Visual Neuroscience

ARTICLE

Open Access

https://doi.org/10.48130/vns-0026-0016

Visual Neuroscience **2026**, 43: e018

Exploring the capabilities of three artificial intelligence chatbots in diagnosis and treatment suggestions for macular hole

Duo Yuan[1#], Xinyu Zhao[2#], Zhenquan Wu[2#], Shaojuan Peng[2], Na Duan[3], Kaixuan Cui[2], Zhen Yu[2], Honglang Zhang[2], Weihua Yang[2], Wenbin Wei[4], Wei Chi[2*] and Guoming Zhang[2*]

[1] *Shenzhen Eye Hospital, Jinan University, Shenzhen 518040, China*

[2] *Shenzhen Eye Hospital, Shenzhen Eye Medical Center, Southern Medical University, Shenzhen 518040, China*

[3] *The First Clinical Medical College of Jinan University, Huizhou Third People's Hospital, Huizhou 516000, China*

[4] *Beijing Tongren Eye Center; Beijing Key Laboratory of Intraocular Tumor Diagnosis and Treatment; Beijing Ophthalmology & Visual Sciences Key Lab; Medical Artificial Intelligence Research and Verification Key Laboratory of the Ministry of Industry and Information Technology, Beijing Tongren Hospital, Capital Medical University, Beijing 100000, China*

[#] Authors contributed equally: Duo Yuan, Xinyu Zhao, Zhenquan Wu

* Correspondence: chiwei@sz-eyes.com (Chi W); zhang-guoming@163.com (Zhang G)

Abstract

To compare diagnostic treatment suggestion, and answer quality of three artificial intelligence (AI) chatbots with an ophthalmology resident for macular hole, we assembled 50 macular hole cases, including lamellar macular hole, full-thickness macular hole (Gass II–IV), and macular hole with rhegmatogenous retinal detachment. Cases with insufficient preoperative information were excluded. Each anonymised record was presented to three AI chatbots (ChatGPT-o3, DeepSeek-R1, Gemini 2.5 Pro) and to an ophthalmology resident with three years of training. The consensus diagnosis and treatment suggestion of two retinal specialists served as the gold standard. Outcomes were diagnosis agreement, treatment suggestion agreement, and Global Quality Score (GQS). Diagnosis agreement was 86% (95% CI 73.3–94.2) for ChatGPT-o3, 82% (68.6–91.4) for DeepSeek-R1, 80% (66.3–89.9) for Gemini 2.5 Pro, and 82% (68.6–91.4) for the resident. Treatment suggestion agreement was 92% (80.8–97.8), 86% (73.3–94.2), 80% (66.3–89.9), and 70% (55.4–82.1), respectively; the resident's agreement was significantly lower than ChatGPT-o3 (p = 0.006). GQS ratings ranked Gemini 2.5 Pro highest, followed by ChatGPT-o3, the resident, and DeepSeek-R1. In conclusion, the three AI chatbots achieved similar diagnosis agreement for macular hole; ChatGPT-o3 most often matched specialist treatment suggestions, and Gemini 2.5 Pro provided the highest answer quality, suggesting that combining their strengths may enhance clinical decision support.

Citation: Yuan D, Zhao X, Wu Z, Peng S, Duan N, et al. 2026. Exploring the capabilities of three artificial intelligence chatbots in diagnosis and treatment suggestions for macular hole. *Visual Neuroscience* 43: e018 https://doi.org/10.48130/vns-0026-0016

Introduction

Macular hole threatens central vision and can cause lasting visual disability if surgical repair is delayed, imposing a serious burden on participants and health systems[1,2]. This condition often affects people in their working years, so delayed treatment carries significant personal and socioeconomic consequences. Access to timely and consistent care is challenging in many regions due to a shortage and uneven distribution of retinal specialists, with referral pathways varying widely[3]. As a result, general ophthalmologists often struggle to accurately diagnose and manage macular holes without specialist input.

Over the past few years, large language models (LLMs) have been rapidly developed and deployed in healthcare, with artificial intelligence (AI) chatbots like ChatGPT and DeepSeek becoming widely accessible by 2025[4,5]. AI chatbots show good performance on board-style examination questions and are now widely used in patient-initiated consultations[6,7]. As access expands, these AI chatbots can assist generalist providers with complex cases, which motivates focused evaluations in specific diseases[8].

Recent ophthalmic studies have applied AI chatbots to glaucoma and retina management questions, professional examination items, patient education materials, and planning and interpretation of clinical images[9]. Across these evaluations, leading systems sometimes approached clinician output, but performance and readability varied by model and task[10]. Evidence for macular hole remains limited, and further studies are needed to verify the performance of AI chatbots as clinical decision support for diagnosis and treatment suggestions.

This study aims to compare three AI chatbots (ChatGPT-o3, Gemini 2.5 Pro, and DeepSeek-R1) for macular hole by assessing agreement for diagnosis and treatment suggestion, and answer quality using the Global Quality Score (GQS).

Methods

Study design and participants

The study adhered to the Declaration of Helsinki and was approved by the Institutional Review Board of Shenzhen Eye Hospital (No. 2025KYPJ120, approved on 10 July 2025). The ethics committee granted a waiver of informed consent because of the retrospective design and the use of deidentified data.

