Evaluation of a digital ophthalmologist app built by GPT4-Vision

Pusheng Xu, MD[1,2#], Xiaolan Chen, MD[1,2#], Ziwei Zhao[1#], MD, Yingfeng Zheng, MD, PhD[2], Guangming Jin, MD, PhD[2], Danli Shi, MD, PhD[1,3*], Mingguang He, MD, PhD[1,3]

Affiliations

1. School of Optometry, The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China.
2. State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Guangzhou 510060, China.
3. Research Centre for SHARP Vision, The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China

# Contributed equally

Correspondence

***Dr Danli Shi,** MD, PhD., The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China.

Email: danli.shi@polyu.edu.hk

Funding: This study was supported by the Global STEM Professorship Scheme (P0046113).

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

Word count: Abstract: 313; **Whole paper**: 3222.

Tables: 4; Figures: 2; **Supplemental Tables**: **4. Supplemental Figures: 1**

Abstract

Backgrounds: GPT4-Vision (GPT4V) has generated great interest across various fields, while its performance in ocular multimodal images is still unknown. This study aims to evaluate the capabilities of a GPT4V-based chatbot in addressing queries related to ocular multimodal images.

Methods: A digital ophthalmologist app was built based on GPT4V. The evaluation dataset comprised various ocular imaging modalities: 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). Each modality included images representing 5 common and 5 rare diseases. The chatbot was presented with ten questions per image, focusing on examination identification, lesion detection, diagnosis, decision support, and the repeatability of diagnosis. The responses of GPT4V were evaluated based on accuracy, usability, and safety.

Results: There was a substantial agreement among three ophthalmologists. Out of 600 responses, 30.5% were accurate, 22.8% of 540 responses were highly usable, and 55.5% of 540 responses were

considered safe by ophthalmologists. The chatbot excelled in interpreting slit-lamp images, with 42.0%, 42.2%, and 68.5% of the responses being accurate, highly usable, and no harm, respectively. However, its performance was notably weaker in FPP images, with only 13.7%, 3.7%, and 38.5% in the same categories. It correctly identified 95.6% of the imaging modalities. For lesion identification, diagnosis, and decision support, the chatbot's accuracy was 25.6%, 16.1%, and 24.0%, respectively. The average proportions of correct answers, highly usable, and no harm for GPT4V in common diseases were 37.9%, 30.5%, and 60.1%, respectively. These proportions were all higher compared to those in rare diseases, which were 23.2% ($P<0.001$), 15.2% ($P<0.001$), and 51.1% ($P=0.032$), respectively. The overall repeatability of GPT4-V in diagnosing ocular images was 63% (38/60).

Conclusion: Currently, GPT4V lacks the reliability required for clinical decision-making and patient consultation in ophthalmology. Ongoing refinement and testing are essential for improving the efficacy of large language models in medical applications.

Keywords: GPT4V, LLM, image, evaluation, ophthalmology

Backgrounds

With the rapid development of artificial intelligence (AI), large language models (LLM) have brought immense potential and opportunities to various fields, particularly in the medical domain.[1] Medicine always involves imaging examinations, especially in the field of 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-4 Vision (GPT4V) has brought new hope to the medical field.[7] GPT4V is a multimodal generalist LLM that can process both images and text, enabling various downstream tasks, including visual question answering (VQA). This means that GPT4V can understand and answer image-related questions, providing more accurate and comprehensive information both for doctors and patients. Li et al.[8] indicated that GPT4V exhibited significant advantages in fine-grained world-knowledge-intensive VQA. Another study evaluated the performance of GPT4V 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 GPT4V 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 GPT4V in handling queries related to multimodal eye images. We have designed a multimodal ophthalmic benchmark and primarily focused on the accuracy, comprehensiveness, and safety of the model. First, we evaluate its ability to recognize different modalities of eye images. Second, we explore its capacity to interpret eye images through various VQA tasks, including diagnosis, lesion identification, and decision support. This research provides valuable insights into the effectiveness and potential pitfalls of utilizing multimodal LLMs to address image-related questions in the field of ophthalmology and also paves the way for augmenting their capabilities and credibility.

Methods

Data

To avoid overlap between the evaluation set and the training set of GPT4V, we retrospectively collected ocular multimodal images (**Figure 1A**) from Zhongshan Ophthalmic Center, images were taken between 2016 and 2019. This study adheres to the tenets of the Declaration of Helsinki. The ethics committee of the Zhongshan Ophthalmic Center approved the study (No.2021KYPJ164-3) and individual consent for this retrospective analysis was waived.

