: A large-scale hierarchical image database. 2009 IEEE conference on computer vision and pattern recognition; 2009. Ieee.
20. Wang C, Cheng S, Xu Z, Ding B, Wang Y, Zhang Y. Evaluating open question answering evaluation. arXiv preprint arXiv:2305.12421 2023
21. Zhou Y, Chia MA, Wagner SK, et al. A foundation model for generalizable disease detection from retinal images. Nature 2023;**622**(7981):156-63 doi: 10.1038/s41586-023-06555-x[published Online First: Epub Date]|.
22. Chen X, Xu P, Li Y, et al. ChatFFA: Interactive Visual Question Answering on Fundus Fluorescein Angiography Image Using ChatGPT. Available at SSRN 4578568
23. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems 2020;**33**:9459-74
24. Liu H, Xue W, Chen Y, et al. A survey on hallucination in large vision-language models. arXiv preprint arXiv:2402.00253 2024
25. Li Y, Liu Y, Wang Z, et al. A Comprehensive Study of GPT-4V's Multimodal Capabilities in Medical Imaging. medRxiv 2023:2023.11. 03.23298067
26. Liu Y, Duan H, Zhang Y, et al. MMBench: Is Your Multi-modal Model an All-around Player? arXiv preprint arXiv:2307.06281 2023

Figure Legend:

Figure 1. Ocular Imaging Modalities (A) and the Overview of This Study (B). SLO = Scanning Laser Ophthalmoscopy; FPP = Fundus Photography of the Posterior Pole; OCT = Optical Coherence Tomography; FFA = Fundus Fluorescein Angiography; OUS = Ocular Ultrasound. Slit-lamp images were captured by a combination of slit-lamp (BQ-900, Haag-Streit) and camera (Canon EOS 6). SLO images were obtained using an OPTOS nonmydriatic widefield camera (OPTOS Daytona, Dunefermline, UK). FPP images were captured using a Topcon fundus camera (TRC NW6; Topcon, Tokyo, Japan) with a 45-degree field of view. OCT images were obtained with a spectral-domain OCT (Spectralis OCT; Heidelberg Engineering, Heidelberg, Germany). FFA images were obtained using Zeiss FF450plus fundus camera. OUS images were obtained using a B-scan ophthalmic ultrasound machine (CineScan A/B; Quantel Medical, Bozeman, MT).

Figure 2. Stacked Bars of Accuracy, Usability and Safety in Ocular Multimodal Images.

Supplemental Figure 1. Screenshot of the process of building a digital ophthalmologist based on GPT-4V(ision).

Table 1. The Intragrader Agreement of three Ophthalmologists in evaluating the Performance of GPT-4V(ision) on Ocular Multimodal Images (600 QA pairs)

	Rater 1, N (%)[†]	Rater 2, N (%)	Rater 3, N (%)	Kappa
Accuracy	179 (29.8)	183 (30.5)	187 (31.2)	0.758
Usability	120 (22.2)	123 (22.7)	105 (19.4)	0.751
Safety	302 (55.9)	304 (56.3)	293 (54.3)	0.767

[†] Number and percentage of Correct/Highly Usable/No Harm were presented.

Table 2. Evaluation of Identification, Diagnosis, and Reasoning Capabilities of GPT4-V(ision)

