BMJ Open Ophthalmology

Can off-the-shelf visual large language models detect and diagnose ocular diseases from retinal photographs?

Sahana Srinivasan ,[1,2,3] Hongwei Ji ,[4] David Ziyou Chen ,[1,5] Wendy Wong ,[1,5] Zhi Da Soh ,[2] Jocelyn Hui Lin Goh ,[1,2,3] Krithi Pushpanathan ,[1,3] Xiaofei Wang ,[6] Weizhi Ma ,[7] Tien Yin Wong ,[2,4,8,9] Ya Xing Wang ,[8,9] Ching-Yu Cheng ,[1,2,3,10] Yih Chung Tham [1,2,3,10]

To cite: Srinivasan S, Ji H, Chen DZ, *et al*. Can off-the-shelf visual large language models detect and diagnose ocular diseases from retinal photographs? *BMJ Open Ophthalmology* 2025;**10**:e002076. doi:10.1136/bmjophth-2024-002076

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjophth-2024-002076).

SS and HJ contributed equally.

SS and HJ are joint first authors.

Received 29 November 2024
Accepted 14 February 2025

For numbered affiliations see end of article.

Correspondence to
Dr Yih Chung Tham; thamyc@nus.edu.sg

ABSTRACT

Background The advent of generative artificial intelligence has led to the emergence of multiple vision large language models (VLLMs). This study aimed to evaluate the capabilities of commonly available VLLMs, such as OpenAI's GPT-4V and Google's Gemini, in detecting and diagnosing ocular diseases from retinal images.

Methods and analysis From the Singapore Epidemiology of Eye Diseases (SEED) study, we selected 44 representative retinal photographs, including 10 healthy and 34 representing six eye diseases (age-related macular degeneration, diabetic retinopathy, glaucoma, visually significant cataract, myopic macular degeneration and retinal vein occlusion). OpenAI's GPT-4V (both default and data analyst modes) and Google Gemini were prompted with each image to determine if the retina was normal or abnormal and to provide diagnostic descriptions if deemed abnormal. The outputs from the VLLMs were evaluated for accuracy by three attending-level ophthalmologists using a three-point scale (poor, borderline, good).

Results GPT-4V default mode demonstrated the highest detection rate, correctly identifying 33 out of 34 detected correctly (97.1%), outperforming its data analyst mode (61.8%) and Google Gemini (41.2%). Despite the relatively high detection rates, the quality of diagnostic descriptions was generally suboptimal—with only 21.2% of GPT-4V's (default) responses, 4.8% of GPT-4V's (data analyst) responses and 28.6% for Google Gemini's responses rated as good.

Conclusions Although GPT-4V default mode showed generally high sensitivity in abnormality detection, all evaluated VLLMs were inadequate in providing accurate diagnoses for ocular diseases. These findings emphasise the need for domain-customised VLLMs and suggest the continued need for human oversight in clinical ophthalmology.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ General purpose visual large language models (VLLMs) have demonstrated potential in ophthalmological disease detection, but their capabilities to diagnose from retinal images remain underexplored.

WHAT THIS STUDY ADDS

⇒ While GPT-4V's default mode achieved high sensitivity for detecting general abnormalities (97.1%), the diagnostic descriptions across all models were generally suboptimal, highlighting significant limitations in their ability to provide accurate and clinically meaningful outputs for ophthalmic imaging tasks.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Our findings reveal the limitations of general-purpose VLLMs in accurately detecting and diagnosing major eye diseases from retinal photographs. These results emphasise the need for task-specific refinement for VLLMs and the continued importance of human oversight in clinical application.