We performed a retrospective comparative review of 50 participants with macular holes identified from medical records between 1 January and 31 March 2025. Records were eligible when the chart contained adequate clinical history, examination findings, and ancillary results, including optical coherence tomography (OCT) report results when available, to construct a standardized participant record. Records lacking core preoperative information were excluded. Atypical or equivocal presentations were not specifically selected, and most eligible records had sufficiently clear

documentation to support a reference diagnosis. For each eligible participant, key clinical variables were abstracted, deidentified, and rewritten as a participant record with harmonized terminology, derived from medical records and OCT report results. The overall workflow is shown in Fig. 1.

Case preparation and prompting protocol

All participant records were rewritten into a fixed text template with harmonized length and terminology. For each participant record, the template included age, sex, affected eye, presenting symptoms and their duration, and preoperative best-corrected visual acuity, as well as lens status, any history of ocular surgery, and major systemic comorbidities when documented. For every participant record, an OCT report was available and had been written in a standardized format by experienced technicians in the imaging department. From these OCT reports, we recorded the minimum hole diameter, the base diameter when reported, and whether vitreomacular traction, epiretinal membrane, intraretinal cysts, and retinal detachment were present. The same order and phrasing were

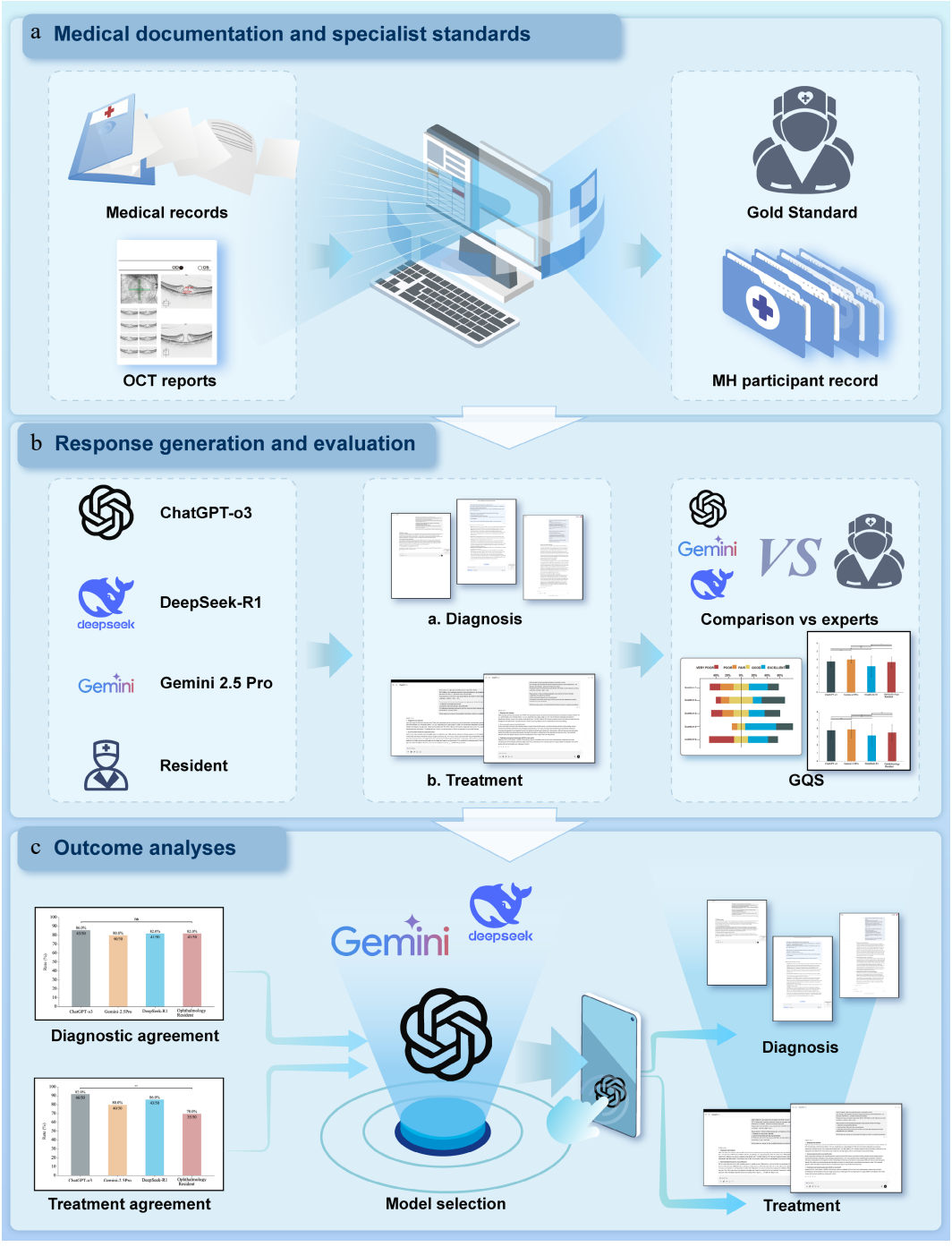

Fig. 1 Study design and workflow. The workflow comprised three steps: (a) Medical documentation and specialist standards used to construct the MH participant record and define the gold standard, (b) response generation and evaluation by three AI chatbots and an ophthalmology resident, and (c) outcome analyses, including diagnostic agreement, treatment suggestion agreement, and GQS. Abbreviations: MH, macular hole; OCT, optical coherence tomography; GQS, Global Quality Score.

used for all templates so that the three AI chatbots and the resident received comparable structured information. The order of participant records was independently randomized for each method. To standardise language, the source participant records were written in Chinese and translated to English through a single, standardized pass with ChatGPT; the same English version was used unchanged for all evaluations. Prompts requested one final diagnosis and one treatment suggestion for each record. External tools and web browsing were disabled.

