MAEDICA – a Journal of Clinical Medicine
2025; 20(4): 765-770
https://doi.org/10.26574/maedica.2025.20.4.765

MÆDICA - a Journal of Clinical Medicine

ORIGINAL PAPER

Artificial Intelligence-Assisted Diagnosis of Retinal and Neuro-Ophthalmic Diseases: a Comparative Evaluation of ChatGPT, Google Gemini and An Ophthalmologist

SHRINKHAL[a], Aparajita SHUKLA[a], Mukesh SHUKLA[b], Pragati GARG[a], Ruchi SHUKLA[a], Swarastra Prakash SINGH[a]

[a]Department of Ophthalmology, All India Institute of Medical Sciences (AIIMS), Raebareli, India

[b]Department of SPM, All India Institute of Medical Sciences (AIIMS), Raebareli, India

ABSTRACT

Objectives: *This study evaluated the diagnostic performance of two large language models (LLMs), ChatGPT and Google Gemini, to identify common retinal and optic nerve diseases benchmarked against an experienced ophthalmologist.*

Methods: *Thirty standardized case vignettes, each comprising a brief clinical history and a high-resolution fundus image, were independently evaluated by ChatGPT, Google Gemini and an ophthalmologist. Ten retinal and optic nerve diseases were included. Diagnostic accuracy was calculated against a gold standard defined by consensus of two retina specialists. Inter-rater agreement was assessed using Cohen's kappa (κ). Secondary outcomes included interpretation time and clarity of explanation.*

Results: *The ophthalmologist achieved the highest diagnostic accuracy (96.7%), followed by ChatGPT (90.0%) and Google Gemini (86.7%). Agreement between ChatGPT and Gemini was moderate (κ = 0.51, p = 0.004). ChatGPT showed moderate agreement with the ophthalmologist (κ = 0.47, p = 0.002), while Gemini demonstrated fair agreement with the ophthalmologist (κ = 0.36, p = 0.01). ChatGPT was the fastest (mean 21.7 seconds), followed by Gemini (25.7 seconds) and the ophthalmologist (149.8 seconds). Clarity of interpretation was highest for the ophthalmologist (mean 4.53/5), followed by ChatGPT (3.60/5) and Gemini (2.96/5), with significant differences between groups.*

Conclusion: *Ophthalmologists remain superior in diagnostic accuracy and clarity. However, ChatGPT and Google Gemini demonstrated strong performance in several retinal conditions. Their rapid evaluation times indicate potential utility as adjunct tools in triage, screening and education.*

Keywords: artificial intelligence, ChatGPT, Google Gemini, retinal and neuro-ophthalmic diseases.

Address for correspondence:
DR. SWARASTRA PRAKASH SINGH, Additional Professor
Postal Address: Department of Ophthalmology, AIIMS, Raebareli, India
Tel.: 9415087772; email: swarastra@gmail.com

Article received on the 9th of September 2025 and accepted for publication on the 9th of December 2025

INTRODUCTION

Retinal diseases are among the leading causes of visual impairment and irreversible blindness worldwide, contributing substantially to the global burden of disease and healthcare costs (1). The retina, as an extension of the central nervous system, provides a unique opportunity to diagnose both ocular and systemic disorders early in their course (2). Common retinal pathologies such as diabetic retinopathy, age-related macular degeneration, myopic degeneration, vascular occlusions, retinal detachment and optic neuropathies account for a significant proportion of preventable blindness (1, 3, 4).

Despite advances in imaging and therapeutics, timely diagnosis remains a challenge in regions with limited access to vitreoretinal specialists (5). Artificial intelligence (AI) has transformed ophthalmology, with deep learning algorithms demonstrating near-human performance in the detection of conditions such as diabetic retinopathy, retinopathy of prematurity and macular degeneration (6, 7). However, these systems typically function as "black boxes", producing probabilistic outputs without integrating clinical history into their reasoning.

