

Original Paper

Comparison of Two AI Chatbots for Diagnosis and Providing Treatment Suggestions in Retinopathy of Prematurity: Retrospective Study

Shaojuan Peng^{1*}, MBBS; Xinyu Zhao^{2*}, MD, PhD; Zhenquan Wu^{2*}, MD, PhD; Duo Yuan², MM; Na Duan³, MM; Kaixuan Cui², MD, PhD; Zhen Yu², MD, PhD; Weihua Yang², MD, PhD; Wenbin Wei⁴, MD, PhD; Wei Chi², MD, PhD; Guoming Zhang², MD, PHD

¹Jinan University, Guangzhou, China

²Shenzhen Eye Medical Center, Shenzhen Eye Hospital, Southern Medical University, Shenzhen, China

³Huizhou Third People's Hospital, First Affiliated Hospital of Jinan University, Guangzhou, China

⁴Beijing Tongren Eye Center, Capital Medical University, Beijing, China

*these authors contributed equally

Corresponding Author:

Guoming Zhang, MD, PHD
Shenzhen Eye Medical Center
Shenzhen Eye Hospital, Southern Medical University
18 Zetian Road, Futian District
Shenzhen 518043
China
Phone: 1 0755-23959600
Email: zhang-guoming@163.com

Abstract

Background: Retinopathy of Prematurity (ROP) is a leading cause of preventable childhood blindness; yet, a global shortage of experienced pediatric ophthalmologists impedes timely diagnosis and treatment. While emerging AI chatbots are promising clinical decision-support tools in some ophthalmic diseases, their performance in ROP diagnosis and providing treatment suggestions remains uncertain.

Objective: This study aimed to compare the performance of Google's Gemini 2.5 Pro and OpenAI's ChatGPT o4-mini in ROP diagnosis and providing treatment suggestions against the gold standard of clinical consensus.

Methods: A retrospective analysis was conducted on 70 infants (140 eyes) with treatment-requiring ROP, each providing structured clinical text data and wide-field fundus images. We adopted a 2-stage prompting strategy for AI chatbots, instructing them first to generate ROP diagnoses (including zone, stage, and presence of plus disease), and subsequently to provide treatment suggestions. After collecting the generated responses, we assessed their performance by comparing the consistency of their diagnosis and treatment suggestions with the consensus gold standard. Furthermore, 2 independent specialists quantitatively assessed the outputs of Gemini 2.5 Pro and ChatGPT o4-mini using the ROP-specific Global Quality Score (GQS), which is a 5-point scale ranging from 1 (unusable) to 5 (excellent). Statistical significance was set at $P < .05$, and all statistical analyses were performed using R (version 4.4.1; R Foundation for Statistical Computing).

Results: For the tasks of ROP zoning and staging, the consistency rates between Gemini 2.5 Pro and ChatGPT o4-mini were 79.3% (111/140) vs 85.7% (120/140; zone), 64.3% (90/140) vs 70.0% (98/140; stage), respectively. For the task of treatment requirement, the rates (also referred to as sensitivity) were 93.6% (131/140) vs 90.7% (127/140), respectively. None of these differences were statistically significant ($P > .05$). However, Gemini 2.5 Pro showed significantly better performance in plus disease identification (consistency: 108/140, 77.1% vs 80/140, 57.1%; $P = .006$), while ChatGPT o4-mini demonstrated significantly higher guideline adherence in treatment modality suggestions based on gold-standard ROP diagnoses (consistency: 91/140, 65.0%; vs 55/140, 39.3%; $P = .01$). Compared to Gemini 2.5 Pro, ChatGPT o4-mini performed better in providing ROP treatment suggestions (GQS score; $P = .001$), while the 2 AI chatbots had comparable GQS scores in diagnostic tasks.

Conclusions: ChatGPT o4-mini shows greater promise in generating evidence-based treatment suggestions based on gold-standard diagnoses, whereas Gemini 2.5 Pro shows advantages in visual interpretation, supporting its potential for targeted

ROP diagnostic screening, particularly in identifying plus disease. As these AI chatbots continue to evolve, their performance merits further validation using larger cohorts.

J Med Internet Res 2026;28:e86726; doi: [10.2196/86726](https://doi.org/10.2196/86726)

Keywords: Artificial intelligence chatbot; retinopathy of prematurity; wide-field fundus image; diagnosis; treatment suggestion

Introduction

Retinopathy of prematurity (ROP) is one of the leading causes of preventable childhood blindness worldwide, and its effective prevention and control have become an urgent global public health challenge [1,2]. Particularly in many middle-income countries [3], the improvement of neonatal intensive care capabilities has significantly increased the survival rate of low-birth-weight infants [2]. However, this increase has been accompanied by a rising incidence of ROP, which exacerbates the burden of childhood blindness [4]. Given the rapidly growing demand for ROP screening and treatment, the global shortage of experienced pediatric ophthalmologists has become increasingly severe, particularly in resource-limited regions [5,6]. In such regions, auxiliary diagnostic and treatment tools can reduce screening costs while supporting ophthalmologists in expanding service coverage through telemedicine [7,8]. This has further driven an unprecedented need for scalable, accurate, and standardized auxiliary tools [9].

