Clinical Ophthalmology

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Assessing the Diagnostic Capabilities of ChatGPT-4 Omni in Grading Diabetic Retinopathy Fundoscopy Using Color Fundus Photographs

Nitin Chetla[1], Sai S Samayamanthula[1], Joseph He Chang[2], Arnold Y Leigh[3], Sinan Akosman[3], Mihir Tandon[4], Tamer R Hage[5], Michael Cusick[1]

[1]University of Virginia School of Medicine, Charlottesville, VA, USA; [2]University of Passau, Passau, Germany; [3]George Washington School of Medicine and Health Sciences, Washington, DC, USA; [4]Albany Medical College, Albany, NY, USA; [5]Virginia Tech, Blacksburg, VA, USA

Correspondence: Sai S Samayamanthula, University of Virginia School of Medicine, 828 Cabell Ave Apt F, Charlottesville, VA, 22903, USA, Tel +1 571 466 7455, Email Sss3tj@virginia.edu

Purpose: Diabetic retinopathy (DR) is a leading cause of vision loss in working-age adults. Despite the importance of early DR detection, only 60% of patients with diabetes receive recommended annual screenings due to limited eye care provider capacity. FDA-approved AI systems were developed to meet the growing demand for DR screening; however, high costs and specialized equipment limit accessibility. More accessible and equally as accurate AI systems need to be evaluated to combat this disparity. This study evaluated the diagnostic accuracy of ChatGPT-4 Omni (GPT-4o) in classifying DR from color fundus photographs (CFPs) to assess its potential as a low-cost alternative screening tool.

Methods: We utilized the publicly available EyePACS DR detection competition dataset from Kaggle, which includes 2,500 CFPs representing no DR, mild DR, moderate DR, severe DR, and proliferative DR. Each image was presented to GPT-4o with 1 of 8 prompts designed to enhance the model's accuracy. The results were analyzed through confusion matrices, and metrics such as accuracy, precision, sensitivity, specificity, and F1 scores were calculated to evaluate performance.

Results: In prompts 1–3, GPT-4o showed a strong bias towards classifying images as no DR, with an average accuracy of 51.0%, while accuracy for other stages ranged from 70% to 80%. GPT-4o struggled with misclassifications, particularly between adjacent DR levels. It performed best in detecting proliferative DR (Level 4), achieving an F1 score above 0.3 and accuracy exceeding 80%. In binary classification tasks (Prompts 4.1–4.4), GPT-4o's performance improved, though it still had difficulty distinguishing mild DR (49.8% accuracy). When compared to FDA-approved AI systems, GPT-4o's sensitivity (47.7%) and specificity (73.8%) were significantly lower.

Conclusion: While GPT-4o shows promise identifying severe DR, limitations in distinguishing early stages exist and highlight the need for further refinement before clinical usage in DR screening. Unlike traditional CNN-based tools like IDx-DR, GPT-4o is a multimodal foundation model with a fundamentally different architecture and training process, which may contribute to its diagnostic limitations. GPT-4o and other LLMs are not designed to learn about important DR features like microaneurysms or hemorrhages using pixel data which is why they may struggle to detect DR compared to CNN models.

Keywords: artificial intelligence, large language model, diabetes, EyePACS, eye screening, multimodal AI

Introduction

Diabetic retinopathy (DR) is the most common microvascular complication of diabetes affecting the eyes and a leading cause of vision loss in working-age adults.[1] In the United States, the number of individuals with DR by the year 2050 is projected to reach 16 million adults older than 40 years, of whom 3.4 million will have vision-threatening retinopathy.[2] The growing prevalence of DR is a significant public health concern due to its socioeconomic impact.[3] People with diabetes fear vision loss and blindness more than any other complication of the disease.[4] However, patients often remain visually asymptomatic until more advanced stages of retinopathy develop. Within 20 years of diagnosis of diabetes,

Received: 13 January 2025
Accepted: 31 July 2025
Published: 31 August 2025

nearly all patients with type 1 diabetes and over 60% of those with type 2 diabetes will develop DR.[5] The impact of worsening DR on patient quality of life is significant, affecting daily activities such as reading, driving, and the ability to work.[6,7] Fortunately, early detection and treatment can prevent up to 98% of blindness caused by DR.[8] Therefore, screening for DR is an important public health issue and cost-effective component of the care for patients with diabetes.[9,10] Despite its importance, only approximately 60% of patients with diabetes receive the recommended annual screenings.[11] The growing prevalence of diabetes, combined with the limited capacity of eye care providers to perform an increasing number of screenings, has created a need for alternative screening options in order to fulfill the clinical recommendations for annual screenings.

Artificial intelligence (AI) systems using color fundus photographs (CFPs) have been developed for DR screening to help address the increased demand for screening. In the United States, two AI systems for detecting DR have been approved by the Food and Drug Administration (FDA). In 2018, the FDA approved the IDx-DR system, which has shown 87.2% sensitivity and 90.7% specificity for detecting more than mild DR (mtmDR) using the Topcon NW400 non-mydriatic fundus camera.[12] Similarly, the EyeArt system, initially approved by the FDA in 2020, demonstrated screening with 96% sensitivity and 88% specificity for detecting mtmDR using the Canon CR-2 AF and Canon CR-2 Plus AF cameras.[13] In 2023, EyeArt v2.2.0 received FDA clearance to use the Topcon NW400 retinal camera with new data revealing a sensitivity of 94.4% sensitivity and specificity of 91.1% for mtmDR.[14] Both the IDx-DR and EyeArt AI systems use deep learning models with multiple neural networks designed for specific classification tasks.