INTRODUCTION

The development of artificial intelligence (AI) driven, image-based diagnostics is a field of growing interest.[1] Generative AI and vision large language models (VLLMs) are advancing AI's capabilities by processing both text and image inputs through transformer-based architectures and image encoders.[2] By tokenising inputs into manageable pieces, VLLMs efficiently analyse data, offering significant potential in medical fields like ophthalmology.[2–4] VLLMs bring a new wave of budding potential for innovations in medical fields which are heavily reliant on images such as ophthalmology.[5] These algorithms may potentially complement the delivery of eye care services in locations with inadequate or limited access to eye care, such as in low and middle-income countries.[6]

In July 2023, Google Gemini began to roll out features which accept images as inputs.[5] In September 2023, this new capability was also extended by OpenAI's GPT-4V.[7] This suggests that VLLM-based chatbots could potentially impact medical diagnostics by assisting with or even automating the analysis

of medical images.[7] One potential application of these VLLMs in ophthalmology would be to assist in detecting and diagnosing major eye diseases from retinal photographs.[8] Retinal photographs, crucial for diagnosing and managing many of these diseases, traditionally require accurate, specialist-driven analysis.[9] The application of VLLMs could potentially help streamline the clinical process, improving patient eye care and access.[8]

Nevertheless, the clinical utility of current commonly available VLLMs in ophthalmology remains relatively underexplored. This study aimed to evaluate the capabilities of three popular VLLMs, namely GPT-4V's default mode, GPT-4V's data analyst mode and Google Gemini in accurately detecting and diagnosing ocular diseases from retinal photographs.[5 10]

MATERIALS AND METHODS

Our study, conducted from 30 September to 20 October 2023, analysed 44 retinal photographs curated from the Singapore Epidemiology of Eye Diseases (SEED) study.[11] The SEED dataset comprises a comprehensive collection of retinal images with robust disease labelling.[11] In this study, we systematically selected representative images to evaluate the performance of VLLMs in detecting and diagnosing six major eye diseases. The evaluation set used consisted of 10 normal retinal images and 34 images with single diagnosis from six major eye diseases: age-related macular degeneration (AMD, n=10), diabetic retinopathy (DR, n=6), glaucoma (n=5), visually significant cataract (VSC, n=5), myopic macular degeneration (MMD, n=5) and retinal vein occlusion (RVO, n=3). This selection provided a balanced representation of major eye diseases and ensured the specificity of the evaluation by focusing on images with single diagnoses to minimise confounding. Given the exploratory, proof-of-concept nature of this study, the sample size was pragmatically chosen to balance the demands of the labour-intensive, consensus-based evaluation framework involving three attending-level ophthalmologists and three rounds of grading.

The images were manually uploaded into GPT-4V's default mode, GPT-4V's data analyst mode and Google Gemini through each chatbot's interfaces, one image at a time accompanied by the following prompt: "Is this retina normal or abnormal?". We then assessed the VLLMs' capability to detect general abnormality and to correctly identify normal images. For images correctly identified as abnormal, we further prompted: "Please identify any potential abnormalities in this retina and provide a list of possible diagnoses". We then evaluated the quality of the descriptions generated by the three VLLMs (figure 1).

The accuracy and quality of the three VLLMs' responses were then evaluated by three attending-level ophthalmologists (WW, DZC, YXW). The image sequences were presented to the graders in a randomised order, and a 1 day washout period was observed before exposing the graders to outputs of the other models. The responses were graded as 'good' when the response was clear, clinically accurate and did not mislead or give potentially harmful advice. They were graded as 'borderline' when the response was partially clear and accurate, partially incomplete, and might indirectly cause unnecessary concern or confusion to the patient. They were graded as 'poor' when it was generally vague and provided misleading or incorrect information that could lead to improper treatment and management. For each case, a majority consensus among the three ophthalmologists was adopted to reach the final grading (see online supplemental table 1). In instances where a common consensus was not reached among the three ophthalmologists (ie, three different ratings were provided), we defaulted to a stringent approach, assigning the lowest score (ie, 'poor') to the VLLM's response.

Performance differences between three VLLMs were denoted in proportions and analysed using the χ^2 test using R version 4.2.1 (R Foundation, Vienna, Austria). Significance was set at a two-sided p value of less than 0.05.