All AI chatbot interactions followed one uniform protocol. If neutral clinical wording triggered safety filters, we applied minimal rephrasing without altering clinical content. If a reply was cut off by the platform's length limit, we sent one continuation request. If an answer drifted from the requested structure, we issued one reminder to return to the predefined format. Supplementary Figs S1–S3 show the standardized interaction format used in this study.

AI chatbots selection

Each participant's record was presented as text to three AI chatbots and to an ophthalmology resident with three years of training. ChatGPT-o3 (OpenAI, USA; released April 2025; knowledge cutoff June 2024) was included as a widely used conversational model with peer-reviewed evidence from clinical settings[11]. Gemini 2.5 Pro (Google, USA; released March 2025; knowledge cutoff January 2025) was included as a contemporary model with a published head-to-head evaluation against ChatGPT on retinal detachment information tasks, covering accuracy, readability, and expert quality grading[12]. DeepSeek-R1 (DeepSeek, China; released January 2025; knowledge cutoff 2024) was included for its clinical reasoning strength and growing use in China, with published evaluations reporting competitive performance on clinical decision-support tasks[13]. All interactions with the AI chatbots were performed in June 2025 using the official web interfaces with default settings. An ophthalmology resident working full-time in our department received the same participant records and provided one final diagnosis and one treatment suggestion per participant.

Outcomes and grading

The primary outcome measures were the agreement of diagnosis and treatment suggestion with the gold standard set by two retinal specialists. All participant records were deidentified and then independently reviewed by two retinal specialists, each with more than ten years of experience in clinical practice. For every participant, each specialist first recorded an initial diagnosis as full-thickness macular hole, lamellar macular hole, macular hole with retinal detachment, or another diagnosis when appropriate. They then recorded an initial treatment suggestion, including whether they would recommend surgery or observation, the main surgical approach they would choose if surgery was planned, the type of intraocular tamponade they would use, whether they would perform combined cataract surgery in the same session, and the postoperative positioning they would advise. After this independent step, the two specialists met to review all participant records together. Whenever their initial opinions differed for the diagnosis or for the treatment suggestion, they discussed the clinical details of that participant until they reached a consistent conclusion. If they could not reach an agreement after discussion, that participant would be excluded from the study. These initial independent assessments were also used to summarise the inter-observer agreement between the two specialists before consensus. The final consensus diagnosis and the final consensus treatment suggestion for each participant were used as the gold standard. A response from a chatbot or from the ophthalmology resident was counted as correct when it matched this consensus diagnosis or treatment suggestion.

Answer quality was graded by two masked graders, each with ten years of ophthalmology clinical experience, using the five-point GQS (Table 1). Graders evaluated each participant's record for clarity, clinical correctness, completeness, and practical usefulness. Graders worked independently and were masked to the responder's identity and to each other's scores. For each grader, GQS was analysed separately.

Statistical analysis

All statistical analyses were performed using R software (version 4.4.1; Posit, USA). Initial inter-observer agreement between the two retinal specialists before consensus was summarised as raw percentage agreement for diagnosis and for treatment suggestions. Paired chi-square test was used to compare accuracy rates, and score comparisons were conducted with the generalized estimating equation (GEE). For the GQS, the mean score of each response was calculated, followed by the computation of the overall mean score. $p < 0.05$ was considered statistically significant.

Results

The study consisted of 50 participants with a mean age of 59.5 ± 9.9 years. Of these participants, 72% were female, and 40% had the right eye affected. Phenotypes included full-thickness macular hole (FTMH) 37 (74%), lamellar macular hole (LMH) 5 (10%), and macular hole with rhegmatogenous retinal detachment (MH-RRD) 8 (16%). For the FTMH subset (n = 37), Gass staging was: stage II, 9 (24%); stage III, 3 (8%); stage IV, 25 (68%). In the gold standard treatment suggested by the retinal specialists, the tamponade choice was gas in 44/50 (88%), silicone oil in 4/50 (8%), and no tamponade in 2/50 (4%). Baseline characteristics are summarised in Table 2.

Diagnosis agreement

Before consensus, the two retinal specialists showed high initial agreement, with raw agreement of 49/50 (98%) for diagnosis and 47/50 (94%) for treatment suggestions. Diagnosis agreement was 43/50 (86%; 73.3–94.2) for ChatGPT-o3, 41/50 (82%; 68.6–91.4) for DeepSeek-R1, 40/50 (80%; 66.3–89.9) for Gemini 2.5 Pro, and 41/50 (82%; 68.6–91.4) for the resident. By macular hole subtype, diagnostic agreement with the reference standard was high for FTMH and LMH and lower for MH-RRD. For FTMH (n = 37), agreement was 31/37 for ChatGPT-o3, 30/37 for DeepSeek-R1, 29/37 for Gemini 2.5 Pro, and 32/37 for the ophthalmology resident. For LMH (n = 5), agreement was 5/5 for ChatGPT-o3, 5/5 for DeepSeek-R1, 5/5 for

Table 1. Global quality score description.