Large language models such as ChatGPT and Google Gemini represent a new paradigm by combining natural language processing with clinical reasoning, enabling them to synthesize patient history and imaging data to generate differential diagnoses in a human-readable format (8, 9). Although promising, few studies have systematically evaluated their diagnostic performance across multiple retinal diseases or compared them with human experts.

This study aimed to prospectively compare the diagnostic accuracy, efficiency and interpretive clarity of ChatGPT and Google Gemini against an experienced retina specialist across ten common retinal and optic nerve diseases. ❑

MATERIALS AND METHODS

Thirty anonymized case vignettes were curated from clinical records. Each vignette included patient demographics, a structured clinical history with presenting complaints and systemic associations and a high-quality fundus photograph (Table 1, Figure 1A and 1B). The conditions studied were diabetic retinopathy, optic atrophy, central retinal vein occlusion, dry age-related macular degeneration, central serous chorioretinopathy, myopic fundus, central retinal artery occlusion, retinitis pigmentosa, retinal detachment and papilledema.

The gold standard diagnosis for each vignette was established by consensus between two senior retina specialists who were not involved in the evaluation process. Cases were excluded if either fundus image quality was inadequate or multiple pathologies were present, if there was

Disease	Clinical features	Fundus image provided
Diabetic retinopathy	History of diabetes	Yes
Optic atrophy	Progressive vision loss following trauma	Yes
Central retinal vein occlusion	Sudden painless loss of vision	Yes
Dry age-related macular degeneration (AMD)	Elderly patient with gradual central vision blurring	Yes
Central serous chorioretinopathy	Young male with acute central scotoma	Yes
Myopic fundus	High myopia	Yes
Central retinal artery occlusion	Acute profound vision loss	Yes
Retinitis pigmentosa	Night blindness with peripheral field constriction	Yes
Retinal detachment	Flashes and floaters followed by a curtain-like shadow	Yes
Papilledema	Headache, nausea, transient visual obscurations	Yes

TABLE 1. Disease-specific prompts used in the study, with all cases being accompanied by representative fundus photographs (Figure 1A and 1B) corresponding to the clinical scenario

