

Vision-Language Models for Feature Detection of Macular Diseases on Optical Coherence Tomography

Fares Antaki MDCM, FRCSC^{1,2,3}, Reena Chopra, PhD^{1,2,4}, Pearse A. Keane MD^{1,2,4}

1. Institute of Ophthalmology, University College London, London, United Kingdom
2. Moorfields Eye Hospital NHS Foundation Trust, London, United Kingdom
3. The CHUM School of Artificial Intelligence in Healthcare, Montreal, Quebec, Canada
4. NIHR Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust, London, United Kingdom

ORCID of the authors: Fares Antaki (0000-0001-6679-7276), Reena Chopra (0000-0002-4264-8329), Pearse A. Keane (0000-0002-9239-745X)

Corresponding author: [Pearse A. Keane, MD](mailto:p.keane@ucl.ac.uk) Institute of Ophthalmology, University College London, London, United Kingdom p.keane@ucl.ac.uk

Word count: 1018

Funding: No funding was obtained for this work.

Ethics Approval: Ethics approval was not required for this work.

Patient Consent: Patient consent was not required as this work did not involve patients.

Keywords: artificial intelligence; foundation models; large language models

Key Points

Question: How well does the Gemini Pro vision-language model (VLM) perform in expert-level diagnostic and management tasks for macular diseases using optical coherence tomography (OCT) scans?

Findings: In a study including 50 expert-labeled OCTs, the VLM showed limited ability in detecting pathological disease features (average F1 score: 10.7%). A correct diagnosis was achieved in 17/50 (34%) of cases, and a correct referral recommendation in 28/50 (56%).

Meaning: A recently-released VLM demonstrated limited capabilities in OCT-based expert-level tasks. As models continue to improve, creating large ophthalmic benchmarking datasets will be important to ascertain the true potential of VLMs in ophthalmology.

Abstract

Importance: Vision-language models (VLMs) are a novel artificial intelligence technology capable of processing image and text inputs. While demonstrating strong generalist capabilities, their performance in ophthalmology has not been extensively studied.

Objective: To assess the performance of the Gemini Pro VLM in expert-level tasks for macular diseases from optical coherence tomography (OCT) scans.

Design: Cross-sectional diagnostic accuracy study

Setting: Evaluation of a generalist VLM on ophthalmology-specific tasks

Participants: The dataset included OCT B-scans from 50 unique patients: normal cases and those with macular hole, diabetic macular edema, central serous chorioretinopathy, and age-related macular degeneration. Each OCT was labeled for 10 key pathological features, referral recommendations, and treatments.

Exposure: We tested Gemini Pro using a standard prompt to extract structured responses on December 15, 2023.

Main Outcome and Measure: Our primary outcome was to assess model responses against expert labels, calculating F1 scores for each pathological feature. Secondary outcomes included accuracy in diagnosis, referral urgency, and treatment recommendation. We also evaluated the model's internal concordance by measuring the alignment between referral and treatment recommendations, independent of diagnostic accuracy.

Results: The average F1 score was 10.7% (95% CI, 2.4 to 19.2). Only macular hole (36.4%; 95% CI, 0 to 71.4), pigment epithelial detachment (26.1%; 95% CI, 0 to 46.2), subretinal hyperreflective material (24.0%; 95% CI, 0 to 45.2) and subretinal fluid (20.0%; 95% CI, 0 to 45.5) had a measurable F1 score. A correct diagnosis was achieved in 17/50 cases (34%; 95% CI, 22 to 48). Referral recommendations varied: 28/50 were correct (56%; 95% CI, 42 to 70), 10/50 were over-cautious (20%; 95% CI, 10 to 32), and 12/50 were under-cautious (24%; 95% CI, 12 to 36). Referral and treatment concordance were very high, with 48/50 (96%; 95% CI, 90 to 100) and 48/49 (98%; 95% CI, 94 to 100) correct answers.

Conclusions and Relevance: A generalist VLM demonstrated limited 'vision' capabilities for feature detection and management of macular disease. However, it showed low self-contradiction, suggesting strong 'language' capabilities. As VLMs continue to improve, validating their performance on large benchmarking datasets will help ascertain their potential in ophthalmology.