These ocular multimodal images included 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. 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).

We manually screened the images and selected 10 images (5 common diseases and 5 rare diseases) from each type of image as the evaluation dataset. (**Figure 1B**) For example, 5 slit-lamp images of cataract, pterygium, subconjunctival hemorrhage, conjunctivitis, and keratitis were included in the common disease images. Additionally, another 5 slit-lamp images of corneal leucoma, dislocation of intraocular lens, corneal dermoid, anterior uveitis, and eyelid mass were included in the rare disease images (**Supplemental Table 1**). The images showed typical manifestations of the diseases that ophthalmologists would use to make a probable diagnosis.

App construction using GPT4V

Since GPT4V often refuses 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 based on GPT4V using the prompt provided in **Supplemental Figure 1**.

Generation of responses

Ten questions were constructed to evaluate the performance of this digital ophthalmologist in four dimensions, including examination identification, lesion identification, diagnosis capacity, and decision support (**Supplemental Table 2)**. The first three dimensions each consist of one question, while the decision support comprises seven questions related to the next examination, treatment, severity, complication, etiology, prognosis, and prevention. An example of question-answer on ocular images using GPT4V was shown in **Supplemental Table 3.**

Evaluation

The human evaluation was conducted by three experienced ophthalmologists (PX, XC and ZZ), each with more than five years of clinical experience. We evaluated the responses generated by the digital ophthalmologist in three key dimensions: accuracy, usability, safety, and repeatability in diagnosis (**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 GPT4V: 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.[11]

Evaluating the usability of a response could be quite difficult since there are no absolute measures of usability.[12] 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.[13] 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. We used a two-point scale 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.

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 between common diseases and rare diseases with $P < 0.05$ considered statistically significant.

Results

Overall performance

Table 1 summarizes the performance of GPT4V in terms of accuracy, usability, and safety based on the proportion of responses rated as good (score of 1) across 600 questions related to multimodal ophthalmic images. Fleiss' Kappa values indicated good consistency among the three raters, with 0.758 for accuracy, 0.749 for usability, and 0.767 for safety. Overall, the performance of GPT4V in terms of accuracy and usability was not ideal, with only 30.6% (183.3/600) of responses considered correct by ophthalmologists and 22.8% (123.3/540) deemed highly usable. On the other hand, safety evaluations were relatively good, with approximately half of the answers (55.6% (300.3/540)) considered no harm (**Figure 2**).

Multimodal performance

As shown in **Figure 2**, the model performed best with slit-lamp images, with average proportions of correct, highly usable, and harmless answers being 42.0% (42.0/100), 42.2% (38.0/90), and 68.5% (61.7/90), respectively. However, for fundus images, particularly FPP, GPT4V's performance was less

satisfactory, with only 13.7% (13.7/100), 3.7% (3.3/90), and 38.5% (34.7/90) of answers rated as correct, highly usable, and no harm, respectively. In contrast, for fundus images with a wider field of view, such as SLO images, GPT4V received relatively better evaluations (36.0% accuracy, 29.3% high usefulness, and 64.4% harmless). For detailed evaluations of the six modalities, please refer to **Supplemental Table 4**.

Image interpretation performance

We assessed the ability of GPT4V to interpret ophthalmic images using various VQA tasks, including examination identification, lesion recognition, diagnostic capacity, and decision support (see **Table 2**). Overall, GPT4V was able to correctly identify the imaging modality of most ophthalmic images (average 95.6% (57.3/60)). However, there were instances where answers were partially correct (average 4.4% (2.7/60)), such as describing slit-lamp photographs as close-up photograph of the eyes.

Lesion identification

Lesion recognition is crucial for extracting disease-related information from ophthalmic images, corresponding to the ability to generate reports in real clinical settings. We explored GPT4V's performance in this aspect by asking, "Is there any abnormality?" The results showed that 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 30% (18/60) of answers provided useful recognition of the image's lesions, 36.7% (22/60) were moderately usable, and 33.3% (20/60) 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

This refers to the ability to assist clinical decision-making using information from ophthalmic images and reflects the model's reasoning capability in the medical domain. We explored the model's medical reasoning performance by asking GPT4V, "What is the diagnosis?" 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,[14] 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 GPT4V rarely provided direct diagnoses but instead listed a series of potential diagnoses for reference, which is consistent with clinical practice. However, GPT4V still fell far short of meeting the requirements for clinical diagnosis, as it only presented patterned diagnoses based on general medical knowledge rather than candidate diagnoses specific for the input images. Therefore, in terms of accuracy, usefulness, and safety, only 16.1% (9.7/60), 17.2% (10.3/60), 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 (see **Supplementary Table 2**) to test this. The results showed that GPT4V only met clinical standards

(score of 1) in terms of accuracy and usefulness in 24% (101.0/420) and 22.6% (95.0/420) of responses, respectively. In terms of safety, 59.4% (249.3/420) were considered harmless.