The clinical management of ROP strictly adheres to evidence-based guidelines, and its diagnosis and treatment suggestions depend highly on the precise interpretation of fundus vascular and lesion features [10]. The third edition of the International Classification of ROP (ICROP3) updated and refined key indicators, including lesion zone, stage, and plus disease [11]. Wide-field fundus imaging is crucial not only for accurate diagnosis but also for providing objective visual evidence. These include the extent and morphology of the avascular retinal area, which are the core basis for guiding interventions such as laser therapy or anti-vascular endothelial growth factor (VEGF) injection [10,12]. In recent years, efficient feature preservation segmentation networks have further optimized the boundary delineation of fundus structures and lesions, thereby improving the reliability of AI-driven image interpretation [13]. To bridge the gap between these sophisticated diagnostic requirements and the global shortage of expert ophthalmologists, innovative auxiliary tools are urgently needed. Accordingly, AI chatbots

that can integrate clinical text data and medical images may offer potential support in ROP diagnosis and providing treatment suggestions [14,15].

In recent years, AI chatbots such as OpenAI's ChatGPT [16], Google's Gemini [17], and DeepSeek (Liang Wenfeng) [18] have demonstrated tangible efficacy in diagnosis and providing treatment suggestions for ophthalmic diseases [19-21]. For fundus disorders such as retinal detachment, age-related macular degeneration, and macular hole, both ChatGPT and Google Gemini exhibited consistent assessments of clinical records, showing high concordance with retinal specialists' evaluations [22,23]. Similarly, for other major fundus diseases, including age-related macular degeneration, glaucoma, and vitreous hemorrhage, ChatGPT showed preliminary diagnostic potential based solely on fundus photographs, although its performance remained variable [24]. The uncertain efficacy of AI chatbots in ROP diagnosis and providing treatment suggestions warrants a comparative analysis of their performance using a text-image fusion approach.

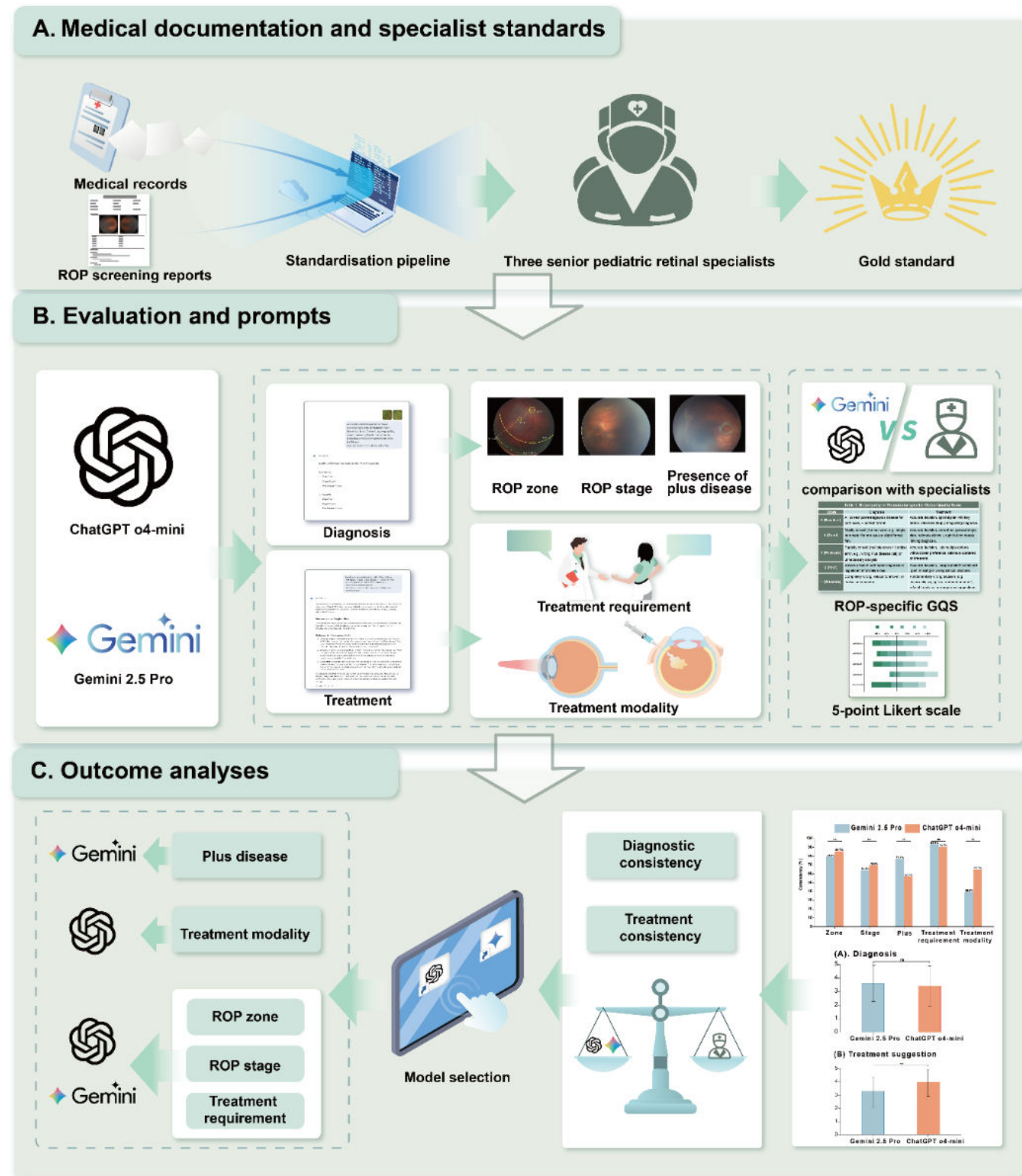
To compare the performance of Gemini 2.5 Pro and ChatGPT o4-mini in ROP diagnosis and providing treatment suggestions, we collected 70 infants with treatment-requiring ROP and evaluated the performance of the two AI chatbots. This study provides key empirical evidence and improvement directions for the application, evaluation, and iterative optimization of the 2 AI chatbots in pediatric ophthalmology and other imaging-driven clinical specialties.