RESULTS

In our assessment of general abnormality detection, GPT-4V's default mode outperformed the other models by correctly identifying 97.1% (33 out of 34) of abnormal retinal images. This was significantly higher than GPT-4V's data analyst mode (61.7%, 21 out of 34) and Google Gemini (41.1%, 14 out of 34) (all p<0.001). On the other hand, when evaluating the 10 normal images, GPT-4V's default mode only correctly identified 5 as normal (50%), while the other two VLLMs correctly identified all as being normal (table 1).

In GPT-4V's default mode, of the 33 abnormal images correctly identified, only 7 (21.2%) of its diagnostic descriptions were deemed as 'good' by the panel of experts (figure 2). In contrast, GPT-4V's data analyst mode only had 1 response deemed as 'good' (4.8%, out of 21 abnormal images correctly identified by it). On the other hand, Google Gemini only had 4 responses with descriptions evaluated as 'good' (28.6%, out of 14 abnormal images correctly identified by it).

On further examination, GPT-4V's default mode performed best in AMD cases. Of the 10 AMD cases, GPT-4V's default mode correctly identified all as abnormal (100% sensitivity) and provided accurate diagnosis and good quality descriptions for 6 of them. For DR cases, GPT-4V's default mode was able to identify all six cases as abnormal but only one yielded 'good' quality description (online supplemental data).

When evaluating the five retinal photos of eyes with VSC, all VLLMS' diagnoses and descriptions were generally rated as 'poor' (online supplemental data). For instance, of the five VSC cases, GPT-4V's default mode yielded three 'poor' quality responses, and GPT-4V's data analyst mode yielded five 'poor' quality responses. On the other hand, Google Gemini misidentified all VSC cases as normal (false negative) in the first place.