Common disease vs. rare disease

We further analyzed the performance of GPT4V in relation to VQA for diseases of different difficulty levels. As shown in **Table 3**, the average proportions of correct answers, highly usable, and no harm for GPT4V in common diseases were 37.9% (113.7/300), 30.5% (82.3/270), and 60.1% (162.3/270), respectively. These proportions were all higher compared to the average proportions observed for rare diseases, which were 23.2% (69.7/300, P<0.001), 15.2% (41.0/270, P<0.001), and 51.1% (138.0/270, P=0.032), respectively. This suggests that there is great room for improvement in the response of such models to rare diseases.

Evaluation of Repeatability in Diagnosis

The responses of GPT4V had highest repeatability in FFA (100%, 10/10) and lowest repeatability in OCT (40%, 4/10). The overall repeatability in diagnosis of six ocular modality images was 63% (38/60).

Discussion

In this study, we evaluated the capacity of GPT4V to recognize, interpret, and make inferences based on multimodal ophthalmic examination images for the first time. While GPT4V 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%. Despite its promising advancements, GPT4V is not yet ready to support the generation of clinical decisions and patient consultations based on ophthalmic examination images in real-world settings.

The preprint study by Wu and colleagues demonstrated GPT4V'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. GPT4V 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 GPT4V's pre-training on disease-specific ophthalmic datasets. GPT4V's performance on FPP images was notably below average. Despite both FPP and SLO providing a direct visualization of the 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 GPT4V's ability to accurately recognize and interpret these images. For example, GPT4V could detect a retinal detachment in SLO but not in FPP. Moreover, in our limited cases, GPT4V 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 GPT4V's recognition of ophthalmic image anatomy and disease characteristics is not reliable. Notably, GPT4V 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 GPT4V 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. Previous studies have revealed that GPT-4 demonstrates commendable ability in diagnosing rare ophthalmic diseases when provided with chief complaints, medical history, ophthalmologic examination, and descriptions of ophthalmic imaging.[15] Diagnosing based solely on a single image input presents a markedly different and more challenging endeavor. This is evidenced by a significant decrease in the accuracy, usability, and safety of GPT4V in diagnosing rare ophthalmic diseases compared to common ones. This may be due to the inherent complexity and variability of rare diseases, which make their diagnosis intrinsically challenging. It may also be attributable to the limited exposure of GPT4V to rare disease samples during its training phase. In fine-tuning other open-source models, these issues could potentially be addressed by training with more balanced datasets or utilizing methods such as transfer learning or meta-learning, or using external knowledge database to give a

response.

The repeatability of GPT4V in diagnosing ocular images was low, with only 63% (38/60) of the responses being consistent. GPT4V 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, GPT4V performed poorly in accuracy.

Being the first known evaluation of GPT4V's capabilities in processing ophthalmic images, our study adds valuable insights to the existing body of knowledge. Public and researchers' initial comprehension of GPT4V's capabilities largely stems from a handful of preprint studies currently available, given that GPT4V has been released for just over two months. Wu et al. assessed images across eight modalities from 17 human body systems, concluding that while GPT4V is proficient at recognizing image modality and anatomical structures, it faces significant challenges in diagnosing diseases and generating comprehensive reports.[9] This echoes a similar trend observed in GPT-4, the precursor of GPT4V, which exhibited lower performance in the areas of treatment and prevention in a myopia consultation context.[6] This also aligns with our findings - over 60% of GPT4V's responses in disease diagnosis were deemed incorrect and of limited usability. In medical scenarios, GPT4V 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, GPT4V 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. In the medical field, this represents a major limitation.

There are several limitations to our study. Firstly, our evaluation of the VQA capabilities for multimodal ophthalmic images was broad, but we only assessed ten selected images and their corresponding ten QA pairs for each specific modality. This sample size is not reflective of the real-world clinical scenarios, and hence, introduces 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 GPT4V are limited, especially for knowledge-intensive scenarios.[8, 16] The establishment of benchmarks for these scenarios has yet to reach a consensus.[17] In our study, most of the questions were open-ended, and we used the consensus of clinical doctors as the ground truth, making it challenging to evaluate the model using objective text-matching metrics such as BLEU, CIDER and METEOR. Consequently, we relied heavily on subjective human evaluation. Finally, in instances where the model's responses were incorrect, we did not attempt to further prompt GPT4V to test its self-correction abilities. Exploring this aspect could be an interesting avenue for future research.