Score	Overall description
1	Poor quality, poor flow of the site, most information missing, not at all useful for patients
2	Generally poor quality and poor flow, some information listed but many important topics missing, of very limited use to patients
3	Moderate quality, suboptimal flow, some important information is adequately discussed but others poorly discussed, somewhat useful for patients
4	Good quality and generally good flow, most of the relevant information is listed, but some topics not covered, useful for patients
5	Excellent quality and excellent flow, very useful for patients

Table 2. Baseline clinical characteristics of participants with macular hole.

Variable	Value	*n* (%)
Number of participants	50	N/A
Age (years)	59.5 ± 9.9	N/A
Sex (male/female)	14/36	28/72
Eye laterality (right/left)	20/30	40/60
Macular hole phenotype		
LMH	5	10
FTMH	37	74
MH-RRD	8	16
Gass stage (FTMH only, *n* = 37)		
Stage II	9	24
Stage III	3	8
Stage IV	25	68
Tamponade in reference plan[a]		
Gas	44	88
Silicone oil	4	8
None	2	4
Ocular comorbidities[b]		
Cataract	8	16
High myopia	6	12
Epiretinal membrane	5	10
Others	3	6

[a] Percentages use the cohort size as denominator. [b] Comorbidities are not mutually exclusive. Abbreviations: LMH, lamellar macular hole; FTMH, full-thickness macular hole; MH-RRD, macular hole with rhegmatogenous retinal detachment.

Gemini 2.5 Pro, and 4/5 for the ophthalmology resident. For MH-RRD (*n* = 8), agreement was 7/8 for ChatGPT-o3, 6/8 for DeepSeek-R1, 6/8 for Gemini 2.5 Pro, and 5/8 for the ophthalmology resident. Paired Chi-square tests versus ChatGPT-o3 showed no significant differences (DeepSeek-R1 $p = 0.617$; Gemini 2.5 Pro $p = 0.248$; resident $p = 0.803$). Table 3 summarises these results, and Fig. 2a shows the diagnosis agreement.

Treatment suggestion agreement

Agreement rates were 46/50 (92%; 80.8–97.8) for ChatGPT-o3, 43/50 (86%; 73.3–94.2) for DeepSeek-R1, 40/50 (80%; 66.3–89.9) for Gemini 2.5 Pro, and 35/50 (70%; 55.4–82.1) for the resident. In comparison with ChatGPT-o3, differences were not significant for DeepSeek-R1 ($p = 0.248$) or Gemini 2.5 Pro ($p = 0.077$), whereas the residents' agreement was significantly lower ($p = 0.006$). Fig. 2b shows the comparison of treatment suggestion agreement.

Answer quality of three AI chatbots

Answer quality was independently rated by two graders. For grader 1, the mean GQS was 4.02 ± 0.43 for Gemini 2.5 Pro, 3.78 ± 0.65 for ChatGPT-o3, 3.70 ± 0.65 for the resident, and 3.18 ± 1.26 for DeepSeek-R1. Gemini 2.5 Pro was the highest, ChatGPT-o3 and the resident were close, and DeepSeek-R1 was the lowest. For grader 2, the mean GQS was 3.88 ± 1.00 for Gemini 2.5 Pro, 3.78 ± 1.00 for ChatGPT-o3, 3.50 ± 1.02 for the resident, and 3.14 ± 1.32 for DeepSeek-R1. The order was the same.

Pairwise comparisons from a GEE model showed the following. For grader 1, Gemini 2.5 Pro was higher than ChatGPT-o3 by 0.24 points ($p = 0.007$) and higher than the resident by 0.32 points ($p < 0.001$). Both Gemini 2.5 Pro and ChatGPT-o3 were higher than DeepSeek-R1 (0.84, $p < 0.001$; 0.60, $p = 0.003$), and ChatGPT-o3 and the resident did not differ ($p = 0.151$). For grader 2, ChatGPT-o3 was higher than DeepSeek-R1 by 0.64 points ($p = 0.006$) and higher than the resident by 0.28 points ($p = 0.007$), and Gemini 2.5 Pro was higher than DeepSeek-R1 by 0.74 points ($p = 0.003$). Gemini 2.5 Pro did not differ from ChatGPT-o3 ($p = 0.587$) and was 0.38 points above the resident ($p = 0.051$). Across graders, DeepSeek-R1 scored consistently lower, whereas Gemini 2.5 Pro and ChatGPT-o3 had similar scores, and their rank varied by grader (Fig. 3a, b).

Table 3. Macular hole diagnosis and treatment suggestion agreement and global quality score.