Introduction

Foundation models are a novel paradigm for building artificial intelligence (AI) systems allowing model training at scale on large amounts of unlabeled multimodal data before fine-tuning them for a variety of downstream tasks.¹ They have demonstrated expert-level capacities in numerous fields in medicine using text-based information, including in ophthalmology.² However, medicine is inherently multimodal and models capable of handling images are potentially poised to be most beneficial. Recently, vision-language models (VLMs) have emerged with the capacity to understand interleaved multimodal inputs, such as combining images with short questions or instructions.³ Their performance in ophthalmology, to our knowledge, has not yet been studied. We assessed the capabilities of a VLM (Gemini Pro Vision, developed by Google DeepMind)⁴ in identifying pathological features, proposing diagnoses, referral recommendations and treatments for macular diseases from optical coherence tomography (OCT) scans.

Methods

We used an open-access OCT dataset of macular diseases and curated a single foveal-centered B-scan from 50 unique eyes.⁵ The collection featured 10 normal OCT scans and 40 abnormal ones including: macular hole (n=10), diabetic macular edema (n=10), central serous chorioretinopathy (n=9), and age-related macular degeneration (n=11).

Based on our previous work, we defined 10 key features for OCT interpretation.⁶ We experimented with several prompting styles, including open-ended questions and a "spot the difference" format with normal/abnormal OCT image pairs (**Supplement 1**).³ Ultimately, we adopted a standard prompt (**Figure 1**) to generate structured answers on the presence or absence of those features. This approach ensured objective performance metrics, mitigating the inherent subjectivity of evaluating long-form answers. Retinal features were jointly

graded by a retina fellow (FA), an experienced optometrist (RC), and overseen by a medical retina consultant (PAK) to set the reference standard.

Our primary outcome was to determine whether the model's responses matched the reference standard. For each feature, we calculated common machine learning metrics, including an F1 score. This metric represents the harmonic mean of precision and recall, and is useful in imbalanced datasets like ours. Accuracy of diagnosis, referral urgency, and treatment recommendation were secondary outcomes. Concordance in referral and treatment recommendations, regardless of diagnostic accuracy, were also evaluated to assess the model's internal concordance. We used a bootstrapping technique to estimate the 95% confidence intervals (CIs) for all metrics. The reporting of study data aligns with the Standards for Reporting of Diagnostic Accuracy (STARD) reporting guideline for elements relevant to generative AI research.

Results

The average F1 score was 10.7% (95%CI, 2.4 to 19.2). Only macular hole, pigment epithelial detachment, subretinal hyperreflective material and subretinal fluid had a measurable F1 score. The scores were generally low, ranging from 20.0 to 36.4 (**Table 1**). A correct diagnosis was provided in 17/50 (34%; 95% CI, 22 to 48) of cases, and a correct referral recommendation in 28/50 (56%; 95% CI, 42 to 70) of cases. There were 10/50 (20%; 95% CI, 10 to 32) over-cautious and 12/50 (24%; 95% CI, 12 to 36) under-cautious referral recommendations. Referral and treatment concordance were very high, with 48/50 (96%; 95% CI, 90 to 100) and 48/49 (98%; 95% CI, 94 to 100) correct answers.

Discussion

A recently-released [VLM](#) achieved low capabilities in detecting expert-level OCT-based features of retinal diseases and proposing accurate diagnoses. Referral accuracy was generally favorable, but up to 24% of scans were under-referred, which may raise safety concerns. Those findings suggest that the 'vision' capabilities of the tested [VLM](#) are not robust enough for expert-level diagnostics in ophthalmology, at this stage. This is likely due to the scarcity of highly detailed ophthalmic image-text pairs in the training corpora of this model. Despite that, the [VLM](#) was highly internally concordant with low self-contradiction for referral and treatment recommendations, which speaks to its 'language' capabilities.

Previously, our group reported strong discriminative performance of AI models for multi-class disease detection using this dataset. [RETFound](#), our bespoke ophthalmic foundation model, achieved a very high area under the precision-recall curve (0.993), suggesting potentially better performance than what we observed with Gemini Pro.⁷ We reported similar capabilities of code-free deep learning models on this dataset with a mean F1 score of 95%.⁸ However, feature detection was not examined in our previous work and may pose distinct challenges for any AI system. Moreover, generative tasks, where the model produces diagnoses and treatments without predefined labels, may be inherently more complex than traditional supervised learning tasks. Nonetheless, the limited performance obtained with Gemini Pro in this study suggests that generalist [VLMs](#) may fall short of the specific vision needs of ophthalmology.