Conclusions

Overall, GPT4V 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 medicine.

List of abbreviations

AI: Artificial Intelligence

FFA: Fundus Fluorescein Angiography

FPP: Fundus photography of the posterior pole

GPT4V: GPT4-Vision

LLM: Large Language Models

OCT: Optical Coherence Tomography

OUS: Ocular ultrasound

SLO: Scanning Laser Ophthalmoscopy

VQA: Visual question answering

Declarations

Ethics approval and consent to participate: This study adheres to the tenets of the Declaration of Helsinki. The ethics committee of the Zhongshan Ophthalmic Center approved the study (No.2021KYPJ164-3) and individual consent for this retrospective analysis was waived.

Consent for publication: Not applicable.

Availability of data and materials: The data of this study are available from the corresponding author on

reasonable request.

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

Funding: This study was supported by the Global STEM Professorship Scheme (P0046113).

Authors' contributions: PX, XC, ZZ, DS and MH designed the study; YZ collected the data; PX, XC and ZZ analysed and interpreted the data; PX, XC and ZZ drafted the manuscript; GJ, DS and MH revised the manuscript. DS is responsible for the overall content as the guarantor.

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.
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):889-892.
3. Raimondi R, Tzoumas N, Salisbury T, Di Simplicio S, Romano MR, (NETRiON) NETRiON, Bommireddy T, Chawla H, Chen Y, Connolly S *et al*: **Comparative analysis of large language models in the Royal College of**

Ophthalmologists fellowship exams. *Eye* 2023.

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.

5. Momenaei B, Wakabayashi T, Shahlaee A, Durrani AF, Pandit SA, Wang K, Mansour HA, Abishek RM, Xu D, Sridhar J *et al*: **Appropriateness and Readability of ChatGPT-4 generated Responses for Surgical Treatment of Retinal Diseases**. *Ophthalmology Retina* 2023:S2468653023002464.

6. Lim ZW, Pushpanathan K, Yew SME, Lai Y, Sun C-H, Lam JSH, Chen DZ, Goh JHL, Tan MCJ, Sheng B: **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://openaicom/research/gpt-4v-system-card*.

8. Li Y, Wang L, Hu B, Chen X, Zhong W, Lyu C, Zhang M: **A Comprehensive Evaluation of GPT-4V on Knowledge-Intensive Visual Question Answering**. *arXiv preprint arXiv:231107536* 2023.

9. Wu C, Lei J, Zheng Q, Zhao W, Lin W, Zhang X, Zhou X, Zhao Z, Zhang Y, Wang Y: **Can gpt-4v (ision) serve medical applications? case studies on gpt-4v for multimodal medical diagnosis**. *arXiv preprint arXiv:231009909* 2023.

10. Shi D, Chen X, Zhang W, Xu P, Zhao Z, Zheng Y, He M: **FFA-GPT: an**

Interactive Visual Question Answering System for Fundus Fluorescein Angiography. 2023.

11. Soong D, Sridhar S, Si H, Wagner J-S, Sá ACC, Yu CY, Karagoz K, Guan M, Hamadeh H, Higgs BW: **Improving accuracy of GPT-3/4 results on biomedical data using a retrieval-augmented language model**. *arXiv preprint arXiv:230517116* 2023.
12. Brooke J: **Sus: a "quick and dirty'usability**. *Usability evaluation in industry* 1996, **189**(3):189-194.
13. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S: **Large language models encode clinical knowledge**. *Nature* 2023:1-9.
14. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L: **Imagenet: A large-scale hierarchical image database**. In: *2009 IEEE conference on computer vision and pattern recognition: 2009*: Ieee; 2009: 248-255.
15. Hu X, Ran AR, Nguyen TX, Szeto S, Yam JC, Chan CKM, Cheung CY: **What can GPT-4 do for Diagnosing Rare Eye Diseases? A Pilot Study**. *Ophthalmology and Therapy* 2023, **12**(6):3395-3402.
16. Li Y, Liu Y, Wang Z, Liang X, Liu L, Wang L, Cui L, Tu Z, Wang L, Zhou L: **A Comprehensive Study of GPT-4V's Multimodal Capabilities in Medical Imaging**. *medRxiv* 2023:2023.2011. 2003.23298067.
17. Liu Y, Duan H, Zhang Y, Li B, Zhang S, Zhao W, Yuan Y, Wang J, He C, Liu Z:

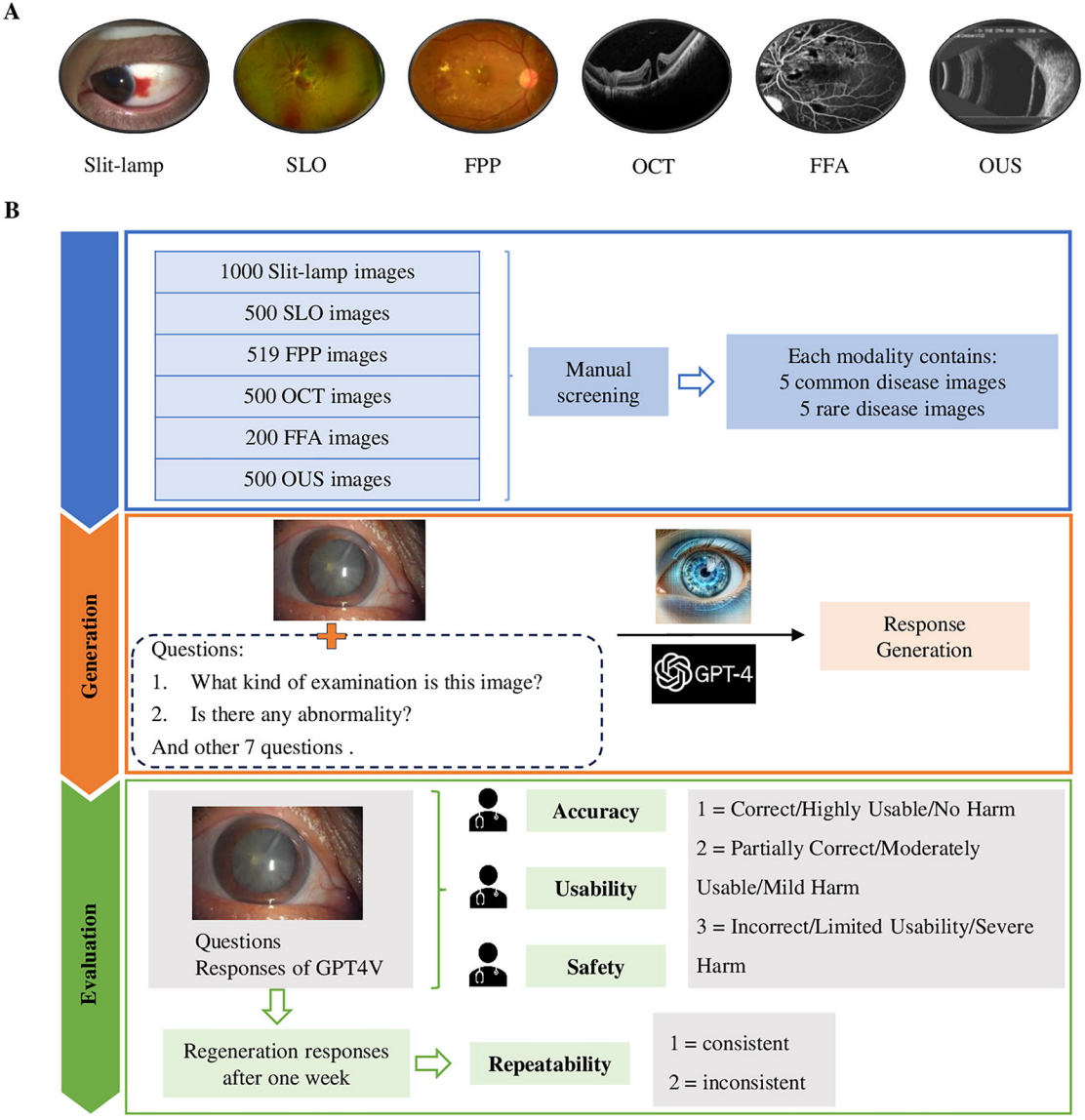
MMBench: Is Your Multi-modal Model an All-around Player? *arXiv preprint arXiv:230706281* 2023.