ChatGPT, an AI system developed by OpenAI, utilizes a large language model (LLM) to answer questions and has demonstrated impressive capabilities in comprehending clinical expertise and providing relevant information, even in tasks for which it was not specifically trained (zero-shot learning).[15] ChatGTP-4 Omni (ChatGPT-4o), the latest version of the AI system, can perform more complex tasks than previous versions with advanced visual capabilities, such as describing photographs and generating captions for images. These enhancements offer promising applications in ophthalmology, particularly in automating diagnostics.[16] Unlike the FDA-approved AI systems for DR detection, which require significant up-front costs and specialized equipment, GPT-4o is freely accessible. While the IDx-DR and EyeArt systems demonstrate high sensitivity and specificity for detecting mtmDR, the costs combined with declining reimbursements for office-based retinopathy screening may make implementing these systems cost-prohibitive for some practices. Therefore, GPT-4o may offer a viable, low-cost alternative for DR screening, provided its diagnostic accuracy proves comparable. Additionally, with over 180 million users and approximately 600 million visits per month, ChatGPT's widespread use makes it a highly influential tool.[17] Its popularity highlights the importance of rigorously testing the model's accuracy, especially in medical contexts where users may rely on it for self-evaluation or interpreting medical images. Therefore, the purpose of this study is to assess the diagnostic accuracy of ChatGPT-4o in identifying the various stages of DR using CFPs.

Materials and Methods

We used the publicly available DR detection competition dataset from Kaggle, which contains 2,500 high-resolution CFPs in JPEG format, provided by EyePACS, a free platform for retinopathy screening.[18] This competition, sponsored by the California Healthcare Foundation in 2015, was launched with the support of a data science team to introduce artificial intelligence in the detection of DR. The images, consisting of both left and right eye photographs taken under various conditions, were graded by a single specialist according to the standards of the International Clinical Diabetic Retinopathy severity scale. The dataset includes 500 images per category, graded on a scale from 0 to 4, representing different stages of DR: 0 indicating no DR, 1 mild DR, 2 moderate DR, 3 severe DR, and 4 proliferative DR.[19] To our knowledge, GPT-4o was not trained on labeled retinal datasets, underscoring the zero-shot nature of the evaluation and highlighting the model's generalizability to previously unseen imaging domains.

Each image was processed and then presented to ChatGPT-4o individually with the following initial prompt (Prompt 1) designed to simulate a examination scenario:

> This is an image on the Step 1 examination, the multiple choice question is as follows. Based on the image, does the patient have A) no Diabetic Retinopathy, B) mild Diabetic Retinopathy, C) moderate Diabetic Retinopathy, D) severe Diabetic Retinopathy, E) proliferative Diabetic Retinopathy? Only output the answer as A, B, C, D, or E.

This format was chosen because of ChatGPT's demonstrated success in performing well on USMLE Step exams. In prior studies, the model achieved accuracy scores at or near passing threshold (60%) for all three USMLE exams (Step 1, Step 2CK, and Step 3).[20] By using a clear, concise prompt in this familiar exam format, we aimed to harness ChatGPT's high accuracy, concordance, and insight demonstrated during its USMLE performance. This approach was intended to reduce potential confusion and encourage the model to attend to relevant features of the image without the added complexity of elaborate instructions.

To explore different angles of analysis and evaluate the robustness of the model's performance, we repeated the process with seven additional prompts. This decision was driven by the principles of prompt engineering, a method of providing precise instructions to large language models (LLMs) that has been shown to improve their performance in various applications. This technique involves carefully crafting prompts that guide the model in a logical manner, ensuring it generates accurate and relevant outputs. By specifying the context, structure, and format of the response, prompt engineering helps LLMs better understand complex queries and produce more useful, precise responses.[21] By refining and varying the prompts, we aimed to test whether we could enhance the model's ability to accurately diagnose DR.

First, the initial prompt was slightly modified (Prompt 2) to simulate a real clinical setting:

> This is an image found on clinical examination. Based on the image, does the patient have A) no Diabetic Retinopathy, B) mild Diabetic Retinopathy, moderate Diabetic Retinopathy, D) severe Diabetic Retinopathy, E) proliferative Diabetic Retinopathy? Only output the answer as A, B, C, D, or E.

We then used a more detailed prompt (Prompt 3) adapted from AlRyalat et al[22] to leverage the model's potential by providing a specific role-playing scenario:

> Hello ChatGPT, you are simulating an ophthalmologist with a specialization in identifying diabetic retinopathy using fundus photographs. Your task is to perform a preliminary analysis of the attached fundus photographs to determine whether they show signs of diabetic retinopathy. Based on the image, does the patient have A) no Diabetic Retinopathy, B) mild Diabetic Retinopathy, C) moderate Diabetic Retinopathy, D) severe Diabetic Retinopathy, E) proliferative Diabetic Retinopathy? Only output the answer as A, B, C, D, or E.

This role-playing approach was intended to guide the model to shift processing from general data analysis to more focused, knowledge-based decision-making and prioritize relevant information for disease diagnosis.