Limitations

Our study has limitations. To ensure objective performance metrics, we limited the model's output to a structured presence/absence format for most elements. [While this approach ensured consistency, it might have neglected potentially valuable nuances the model could](#)

reveal in open-ended responses. Furthermore, while we used a single prompt and no clinical vignettes, we recognize that this might not capture the full spectrum of the model's abilities due to potential bias towards specific aspects of the task. In initial testing, the lead-in prompt 'you are an expert ophthalmologist' increased the average F1 score by 5% (Supplement 2). This highlights the potential of prompt engineering.⁹ We suspect that with extensive prompt engineering and empiric testing, the VLM may achieve improved performances. Due to the small sample size in our study, additional research with larger datasets will be needed to ascertain the performance of VLMs in identifying pathological features on OCT and their broader application in retinal disease.

Conclusion

VLMs have the potential to transform ophthalmology. However, without domain-specific fine-tuning or retrieval capabilities to augment their internal representation of our specialty, generalist VLMs may have limited capabilities. Combining ophthalmology-specific foundation models with fine-tuned text-based large language models (LLMs) for medicine may maximize patient benefits by leveraging the specialized vision capabilities of ophthalmic models and the reasoning abilities of LLMs.^{7,10} Further investigation is required to conclusively determine the potential of VLMs. It will be worthwhile to observe how performance develops alongside the release of larger, more advanced models.

Acknowledgment

Funding: None

Role of funder/sponsor: Not applicable

Competing interests: Dr. Keane has acted as a consultant for Google, DeepMind, Roche, Novartis, Apellis, and Bitfount and is an equity owner in Big Picture Medical. He has received speaker fees from Heidelberg Engineering, Topcon, Allergan, and Bayer, outside the submitted work. No other disclosures were reported.

Access to data and data analysis: Dr Antaki and Dr Chopra had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Meeting presentation: None