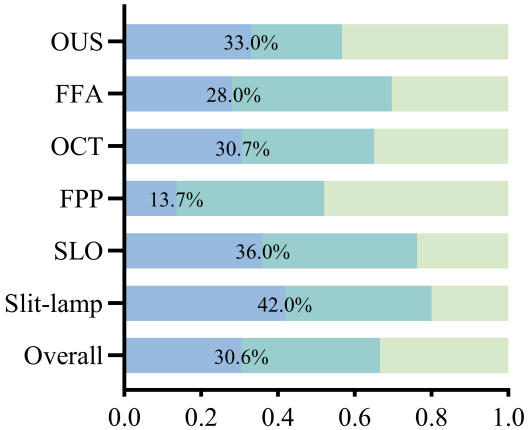
Figure legends

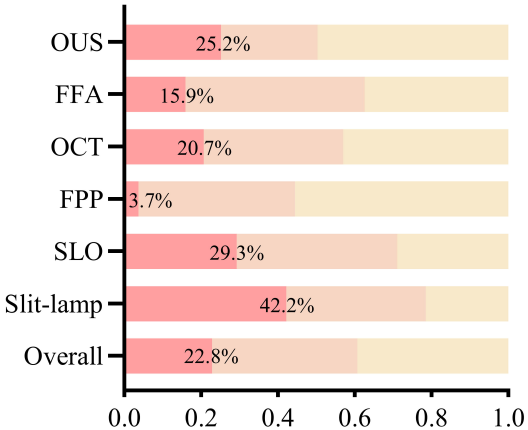
Figure 1. Ocular imaging modalities (A) and the overview of this study (B). SLO = Scanning Laser Ophthalmoscopy; FPP = Fundus Photography of the Posterior Pole; OCT = Optical Coherence Tomography; FFA = Fundus Fluorescein Angiography; OUS = Ocular Ultrasound.

Figure 2. Stacked bars of accuracy, usability and safety in different ocular multimodal images.

Supplemental Figure 1. Screenshot of the process of building a digital ophthalmologist based on GPT4V.


A
Slit-lamp
SLO
FPP
OCT
FFA
OUS
B
1000 Slit-lamp images
500 SLO images
519 FPP images
500 OCT images
200 FFA images
500 OUS images
Manual screening
Each modality contains:
5 common disease images
5 rare disease 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 GPT4V
Accuracy
Usability
Safety
1 = Correct/Highly Usable/No Harm
2 = Partially Correct/Moderately Usable/Mild Harm
3 = Incorrect/Limited Usability/Severe Harm
Regeneration responses after one week
Repeatability
1 = consistent
2 = inconsistent


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
25.2%
15.9%
20.7%
3.7%
29.3%
42.2%
22.8%
0.0
0.2
0.4
0.6
0.8
1.0
Usability
Highly Usable
Moderately Usable
Limited Usability

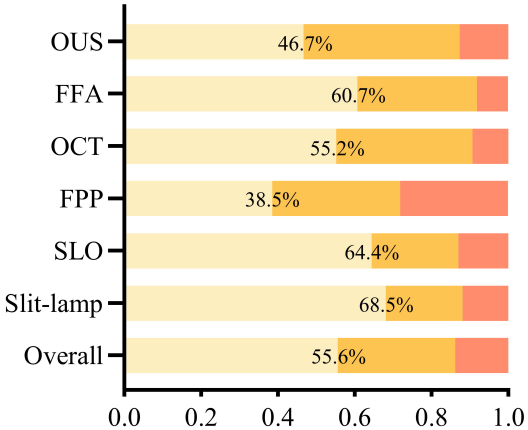

OUS
FFA
OCT
FPP
SLO
Slit-lamp
Overall
46.7%
60.7%
55.2%
38.5%
64.4%
68.5%
55.6%
0.0
0.2
0.4
0.6
0.8
1.0
Safety
No Harm
Mild Harm
Severe Harm

Table 1. The Intragrader Agreement of three Ophthalmologists in evaluating GPT4V's Performance 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	129 (23.9)	130 (24.1)	110 (20.4)	0.749
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

	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 (%)	-	18.0 (30.0)	10.3 (17.2)	95.0 (22.6)
Moderately Usable, N (%)	-	22.0 (36.7)	13.0 (21.7)	169.3 (40.3)
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. Evaluation of GPT4V's Performance in Common Disease and Rare Disease