Next, we employed the following four comparative prompts (Prompts 4.1–4.4) to determine if simplifying the decision-making process could improve accuracy:

> This is an image found on examination, the multiple choice question is as follows. Based on the image, does the patient have A) no Diabetic Retinopathy, B) has mild Diabetic Retinopathy. Only output the answer as A or B.

> This is an image found on examination, the multiple choice question is as follows. Based on the image, does the patient have A) no Diabetic Retinopathy, C) has moderate Diabetic Retinopathy. Only output the answer as A or C.

> This is an image found on examination, the multiple choice question is as follows. Based on the image, does the patient have A) no Diabetic Retinopathy, D) has severe Diabetic Retinopathy. Only output the answer as A or D.

> This is an image found on examination, the multiple choice question is as follows. Based on the image, does the patient have A) no Diabetic Retinopathy, E) has proliferative Diabetic Retinopathy. Only output the answer as A or E.

To directly compare ChatGPT's accuracy with the previously mentioned IDx-DR and EyeArt AI systems, both specifically approved for identifying mtmDR, we designed the following final prompt (Prompt 5):

This is a color fundoscopy image from an exam. Based on image, is this A) moderate Diabetic Retinopathy, B) severe Diabetic Retinopathy, or C) proliferative Diabetic Retinopathy? Only output the answer as A, B, or C.

By limiting the answer choices to moderate, severe, and proliferative DR, we aimed to evaluate whether GPT could match the performance of these specialized AI tools when analyzing images of this particular severity level.

For each image, the diagnostic accuracy of ChatGPT-4o was compared against the provided labels. The image analysis was conducted between July 26, 2024, and October 8, 2024. To visualize ChatGPT-4o's predictions for each image and prompt, a confusion matrix was constructed. As the name suggests, a confusion matrix provides a representation of where the model's predictions align with or deviate from the actual categories. It consists of four key components: true positives, true negatives, false positives, and false negatives. Using these matrices, accuracy (Acc), precision (Pre), recall (TPR), sensitivity (Sen), specificity (Spe), positive predictive value (PPV), negative predictive value (NPV), and F1 score were calculated. Instances where ChatGPT-4o was indecisive, refused to diagnose, or left responses blank were considered null values and excluded from the study. This research has been reviewed by the University of Virginia Institutional Review Board (IRB) and was deemed as non-human subject research and exempt from IRB oversight.

Results

In both Prompts 1 and 2, where the entire dataset was utilized, ChatGPT-4o exhibited a strong bias towards classifying images as no DR (Level 0). The accuracy for Prompts 1 through 3 for no DR was the lowest among all categories, averaging 51.0%, whereas the other categories showed accuracies around 70–80%. This was largely due to a high number of false positives, despite having the highest number of true positives for each prompt. This is visually represented by the light-colored squares in the "predicted 0" column of the confusion matrices for these prompts, indicating that ChatGPT-4o frequently misclassified images from other categories as Level 0 (Figure 1).

Nevertheless, ChatGPT was still able to identify or at least attempt to identify some images from higher stages of retinopathy. For severe DR (Level 3), it began distinguishing between Level 3 and Level 0, but many images at this level were incorrectly classified as Level 2, which represents the midpoint between stages.

At proliferative DR (Level 4), the model demonstrated its best performance (excluding Level 0) across all prompts, achieving a higher F1 score and accuracy. In all three prompts, the F1 score exceeded 0.3, and accuracy was above 0.8 (Table 1) (Figure 2).

Interestingly, ChatGPT performed significantly better in binary classification than in multi-classification. In Prompts 4.1–4.4, it achieved much higher success in distinguishing between no DR and all other stages of retinopathy, except for mild DR (Level 1), where it showed a 49.8% accuracy - similar to its accuracy for no DR in Prompts 1 through 3 (Table 2A).

When comparing ChatGPT-4o's performance in detecting mtmDR to current FDA-approved AI systems, it showed much lower sensitivity and specificity. In Prompt 5, ChatGPT-4o's overall sensitivity was 47.7% and its specificity 73.8%, both significantly lower than those of IDx-DR and EyeArt (Table 2B). The IDx-DR system achieved 87% sensitivity and 90% specificity, while the original EyeArt system demonstrated 96% sensitivity and 88% specificity. EyeArt v2.2.0 further improved with 94.4% sensitivity and 91.1% specificity (Figure 3).

Discussion

Our findings highlight the significant potential of ChatGPT-4o in assisting with the classification of diabetic retinopathy (DR), especially in detecting more severe cases. In Prompts 4.1 through 4.4, GPT-4o showed increased accuracy, precision, sensitivity, and specificity when tasked with binary comparisons, particularly for proliferative DR, where values reached 75.6%, 92.2%, 56.4%, and 95.2%, respectively.

