

Multimodal large language models for use in diabetic retinopathy screening

S. Saeed Mohammadi¹, Sahana Aggarwal¹, Kavina Aggarwal¹, Grant Wiarda¹, Kayla Nguyen¹, Emmanuel A. Sarmiento¹, Quan Dong Nguyen², Manjot K. Gill¹

¹Department of Ophthalmology, Northwestern University, Feinberg School of Medicine, Chicago, IL, USA; ²Spencer Center for Vision Research, Byers Eye Institute, Stanford University School of Medicine, Palo Alto, CA, USA

Abstract

Purpose: To evaluate the performance of ChatGPT-4o and Gemini 2.5 Pro in detecting more-than-mild diabetic retinopathy (mtmDR) from fundus photography (FP) and diabetic macular edema (DME) from optical coherence tomography (OCT) using publicly available datasets.

Methods: A custom GPT (powered by ChatGPT-4o) was created and instructed to follow the LumineticsCore™ (IDx-DR) screening criteria for mtmDR, defined as an ETDRS level ≥ 35 and/or clinically significant diabetic macular edema (CSDME). Gemini 2.5 Pro was evaluated with the same criteria. Performance on FPs was assessed using 2 publicly available datasets: MESSIDOR-2 ($n = 106$; 66 mtmDR, 40 no mtmDR) and EyePACS ($n = 99$; 56 mtmDR, 43 non-mtmDR). To assess detection of DME, a separate OCT dataset ($n = 48$; 24 DME, 24 normal) was used to evaluate identification of intraretinal and/or subretinal fluid. Sensitivity, specificity, positive predictive value, and negative predictive value for detecting mtmDR on FP and DME on OCT were calculated for each multimodal large language model (LLM).

Results: On MESSIDOR-2 ($n = 106$), ChatGPT-4o achieved a sensitivity of 90.77%, specificity of 97.50%, PPV of 98.33%, and NPV of 86.67% for mtmDR detection. Gemini 2.5 Pro achieved a sensitivity of 80.30%, specificity of 97.50%, PPV of 98.15%, and NPV of 75.00%. On EyePACS ($n = 99$), ChatGPT-4o demonstrated a sensitivity of 94.64%, specificity of 86.05%, PPV of 89.83%, and NPV of 92.50%,

Correspondence: Manjot K. Gill, MD, MS, Department of Ophthalmology, Northwestern University, Feinberg School of Medicine, 645 North Michigan Avenue, Ste 440, Chicago, IL 60611, USA.

E-mail: mgill@nm.org

while Gemini 2.5 Pro achieved a sensitivity of 89.29%, specificity of 88.37%, PPV of 90.91%, and NPV of 86.36%. For OCT-based DME detection ($n = 48$), ChatGPT-4o achieved a sensitivity of 95.83%, specificity of 100%, and PPV of 100%, while Gemini 2.5 Pro achieved a sensitivity of 95.83%, specificity of 95.65%, PPV of 95.83%, and NPV of 95.65%.

Conclusion: ChatGPT-4o and Gemini 2.5 Pro demonstrated high performance in detecting mtmDR and DME across multiple publicly available datasets. These findings support the potential of multimodal LLMs as cost-effective and accessible tools for diabetic retinopathy screening. Further validation in larger, more diverse real-world datasets is warranted.

Keywords: artificial intelligence, ChatGPT, diabetic retinopathy, Gemini, multimodal large language model

Introduction

Diabetic retinopathy (DR) is a microvascular complication of diabetes mellitus characterized by progressive damage to the retinal vasculature, manifesting as microaneurysms, intraretinal hemorrhages, capillary nonperfusion, retinal edema, and, in advanced stages, neovascularization and fibrous tissue proliferation.^{1,2} These pathologic changes can lead to sight-threatening complications, including diabetic macular edema (DME), tractional retinal detachment, and neovascular glaucoma, which may result in irreversible vision loss if left untreated. Screening for DR is essential as the condition is often asymptomatic in its early stages allowing for timely therapeutic intervention, which can prevent or significantly reduce the risk of vision loss. Studies have shown that up to 90% of severe vision impairment can be avoided when retinopathy is identified and treated promptly.³⁻⁵

Artificial intelligence (AI) is playing an increasingly prominent role in the screening of DR, primarily through the automated analysis of retinal fundus images to identify referable or vision-threatening stages of the disease. Several AI-based systems, such as LumineticsCore™ (Digital Diagnostics, Coralville, IA, USA; formerly IDx-DR), have demonstrated high diagnostic accuracy and are now recognized as viable alternatives to conventional manual grading performed by ophthalmologists.³

Multimodal large language models (LLMs) are also more frequently integrated into various aspects of medicine, including the management of DR, where they function as adjunct tools for clinical decision support, medical documentation, and patient care. In the context of DR, LLMs have been combined with image-based deep learning systems to aid primary care providers in screening and disease management. Such integration has been associated with improved diagnostic accuracy for referable DR as well as greater patient adherence to follow-up referrals

and self-management practices in real-world settings.⁶ Additionally, LLMs facilitate the automated extraction and classification of DR severity from unstructured clinical texts, thereby streamlining data annotation processes and supporting the development of more efficient and adaptable image classification models.⁷ Although multimodal LLMs such as ChatGPT and Gemini were originally designed for text-based tasks, recent advancements have enabled analysis and interpretation of images alongside textual data.