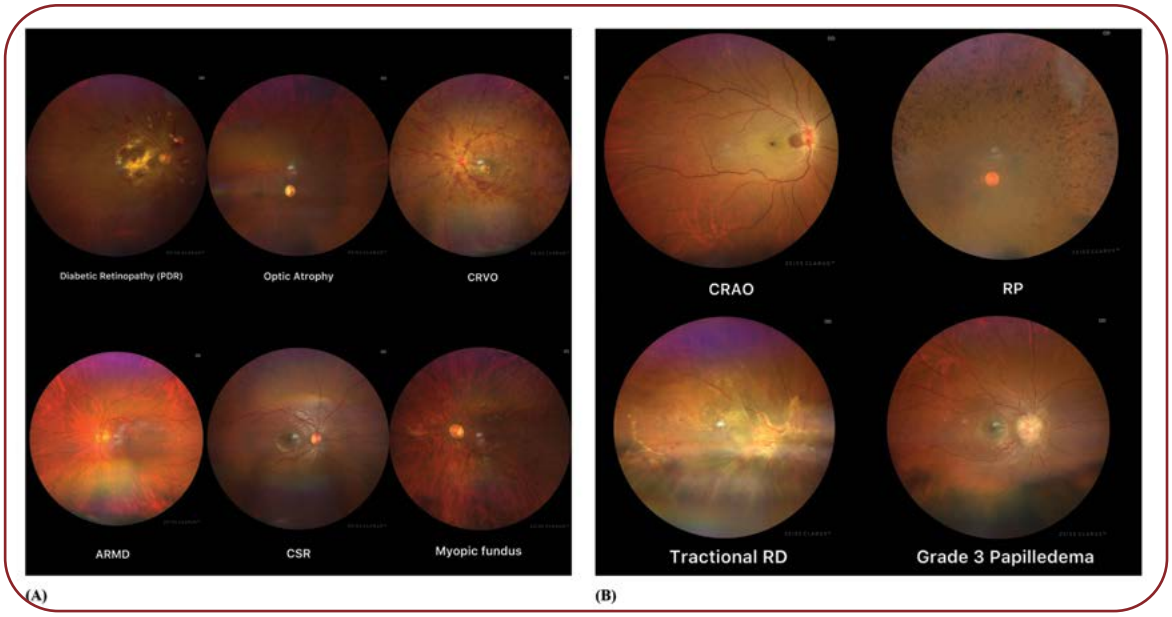

FIGURE 1. Representative fundus photographs used in disease-specific prompts: (A) proliferative diabetic retinopathy, optic atrophy, central retinal vein occlusion, age-related macular degeneration, central serous chorioretinopathy and myopic fundus; and (B) central retinal artery occlusion, retinitis pigmentosa, tractional retinal detachment and Grade 3 papilledema

prior ocular surgery or trauma that altered fundus findings, or if essential clinical history was incomplete.

ChatGPT (version 4.0) and Google Gemini were tested in freshly initiated, independent sessions to avoid contextual bias. Each vignette was presented as a structured prompt including clinical history and the fundus image. Both models were instructed to provide the most likely diagnosis and, when appropriate, differential diagnoses. The ophthalmologist evaluated each vignette independently under similar conditions.

The primary outcome was diagnostic accuracy, defined as the proportion of correct diagnoses compared with the gold standard. Inter-rater agreement between evaluators was measured using Cohen's kappa (κ), interpreted as poor (<0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80) and almost perfect (>0.81). Secondary outcomes included time to diagnosis, which was measured in seconds from case presentation to output, and clarity of interpretation, which was scored on a five-point Likert scale (1 = poor, 5 = excellent).

Statistical analyses were performed using SPSS version 25. Accuracy was expressed as percentages. Continuous variables were summarized as means with standard deviations and medians with interquartile ranges. Group comparisons were conducted using the chi-square test or Fisher's exact test for categorical variables and the Kruskal–Wallis H test for continuous variables. A p-value <0.05 was considered statistically significant.

Inter-rater agreement between ChatGPT, Google Gemini and the ophthalmologist was assessed using Cohen's kappa statistic. Kappa values range from -1 to $+1$, where $+1$ indicates perfect agreement, 0 denotes agreement equal to chance and negative values indicate disagreement worse than chance. Values between 0.20–0.40 represent slight agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement and >0.80 almost perfect agreement. In this study, kappa values ranged between 0.36 and 0.51, indicating slight to moderate agreement, all of which were statistically significant ($p < 0.05$).

Clarity of interpretation was evaluated using a five-point Likert scale (1–5), independently scored by a retinal specialist, with higher scores reflecting greater clarity, coherence and clinical usefulness of the explanation. ❑

RESULTS

The ophthalmologist achieved the highest diagnostic accuracy, correctly identifying 29 of

Retinal morbidities (n=3 each)	Correctly diagnosing the retinal morbidities		
	ChatGPT n (%)	Google Gemini n (%)	Ophthalmologist n (%)
Diabetic retinopathy	3 (100.0)	3 (100.0)	3 (100.0)
Optic atrophy	3 (100.0)	3 (100.0)	3 (100.0)
Central retinal vein occlusion	2 (66.6)	3 (100.0)	3 (100.0)
Dry age-related macular degeneration	2 (66.6)	2 (66.6)	2 (66.6)
Central serous chorioretinopathy	3 (100.0)	2 (66.6)	3 (100.0)
Myopic fundus	3 (100.0)	3 (100.0)	3 (100.0)
Central retinal artery occlusion	3 (100.0)	2 (66.6)	3 (100.0)
Retinitis pigmentosa	2 (66.6)	2 (66.6)	3 (100.0)
Retinal detachment	3 (100.0)	3 (100.0)	3 (100.0)
Papilledema	3 (100.0)	3 (100.0)	3 (100.0)
Overall correct diagnosis	**27 (90.0)**	**26 (86.7)**	**29 (96.7)**