Evaluator	Diagnosis (95% CI)		Treatment (95% CI)		GQS	
	Agreement (%)	*p*	Agreement (%)	*p*	Grader 1	Grader 2
ChatGPT-o3	0.86 (73.3-94.2)	N/A	0.92 (80.8-97.8)	N/A	3.78 ± 0.65	3.78 ± 1.00
Gemini 2.5 Pro	0.80 (66.3-89.9)	0.248	0.80 (66.3-89.9)	0.077	4.02 ± 0.43	3.88 ± 1.00
DeepSeek-R1	0.82 (68.6-91.4)	0.617	0.86 (73.3-94.2)	0.248	3.18 ± 1.26	3.14 ± 1.32
Resident	0.82 (68.6-91.4)	0.803	0.70 (55.4-82.1)	**0.006**	3.70 ± 0.65	3.50 ± 1.02

CI, confidence interval; GQS, Global Quality Score. Agreement *p*-values are from the paired chi-square test versus ChatGPT-o3. Both masked graders had ten years of ophthalmology clinical experience.

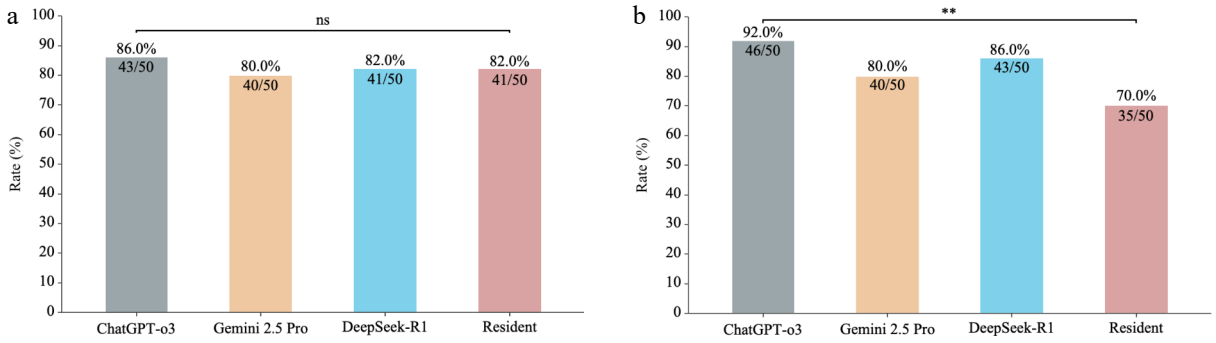

Fig. 2 Diagnosis and treatment suggestion agreement for macular hole across ChatGPT-o3, Gemini 2.5 Pro, DeepSeek-R1, and an ophthalmology resident. (a) Diagnosis agreement. (b) Treatment suggestion agreement. Bars indicate the agreement rate, with exact Clopper–Pearson 95% confidence intervals (*n* = 50). Pairwise comparisons were performed using paired chi-square tests. ** $p < 0.01$; ns, not significant. Abbreviations: CI, confidence interval.

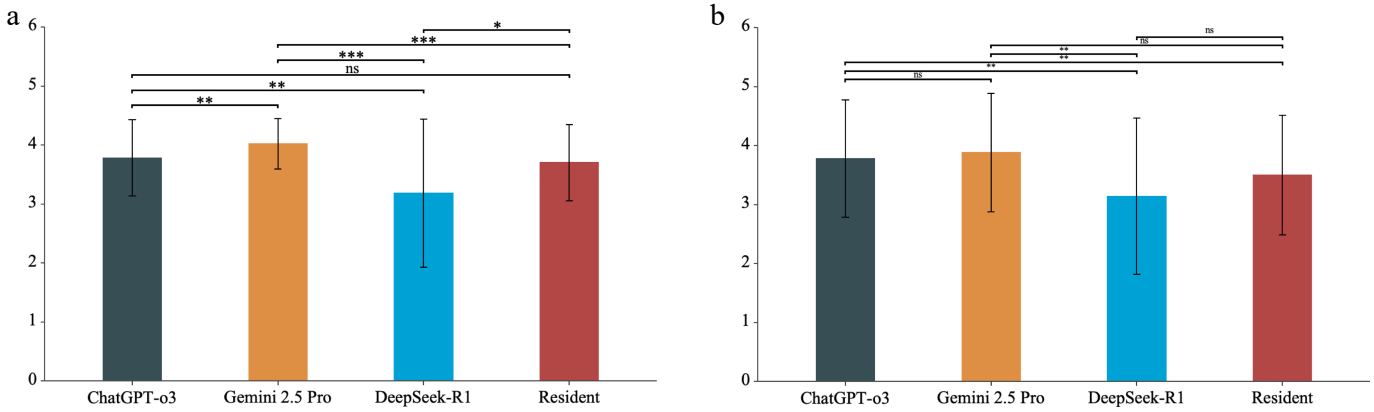

Fig. 3 Global quality score across ChatGPT-o3, Gemini 2.5 Pro, DeepSeek-R1, and an ophthalmology resident. (a) Grader 1. (b) Grader 2. Bars show mean ± SD. Brackets indicate pairwise comparisons based on generalized estimating equations. *$p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; ns, not significant. Abbreviations: GQS, Global Quality Score; SD, standard deviation.

Error analysis

We grouped errors into three main categories: diagnostic errors, treatment planning errors, and format deviations. All three AI chatbots provided a diagnosis and a treatment suggestion for every participant, so errors rarely involved a missing answer. Diagnostic errors mainly reflected mislabelling rather than completely different disease types. The most typical pattern was that the written description mentioned high myopia, but the macular hole was still labelled as idiopathic (ChatGPT-o3 $n = 3$, DeepSeek-R1 $n = 2$, Gemini 2.5 Pro $n = 1$), which can be regarded as misclassification of etiology.