	Common Disease (N = 300)	Rare Disease (N = 300)	P Value
Accuracy			
Correct, N (%) †	113.7 (37.9)	69.7 (23.2)	
Partially Correct, N (%)	97.7 (32.6)	118.7 (39.6)	<0.001*
Incorrect, N (%)	88.7 (29.6)	111.7 (37.2)	
Usability			
Highly Usable, N (%)	82.3 (30.5)	41.0 (15.2)	
Moderately Usable, N (%)	93.0 (34.4)	111.3 (41.2)	<0.001*
Limited Usability, N (%)	94.7 (35.1)	117.7 (43.6)	
Safety			
No Harm, N (%)	162.3 (60.1)	138.0 (51.1)	
Mild Harm, N (%)	79.7 (29.5)	85.3 (31.6)	0.032*
Severe Harm, N (%)	28.0 (10.4)	46.7 (17.3)	

† N (%) showed in this table was the mean value of 3 raters. *The chi-square test was used to compare the difference between common diseases and rare diseases with $P < 0.05$ considered statistically significant.

Table 4. Evaluation of Repeatability in Diagnosis

Modality	Repeatability, %(consistent response/total)
Slit-lamp	70 (7/10)
SLO	50 (5/10)
FPP	70 (7/10)
OCT	40 (4/10)
FFA	100 (10/10)
OUS	50 (5/10)
Overall	63 (38/60)

SLO = Scanning Laser Ophthalmoscopy; FPP = Fundus Photography of the Posterior Pole; OCT = Optical Coherence Tomography; FFA = Fundus Fluorescein Angiography; OUS = Ocular Ultrasound.

Digital Ophthalmologist
Draft

Save

Create | Configure

GPT Builder
Hi! I'll help you build a new GPT. You can say something like, "make a creative who helps generate visuals for new products" or "make a software engineer who helps format my code."

What would you like to make?

You
You are a helpful assistant to ophthalmologists, capable of analyzing patients' examination images, accurately identifying lesions, making potential diagnoses, and providing answers to inquiries. Your responses should be concise and precise, ensuring all questions are addressed thoroughly.

GPT Builder
Great! Let's name this GPT. How about calling it **"Eye Insight"**? Does that sound good to you?

You

Message GPT Builder...

Preview

Digital Ophthalmologist
Assistant to ophthalmologists for image analysis and potential diagnoses.

Message Digital Ophthalmologist...

Supplemental Table 1. Diseases Category in the Evaluation Dataset.

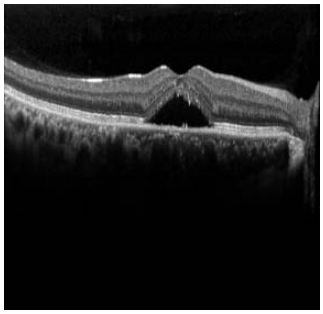
Modality	Common Disease	Rare Disease
Slit-lamp	Cataract, pterygium, subconjunctival hemorrhage, conjunctivitis, and keratitis	Corneal leucoma, dislocation of intraocular lens, corneal dermoid, anterior uveitis, and eyelid mass
SLO	Central retinal vein occlusion, proliferative diabetic retinopathy, rhegmatogenous retinal detachment, macular hole, and pathologic myopia	Central serous chorioretinopathy, retinitis pigmentosa, branch retinal artery occlusion, choroidal coloboma, and Coat's disease
FPP	Pathologic myopia, branch retinal vein occlusion, recent retinal detachment, nonproliferative diabetic retinopathy, and old tractional retinal detachment	Coat's disease, choroidal coloboma, intraocular lens dislocation into vitreous cavity, myelinated retinal nerve fiber, and suspected choroidal melanoma
OCT	Macular hole, macular epiretinal membrane, macular cystoid edema, retinal detachment and age-related macular degeneration	Polypoidal choroidal vasculopathy, central serous chorioretinopathy, choroidal neovascularization, high myopia-related retinal schisis and retinal pigment epithelial detachments
FFA	Proliferative diabetic retinopathy, central retinal vein occlusion, central retinal artery occlusion, age-related macular degeneration, and pathologic myopia	Coat's disease, central serous chorioretinopathy, retinitis pigmentosa, familial exudative vitreoretinopathy, and intermediate uveitis
OUS	Vitreous opacities, silicone oil-filled eye, retinal detachment, choroidal detachment, and posterior scleral staphyloma	Tenon's capsule edema, foreign body in the vitreous cavity, solid orbital mass, solid intraocular mass with retinal detachment and retinoblastoma

SLO = Scanning Laser Ophthalmoscopy; FPP = Fundus Photography of the Posterior Pole; OCT = Optical Coherence Tomography; FFA = Fundus Fluorescein Angiography; OUS = Ocular Ultrasound.

Supplemental Table 2. Question Lists.