In this study, we evaluated the performance of ChatGPT-4o and Gemini 2.5 Pro in DR screening using publicly available datasets of fundus photographs (FPs) and optical coherence tomography (OCT) images.

Methods

This study was conducted in January 2025 in accordance with the principles of the Declaration of Helsinki. Institutional review board approval was not required because only publicly available datasets were used.

Datasets

Three publicly available retinal imaging datasets were utilized:

1. The Messidor-2 dataset (Methods to Evaluate Segmentation and Indexing Techniques in the Field of Retinal Ophthalmology) comprises 1,748 FPs collected from 3 ophthalmic departments in France. Images were acquired using a Topcon TRC NW6 non-mydratic camera with a digital video recording system, with image resolution ranging from 1440×960 pixels to 2304×1536 pixels. Each image was graded for DR severity on a 5-point scale (0 = no DR, 1 = mild NPDR, 2 = moderate NPDR, 3 = severe NPDR, 4 = PDR).^{8,9}
2. The EyePACS dataset was obtained from the Kaggle Diabetic Retinopathy Detection challenge and contains 71,054 color FPs collected in the United States. Images were captured using various devices, including Centervue DRS, Optovue iCam, Canon CR1/DGi/CR2, and Topcon NW cameras, with resolutions ranging from 433×289 pixels to 5184×3456 pixels. All images were labeled according to DR severity using the same 5-grade scale as Messidor-2.¹⁰⁻¹²
3. A publicly available OCT dataset¹³ was also used, consisting of 108,312 images categorized into choroidal neovascularization (CNV, $n = 37,206$), diabetic macular edema (DME, $n = 11,349$), drusen ($n = 8,617$), and normal ($n = 51,140$) classes. For this study, only images from the DME and normal groups were included. An equal number of images were randomly selected from each group for evaluation. To minimize selection bias, we used the RAND function in Excel to randomly select the images from the datasets included in the study.

Screening criteria

LumineticsCore™ is the first AI system authorized by the US Food and Drug Administration (FDA) and awarded the Conformité Européenne (CE) mark for the early detection of DR. It automatically identifies more-than-mild diabetic retinopathy (mtmDR) in adults without a prior diagnosis of the condition.

MtmDR was defined as:

1. Early Treatment Diabetic Retinopathy Study (ETDRS) level ≥ 35 on FP, and/or
2. clinically significant diabetic macular edema (CSDME) on FP.

Multimodal mtmDR was defined as:

1. ETDRS level ≥ 35 on FP, and/or
2. CSDME on FP, and/or
3. center-involved diabetic macular edema (CI-DME) on OCT.¹⁴

In this study, we applied the same diagnostic criteria to evaluate the ability of multimodal LLMs to screen for DR.

As there is currently no publicly available dataset containing both FPs and corresponding OCT images from the same patients, we used a separate OCT dataset to assess the ability of multimodal LLMs to detect the presence of intraretinal or subretinal fluid.

Multimodal LLMs

A ChatGPT Plus subscription was used in this study. Under the “GPTs” section, a custom GPT (powered by ChatGPT-4o) was created using the following instruction: “ONLY USE ORIGINAL image WITHOUT any Conversion of image to RGB/Black & White Images/edge detection for feature extraction or determining the presence of mtmDR. Screen DR based on the fundus photos that are uploaded. Determine if the image has ‘mtmDR or more than mild DR’ based on the below criteria: (no need to tell the exact stage). mtmDR is defined as: 1. ETDRS (DRSS) level \geq (equal or more than) 35 (determined from fundus photographs)* and/or 2. CSDME (determined from fundus photographs) defined as: Retinal thickening at or within 500 μm of the center of the fovea, Hard exudates at or within 500 μm of the center of the fovea if adjacent to an area of retinal thickening, Retinal thickening of at least 1 disc area any portion of which is within 1500 μm (approximately 1 disc diameter) from the center of the fovea. Use the uploaded spreadsheet for the exact ETDRS staging.”

An Excel spreadsheet detailing the exact ETDRS DR severity scale¹⁵ was prepared and uploaded under the Knowledge section of the custom GPT (Fig. 1).

A separate prompt, “Determine the presence of intraretinal fluid and/or subretinal fluid on the uploaded OCT”, was used to assess the presence or absence of intraretinal or subretinal fluid on OCT images.