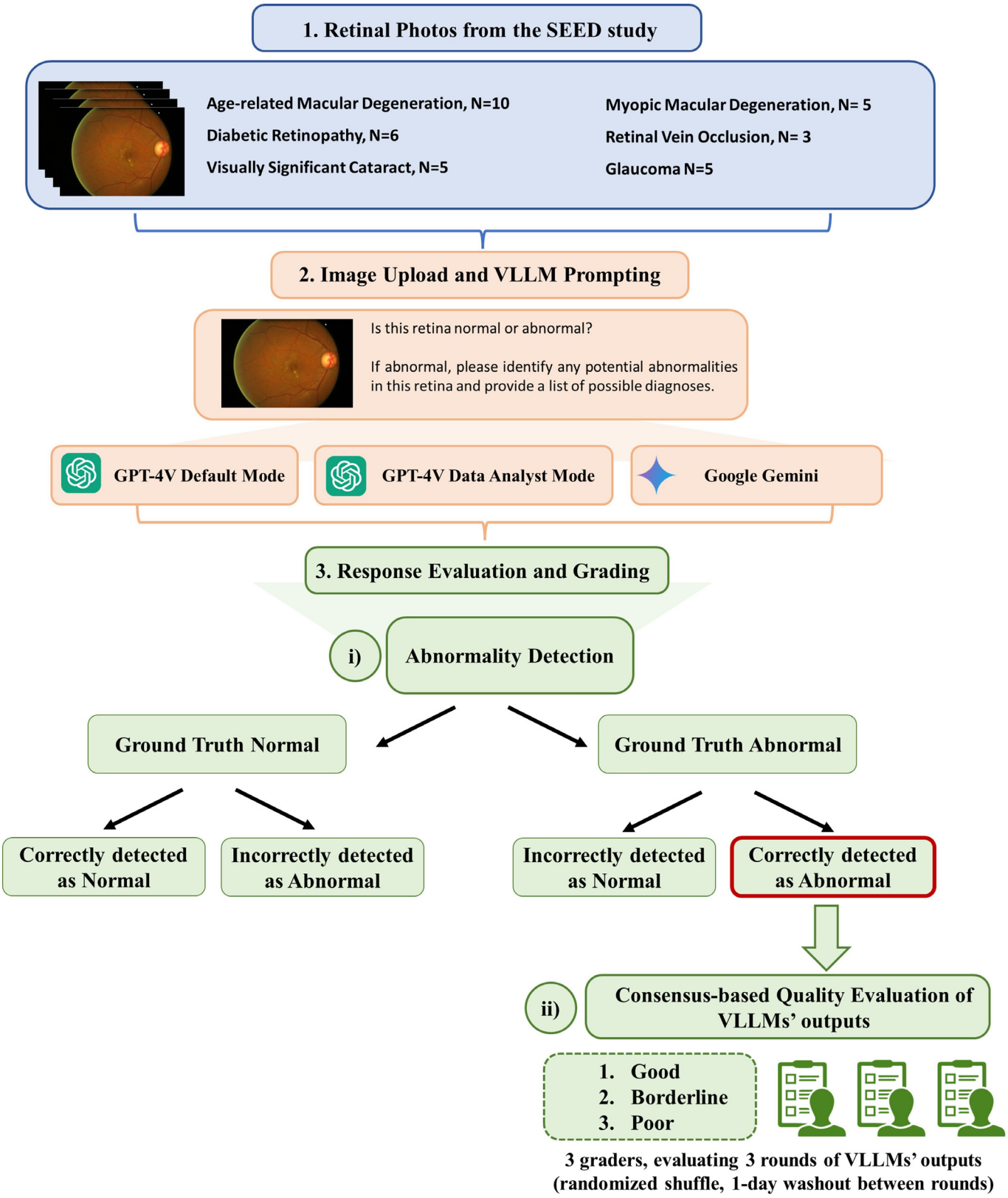

Figure 1 Flowchart of overall study design.

DISCUSSION

Our study presents a rigorous head-to-head comparison of three common general VLLMS in detecting and diagnosing common eye diseases using retinal photographs. Prior studies primarily focused on evaluating single models such as those evaluating language capabilities alone or detection capabilities in isolation.[3 4 12–14] Our study provides a more comprehensive assessment by evaluating three prominent VLLMs (GPT-4V default, GPT-4V data

Table 1 Sensitivity and specificity of respective vision large language models in detecting ocular diseases

		GPT-4V default mode		
		Abnormal	**Normal**	
Ground truth	Abnormal (n=34)	33	1	Sensitivity=97.0%
	Normal (n=10)	5	5	Specificity=50.0%
		GPT-4V data analyst mode		
		Abnormal	**Normal**	
Ground truth	Abnormal (n=34)	21	13	Sensitivity=61.7%
	Normal (n=10)	0	10	Specificity=100%
		Google Gemini (October 2023 version)		
		Abnormal	**Normal**	
Ground truth	Abnormal (n=34)	14	20	Sensitivity=41.1%
	Normal (n=10)	0	10	Specificity=100%

analyst mode and Google Gemini) on their ability to both detect and diagnose six major eye diseases using colour fundus photographs.[3 4 12–16] Additionally, unlike previous research that often required specific prompts or pre-structured inputs, our study uniquely evaluated the models' capabilities to generate diagnostic outputs directly from retinal images without overly elaborated input guidance which may inadvertently inflate the performance of the VLLMs.[14] A further distinction lies in our robust evaluation framework, which used majority consensus evaluations based on three attending-level ophthalmologists, ensuring reliability and clinical relevance.[14] These methodical advancements fill critical gaps in existing literature and offer greater insights into the potential and limitations of VLLMs as decision-support tools in ophthalmology. Notably, the default mode of GPT-4V showcased superior sensitivity in detecting general abnormalities (97.1%), compared with its data analyst mode and