However, the model exhibits critical limitations, particularly in distinguishing between milder forms of the disease as seen in Prompts 1–5. In the context of AI-based DR screening, systems like IDx-DR and EyeArt have set a high benchmark for sensitivity and specificity. In a study using the Messidor-2 dataset, IDx-DR achieved a sensitivity of 96.8% and a specificity of 87.0% for detecting referable DR. In a retrospective analysis of 78,685 patient encounters,

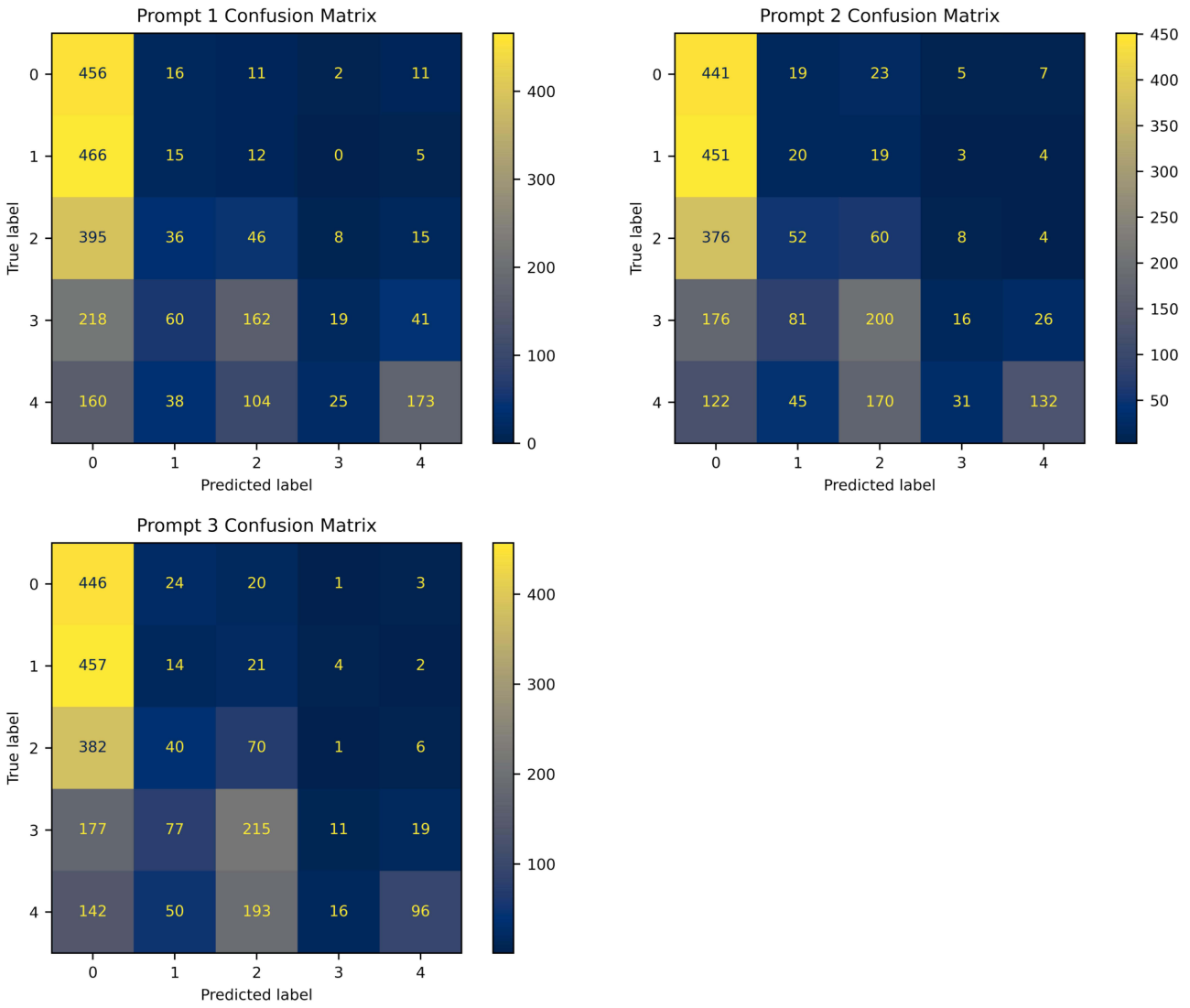

Figure 1 Confusion matrix for Prompts 1, 2, and 3. Prompt 1) ChatGPT-4o showed a strong bias toward predicting no DR, resulting in a high number of false negatives for the other stages of DR. The model struggled with distinguishing between adjacent DR levels, particularly between Level 2, 3, and 4, leading to frequent misclassification across these categories. Prompt 2) ChatGPT-4o continued to exhibit a strong bias towards predicting no DR, with 441 true positives. A significant number of images from other levels, particularly Level 1 (451) and Level 2 (376), were misclassified as Level 0. While there was a slight improvement in correctly identifying images across various stages compared to Prompt 1, the model still struggled to distinguish between adjacent stages of DR, especially between Levels 2, 3, and 4. A notable portion of Level 3 images were misclassified as Level 2 and Level 4 images were frequently split between being identified as Level 3 and Level 4. This suggests that, although the modified prompt led to some improvement, the model's diagnostic accuracy remains limited, particularly in differentiating between mild and moderate cases. Prompt 3) ChatGPT-4o continued to exhibit a strong bias towards predicting no DR, with the vast majority of images (1,604) being classified as no DR. Similar to Prompts 1 and 2, the model struggled to distinguish between adjacent stages of DR, notably misclassifying 457 Level 1 images as Level 0 and 215 Level 3 images as Level 2. While the model performed relatively better in identifying images from Level 4, it interestingly performed worse in Prompt 3 compared to Prompts 1 and 2. In this case, it correctly identified 96 Level 4 images, a decrease from 173 in Prompt 1 and 132 in Prompt 2.