A Gemini Pro subscription was also utilized for this study. Gemini 2.5 Pro was tested with a similar prompt, but without the Excel spreadsheet detailing the exact ETDRS staging, as custom GPT creation was not supported at the time of

The screenshot shows the 'Configure' page for a custom GPT named 'Diabetic Retinopathy'. The interface includes a 'Create' button and a 'Configure' button. A central dashed circle with a plus sign indicates where to add an image. Below this are several sections:

- Name:** Diabetic Retinopathy
- Description:** Screening DR
- Instructions:**

ONLY USE ORIGINAL image WITHOUT any Conversion of image to RGB/Black & White Images/edge detection for feature extraction or determining the presence of mtmDR. Screen diabetic retinopathy based on the fundus photos that are uploaded.

Determine if the image has 'mtmDR or more than mild DR' based on the below criteria: (no need to tell the exact stage). mtmDR is defined as:

Conversations with your GPT can potentially include part or all of the instructions provided.
- Conversation starters:**
 - Please upload the fundus photo for screening.
- Knowledge:**

Conversations with your GPT can potentially reveal part or all of the files uploaded.

The following files are only available for Code Interpreter:

 - ETDRS.xlsx Spreadsheet

Fig. 1. Custom GPT Configurations on ChatGPT-4o. An Excel spreadsheet detailing the exact ETDRS DR severity scale¹⁵ was prepared and uploaded under the Knowledge section of the custom GPT.

evaluation. The same prompt was also used with Gemini 2.5 Pro to assess for the presence of fluid on OCT images.

Statistical analysis

Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated by comparing each multimodal LLM classification of FP as mtmDR-positive or mtmDR-negative to the ground truth labels. True positives (TP) were defined as correctly identified mtmDR-positive cases, true negatives (TN) as correctly identified mtmDR-negative cases, false positives (FP) as cases incorrectly labeled positive, and false negatives (FN) as cases incorrectly labeled negative. Sensitivity was calculated as $TP / (TP + FN)$, specificity as $TN / (TN + FP)$, PPV as $TP /$

(TP + FP), and NPV as $TN / (TN + FN)$. All calculations were conducted using Microsoft Excel (Microsoft Corporation, Redmond, WA, USA).

Results

In the MESSIDOR-2 dataset, 40 FPs with no DR or mild NPDR and 66 photos with mtmDR were randomly selected and evaluated using a custom GPT model on ChatGPT-4o. On this dataset, ChatGPT-4o achieved a sensitivity of 90.77%, specificity of 97.50%, PPV of 98.33%, and NPV of 86.67% for detecting mtmDR. Figure 2 depicts ChatGPT-4o's response to screening of a FP from the MESSIDOR-2 dataset. In comparison, Gemini 2.5 Pro achieved a sensitivity of 80.30%, specificity of 97.50%, PPV of 98.15%, and NPV of 75.00%. Figure 3 shows Gemini 2.5 Pro's response to screening of a FP from the MESSIDOR-2 dataset.

A total of 99 FPs from the EyePACS dataset were analyzed, including 43 images without mtmDR and 56 with mtmDR. ChatGPT-4o demonstrated a sensitivity of 94.64%, specificity of 86.05%, PPV of 89.83%, and NPV of 92.50%. In comparison, Gemini 2.5 Pro achieved a sensitivity of 89.29%, specificity of 88.37%, PPV of 90.91%, and NPV of 86.36%.

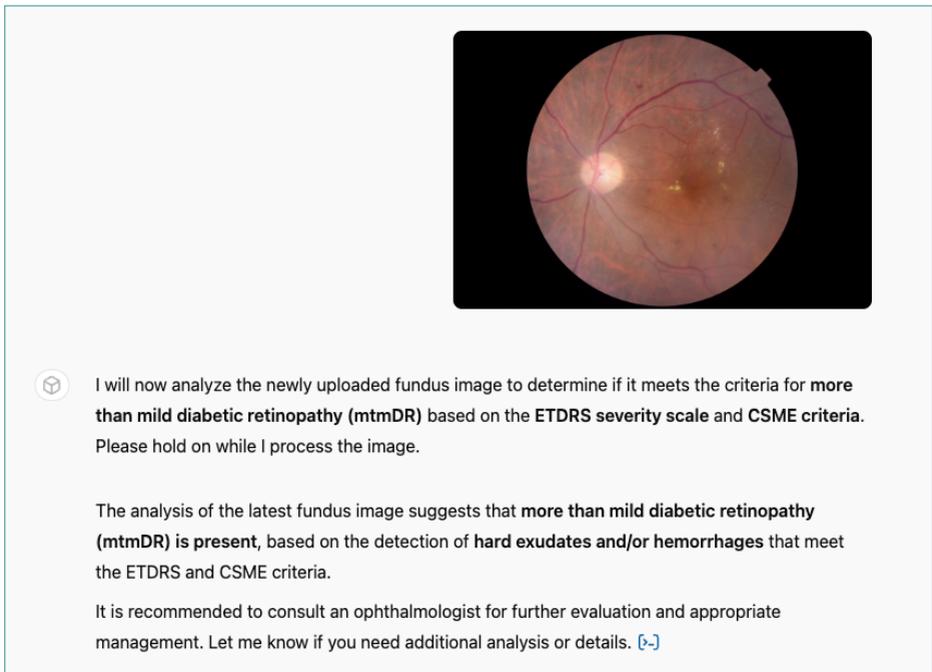
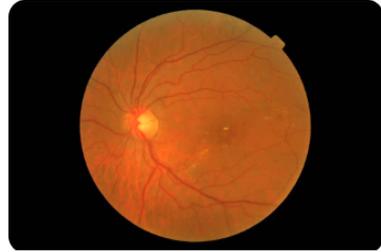