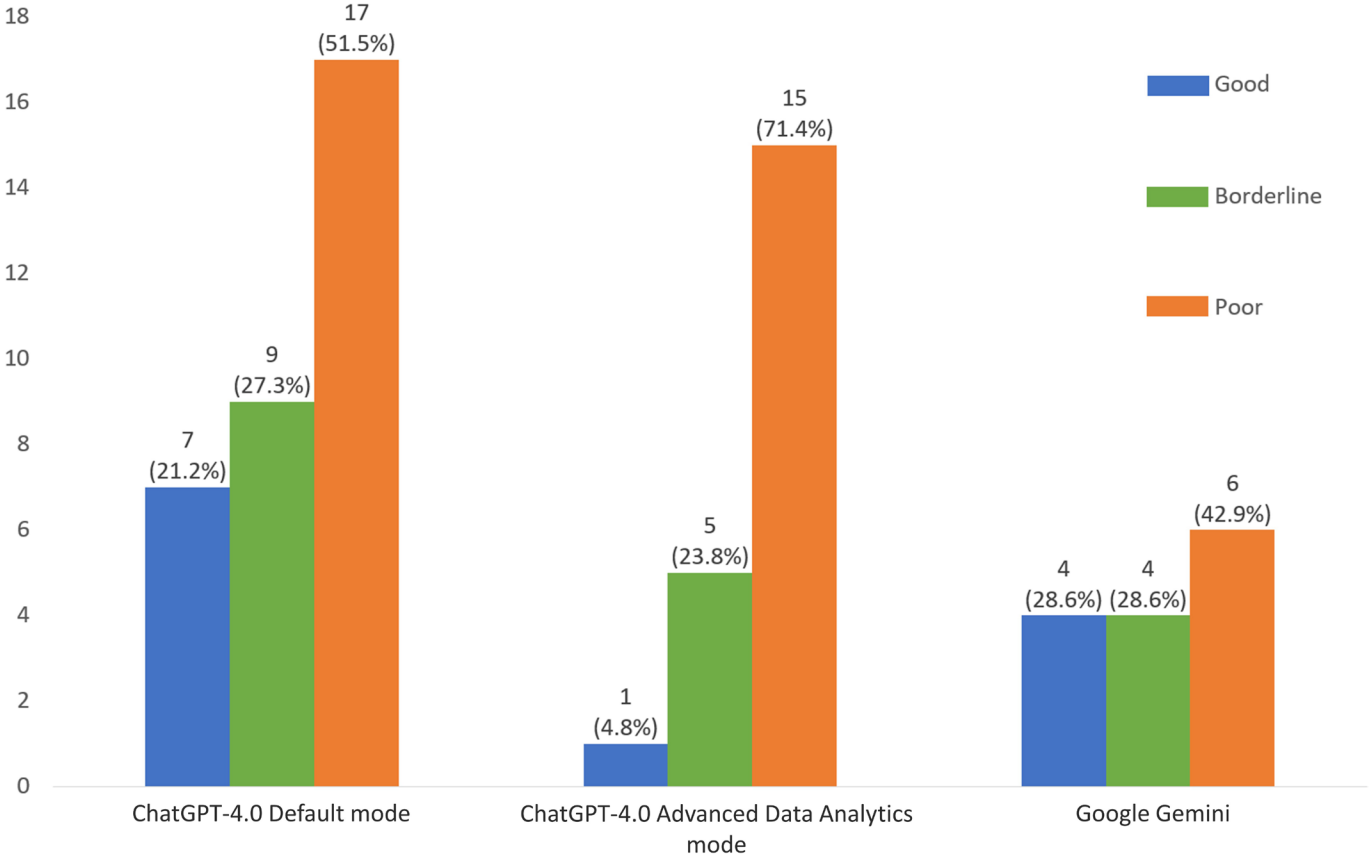

Figure 2 Consensus-based quality ratings of VLLMs' outputs, as determined by three consultant-level ophthalmologists.

Google Gemini. However, the diagnostic descriptions of these three VLLMs were generally suboptimal. Furthermore, the quality of disease descriptions was largely inadequate. These 'off-the-shelf' VLLMs are not currently not tailored for ophthalmic imaging tasks, as a result, struggle to accurately detect and diagnose common eye diseases. Our findings highlight some promising potential and the significant limitations of these general VLLMs in the realm of eye disease detection and diagnosis. Altogether, this underscores the need for domain-specific customisations to enhance their clinical applicability.