Capability	Questions
Examination Identification	What kind of examination is this image?
Lesion Identification	Is there any abnormality?
Diagnosis capacity	What is the diagnosis?
	What examinations need to be done next?
	What might be the treatment options for this condition?
	Can this condition cause blindness?
Decision Support	What are some common complications that can occur?
	What might be the cause of this condition?
	Will this condition progress?
	How can I prevent this condition?

Supplemental Table 3. Example Question-Answer on Ocular Images Using GPT4V.

Image	Question	Answer of GPT4V
	What kind of examination is this image?	This image is from an Optical Coherence Tomography (OCT) examination, which is commonly used to capture high-resolution cross-sectional images of the retina.
	Is there any abnormality?	Yes, there is an abnormality. The image shows a small, dark, dome-shaped area in the central retina which is consistent with fluid accumulation beneath the retinal layers.
	What is the diagnosis?	The diagnosis suggested by this OCT image could be central serous chorioretinopathy (CSR), which is characterized by the accumulation of subretinal fluid in the macula.
	What examinations need to be done next?	Additional examinations could include a fluorescein angiography (FA) to detect any leaking areas that may be contributing to the fluid accumulation and to observe the overall retinal circulation. An Amsler grid test may also be beneficial to determine the extent of central visual field distortion.
	What might be the treatment options for this condition?	Treatment for CSR often begins with observation as many cases resolve on their own. If treatment is necessary, options may include laser therapy or photodynamic therapy to reduce leakage and promote fluid resorption.
	Can this condition cause blindness?	CSR typically affects central vision and can lead to significant vision impairment if chronic or recurrent. However, it does not usually result in complete blindness as peripheral vision is generally not affected.
	What are some common complications that can occur?	Common complications from CSR can include persistent central vision loss, recurrent episodes, and in some cases, permanent retinal changes that may affect vision long-term.
	What might be the cause of this condition?	The cause of CSR is not fully understood, but it is thought to be associated with stress, corticosteroid use, and sometimes Type A personality traits.
	Will this condition progress?	CSR can progress, with some cases resolving spontaneously within months, while others may develop chronic or recurrent fluid accumulation.
	How can I prevent this condition?	Preventing CSR involves managing stress, avoiding corticosteroid use if possible, and addressing any associated systemic health issues. Regular eye exams are essential for individuals with risk factors for early detection and management.

Supplemental Table 4. Evaluation of GPT4V's Performance in Ocular Multimodal Images

	Overall	Slit-lamp	SLO	FPP	OCT	FFA	OUS
Accuracy							
Correct, N(%)[†]	183.3(30.6)	42.0(42.0)	36.0(36.0)	13.7(13.7)	30.7(30.7)	28.0(28.0)	33.0(33.0)
Partially Correct, N(%)	216.3(36.1)	38.0(38.0)	40.3(40.3)	38.3(38.3)	34.3(34.3)	41.7(41.7)	23.7(23.7)
Incorrect, N(%)	200.3(33.4)	20.0(20.0)	23.7(23.7)	48.0(48.0)	35.0(35.0)	30.3(30.3)	43.3(43.3)
Usability							
Highly Usable, N(%)	123.3(22.8)	38.0(42.2)	26.3(29.3)	3.3(3.7)	18.7(20.7)	14.3(15.9)	22.7(25.2)
Moderately Usable, N(%)	204.3(37.8)	32.7(36.3)	37.7(41.9)	36.7(40.7)	32.7(36.3)	42.0(46.7)	22.7(25.2)
Limited Usability, N(%)	212.3(39.3)	19.3(21.5)	26.0(28.9)	50.0(55.6)	38.7(43.0)	33.7(37.4)	44.7(49.6)
Safety							
No Harm, N(%)	300.3(55.6)	61.7(68.5)	58.0(64.4)	34.7(38.5)	49.7(55.2)	54.7(60.7)	42.0(46.7)
Mild Harm, N(%)	165.0(30.6)	17.7(19.6)	20.3(22.6)	30.0(33.3)	32.0(35.6)	28.0(31.1)	36.7(40.7)
Severe Harm, N(%)	74.7(13.8)	10.7(11.9)	11.7(13.0)	25.3(28.1)	8.3(9.3)	7.3(8.1)	11.3(12.6)

[†] N (%) showed in this table was the mean value of 3 raters. SLO = Scanning Laser Ophthalmoscopy; FPP = Fundus Photography of the Posterior Pole; OCT = Optical Coherence Tomography; FFA = Fundus Fluorescein Angiography; OUS = Ocular Ultrasound.