Fig. 2. ChatGPT-4o response to FP screening from the Messidor-2 dataset.



1. ONLY USE ORIGINAL image WITHOUT any Conversion of image to RGB/Black & White Images/edge detection for feature extraction or determining the presence of...

◆ Show thinking ▾

Based on the analysis of the provided fundus photograph and the criteria given:

1. **ETDRS (DRSS) Level:** The image shows evidence of diabetic retinopathy, including what appear to be microaneurysms/small hemorrhages and distinct hard exudates. While precisely grading the ETDRS level from a single 2D image is challenging, the findings suggest at least mild non-proliferative diabetic retinopathy (NPDR).
2. **CSME:** There are noticeable hard exudates (yellowish deposits) present, particularly clustered near the center of the macula (the foveal area). Several of these exudates appear to be located at or very close to (likely within 500 μm of) the center of the fovea. According to the criteria, hard exudates in this location, if associated with retinal thickening, constitute CSME. While retinal thickening cannot be directly measured from this 2D image, the presence of hard exudates in this pattern strongly suggests associated leakage and thickening.

Conclusion:

The presence of hard exudates near the center of the fovea strongly suggests **Clinically Significant Macular Edema (CSME)**. Since mtmDR is defined as having an ETDRS level ≥ 35 and/or CSME, the findings consistent with CSME in this image indicate the presence of **mtmDR (more than mild Diabetic Retinopathy)**.

Disclaimer: This analysis is based on a visual interpretation of a single fundus photograph according to the provided criteria. It is not a substitute for a comprehensive eye examination and diagnosis by a qualified ophthalmologist or optometrist. Accurate assessment of diabetic retinopathy and macular edema often requires specialized imaging (like OCT) and clinical examination.

Fig. 3. Gemini 2.5 Pro response to FP screening from the Messidor-2 dataset.

For the OCT dataset, 24 OCT images with DME—defined as the presence of subretinal or intraretinal fluid—and 24 normal OCT images were randomly selected and evaluated by both ChatGPT-4o and Gemini 2.5 Pro. ChatGPT-4o detected DME with a sensitivity of 95.83%, specificity of 100%, and PPV of 100% (Fig. 4), while Gemini 2.5 Pro achieved a sensitivity of 95.83%, specificity of 95.65%, PPV of 95.83%, and NPV of 95.65% for the detection of DME (Fig. 5).

Table 1 summarizes the screening performance of ChatGPT-4o and Gemini 2.5 Pro for DR utilizing the 3 different datasets.

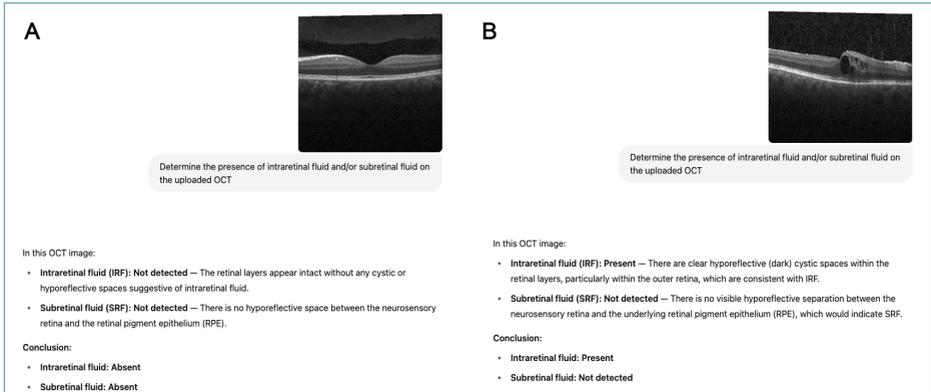


Fig. 4. (A) Evaluation of a normal OCT by ChatGPT-4o. (B) Evaluation of an OCT with IRF and SRF by ChatGPT-4o.