EyeArt achieved a sensitivity of 91.7% and a specificity of 91.5% for referable DR. EyeArt was also tested using smartphone-based fundus photography in a study of 296 patients, achieving a sensitivity of 95.8% for any DR, 99.3% for referable DR, and 99.1% for vision-threatening DR.[23]

Unlike IDx-DR and EyeArt, which were designed specifically for DR screening and leverage training from extensive image datasets, ChatGPT-4o's foundation as a LLM inherently limits its accuracy in image analysis. Even though IDx-DR and EyeArt excel at telling the severity of DR, the way they do so is not transparent because of deep learning, so it is hard to say if they depend on features like microaneurysms, hemorrhages and neovascularization. As a result, the model is more prone to underestimating the severity of retinopathy and experiences difficulty in handling subtle gradations between DR stages. This performance gap highlights a critical difference between general-purpose AI and specialized,

https://doi.org/10.2147/OPTH.S517238

Table 1 Statistical Measurements of Prompts 1–3

	Sensitivity	Specificity	PPV	NPV	F1	Accuracy
Prompt 1						
Level 0	0.919	0.38	0.269	0.95	0.416	0.487
Level 1	0.03	0.925	0.091	0.793	0.045	0.746
Level 2	0.092	0.855	0.137	0.79	0.11	0.702
Level 3	0.038	0.982	0.352	0.803	0.069	0.793
Level 4	0.346	0.964	0.706	0.855	0.464	0.84
Prompt 2						
Level 0	0.891	0.436	0.282	0.942	0.428	0.527
Level 1	0.04	0.901	0.092	0.79	0.056	0.729
Level 2	0.12	0.793	0.127	0.782	0.123	0.658
Level 3	0.032	0.976	0.254	0.801	0.057	0.787
Level 4	0.264	0.979	0.763	0.841	0.392	0.836
Prompt 3						
Level 0	0.903	0.419	0.278	0.946	0.425	0.515
Level 1	0.028	0.904	0.068	0.788	0.04	0.729
Level 2	0.14	0.774	0.135	0.782	0.138	0.647
Level 3	0.022	0.989	0.333	0.801	0.041	0.795
Level 4	0.193	0.985	0.762	0.83	0.308	0.827

Notes: Sensitivity, specificity, PPV (positive predictive value), NPV (negative predictive value), F1 score, and accuracy metrics were calculated for different levels of diabetic retinopathy (DR). There is a slight improvement in ChatGPT-4o's diagnostic performance, particularly for proliferative DR (Level 4). Sensitivity, specificity, and F1 scores showed consistent increases for this stage across the prompts. However, the model struggled with lower levels of DR, especially mild DR (Level 1), where sensitivity and F1 scores remained exceptionally low, highlighting a persistent difficulty in detecting early-stage DR. The accuracy for no DR (Level 0) is relatively high compared to other levels, but this is coupled with low specificity, leading to frequent false positives. These trends suggest that while the model performs better in identifying late-stage DR, it requires further refinement to enhance detection accuracy for earlier stages of the disease.

clinically validated AI systems.[24] This limitation arises because GPT-4o, unlike convolutional neural networks (CNNs), is not trained to recognize pixel-level features such as microaneurysms or hemorrhages, which are essential for accurate DR diagnosis.

The lower upfront cost and accessibility of ChatGPT-4o present a compelling argument for its continued development as a potential screening tool. However, the trade-offs in accuracy would need to be addressed before it can be considered a suitable replacement. Looking forward, improving the model's diagnostic accuracy for DR classification could involve integrating its language-based capabilities with specialized image recognition models like IDx-DR and EyeArt. Additionally, training ChatGPT-4o on large, diverse datasets of color fundus images may help improve its ability to recognize early-stage DR. Importantly, the dataset used in this study, sourced from Kaggle, is curated for image quality, which may not reflect the variability and imperfections found in real-world clinical data. Furthermore, the lack of accompanying clinical demographic information limits the model's ability to contextualize findings, which is particularly relevant given that ChatGPT-4o is fundamentally a language model