We observed that GPT-4V's data analyst mode often produced vague descriptions and non-specific diagnoses which were unrelated to the actual condition. This phenomenon may be attributed to the underlying training methodology employed for each mode, particularly the use of user-intent-based reinforcement learning with human feedback (RLHF).[17] RLHF, a technique where models are fine-tuned based on their ability to align with human preferences, might have been applied differently across GPT-4V's default mode and data analyst mode.[17] While the default mode was fine-tuned with a focus on understanding and generating human-like text to broadly match a variety of user intents, the data analyst mode was specifically targeted towards data analysis tasks.[17] This divergence in training objectives likely led to the default mode producing more nuanced, human-like and contextually appropriate responses, aligning close to what a clinician might say. In contrast, the data analyst mode, adhering strictly to its analytical approach, might list several possible conditions without prioritising or contextualising them effectively for a clinical setting, thus, appearing 'vague' with its descriptions.

Both the GPT-4V's default and data analyst modes often included cautious advisory statements, likely reflecting the implementation of medical domain-specific constraints within the GPT-4V models.[10] Such constraints are designed to prevent the model from making unfounded medical assertions and to better align the model outputs with ethical guidelines and safety considerations. These constraints, while designed to prevent the model from making unfounded assertions, might inadvertently lead to overly cautious or ambiguous responses that lack the specificity required for accurate medical diagnosis. However, despite these constraints, the inaccurate medical information in the outputs may still cause unnecessary stress or worry among users who are not medically trained.

On the other hand, Google Gemini's outputs, despite the majority of them being rated as poor (42.8%), often displayed a misleading confident tone. Such tone and delivery style could potentially mislead end-users or non-specialist practitioners (online supplemental table 2, images from Gemini with poor rated score). The limitations in outputs observed in these VLLMs further emphasise the importance of continuous refinement and testing to align with clinical standards and patient safety considerations.

The overall low proportion of 'good' quality descriptions across the evaluated VLLMs highlights a significant limitation in their application for clinical diagnosis, reinforcing the need to fine-tune the models' capability.[18] Altogether, findings from our analysis indicate that current VLLMs may not yet be fully equipped for detecting and diagnosing abnormalities from retinal images. This observation aligns with previous research which also highlights VLLMs' limited diagnostic capability across a range of medical specialties, including dentistry, endocrinology and gynaecology.[19 20] Given these insights, we advocate for prudent use of VLLMs, one which involves human oversight.[6] These VLLMs should be viewed as supplementary decision support tools that aid in the evaluation process rather than as replacements for comprehensive evaluations conducted by ophthalmologists or optometrists.[13]

The strengths of this study include the disease representation from major eye diseases. Furthermore, this study employs a robust, clinically relevant evaluation framework, with a consensus-based evaluation involving from three attending-level experts, and comprehensively assesses the outputs by the VLLMs. Additionally, we used data exclusively from our 'private' SEED study, which the VLLMs could not have been exposed to in their training, thus, mitigating potential model bias, improving the fairness and reliability of our assessment. However, this proof-of-concept study is limited by its small sample size and the selection of major eye diseases, which may restrict the generalisability of our findings to other eye diseases. The choice of this relatively small sample size was partly due to the demand of labour-intensive, majority consensus evaluation which involved three attending-level ophthalmologists and three rounds of grading. Nevertheless, to mitigate the impact of small sample size, we used non-parametric χ^2 tests, which are robust to smaller sample sizes and do not rely on assumptions of normality. Future research with larger, more representative datasets is still warranted.

Future research should prioritise fine-tuning VLLMs using ophthalmology-specific datasets, including tasks such as visual question-answering (VQA), analysis of textual patient notes and de-identified clinical data. This targeted fine-tuning could significantly enhance the models' ability to interpret complex ophthalmological inputs. Additionally, integrating fine-tuned VLLMs with retrieval-augmented generation (RAG) systems tailored for ophthalmology would improve their diagnostic accuracy and reasoning capabilities, particularly in handling nuanced cases.[21] Further efforts could focus on addressing challenges in data scalability, such as the development of diverse and representative datasets that better capture the variability in ocular diseases across different settings and populations.[22] Improving model interpretability

through explainability frameworks is another critical avenue, ensuring that clinicians can trust and validate the outputs of these systems. Advancements in these areas could pave the way for VLLMs to be more reliable and clinically applicable as decision-support tools in ophthalmology.