TABLE 2. Performance of AI tools and ophthalmologist in diagnosing retinal morbidities across ten retinal conditions, with values representing the number of correct diagnoses per three cases for each disease

AI models	Agreement					
	Chat Gpt		Google Gemini		Ophthalmologist	
	Kappa-statistics	p-value	Kappa-statistics	p-value	Kappa-statistics	p-value
Chat GPT	#NA		0.51	*0.004	0.47	*0.002
Google Gemini	0.51	*0.004	#NA		0.36	*0.01
Ophthalmologist	0.47	*0.002	0.36	*0.01	#NA	

*p<0.05 considered significant

TABLE 3. Agreement between ChaGPT, Google Gemini and ophthalmologist in diagnosing retinal morbidities

Interpretation	Time consumed mean (SD); median (IQR)	#p-value
Chat GPT	21.67 (3.92); 22.00 (20.00-25.00)	**<0.001**
Google Gemini	25.67 (4.13); 25.00 (25.00-30.00)	
Ophthalmologist	149.83 (30.17); 157.50 (121.50-175.00)	

SD: standard deviation; IQR: interquartile range; #Kruskal-Wallis H test; significant difference (p<0.05) in time consumed between ophthalmologist with ChatGPT and Google Gemini

TABLE 4. Comparison of time required by ChatGPT, Gemini and ophthalmologist in giving interpretation

TABLE 5. Comparison of clarity of ChatGPT, Gemini and an ophthalmologist in giving interpretation

Interpretation	Clarity of interpretation* mean (SD); median (IQR)	#p-value
Chat GPT	3.60 (1.06); 4.00 (3.00-4.00)	**<0.001**
Google Gemini	2.96 (0.71); 3.00 (2.00-3.25)	
Ophthalmologist	4.53 (0.73); 5.00 (4.00-5.00)	

*Based on Likert scale; #Kruskal-Wallis H test; significant difference (p<0.05) in clarity scores between ophthalmologist with ChatGPT and ophthalmologist with Google Gemini

30 cases (96.7%). ChatGPT correctly diagnosed 27 cases (90.0%), while Google Gemini correctly diagnosed 26 cases (86.7%). All three evaluators demonstrated perfect accuracy in conditions such as diabetic retinopathy, optic atrophy, myopic fundus, retinal detachment and papilledema.

Performance diverged in vascular occlusions and specific macular conditions. Gemini outperformed ChatGPT in diagnosing central retinal vein occlusion, while ChatGPT performed better in central retinal artery occlusion. ChatGPT also showed superior accuracy compared to Gemini in central serous chorioretinopathy. Both AI tools matched the ophthalmologist in several non-vascular, pattern-based retinal diseases (Table 2).

In cases where misclassification occurred, the incorrect outputs were generally closely related differential diagnoses of the uploaded fundus images. Variability in image quality, resolution and pixel clarity may have also influenced AI interpretation accuracy. Details of the conditions correctly and incorrectly diagnosed by each evaluator are presented in Table 2.

Agreement analysis demonstrated moderate concordance between ChatGPT and Google

Gemini ($\kappa = 0.51$, $p = 0.004$). ChatGPT also showed moderate agreement with the ophthalmologist ($\kappa = 0.47$, $p = 0.002$). In comparison, Gemini showed fair agreement with the ophthalmologist ($\kappa = 0.36$, $p = 0.01$). All agreement values were statistically significant, indicating measurable but incomplete agreement between evaluators (Table 3).

Regarding efficiency, ChatGPT was the fastest, requiring a mean of 21.7 seconds per interpretation. Gemini required 25.7 seconds and the ophthalmologist required 149.8 seconds, reflecting greater depth and manual evaluation time (Table 4).