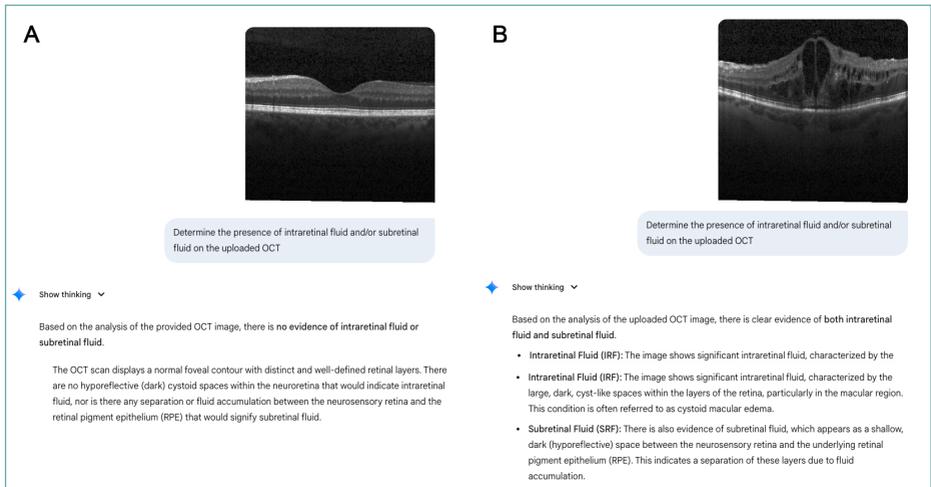


Fig. 5. (A) Evaluation of a normal OCT by Gemini 2.5 Pro. (B) Evaluation of an OCT with intraretinal fluid and subretinal fluid by Gemini 2.5 Pro.

Table 1. Screening performance of custom GPT and Gemini 2.5 Pro

| | MESSIDOR 2 | | | | | EYEPACS | | | | | OCT | |
|----------------|------------|-----------|----------|----------|-------|---------|-----------|----------|----------|-------|--------|---------|
| | No DR | Mild NPDR | Mod NPDR | Sev NPDR | PDR | No DR | Mild NPDR | Mod NPDR | Sev NPDR | PDR | No DME | IRF/SRF |
| Custom GPT | 30/30 | 9/10 | 21/25 | 18/20 | 21/21 | 16/20 | 21/23 | 11/11 | 22/25 | 20/20 | 24/24 | 23/24 |
| Gemini 2.5 Pro | 29/30 | 10/10 | 18/25 | 18/20 | 18/21 | 16/20 | 22/23 | 9/11 | 22/25 | 19/20 | 23/24 | 23/24 |

DR: diabetic retinopathy; NPDR: non-proliferative diabetic retinopathy; Mod: moderate; Sev: severe; IRF: intraretinal fluid; SRF: subretinal fluid

Discussion

The global prevalence of diabetes has risen dramatically over recent decades, with an estimated 462 million adults affected in 2019 projected to reach 700 million by 2045.¹⁶ The prevalence of DR closely parallels the increasing incidence of diabetes worldwide with estimates that 22%–35% of individuals with diabetes have DR, and 6%–10% have vision-threatening DR.^{1,16-18} In 2020, over 100 million adults globally were living with DR, a figure expected to surpass 160 million by 2045, positioning DR as a major global public health concern.¹⁷ The highest prevalence of DR has been reported in Africa and the Caribbean, while the lowest rates are seen in South and Central America, highlighting significant disparities by ethnicity and region.^{16,17}

DR remains a leading cause of vision loss among working-age adults worldwide, underscoring the critical need for regular screening to mitigate irreversible visual impairment.¹ As DR can be asymptomatic until advanced stages, annual screening is recommended by both the American Academy of Ophthalmology and the American Diabetes Association to prevent or reduce the risk of vision loss.¹

Despite these well-established guidelines, widespread and timely DR screening remains a significant challenge, particularly in underserved and resource-limited settings. Barriers such as limited access to ophthalmologists, high costs, and logistical and structural barriers contribute to suboptimal screening rates. In the United States, only approximately 60% of patients receive annual screening, with even lower rates observed in low-income and minority populations.^{19,20}

AI has become an important tool in DR screening, enabling automated and consistent analysis of retinal images. The American Diabetes Association recognizes FDA-approved AI systems, including LumineticsCore™, EyeArt, and AEYE-DS as alternatives to traditional screening, if image quality is sufficient and abnormal findings are referred appropriately.¹

Development of these systems requires extensive data collection, algorithm training, and clinical validation, resulting in high initial and ongoing costs.²¹

Performance limitations also exist, as most algorithms are trained and validated on FPs from specific imaging devices, which can reduce diagnostic accuracy when applied to images from other devices or in different clinical settings. Real-world studies have demonstrated that algorithm sensitivity and specificity can vary across populations and imaging platforms, highlighting the need for rigorous external validation before widespread adoption.^{22,23}