In conclusion, our findings reveal the limitations of general-purpose VLLMs in accurately detecting and diagnosing major eye diseases from retinal photographs. These results emphasise the need for task-specific refinement for VLLMs and the continued importance of human oversight in clinical application.

Author affiliations
[1]Centre for Innovation and Precision Eye Health, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore
[2]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore
[3]Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[4]Tsinghua Medicine, Tsinghua University, Beijing, China
[5]Department of Ophthalmology, National University Hospital, Singapore
[6]Beihang University, Beijing, China
[7]Institute for AI Industry Research, Tsinghua University, Beijing, China
[8]Beijing Visual Science and Translational Eye Research Institute (BERI), Eye Center of Beijing Tsinghua Changgung Hospital, School of Clinical Medicine, Tsinghua Medicine, Tsinghua University, Beijing, China
[9]Beijing Key Laboratory of Intelligent Diagnostic Technology and Devices for Major Blinding Eye Diseases, Tsinghua Medicine, Tsinghua University, Beijing, China
[10]Eye Academic Clinical Program (Eye ACP), Duke NUS Medical School, Singapore

X Yih Chung Tham @Yihtham

Acknowledgements Role of funders: The funder of the study had no role in study design, data analysis, data interpretation or writing of the report. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Contributors SS, HJ and Y-CT contributed to conception of the study. SS, HJ, Y-CT, DZC, WW and YXW contributed to study design. SS, JHLG and Y-CT contributed to acquisition of the data. SS and HJ contributed to statistical analysis of data. SS, HJ and Y-CT contributed to analysis and interpretation of the data. SS, HJ, JHLG and Y-CT accessed and verified each dataset during the course of the study. Supervision of this research which includes responsibility for the research activity planning and execution was oversighted by Y-CT. SS and HJ contributed to visualisation which includes figure, charts and tables of the data. All authors had access to all the data and SS, HJ and Y-CT were responsible for the decision to submit the manuscript. DZC, WW and YXW contributed to clinical validations. SS, HJ and Y-CT drafted the manuscript. ZDS, KP, XW, WM, TYW, CYC and Y-CT contributed to editing the manuscript. All authors read and approved the final version of the manuscript. Y-CT is the guarantor. AI was used to improve the grammar.

Funding This work is supported by the National University of Singapore and Tsinghua University's Joint Initiative Scientific Research Program. Dr Yih-Chung Tham was funded by the National Medical Research Council of Singapore (NMRC/MOH/ HCSAINV21nov-000). Dr Hong Wei JI was funded by the Shuimu Scholar Program of Tsinghua University, the National Postdoctoral Innovative Talent Support Program (BX20230189). The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Part of a Topic Collection; Not commissioned; externally peer reviewed.

Data availability statement No data are available.

ORCID iDs
Sahana Srinivasan http://orcid.org/0009-0004-9355-6717
Hongwei Ji http://orcid.org/0000-0003-3657-4666
David Ziyou Chen http://orcid.org/0000-0002-2153-3100
Wendy Wong http://orcid.org/0000-0002-4514-250X
Zhi Da Soh http://orcid.org/0000-0002-1182-3489
Jocelyn Hui Lin Goh http://orcid.org/0000-0002-5052-6081
Krithi Pushpanathan http://orcid.org/0000-0001-6263-8489
Xiaofei Wang http://orcid.org/0000-0002-3175-0344
Weizhi Ma http://orcid.org/0000-0001-5604-7527
Tien Yin Wong http://orcid.org/0000-0002-8448-1264
Ya Xing Wang http://orcid.org/0000-0003-2749-7793
Ching-Yu Cheng http://orcid.org/0000-0003-0655-885X
Yih Chung Tham http://orcid.org/0000-0002-6752-797X