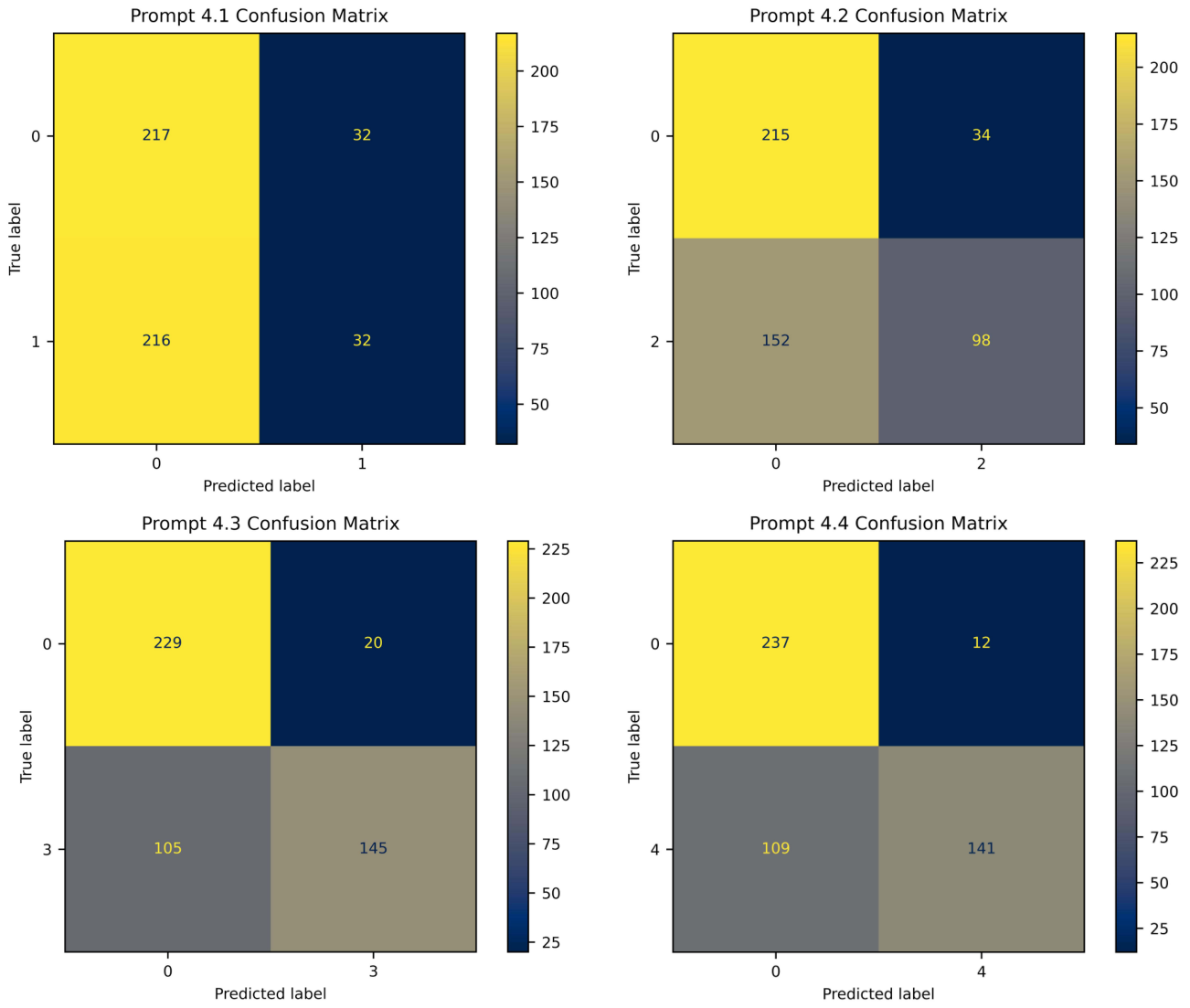

Figure 2 Confusion matrix for Prompts 4.1–4.4. Prompt 4.1) ChatGPT-4o exhibited a strong bias toward classifying images as no DR (Level 0) with the vast majority of images (433) being classified as no DR. The model struggled to correctly identify mild DR (Level 1), with only 32 true positives. This significant misclassification rate between Levels 0 and 1 suggests that the model struggles to differentiate between early stages of DR. Prompt 4.2). ChatGPT-4o demonstrated moderate accuracy in distinguishing between no DR (Level 0) and moderate DR (Level 2). It correctly identified 215 instances of Level 0, but there were still 34 instances where Level 0 was misclassified as Level 2. Conversely, while 98 images were correctly identified as Level 2, a significant number of images (152) were misclassified as Level 0. This indicates that the model struggles with sensitivity, often underestimating the severity of the condition and incorrectly labeling moderate retinopathy as no retinopathy. Prompt 4.3) ChatGPT-4o showed improved performance in identifying severe DR (Level 3). It correctly classified 229 images as no DR (Level 0), with only 20 misclassified as Level 3. However, 105 images of severe DR were still misclassified as no DR, indicating the model still underestimates severity in many cases. Prompt 4.4) ChatGPT-4o showed reasonable performance in identifying proliferative DR (Level 4), correctly classifying 141 images. However, with a sensitivity of only 0.564, the model still misclassified 109 Level 4 images as no DR (Level 0). This indicates that while the model has moderate ability to detect severe cases, its reliability in accurately identifying proliferative DR remains limited.

rather than a dedicated image analysis tool. Only basic prompt engineering strategies were used, suggesting that more sophisticated prompting techniques may yield better performance. Future work should also focus on refining the model's ability to detect subtle changes between DR stages by using prompt engineering techniques that guide the model toward more precise image interpretation.

As ChatGPT becomes more and more accessible, its potential role in clinical settings raises important questions. Like how both physicians and patients increasingly rely on Google for medical information, ChatGPT's ability to process complex data, including diagnostic images, offers the potential to become a valuable point-of-care resource. It could assist clinicians by providing immediate insights during consultations. Rather than simply serving as a screening tool, AI could shape how information is accessed and interpreted during clinical encounters, potentially affecting patient trust and

https://doi.org/10.2147/OPTH.S517238

Table 2 (A,B) Statistical Measurements of Prompts 4.1–4.4 and Prompt 5

	Accuracy	Precision	Recall	F1	Specificity
A. Prompts 4.1–4.4					
Prompt 4.1 (0 vs 1)	0.498	0.5	0.129	0.205	0.871
Prompt 4.2 (0 vs 2)	0.627	0.742	0.392	0.513	0.863
Prompt 4.3 (0 vs 3)	0.749	0.879	0.58	0.699	0.92
Prompt 4.4 (0 vs 4)	0.756	0.922	0.564	0.7	0.952
B. Prompt 5					
Level 2	0.562	0.414	0.797	0.545	0.446
Level 3	0.661	0.466	0.084	0.142	0.952
Level 4	0.727	0.604	0.55	0.575	0.817