Further, financial barriers are substantial. The LumineticsCore™ system, for example, is available via a considerable one-time purchase cost (~US\$13,000), capital leasing, or a no-capital expenditure model requiring a minimum number of quarterly exams. Regardless of the model, a fixed US\$25 fee is charged per patient screened. The 2020 national average Medicare reimbursement was US\$23.82, resulting in a negative margin for practices.^{21,24} This cost misalignment, combined with the need for compatible imaging hardware and staff training, limits widespread adoption, especially in primary care and resource-limited settings.^{21,25}

Multimodal LLMs are increasingly being explored for medical image analysis, including diabetic DR screening. Unlike current AI systems, which are primarily based on deep learning models trained to classify FPs, multimodal LLMs offer the potential to integrate multimodal data, combining image interpretation with clinical context, patient history, and natural language explanations. Recent work demonstrates that, when paired with image-based deep learning modules, LLMs cannot only match or exceed the diagnostic accuracy of primary care physicians for referable DR, but also generate individualized management recommendations and improve patient adherence to follow-up, as shown in the DeepDR-LLM system.⁶

Multimodal LLMs can provide flexible, context-aware outputs, such as generating patient-friendly summaries, and tailored recommendations for different users, which current image-only AI systems cannot achieve. Furthermore, multimodal LLMs may be more generalizable to new populations, imaging devices, and clinical settings, potentially overcoming the scalability and device-dependence limitations of current AI-based image analysis systems.²⁶ LumineticsCore™ is approved for use exclusively with the Topcon NW400 fundus camera while the fundus image datasets used in our study were acquired using a variety of camera models, suggesting that multimodal LLMs may be inherently more flexible and not restricted to a particular imaging device.

Multimodal LLMs, such as ChatGPT and Gemini, have become increasingly accessible, with their advanced capabilities available through subscriptions for nominal cost (approximately US\$20 per month). These models support multimodal inputs, enabling the interpretation of both textual and visual data, including medical images. Although their use in language tasks has been extensively studied, their application to ophthalmic image interpretation, especially in fundus photography, remains underexplored.

In the current study, we evaluated the performance of ChatGPT-4o and Gemini 2.5 Pro using 2 publicly available datasets of FP and 1 open-access dataset of

optical OCT images. For context, the pivotal trial leading to the FDA approval of LumineticsCore™ reported a sensitivity of 87.2% (95% CI: 81.8–91.2%) and a specificity of 90.7% (95% CI: 88.3–92.7%) for detecting mtmDR.²⁷ In evaluating the Messidor-2 dataset, ChatGPT-4o demonstrated a sensitivity of 90.77% with a specificity of 97.50%, while Gemini 2.5 Pro achieved a sensitivity of 80.30% and a similar specificity of 97.50%. In utilizing the EYEPACS dataset, ChatGPT-4o achieved a sensitivity of 94.64% and a specificity of 86.05%, whereas Gemini 2.5 Pro demonstrated a sensitivity of 89.29% and a specificity of 88.37%. A prior study by Aftab *et al.* reported that the general model of ChatGPT-4 achieved a high sensitivity of 96.2% but a low specificity of 19.1% in detecting mtmDR. Their analysis was based on a publicly available retinal image dataset (PACS, Joint Shantou International Eye Centre) hosted on Kaggle.²⁸ In contrast, the high sensitivity and specificity observed in our study may be attributable to differences in prompt engineering, made possible by a custom GPT model instructed with detailed diagnostic criteria aligned with those used by LumineticsCore™, including the ETDRS DR severity scale.

In addition to FPs, LLM performance was evaluated for the detection of DME using OCT images, defined by the presence of any subretinal or intraretinal fluid. Among 24 DME-positive and 24 normal OCT images, ChatGPT-4o achieved a sensitivity of 95.83% and specificity of 100%, while Gemini 2.5 Pro achieved a sensitivity of 95.83% and specificity of 95.65%.

In our findings, one important factor that likely influenced performance was the way the models were instructed. The custom ChatGPT-4o setup used a detailed prompt along with the ETDRS staging spreadsheet, which provided specific guidance on how to judge disease severity. Gemini 2.5 Pro, in contrast, was limited to the prompt alone. This difference suggests that the way a model is prompted, and whether it has access to structured clinical information, can meaningfully affect results rather than the model itself being the only determinant.

Recent work has shown similar patterns, with studies reporting that more structured and clinically focused prompts improve accuracy compared with simple or generic instructions. This supports the idea that prompt design and the inclusion of relevant clinical criteria may help LLMs perform more reliably in medical screening tasks.²⁹ As LLMs continue to develop, prompt optimization and access to clear reference criteria may become an important part of their use in clinical workflows. Future studies should explore standardized prompts, knowledge-based support tools, and other approaches to make performance more consistent across different platforms.