For treatment planning, errors were mostly omissions or inconsistencies in key perioperative details, such as whether to recommend face-down positioning, whether to combine cataract surgery, and how to describe the planned procedure. Among participants for whom gas tamponade was recommended, face-down positioning was recorded in 44/49 cases by ChatGPT-o3, 13/45 by DeepSeek-R1, and 3/47 by Gemini 2.5 Pro; the remaining suggestions omitted positioning and were counted as errors. The AI chatbots also rarely suggested combined cataract extraction and vitrectomy when a clinically significant cataract was described in the participant record; this combination appeared once with Gemini 2.5 Pro and not at all with ChatGPT-o3 or DeepSeek-R1. These treatment planning omissions tended to occur in records where macular findings were described in detail, but cataract symptoms or postoperative instructions were not emphasized in the text.

Format deviations were uncommon and mainly involved giving more than one diagnosis or listing several possible procedures instead of one clear treatment suggestion. When this happened, the responses were re-generated after a brief repeat prompt, and the clinical content did not change. For the ophthalmology resident, discrepancies with the reference standard also mainly concerned perioperative details, such as whether to recommend face-down positioning or whether to combine cataract surgery, rather than opposite decisions on whether to operate or completely different diagnostic categories.

Discussion

Using participant data from medical records and OCT reports, we compared the performance of three AI chatbots on agreement in diagnosis, treatment suggestion, and answer quality. An ophthalmology resident was also included as a clinical comparator. Diagnosis agreement was similar across three AI chatbots and the resident. Treatment suggestion agreement was highest for ChatGPT-o3 and lowest for the resident. For answer quality, both graders ranked Gemini 2.5 Pro highest, ChatGPT-o3 in the middle, and DeepSeek-R1 lowest.

Diagnosis agreement was similar across three AI chatbots and the resident. Tao et al. reported moderate to good performance for ChatGPT-3.5 and Bing Chat on well-defined ophthalmic questions, and Fowler et al. found that chatbots performed best when prompts were explicit and narrowly scoped[14,15]. Carlà et al. analysed 50 retinal detachment cases and reported the diagnostic performance: agreement was 80% for ChatGPT-3.5, 84% for ChatGPT-4, and 70% for Google Gemini, and ChatGPT-4 was higher than Gemini ($p = 0.03$)[16]. Studies outside ophthalmology also describe strong accuracy when instructions are clear and information is structured[17]. In our subtype analysis, agreement remained high for FTMH and LMH but was lower for MH-RRD, consistent with the added complexity of detachment-related cases. Overall, the current chatbots aligned best with conventional macular hole diagnosis and standard vitrectomy decisions, rather than newer refinements in surgical technique[18].

Differences among the three AI chatbots became significant when providing treatment suggestions. ChatGPT-o3 most often matched the reference treatment suggestions for whether to operate and for tamponade choice, making it useful for cross-checking core decisions. Gemini 2.5 Pro achieved the highest GQS and usually provided clearer explanations, which may be helpful for clinical notes and patient instructions, although some aetiology labels did not match the reference. DeepSeek-R1 showed similar agreement to the other models, but it had the lowest GQS and more often omitted face-down positioning. These differences show that the main gaps were in perioperative details rather than in the core treatment choice. The resident showed the same pattern, with lower treatment suggestion agreement, so combining ChatGPT-o3 for checking treatment suggestions and Gemini 2.5 Pro for clear documentation can improve clinical decision support for macular holes.

The treatment suggestion agreement was lower than the diagnosis agreement. This is consistent with clinical practice for macular hole, where decisions are based not only on the macular diagnosis but also on patient-specific factors. Published studies report that the value of face down positioning depends on hole size and related factors, and that centres differ in how they prescribe it, including whether they use it routinely or selectively and how long patients are asked to maintain it[19]. For large macular holes, surgeons may choose internal limiting membrane peeling or the inverted internal limiting membrane flap, and both are used in practice[20,21]. Retina

specialists also consider coexisting cataract, patient age, systemic comorbidities, visual needs, the status of the fellow eye, and the patient's ability to comply with postoperative positioning. These clinical variations and patient factors make modest disagreement in treatment suggestions plausible even when diagnosis agreement is high. Our findings, therefore, support using AI chatbots as adjunct tools to check core treatment suggestions and to generate clear written explanations, rather than as standalone decision makers, with the treating retinal specialist remaining responsible for integrating all patient-specific factors. This is especially important when clinical information is uncertain. If the clinical record or OCT report is ambiguous or incomplete, AI chatbots may still answer with high confidence, which can be misleading in borderline cases. These patient-related factors can influence both the treatment suggestion chosen as the reference standard and the agreement rates observed for different evaluators. Studies outside ophthalmology report the same pattern, with lower agreement for treatment suggestions than for diagnosis[22].