Notes: ChatGPT-4o's performance improved as the severity of diabetic retinopathy (DR) increased. Accuracy, precision, sensitivity, and F1 scores were lowest for Prompt 4.1 (comparing no DR to mild DR), indicating a significant challenge for the model in distinguishing early stages of the disease. In contrast, the highest performance metrics are seen in Prompt 4.4 (comparing no DR to proliferative DR), with accuracy and F1 scores reaching 0.756 and 0.7, respectively. This trend suggests that while the model is more reliable in identifying severe DR, it struggles with early detection, underscoring the need for improvements in recognizing subtle differences between early-stage DR and healthy images. Statistical measurements of Prompt 5. ChatGPT-4o struggled the most with early stages of diabetic retinopathy (DR), particularly moderate DR (Level 2), where it showed a high sensitivity but poor precision and specificity, leading to many false positives. For severe DR (Level 3), the model demonstrated significant difficulty, reflected in the low recall and F1 scores, indicating it fails to identify severe DR accurately. The model performed best for proliferative DR (Level 4), with relatively balanced and higher accuracy, precision, sensitivity, and F1 scores, suggesting it is more reliable for detecting advanced stages of DR.

the dynamics of the physician-patient relationship. While the model offers promise, its limitations must be carefully addressed before it can be integrated into routine clinical practice. Further studies are essential to continue to evaluate its performance as newer versions of the model are released and define the conditions for its safe use in healthcare.

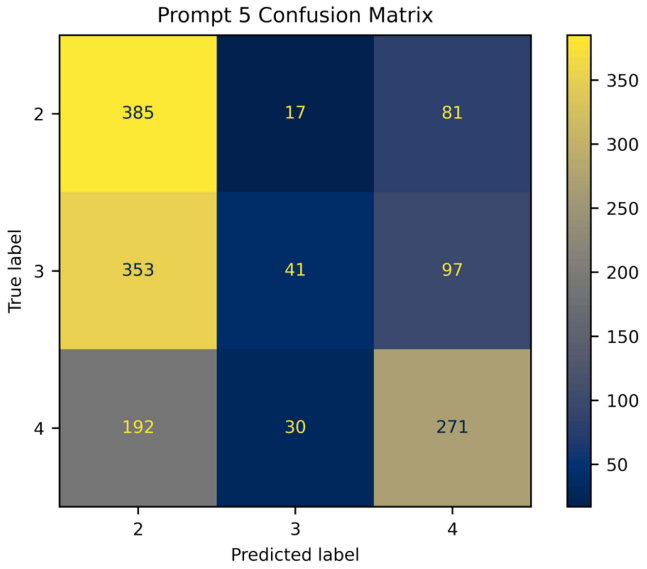

Figure 3 Confusion matrix for Prompt 5. The model showed a strong performance in identifying moderate DR (Level 3) and proliferative DR (Level 4), with a significant number of true positives. The model correctly classified 385 images as moderate DR and 271 images as proliferative DR. However, the model severely underperforms in identifying severe DR, as reflected by the low number of true positives (41) and the large number of misclassifications, with many severe DR cases being mistaken for moderate DR or proliferative DR.

Conclusion

In this paper, we examined the degree of accuracy in which ChatGPT-4 Omni can correctly grade and classify diabetic retinopathy fundoscopy, aiming to provide insight into low-cost and readily accessible alternatives to FDA-approved AI systems currently on the market that can assist in diagnosing diabetic retinopathy. This study analyzed 2,500 high-resolution color fundus photographs, revealing that ChatGPT-4 Omni achieved improved diagnostic metrics—accuracy (75.6%), precision (92.2%), sensitivity (56.4%), and specificity (95.2%)—in binary comparisons involving proliferative diabetic retinopathy. Ultimately, ChatGPT-4 Omni continues to show promise in one day becoming a diagnostic tool or support device in adequately evaluating for severe diabetic retinopathy, but several limitations and ethical considerations remain significant compared to FDA-approved AI systems.

Abbreviations

DR, Diabetic retinopathy; GPT-4o, ChatGPT-4 Omni; CFPs, color fundus photographs; AI, Artificial intelligence; FDA, Food and Drug Administration; mtmDR, mild DR; LLM, large language model; Acc, accuracy; Pre, precision; TPR, recall; Sen, sensitivity; Spe, specificity; PPV, positive predictive value; NPV, negative predictive value.

Acknowledgement

The abstract of this paper was published in "Poster Abstracts" in Investigative Ophthalmology and Visual Science: https://iovs.arvojournals.org/article.aspx?articleid=2805764#:~:text=Conclusions%20%3A%20While%20GPT%2D4o%20shows,clinical%20use%20in%20DR%20screening.

Funding

There is no funding to report.

Disclosure

The author(s) report no conflicts of interest in this work.