Our study has several limitations. Because of uncertainty around HIPAA compliance and data security with commercial LLMs, we restricted our analysis to publicly available datasets and did not evaluate the models on internal, real-world clinical images. As a result, it is unclear how well these findings would translate to our own patient population, where image quality, disease severity, and imaging

devices may differ. The relatively small sample size further limits generalizability, and we did not assess performance across different stages of DR or DME, which could influence accuracy. In addition, the FPs and OCT images were drawn from separate datasets rather than paired images from the same patients, so we could only evaluate each modality independently rather than reflecting true multimodal use in clinical practice. Another limitation is that we did not perform a true head-to-head comparison between Gemini 2.5 Pro and ChatGPT-4o, as ChatGPT-4o supported custom GPT creation with external knowledge files, including the ETDRS spreadsheet, while Gemini 2.5 Pro did not offer an equivalent feature at the time this study was conducted. It is also worth noting that the reported performance of LumineticsCore was based on a different dataset, so direct comparison with our results should be interpreted cautiously. Finally, LLM performance can vary based on prompt design, highlighting the need for more standardized approaches in future studies. Larger, more diverse, and paired FP-OCT datasets, ideally from real clinical settings, will be important to better validate these findings and assess practical implementation.

Conclusions and future perspectives

In conclusion, the results of this study suggest that LLMs, if compliant with HIPAA regulations and data security standards, may provide a cost-effective solution for DR screening, with diagnostic accuracy comparable to that of FDA-approved AI algorithms. Future directions include integrating these models with handheld or smartphone-based retinal imaging devices, which hold promise for expanding access to DR screening, particularly in underserved or resource-limited settings.

Declarations

Ethics approval and consent to participate

This study was conducted in accordance with the principles of the Declaration of Helsinki. Institutional Review Board approval was not required because only publicly available datasets were used.

Competing interests

QDN serves as Section Editor for AIVO; he was not involved at any stage of the editorial process prior to article acceptance. AIVO conducts a single-blinded review process. The remaining authors declare no competing interests.

Funding

This study was supported in part by an unrestricted grant from Research to Prevent Blindness.

Acknowledgements

None to declare.

References

1. Lim JI, Kim SJ, Bailey ST, et al. Diabetic Retinopathy Preferred Practice Pattern(R). *Ophthalmology*. 2025;132(4):P75-P162. <https://doi.org/10.1016/j.ophtha.2024.12.020>
2. Wong TY, Cheung CM, Larsen M, Sharma S, Simo R. Diabetic retinopathy. *Nat Rev Dis Primers*. 2016;2:16012. <https://doi.org/10.1038/nrdp.2016.12>
3. American Diabetes Association Professional Practice C. 12. Retinopathy, Neuropathy, and Foot Care: Standards of Care in Diabetes-2025. *Diabetes Care*. 2025;48(1 Suppl 1):S252-S265. <https://doi.org/10.2337/dc25-S012>
4. Lundeen EA, Andes LJ, Rein DB, et al. Trends in Prevalence and Treatment of Diabetic Macular Edema and Vision-Threatening Diabetic Retinopathy Among Medicare Part B Fee-for-Service Beneficiaries. *JAMA Ophthalmol*. 2022;140(4):345-353. <https://doi.org/10.1001/jamaophthalmol.2022.0052>
5. Solomon SD, Chew E, Duh EJ, et al. Diabetic Retinopathy: A Position Statement by the American Diabetes Association. *Diabetes Care*. 2017;40(3):412-418. <https://doi.org/10.2337/dc16-2641>
6. Li J, Guan Z, Wang J, et al. Integrated image-based deep learning and language models for primary diabetes care. *Nat Med*. 2024;30(10):2886-2896. <https://doi.org/10.1038/s41591-024-03139-8>
7. Jaskari J, Sahlsten J, Summanen P, et al. DR-GPT: A large language model for medical report analysis of diabetic retinopathy patients. *PLoS One*. 2024;19(10):e0297706. <https://doi.org/10.1371/journal.pone.0297706>
8. Abramoff MD, Folk JC, Han DP, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA Ophthalmol*. 2013;131(3):351-357. <https://doi.org/10.1001/jamaophthalmol.2013.1743>
9. Decenière E, Zhang X, Cazuguel G, et al. FEEDBACK ON A PUBLICLY DISTRIBUTED IMAGE DATABASE: THE MESSIDOR DATABASE. 2014. 2014;33(3):4. <https://doi.org/10.5566/ias.1155>
10. Zhou Y, Chia MA, Wagner SK, et al. A foundation model for generalizable disease detection from retinal images. *Nature*. 2023;622(7981):156-163. <https://doi.org/10.1038/s41586-023-06555-x>
11. Men Y, Fhima J, Celi LA, Ribeiro LZ, Nakayama LF, Behar JA. DRStageNet: Deep learning for diabetic retinopathy staging from fundus images. *arXiv preprint arXiv:231214891*. 2023.
12. Dugas E, Jared J, Cukierski W. Diabetic retinopathy detection (2015). URL <https://kaggle.com/competitions/diabetic-retinopathy-detection.7>.
13. Kermany DS, Goldbaum M, Cai W, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*. 2018;172(5):1122-1131 e1129. <https://doi.org/10.1016/j.cell.2018.02.010>
14. FDA. De Novo Classification Request for IDx-DR, DEN180001, Decision Summary. [Available from: https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN180001.pdf.
15. Grading diabetic retinopathy from stereoscopic color fundus photographs--an extension of the modified Airlie House classification. ETDRS report number 10. *Early Treatment Diabetic Retinopathy*