Clarity of interpretation scores showed the highest clarity for the ophthalmologist (mean 4.53/5). ChatGPT scored 3.60 and Gemini 2.96, with statistically significant differences between groups (Table 5). ❑

DISCUSSION

This study provides one of the earliest structured comparisons of ChatGPT, Google Gemini, and an ophthalmologist across ten retinal and optic nerve diseases. The findings highlight that while the ophthalmologist demonstrated the highest diagnostic accuracy and interpretive clarity, both LLMs showed substantial diagnostic capability. ChatGPT achieved an accuracy of 90.0%, slightly outperforming Google Gemini (86.7%) but remaining lower than the ophthalmologist (96.7%). Both models showed excellent performance in diseases with well-defined, pattern-based fundus features such as diabetic retinopathy, myopic fundus, retinal detachment and papilledema, consistent with previous research emphasizing the strong visual-pattern recognition abilities of AI systems (10, 11).

In comparison, both AI tools showed more limited performance in vascular occlusions and subtler macular conditions. Differences between ChatGPT and Gemini were also evident, with Gemini performing better in central retinal vein occlusion and ChatGPT performing better in central retinal artery occlusion and central serous chorioretinopathy. These discrepancies may reflect differences in model architecture, training data and reasoning behavior (11, 12).

Agreement analysis further revealed meaningful but incomplete concordance across evaluators. ChatGPT and Gemini demonstrated moderate agreement, while agreement between Gemini and the ophthalmologist was only fair, emphasizing the continuing importance of human oversight. Importantly, both AI models produced interpretations significantly faster than the ophthalmologist, with ChatGPT taking approximately 22 seconds and Gemini 26 seconds *per* case, compared to nearly 150 seconds for human evaluation. This efficiency underscores potential applications in high-volume screening, teleophthalmology and emergency triage.

Clarity of interpretation varied markedly, with the ophthalmologist providing the most comprehensive and clinically coherent explanations. ChatGPT performed moderately well in clarity, while Gemini lagged behind. These findings suggest that although LLM-generated diagnoses may be accurate, their explanations still lack the depth required for independent clinical use.

This study has several limitations. First, the sample size was relatively small, which may limit generalizability. Second, only fundus photographs were evaluated, without multimodal imaging such as OCT or angiography. Third, variability in image quality and resolution may have influenced AI performance. Fourth, the study assessed static images rather than real-world clinical scenarios, where additional contextual information plays a major diagnostic role.

Future research should explore integrating LLMs with multimodal ophthalmic imaging such as OCT and angiography, as well as domain-specific fine-tuning to improve performance in subtle or acute conditions that require nuanced interpretation. Larger prospective studies, including real-world diagnostic scenarios with multiple comorbidities, are warranted to assess their true clinical applicability (13, 14). Recent reviews have highlighted that LLMs like ChatGPT-3 and ChatGPT-4 demonstrate significant promise in ophthalmology, supporting medical education, clinical assistance, research and patient education, although challenges such as performance variability, bias, and hallucinations remain (15). Additionally, incorporating multimodal AI capabilities may further enhance diagnostic performance by combining textual and image-based data (15, 16).

In summary, while ophthalmologists remain the gold standard for retinal diagnosis, ChatGPT and Google Gemini demonstrated promising ac-

curacy and efficiency, particularly for pattern-based diseases. With appropriate refinement and integration, these models may serve as valuable adjunct tools for triage, patient education and teleophthalmic services in resource-limited settings. ❑

CONCLUSION

Ophthalmologists continue to provide superior diagnostic accuracy and interpretive clarity. However, ChatGPT and Google Gemini demonstrated strong performance across several retinal conditions, with ChatGPT showing modest advantages in diagnostic accuracy and clarity. Their rapid assessment capabilities highlight potential roles in triage, teleophthalmology and educational workflows. With further refinement, multimodal integration and broader validation, LLMs may evolve into valuable tools that augment but do not replace specialist expertise. ❑

Financial support: none declared.
Conflicts of interest: none declared.