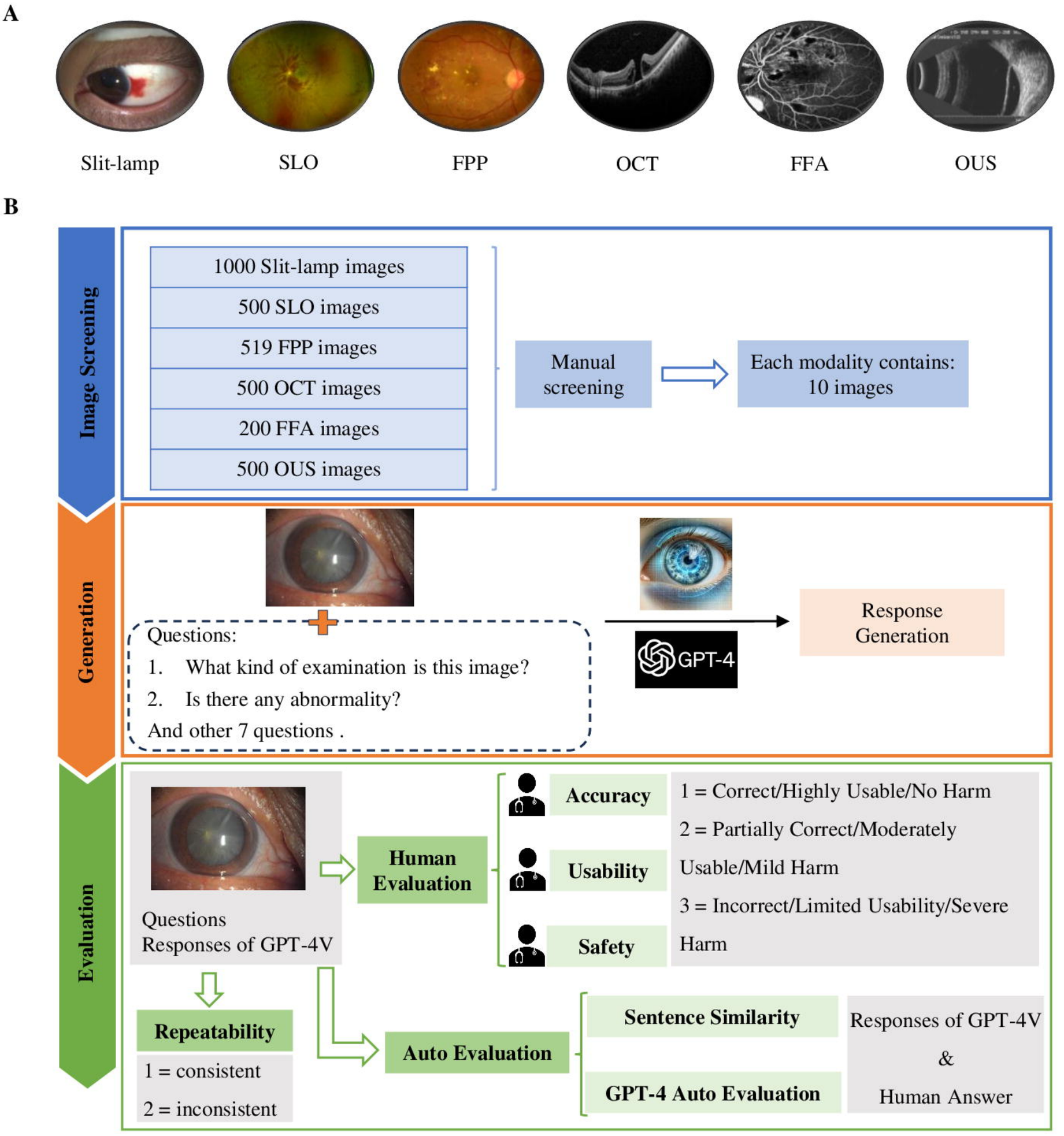
	Examination Identification (N = 60)	Lesion Identification (N = 60)	Diagnosis Capacity (N = 60)	Decision Support (N = 420)
Accuracy				
Correct, N (%)[†]	57.3 (95.6)	15.3 (25.6)	9.7 (16.1)	101.0 (24.0)
Partially Correct, N (%)	2.7 (4.4)	28.3 (47.2)	14.3 (23.9)	171.0 (40.7)
Incorrect, N (%)	0 (0)	16.3 (27.2)	36.0 (60.0)	148.0 (35.2)
Usability				
Highly Usable, N (%)	-	14.0 (23.3)	9.0 (15.0)	93.0 (22.1)
Moderately Usable, N (%)	-	26.0 (43.3)	14.3 (23.8)	171.3 (40.8)
Limited Usability, N (%)	-	20.0 (33.3)	36.7 (61.1)	155.7 (37.1)
Safety				
No Harm, N (%)	-	32.3 (53.9)	18.7 (31.1)	249.3 (59.4)
Mild Harm, N (%)	-	18.3 (30.6)	19.7 (32.8)	127.0 (30.2)
Severe Harm, N (%)	-	9.3 (15.6)	21.7 (36.1)	43.7 (10.4)