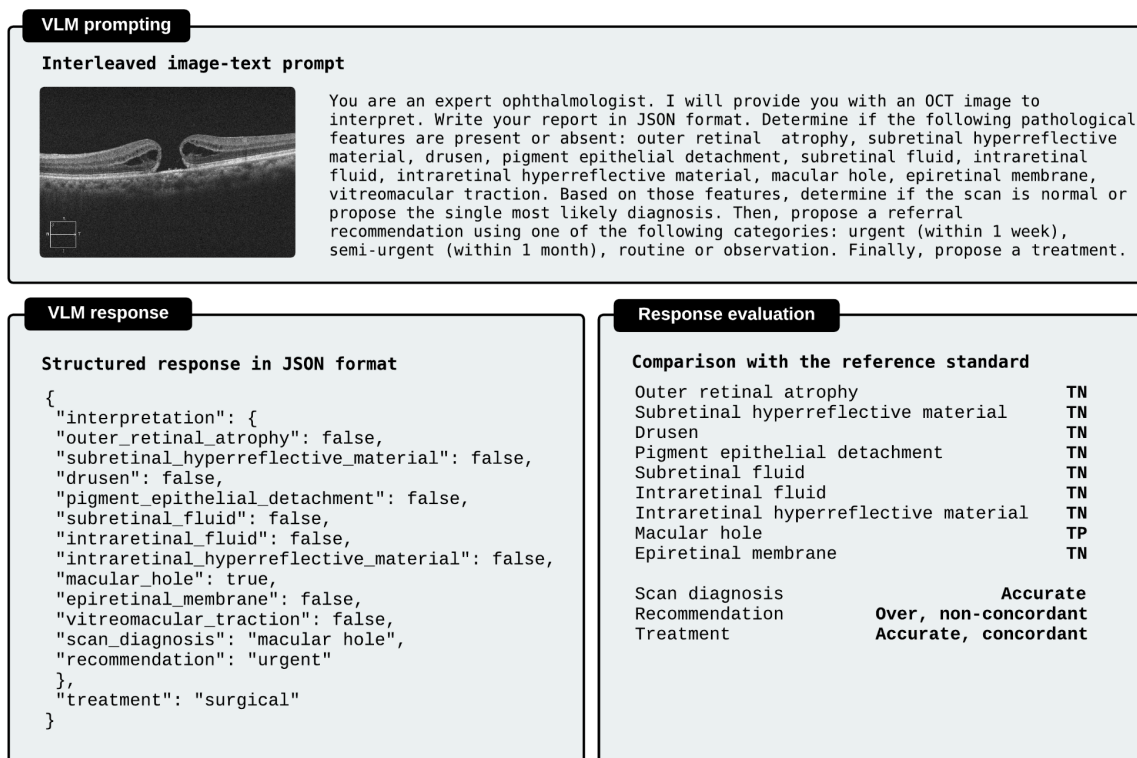
Non-author contributions: None

References

1. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of Its Successes and Shortcomings. *Ophthalmol Sci.* 2023;3(4):100324.
2. Antaki F, Milad D, Chia MA, et al. Capabilities of GPT-4 in ophthalmology: an analysis of model entropy and progress towards human-level medical question answering. *Br J Ophthalmol.* Published online November 3, 2023. doi:10.1136/bjo-2023-324438
3. Yang Z, Li L, Lin K, et al. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision). *arXiv [csCV]*. Published online September 29, 2023. <http://arxiv.org/abs/2309.17421>
4. Gemini Team, Anil R, Borgeaud S, et al. Gemini: A Family of Highly Capable Multimodal Models. *arXiv [csCL]*. Published online December 19, 2023. <http://arxiv.org/abs/2312.11805>
5. Gholami P, Roy P, Parthasarathy MK, Lakshminarayanan V. OCTID: Optical coherence tomography image database. *Comput Electr Eng.* 2020;81:106532.
6. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med.* 2018;24(9):1342-1350.
7. Zhou Y, Chia MA, Wagner SK, et al. A foundation model for generalizable disease detection from retinal images. *Nature.* 2023;622(7981):156-163.
8. Korot E, Guan Z, Ferraz D, et al. Code-free deep learning for multi-modality medical image classification. *Springer Nature.* Published online March 1, 2021:288-298.
9. Gu J, Han Z, Chen S, et al. A Systematic Survey of Prompt Engineering on Vision-Language Foundation Models. *arXiv [csCV]*. Published online July 24, 2023. <http://arxiv.org/abs/2307.12980>
10. MedLM models overview. Google Cloud. Accessed December 19, 2023. <https://cloud.google.com/vertex-ai/docs/generative-ai/medlm/overview>

Figure and Table

Figure 1. Overview of the prompting strategy and response evaluation



Abbreviations: JSON, JavaScript Object Notation; TN, true negative; TP, true positive

Table 1. Summary of the predictive performance for each optical coherence tomography feature

Feature	F1 (95%CI)	SEN (95%CI)	SPE (95%CI)	PPV (95%CI)	NPV (95%CI)	ACC (95%CI)
Macular hole	36.4 (0.0-71.4)	22.2 (0.0-55.6)	100 (100-100)	100 (0.0-100)	85.4 (75.0-95.7)	86.0 (76.0-96.0)
Pigment epithelial detachment	26.1 (0.0-46.2)	21.4 (0.0-43.8)	83.3 (70.6-94.6)	33.3 (0.0-66.7)	73.2 (59.0-86.9)	66.0 (52.0-78.0)
Subretinal hyperreflective material	24.0 (0.0-45.2)	75.0 (0.0-100)	60.9 (45.8-73.9)	14.3 (0.0-30.4)	96.6 (88.9-100)	62.0 (48.0-74.0)
Subretinal fluid	20.0 (0.0-45.5)	18.2 (0.0-45.5)	82.1 (68.4-93.8)	22.2 (0.0-57.1)	78.0 (64.1-90.2)	68.0 (54.0-80.0)
Outer retinal atrophy	0.0 (0.0-0.0)	0.0 (0.0-0.0)	93.3 (85.4-100)	0.0 (0.0-0.0)	89.4 (80.0-97.8)	84.0 (74.0-94.0)
Drusen	0.0 (0.0-0.0)	0.0 (0.0-0.0)	97.8 (93.0-100)	0.0 (0.0-0.0)	89.8 (80.0-98.0)	88.0 (78.0-96.0)
Intraretinal fluid	0.0 (0.0-0.0)	0.0 (0.0-0.0)	100 (100-100)	0.0 (0.0-0.0)	60.0 (46.0-74.0)	60.0 (46.0-74.0)
Intraretinal hyperreflective material	0.0 (0.0-0.0)	0.0 (0.0-0.0)	100 (100-100)	0.0 (0.0-0.0)	68.0 (56.0-82.0)	68.0 (56.0-82.0)
Epiretinal membrane	0.0 (0.0-0.0)	0.0 (0.0-0.0)	100 (100-100)	0.0 (0.0-0.0)	84.0 (74.0-94.0)	84.0 (74.0-94.0)
Vitreomacular traction	0.0 (0.0-0.0)	0.0 (0.0-0.0)	100 (100-100)	0.0 (0.0-0.0)	96.0 (90.0-100)	96.0 (90.0-100)

F1: F1 score, SEN: sensitivity or recall, SPE: specificity, PPV: positive predictive value or precision, NPV: negative predictive value, ACC: accuracy