References

1. **Zhou C, Li S, Ye L, et al.** Visual impairment and blindness caused by retinal diseases: a nationwide register-based study. *J Glob Health* 2023;13:04126. doi: 10.7189/jogh.13.04126.
2. **Alhassan E, Gendelman HK, Sabha MM, et al.** Bilateral retinal vasculitis as the first presentation of systemic lupus erythematosus. *Am J Case Rep* 2021;22:e930650. doi: 10.12659/AJCR.930650.
3. **Flaxman SR, Bourne RRA, Resnikoff S, et al.** Global causes of blindness and distance vision impairment 1990-2020: a systematic review and meta-analysis. *Lancet Glob Health* 2017;5:e1221-e1234. doi: 10.1016/S2214-109X(17)30393-5.
4. GBD 2019 Blindness and Vision Impairment Collaborators; Vision Loss Expert Group of the Global Burden of Disease Study: Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the right to sight: an analysis for the Global Burden of Disease Study. *Lancet Glob Health* 2021;9:e144-e160. doi: 10.1016/S2214-109X(20)30489-7.
5. **Işık MU, Akmaz B, Akay F, et al.** Evaluation of subclinical retinopathy and angiopathy with OCT and OCTA in patients with systemic lupus erythematosus. *Int Ophthalmol* 2021;41:143-150. doi: 10.1007/s10792-020-01561-8.
6. **Gulshan V, Peng L, Coram M, et al.** Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016;316:2402-2410. doi: 10.1001/jama.2016.17216.
7. **Lim JI, Rachitskaya AV, Hallak JA, et al.** Artificial intelligence for retinal diseases. *Asia Pac J Ophthalmol* (Phila) 2024;13:100096. doi: 10.1016/j.apjo.2024.100096.
8. **Chiarello F, Giordano V, Spada I, et al.** Future applications of generative large language models: a data-driven case study on ChatGPT. *Technovation* 2024;133:103002. doi: 10.1016/j.technovation.2024.103002.
9. **Tey KY, Cheong EZK, Ang M.** Potential applications of artificial intelligence in image analysis in cornea diseases: a review. *Eye Vis* (Lond) 2024;11:10. doi: 10.1186/s40662-024-00376-3.
10. **Visibelli A, Roncaglia B, Spiga O, Santucci A.** The impact of artificial intelligence in the odyssey of rare diseases. *Biomedicines* 2023;11:887. doi: 10.3390/biomedicines11030887.
11. **Li JO, Liu H, Ting DSJ, et al.** Digital technology, tele-medicine and artificial intelligence in ophthalmology: a global perspective. *Prog Retin Eye Res* 2021;82:100900. doi: 10.1016/j.preteyeres.2020.100900.
12. **Oshika T.** Artificial intelligence applications in ophthalmology. *JMA J* 2025;8:66-75. doi: 10.31662/jmaj.2024-0139.
13. **Li Z, Wang L, Wu X, et al.** Artificial intelligence in ophthalmology: the path to the real-world clinic. *Cell Rep Med* 2023;4:101095. doi: 10.1016/j.xcrm.2023.101095.
14. **Srivastava O, Tennant M, Grewal P, et al.** Artificial intelligence and machine learning in ophthalmology: a review. *Indian J Ophthalmol* 2023;71:11-17. doi: 10.4103/ijo.IJO_1569_22.
15. **Agnihotri AP, Nagel ID, Artiaga JCM, et al.** Large Language Models in Ophthalmology: A Review of Publications from Top Ophthalmology Journals. *Ophthalmol Sci* 2024;5:100681. doi: 10.1016/j.xops.2024.100681.
16. **Mihalache A, Huang RS, Cruz-Pimentel M, et al.** Artificial intelligence chatbot interpretation of ophthalmic multimodal imaging cases. *Eye* (Lond) 2024;38:2491 2493. doi: 10.1038/s41433-024-03074-5.