REFERENCES

1 Tan TF, Thirunavukarasu AJ, Jin L, *et al*. Artificial intelligence and digital health in global eye health: opportunities and challenges. *Lancet Glob Health* 2023;11:e1432–43.
2 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, eds. *Advances in Neural Information Processing Systems*. Curran Associates, Inc, 2017.
3 Antaki F, Touma S, Milad D, *et al*. Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of Its Successes and Shortcomings. *Ophthalmol Sci* 2023;3:100324.
4 Antaki F, Chopra R, Keane PA. Vision-Language Models for Feature Detection of Macular Diseases on Optical Coherence Tomography. *JAMA Ophthalmol* 2024;142:573–6.
5 Bard. Google bard new features update july 2023. 2023.
6 Wang X, Sanders HM, Liu Y, *et al*. ChatGPT: promise and challenges for deployment in low- and middle-income countries. *Lancet Reg Health West Pac* 2023;41:100905.
7 ChatGPT. ChatGPT — Release Notes, 2023. Available: https://help.openai.com/en/articles/6825453-chatgpt-release-notes
8 Betzler BK, Chen H, Cheng C-Y, *et al*. Large language models and their impact in ophthalmology. *The Lancet Digital Health* 2023;5:e917–24.
9 Korot E, Gonçalves MB, Huemer J, *et al*. Clinician-Driven AI: Code-Free Self-Training on Public Data for Diabetic Retinopathy Referral. *JAMA Ophthalmol* 2023;141:1029–36.
10 OpenAI. GPT-4 system card. 2023.
11 Majithia S, Tham Y-C, Chee M-L, *et al*. Cohort Profile: The Singapore Epidemiology of Eye Diseases study (SEED). *Int J Epidemiol* 2021;50:41–52.
12 Singh S, Djalilian A, Ali MJ. ChatGPT and Ophthalmology: Exploring Its Potential with Discharge Summaries and Operative Notes. *Semin Ophthalmol* 2023;38:503–7.
13 Chen JS, Reddy AJ, Al-Sharif E, *et al*. Analysis of ChatGPT Responses to Ophthalmic Cases: Can ChatGPT Think like an Ophthalmologist? *Ophthalmol Sci* 2025;5:100600.
14 Xu P, Chen X, Zhao Z, *et al*. Unveiling the clinical incapabilities: a benchmarking study of GPT-4V(ision) for ophthalmic multimodal image analysis. *Br J Ophthalmol* 2024;108:1384–9.
15 Haddad F, Saade JS. Performance of ChatGPT on Ophthalmology-Related Questions Across Various Examination Levels: Observational Study. *JMIR Med Educ* 2024;10:e50842.

16 Ming S, Yao X, Guo X, *et al*. Performance of ChatGPT in Ophthalmic Registration and Clinical Diagnosis: Cross-Sectional Study. *J Med Internet Res* 2024;26:e60226.
17 OpenAI. GPT-4 technical report. 2023.
18 Zini JE, Awad M. On the Explainability of Natural Language Processing Deep Models. *ACM Comput Surv* 2023;55:1–31.
19 Wu C, Zheng Q, Zhao W, *et al*. Can GPT-4V(ision) Serve Medical Applications Case Studies on GPT-4V for Multimodal Medical Diagnosis. *arXiv* 2023.:arXiv231009909v2.
20 Thirunavukarasu AJ, Ting DSJ, Elangovan K, *et al*. Large language models in medicine. *Nat Med* 2023;29:1930–40.
21 Luo M-J, Pang J, Bi S, *et al*. Development and Evaluation of a Retrieval-Augmented Large Language Model Framework for Ophthalmology. *JAMA Ophthalmol* 2024;142:798–805.
22 Chen X, Chen R, Xu P, *et al*. Visual Question Answering in Ophthalmology: A Progressive and Practical Perspective. *arXiv* 2024.