Answer quality differed across the three AI chatbots, consistent with prior ophthalmic reports. Work on cataract education found that chatbot materials were generally appropriate yet differed in readability and factual accuracy, which is in line with the variability we observed in GQS[23]. Eid et al. also noted that patient-facing materials from AI chatbots often needed improvements in readability[24]. Even AI chatbots developed for ophthalmology still show differences in clarity and factual accuracy across systems and with changes in prompt wording[25,26]. When questions require interpreting images, performance is lower than when answers can be drawn from the participant record, so clear written reasoning remains important when only medical records and OCT report results are available[27]. Together, these findings explain why AI chatbots can reach similar diagnosis agreement while diverging on GQS and support using answer quality as a complementary metric.

In this study, we evaluated the AI chatbots using case records compiled from medical records and OCT reports, reflecting a common workflow in which clinicians review written information before deciding. This text only design is consistent with recent retinal detachment work that applied GPT-based platforms to standardized clinical records without raw imaging and reported encouraging agreement with clinical decisions[28]. A fixed template and a uniform interaction protocol enabled consistent comparison across evaluators. Guidance on evaluation and deployment emphasises clear prompts, transparent procedures, and ongoing monitoring when chatbots are used to support documentation and triage[29,30].

Limitations

Our study has several limitations. First, the participant records were compiled retrospectively from routine charts, so some clinical details may have been missing or simplified, which could have influenced how the AI chatbots and the ophthalmology resident stated the diagnosis and the treatment suggestion. Second, all evaluations used a standard case template prepared from medical records and OCT reports, and neither the chatbots nor the resident saw the OCT images themselves; this cleaner text format may overestimate performance compared with daily practice, where clinicians interpret imaging together with incomplete or sometimes inconsistent notes. Finally, answer quality was graded with the five-point Global Quality Score, a subjective measure, and the AI chatbots were evaluated at a single time point, so their recommendations may not fully reflect later changes in macular hole surgery or in model updates.

Conclusions

Three AI chatbots demonstrated comparable diagnosis agreement for macular holes. ChatGPT-o3 achieved the highest agreement in treatment suggestions, while Gemini 2.5 Pro received the highest ratings for overall answer quality. Combining the strengths of different AI chatbots may enhance clinical decision support in macular hole management.

Ethical statements

The study was approved by the Institutional Review Board of Shenzhen Eye Hospital (No. 2025KYPJ120, July 10, 2025) and adhered to the Declaration of Helsinki. The requirement for informed consent was waived by the ethics committee due to the retrospective nature of the study and the use of anonymized data.

Author contributions

The authors confirm contribution to the paper as follows: study conception and design: Zhang G, Chi W; methodology: Yuan D, Zhao X, Wu Z; formal analysis: Yuan D; data curation: Peng S, Duan N, Cui K, Yu Z, Zhang H, Yang W; visualization, draft manuscript preparation: Yuan D; writing – review & editing: Yuan D, Zhao X, Wu Z, Peng S, Duan N, Cui K, Yu Z, Zhang H, Yang W, Wei W, Chi W, Zhang G; supervision: Wei W, Chi W, Zhang G; funding acquisition, guarantor: Zhang G. All authors reviewed the results and approved the final version of the manuscript.

Data availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos 82271103, 82301269, 82401315, 82401272, 82301226), Shenzhen Medical Research Fund (Grant Nos C2301005, C2501034, A2403020), Shenzhen Science and Technology R&D Fund Program (Grant Nos JCYJ20240813152703005, JCYJ20250604184009011, JCYJ20220530153607015), Guangdong Basic and Applied Basic Research Foundation (Grant Nos 2026A1515012558, 2022A1515110865), and Sanming Project of Medicine in Shenzhen (Grant No. SZSM202311018).

Conflict interest

The authors declare that they have no conflict of interest.

Supplementary information accompanies this paper online at: https://doi.org/10.48130/vns-0026-0016.

Dates

Received 27 October 2025; Revised 5 February 2026; Accepted 9 February 2026; Published online 27 April 2026

References

[1] Riding G, Teh BL, Yorston D, Steel DH. 2024. Comparison of the use of internal limiting membrane flaps versus conventional ILM peeling on

post-operative anatomical and visual outcomes in large macular holes. *Eye* 38:1876–1881

[2] Chen J, Tao J, Zhang Y. 2024. The inverted internal limiting membrane flap technique is not recommended for the treatment of large macular holes smaller than 650 μm. *Retina* 44(12):2086–2090

[3] Burton MJ, Ramke J, Marques AP, Bourne RRA, Congdon N, et al. 2021. The Lancet Global Health Commission on Global Eye Health: vision beyond 2020. *The Lancet Global Health* 9(4):e489–e551

[4] Thirunavukarasu AJ, Mahmood S, Malem A, Foster WP, Sanghera R, et al. 2024. Large language models approach expert-level clinical knowledge and reasoning in ophthalmology: A head-to-head cross-sectional study. *PLoS Digit Health* 3(4):e0000341

[5] Moëll B, Sand Aronsson F, Akbar S. 2025. Medical reasoning in LLMs: an in-depth analysis of DeepSeek R1. *Frontiers in Artificial Intelligence* 8:1616145

[6] Wei J, Wang X, Huang M, Xu Y, Yang W. 2025. Evaluating the performance of ChatGPT on board style examination questions in ophthalmology: a meta analysis. *Journal of Medical Systems* 49:94