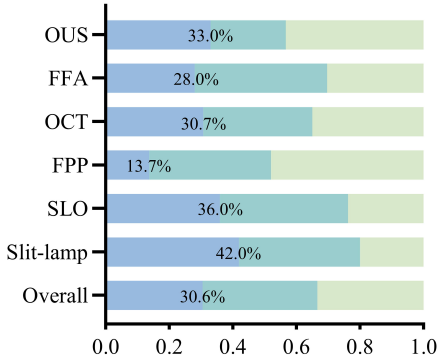
[†] N (%) showed in this table was the mean value of 3 raters.

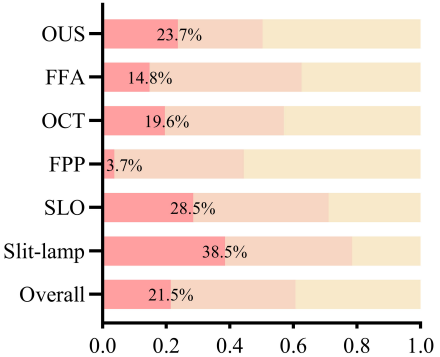
Table 3. Auto-evaluation by GPT-based method.

	Human Evaluation	GPT-4-based Auto Evaluation*	Spearman correlation	Kendall-Tau correlation
Accuracy			0.569	0.456
Correct, %†	30.6	15.9	-	-
Partially Correct, %	36.1	33.9	-	-
Incorrect, %	33.4	50.2	-	-
Usability			0.576	0.474
Highly Usable, %	21.5	12.4	-	-
Moderately Usable, %	39.2	25.6	-	-
Limited Usability, %	39.3	62.0	-	-

†The percentage shown in this table represents the average value of 3 raters in human evaluation and the average value of 20 rounds of auto-evaluation based on GPT-4. * The responses of GPT-4V and human were auto-rated by GPT-4-0613 using the same evaluation criteria in regards to accuracy and usability, and the resulting ratings were compared using the Spearman and Kendall-Tau correlations.


A
Slit-lamp
SLO
FPP
OCT
FFA
OUS
B
Image Screening
1000 Slit-lamp images
500 SLO images
519 FPP images
500 OCT images
200 FFA images
500 OUS images
Manual screening
Each modality contains: 10 images
Generation
Questions:
1. What kind of examination is this image?
2. Is there any abnormality?
And other 7 questions .
GPT-4
Response Generation
Evaluation
Questions
Responses of GPT-4V
Human Evaluation
Accuracy
Usability
Safety
1 = Correct/Highly Usable/No Harm
2 = Partially Correct/Moderately Usable/Mild Harm
3 = Incorrect/Limited Usability/Severe Harm
Repeatability
1 = consistent
2 = inconsistent
Auto Evaluation
Sentence Similarity
GPT-4 Auto Evaluation
Responses of GPT-4V
&
Human Answer


OUS
FFA
OCT
FPP
SLO
Slit-lamp
Overall
33.0%
28.0%
30.7%
13.7%
36.0%
42.0%
30.6%
0.0
0.2
0.4
0.6
0.8
1.0
Accuracy
Correct
Partially Correct
Incorrect