References

1. Fong DS, Aiello L, Gardner TW, et al. Retinopathy in diabetes. *Diabetes Care*. 2004;27(Suppl 1):S84–S87. doi:10.2337/diacare.27.2007.S84
2. Saaddine JB, Honeycutt AA, Narayan KM, Zhang X, Klein R, Boyle JP. Projection of diabetic retinopathy and other major eye diseases among people with diabetes mellitus: united States, 2005–2050. *Arch Ophthalmol*. 2008;126(12):1740–1747. doi:10.1001/archopht.126.12.1740
3. Wittenborn JS, Zhang X, Feagan CW, et al. The economic burden of vision loss and eye disorders among the United States population younger than 40 years. *Ophthalmology*. 2013;120(8):1728–1735. doi:10.1016/j.ophtha.2013.01.068
4. Hendricks LE, Hendricks RT. Greatest fears of type 1 and type 2 patients about having diabetes: implications for diabetes educators. *Diabetes Educator*. 1998;24(2):168–173. doi:10.1177/014572179802400206
5. Fong DS, Aiello L, Gardner TW, et al. Diabetic retinopathy. *Diabetes Care*. 2003;26(Suppl 1):S99–S102. doi:10.2337/diacare.26.2007.S99
6. Mazhar K, Varma R, Choudhury F, et al. Severity of diabetic retinopathy and health-related quality of life: the Los Angeles Latino Eye Study. *Ophthalmology*. 2011;118(4):649–655. doi:10.1016/j.ophtha.2010.08.003
7. Willis JR, Doan QV, Gleeson M, et al. Vision-related functional burden of diabetic retinopathy across severity levels in the United States. *JAMA Ophthalmol*. 2017;135(9):926–932. doi:10.1001/jamaophthalmol.2017.2553
8. Wong TY, Sun J, Kawasaki R, et al. Guidelines on diabetic eye care: the International Council of Ophthalmology recommendations for screening, follow-up, referral, and treatment based on resource settings. *Ophthalmology*. 2018;125(10):1608–1622. doi:10.1016/j.ophtha.2018.04.007
9. Javitt JC, Aiello LP. Cost-effectiveness of detecting and treating diabetic retinopathy. *Ann Internal Med*. 1996;124(1 Pt 2):164–169. doi:10.7326/0003-4819-124-1_Part_2-199601011-00017
10. Rohan TE, Frost CD, Wald NJ. Prevention of blindness by screening for diabetic retinopathy: a quantitative assessment. *BMJ*. 1989;299(6709):1198–1201. doi:10.1136/bmj.299.6709.1198
11. American Academy of Ophthalmology Retina/Vitreous Panel. Preferred practice pattern® guidelines: diabetic retinopathy. San Francisco, CA: American Academy of Ophthalmology; 2019. Available from: www.aao.org/ppp. Accessed August 12, 2025.
12. Abràmoff 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 Digital Med*. 2018;1(39). doi:10.1038/s41746-018-0040-6
13. 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 [published correction appears in JAMA. *Network Open*. 2021;4(12):e2144317.
14. Eyenuk. New FDA clearance makes Eyenuk the first company with multiple cameras for autonomous AI detection of diabetic retinopathy; 2023. Available from https://www.eyenuk.com/us-en/articles/news/eyenuk-fda-multiple-cameras/. Accessed August 12, 2025.
15. Ziegelmayer S, Marka AW, Lenhart N, et al. Evaluation of GPT-4's chest X-ray impression generation: a reader study on performance and perception. *J Med Internet Res*. 2023;25:e50865. doi:10.2196/50865

16. Ting DSJ, Tan TF, Ting DSW. ChatGPT in ophthalmology: the dawn of a new era? *Eye*. 2023;38(1):4–7. doi:10.1038/s41433-023-02619-4
17. Duarte F. Number of ChatGPT users (Aug 2024). Exploding Topics; 2024. Available from: https://explodingtopics.com/blog/chatgpt-users. Accessed August 12, 2025.
18. Dugas E, Jared G, Cukierski W. Diabetic retinopathy detection. Kaggle; 2015. Available from: https://www.kaggle.com/c/diabetic-retinopathy-detection. Accessed August 12, 2025.
19. Tariq M, Palade V, Ma Y. Transfer learning based classification of diabetic retinopathy on the Kaggle EyePACS dataset; 2023. doi:10.13140/RG.2.2.14506.29126.
20. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digital Health*. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198
21. Wada A, Akashi T, Shih G, et al. Optimizing GPT-4 turbo diagnostic accuracy in neuroradiology through prompt engineering and confidence thresholds. *Diagnostics*. 2024;14(14):1541. doi:10.3390/diagnostics14141541
22. AlRyalat SA, Musleh AM, Kahook MY. Evaluating the strengths and limitations of multimodal ChatGPT-4 in detecting glaucoma using fundus images. *Front Ophthalmol*. 2024;4. doi:10.3389/fopht.2024.138719
23. Grzybowski A, Brona P, Lim G, et al. Artificial intelligence for diabetic retinopathy screening: a review. *Eye*. 2020;34(3):451–460. doi:10.1038/s41433-019-0566-0
24. Lim JI, Regillo CD, Sadda SR, et al. Artificial intelligence detection of diabetic retinopathy. *Ophthalmol Sci*. 2023;3(1):100228. doi:10.1016/j.xops.2022.100228