- Study Research Group. *Ophthalmology*. 1991;98(5 Suppl):786-806. <https://www.ncbi.nlm.nih.gov/pubmed/2062513>
16. Wong TY, Sabanayagam C. Strategies to Tackle the Global Burden of Diabetic Retinopathy: From Epidemiology to Artificial Intelligence. *Ophthalmologica*. 2020;243(1):9-20. <https://doi.org/10.1159/000502387>
 17. Teo ZL, Tham YC, Yu M, et al. Global Prevalence of Diabetic Retinopathy and Projection of Burden through 2045: Systematic Review and Meta-analysis. *Ophthalmology*. 2021;128(11):1580-1591. <https://doi.org/10.1016/j.ophtha.2021.04.027>
 18. Yau JW, Rogers SL, Kawasaki R, et al. Global prevalence and major risk factors of diabetic retinopathy. *Diabetes Care*. 2012;35(3):556-564. <https://doi.org/10.2337/dc11-1909>
 19. Ipp E, Liljenquist D, Bode B, et al. Pivotal Evaluation of an Artificial Intelligence System for Autonomous Detection of Referrable and Vision-Threatening Diabetic Retinopathy. *JAMA Netw Open*. 2021;4(11):e2134254. <https://doi.org/10.1001/jamanetworkopen.2021.34254>
 20. Owsley C, McGwin G, Jr., Lee DJ, et al. Diabetes eye screening in urban settings serving minority populations: detection of diabetic retinopathy and other ocular findings using telemedicine. *JAMA Ophthalmol*. 2015;133(2):174-181. <https://doi.org/10.1001/jamaophthalmol.2014.4652>
 21. Rajesh AE, Davidson OQ, Lee CS, Lee AY. Artificial Intelligence and Diabetic Retinopathy: AI Framework, Prospective Studies, Head-to-head Validation, and Cost-effectiveness. *Diabetes Care*. 2023;46(10):1728-1739. <https://doi.org/10.2337/dci23-0032>
 22. Lee AY, Yanagihara RT, Lee CS, et al. Multicenter, Head-to-Head, Real-World Validation Study of Seven Automated Artificial Intelligence Diabetic Retinopathy Screening Systems. *Diabetes Care*. 2021;44(5):1168-1175. <https://doi.org/10.2337/dc20-1877>
 23. Riotta E, Gasser S, Potic J, et al. Accuracy of Autonomous Artificial Intelligence-Based Diabetic Retinopathy Screening in Real-Life Clinical Practice. *J Clin Med*. 2024;13(16). <https://doi.org/10.3390/jcm13164776>
 24. Chen EM, Chen D, Chilakamarri P, Lopez R, Parikh R. Economic Challenges of Artificial Intelligence Adoption for Diabetic Retinopathy. *Ophthalmology*. 2021;128(3):475-477. <https://doi.org/10.1016/j.ophtha.2020.07.043>
 25. Mehra AA, Softing A, Guner MK, Hodge DO, Barkmeier AJ. Diabetic Retinopathy Telemedicine Outcomes With Artificial Intelligence-Based Image Analysis, Reflex Dilation, and Image Overread. *Am J Ophthalmol*. 2022;244:125-132. <https://doi.org/10.1016/j.ajo.2022.08.008>
 26. Rajpurkar P, Lungren MP. The Current and Future State of AI Interpretation of Medical Images. *N Engl J Med*. 2023;388(21):1981-1990. <https://doi.org/10.1056/NEJMr2301725>
 27. Abramoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med*. 2018;1:39. <https://doi.org/10.1038/s41746-018-0040-6>
 28. Aftab O, Khan H, VanderBeek BL, Scoles D, Kim BJ, Tsui JC. Evaluation of ChatGPT-4 in detecting referable diabetic retinopathy using single fundus images. *AJO International*. 2025;2(2):100111. <https://doi.org/https://doi.org/10.1016/j.ajoint.2025.100111>
 29. Chen B, Zhang Z, Langrené N, Zhu S. Unleashing the potential of prompt engineering for large language models. *Patterns*. 2025;6(6):101260. <https://doi.org/https://doi.org/10.1016/j.patter.2025.101260>