OUS
FFA
OCT
FPP
SLO
Slit-lamp
Overall
23.7%
14.8%
19.6%
3.7%
28.5%
38.5%
21.5%
0.0
0.2
0.4
0.6
0.8
1.0
Usability
Highly Usable
Moderately Usable
Limited Usability

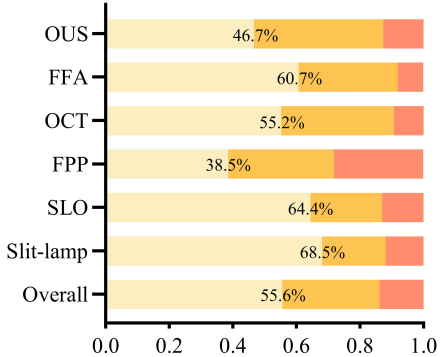

OUS
FFA
OCT
FPP
SLO
Slit-lamp
Overall
46.7%
60.7%
55.2%
38.5%
64.4%
68.5%
55.6%
0.0
0.2
0.4
0.6
0.8
1.0
Safety
No Harm
Mild Harm
Severe Harm

Table 1. The Intragrader Agreement of three Ophthalmologists in evaluating the Performance of GPT-4V(ision) on Ocular Multimodal Images (600 QA pairs)

	Rater 1, N (%)[†]	Rater 2, N (%)	Rater 3, N (%)	Kappa
Accuracy	179 (29.8)	183 (30.5)	187 (31.2)	0.758
Usability	120 (22.2)	123 (22.7)	105 (19.4)	0.751
Safety	302 (55.9)	304 (56.3)	293 (54.3)	0.767

[†] Number and percentage of Correct/Highly Usable/No Harm were presented.

Table 2. Evaluation of Identification, Diagnosis, and Reasoning Capabilities of GPT4-V(ision)

	Examination Identification (N = 60)	Lesion Identification (N = 60)	Diagnosis Capacity (N = 60)	Decision Support (N = 420)
Accuracy				
Correct, N (%)[†]	57.3 (95.6)	15.3 (25.6)	9.7 (16.1)	101.0 (24.0)
Partially Correct, N (%)	2.7 (4.4)	28.3 (47.2)	14.3 (23.9)	171.0 (40.7)
Incorrect, N (%)	0 (0)	16.3 (27.2)	36.0 (60.0)	148.0 (35.2)
Usability				
Highly Usable, N (%)	-	14.0 (23.3)	9.0 (15.0)	93.0 (22.1)
Moderately Usable, N (%)	-	26.0 (43.3)	14.3 (23.8)	171.3 (40.8)
Limited Usability, N (%)	-	20.0 (33.3)	36.7 (61.1)	155.7 (37.1)
Safety				
No Harm, N (%)	-	32.3 (53.9)	18.7 (31.1)	249.3 (59.4)
Mild Harm, N (%)	-	18.3 (30.6)	19.7 (32.8)	127.0 (30.2)
Severe Harm, N (%)	-	9.3 (15.6)	21.7 (36.1)	43.7 (10.4)

[†] N (%) showed in this table was the mean value of 3 raters.

Table 3. Auto-evaluation by GPT-based method.

	Human Evaluation	GPT-4-based Auto Evaluation*	Spearman correlation	Kendall-Tau correlation
Accuracy			0.569	0.456
Correct, %†	30.6	15.9	-	-
Partially Correct, %	36.1	33.9	-	-
Incorrect, %	33.4	50.2	-	-
Usability			0.576	0.474
Highly Usable, %	21.5	12.4	-	-
Moderately Usable, %	39.2	25.6	-	-
Limited Usability, %	39.3	62.0	-	-

†The percentage shown in this table represents the average value of 3 raters in human evaluation and the average value of 20 rounds of auto-evaluation based on GPT-4. * The responses of GPT-4V and human were auto-rated by GPT-4-0613 using the same evaluation criteria in regards to accuracy and usability, and the resulting ratings were compared using the Spearman and Kendall-Tau correlations.