Supplement 1 - Prompting techniques

Open-ended request

Prompt



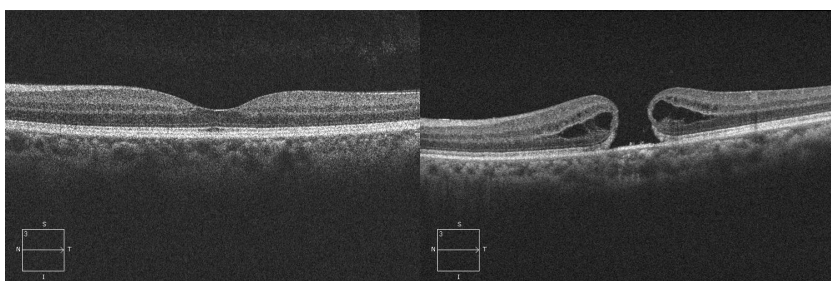
Write a detailed report for this ophthalmic image.

Response

This is an optical coherence tomography (OCT) image of the macula. The OCT image shows a macular hole. A macular hole is a full-thickness defect in the center of the macula. It is a relatively rare condition that can occur in people of all ages, but it is most common in older adults. Macular holes can be caused by a variety of factors, including trauma, inflammation, and age-related changes. Symptoms of a macular hole may include blurred vision, distorted vision, and a dark spot in the center of the field of vision. Treatment for a macular hole may include observation, surgery, or a combination of both.

Spot the difference

Prompt



Look at these two images, compare them and tell me what is different, and write a report about the abnormal one.

Response

The first image is a normal OCT of the macula. The second image is an OCT of a macular hole. A macular hole is a full-thickness defect in the center of the macula. It can cause a decrease in central vision and metamorphopsia. Macular holes are often idiopathic, but they can also be caused by trauma, inflammation, or high myopia. Treatment for macular holes typically involves surgery.

Supplement 2

Summary of the predictive performance for each optical coherence tomography feature using the prompt without the lead in “you are an expert ophthalmologist”

Feature	F1 (95%CI)	SEN (95%CI)	SPE (95%CI)	PPV (95%CI)	NPV (95%CI)	ACC (95%CI)
Macular hole	33.3 (0.0-66.7)	22.2 (0.0-57.1)	97.6 (92.3-100)	66.7 (0.0-100)	85.1 (74.0-94.0)	84.0 (74.0-94.0)
Subretinal hyperreflective material	22.2 (0.0-60.0)	25.0 (0.0-100)	91.3 (82.2-97.9)	20.0 (0.0-66.7)	93.3 (85.1-100)	86.0 (76.0-96.0)
Outer retinal atrophy	0.0 (0.0-0.0)	0.0 (0.0-0.0)	97.8 (93.0-100)	0.0 (0.0-0.0)	89.8 (81.2-97.9)	88.0 (78.0-96.0)
Drusen	0.0 (0.0-0.0)	0.0 (0.0-0.0)	100 (100-100)	0.0 (0.0-0.0)	90.0 (82.0-98.0)	90.0 (82.0-98.0)
Pigment epithelial detachment	0.0 (0.0-0.0)	0.0 (0.0-0.0)	97.2 (90.9-100)	0.0 (0.0-0.0)	71.4 (59.2-83.7)	70.0 (58.0-82.0)
Subretinal fluid	0.0 (0.0-0.0)	0.0 (0.0-0.0)	97.4 (92.1-100)	0.0 (0.0-0.0)	77.6 (65.9-89.4)	76.0 (64.0-88.0)
Intraretinal fluid	0.0 (0.0-0.0)	0.0 (0.0-0.0)	100 (100-100)	0.0 (0.0-0.0)	60.0 (46.0-72.0)	60.0 (46.0-72.0)
Intraretinal hyperreflective material	0.0 (0.0-0.0)	0.0 (0.0-0.0)	100 (100-100)	0.0 (0.0-0.0)	68.0 (54.0-80.0)	68.0 (54.0-80.0)
Epiretinal membrane	0.0 (0.0-0.0)	0.0 (0.0-0.0)	97.6 (92.7-100)	0.0 (0.0-0.0)	83.7 (73.5-92.0)	82.0 (72.0-92.0)
Vitreomacular traction	0.0 (0.0-0.0)	0.0 (0.0-0.0)	100 (100-100)	0.0 (0.0-0.0)	96.0 (90.0-100)	96.0 (90.0-100)

F1: F1 score, SEN: sensitivity or recall, SPE: specificity, PPV: positive predictive value or precision, NPV: negative predictive value, ACC: accuracy