[7] Huang M, Wang X, Zhou S, Cui X, Zhang Z, et al. 2025. Comparative performance of large language models for patient initiated ophthalmology consultations. *Frontiers in Public Health* 13:1673045

[8] Goh E, Gallo R, Hom J, Strong E, Weng Y, et al. 2024. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Network Open* 7(10):e2440969

[9] Li Z, Wang Z, Xiu L, Zhang P, Wang W, et al. 2025. Large language model based multimodal system for detecting and grading ocular surface diseases from smartphone images. *Frontiers in Cell and Developmental Biology* 13:1600202

[10] Huang AS, Hirabayashi K, Barna L, Parikh D, Pasquale LR. 2024. Assessment of a large language model's responses to questions and cases about glaucoma and retina management. *JAMA Ophthalmol* 142(4):371–375

[11] Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine* 183:589–596

[12] Strzalkowski P, Strzalkowska A, Chhablani J, Pfau K, Errera MH, et al. 2024. Evaluation of the accuracy and readability of ChatGPT-4 and Google Gemini in providing information on retinal detachment: a multicenter expert comparative study. *International Journal of Retina and Vitreous* 10:61

[13] Sandmann S, Hegselmann S, Fujarski M, Bickmann L, Wild B, et al. 2025. Benchmark evaluation of DeepSeek large language models in clinical decision-making. *Nat Med* 31(8):2546–2549

[14] Tao BK, Hua N, Milkovich J, Micieli JA. 2024. ChatGPT-3.5 and Bing Chat in ophthalmology: an updated evaluation of performance, readability, and informative sources. *Eye* 38:1897–1902

[15] Fowler T, Pullen S, Birkett L. 2024. Performance of ChatGPT and Google Bard on the official Part 1 FRCOphth practice questions. *The British Journal of Ophthalmology* 108(10):1379–1383

[16] Carlà MM, Gambini G, Baldascino A, Giannuzzi F, Boselli F, et al. 2024. Exploring AI-chatbots' capability to suggest surgical planning in ophthalmology: ChatGPT versus Google Gemini analysis of retinal detachment cases. *The British Journal of Ophthalmology* 108(10):1457–1469

[17] Mehandru N, Miao BY, Almaraz ER, Sushil M, Butte AJ, et al. 2024. Evaluating large language models as agents in the clinic. *NPJ digital medicine* 7:84

[18] Radke NV, Ruamviboonsuk P, Steel DH, Tian T, Hunyor AP, et al. 2025. Controversies, consensuses, and guidelines on macular hole surgery by the Asia–Pacific Vitreo-retina Society (APVRS) and the Asia–Pacific Academy of Professors in Ophthalmology (AAPPO). *Eye and Vision* 12:30

[19] Chaudhary V, Sarohia GS, Phillips MR, Zeraatkar D, Xie JS, et al. 2023. Role of positioning after full-thickness macular hole surgery: a systematic review and meta-analysis. *Ophthalmology Retina* 7:33–43

[20] Chen G, Tzekov R, Jiang F, Mao S, Tong Y, et al. 2020. Inverted ILM flap technique versus conventional ILM peeling for idiopathic large macular holes: a meta-analysis of randomized controlled trials. *PLoS ONE* 15(7):e0236431

[21] Manasa S, Kakkar P, Kumar A, et al. 2018. Comparative evaluation of standard ILM peel with inverted ILM flap technique in large macular holes: a prospective randomized study. *Ophthalmic Surgery, Lasers & Imaging Retina* 49(4):236–240

[22] Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, et al. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine* 30(9):2613–2622

[23] Azzopardi M, Ng B, Logeswaran A, Loizou C, Cheong RCT, Gireesh P, et al. 2024. Artificial intelligence chatbots as sources of patient-education material for cataract surgery: ChatGPT-4 versus Google Bard. *BMJ Open Ophthalmology* 9(1):e001824

[24] Eid K, Eid A, Wang D, Raiker RS, Chen S, et al. 2024. Optimizing ophthalmology patient education via chatbot-generated materials: readability analysis of AI-generated patient education materials and the American society of ophthalmic plastic and reconstructive surgery patient brochures. *Ophthalmic Plastic and Reconstructive Surgery* 40(2):212–216

[25] Chen X, Zhao Z, Zhang W, Xu P, Wu Y, et al. 2024. EyeGPT for patient inquiries and medical education: development and validation of an ophthalmology large language model. *Journal of Medical Internet Research* 26:e60063

[26] Templin T, Perez MW, Sylvia S, Leek J, Sinnott-Armstrong N. 2024. Addressing 6 challenges in generative AI for digital health. *PLOS Digital Health* 3(8):e0000503

[27] Mihalache A, Huang RS, Popovic MM, Patil NS, Pandya BU, et al. 2024. Accuracy of an artificial intelligence chatbot's interpretation of clinical ophthalmic images. *JAMA Ophthalmology* 142(4):321–326

[28] Agin A, Ozturk Y, Kivrak U. 2025. Harnessing generative pre-trained transformer technology for clinical decision support in retinal detachment. *Medical Bulletin of Haseki* 63(3):128–134

[29] Topol EJ. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine* 25:44–56

[30] He J, Baxter SL, Xu J, Xu J, Zhou X, et al. 2019. The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine* 25:30–36