Methods

Overview

This retrospective study was conducted at Shenzhen Eye Hospital (SZEH) and ultimately enrolled 70 infants (140 eyes in total) with ROP who underwent initial treatment for analysis in 2024. The entire process is illustrated in [Figure 1](#).

Figure 1. Study design and process for comparing the performance of two AI chatbots in the diagnosis and providing treatment suggestions for retinopathy of prematurity. The process consisted of 3 key steps: (A) medical documentation and specialist standards; (B) evaluation and prompts; (C) outcome analyses. GQS: Global Quality Score; ROP: retinopathy of prematurity.



Medical Documentation and Specialist Standards

We collected demographic and clinical data for each infant using a standardized structured template, including sex, gestational age (GA), birth weight (BW), pregnancy type, and

mode of delivery [25]. All structured fields were presented to both AI chatbots in an identical format to ensure fair comparisons. GA was recorded in weeks, and BW in grams, using a uniform numerical format. Clinical data comprised preoperative, deidentified bilateral fundus images obtained from ROP screening reports.

To ensure the reliability of the gold standard used for evaluating AI chatbot performance, we invited 3 senior pediatric retinal specialists (XZ, ZW, and GZ) to independently review the same ROP cases. Each specialist was provided with clinical text records and corresponding deidentified bilateral fundus images. Their tasks included (1) confirming the “gold standard” diagnosis for each case, which involved defining the ROP zone, stage, and presence of plus disease in accordance with the ICROP3 guidelines [11]; (2) establishing the optimal treatment suggestion, including the requirement for treatment and the selection of treatment modality (laser therapy or anti-VEGF injection). Fleiss kappa coefficient was used to assess interrater reliability. If all 3 specialists (XZ, ZW, and GZ) reached a unanimous agreement on the diagnosis and treatment, this consensus was finalized as the gold standard. Any disagreements were resolved through group discussion until a consensus was reached.

Selection of Infants With ROP

Inclusion criteria were applied to ensure data integrity and quality: (1) preterm infants with GA <32 weeks or BW <2000 g, (2) ROP diagnosed according to the ICROP3 guidelines [11] and who underwent initial laser therapy or anti-VEGF injection, (3) complete clinical data available, including ROP screening reports (binocular ROP zone, stage, and plus disease assessment), preoperative wide-field fundus images, and treatment records.

Exclusion criteria were as follows: (1) incomplete medical history or clinical information, (2) absence of key preoperative or postoperative reports, (3) previous ROP treatment at other hospitals, (4) ROP stage ≥ 4 (due to more complex treatment regimens and limited feasibility of standardized surgical intervention).

Evaluation and Prompt

Selection of AI Chatbots

The rapid expansion of AI chatbot research in health care [26–28] has witnessed the emergence of mainstream models such as OpenAI’s ChatGPT [16], Google’s Gemini [17], and DeepSeek [18], along with their subversions. Early versions such as ChatGPT-3.5 only supported text interaction and were unable to process medical images [29,30]. Among available Gemini subversions, we selected Gemini 2.5 Pro because complex ROP assessment requires detailed ophthalmic image interpretation and task-specific validation of multimodal reasoning in ophthalmology [31,32]. Specifically, Gemini 2.5 Pro supports text and image inputs, and provides a long context window, making it suitable for processing structured clinical records and fundus images in this text-image fusion task [33]. The development of medical vision-language models and foundation models in ophthalmology further supports the potential value of multimodal AI systems for ophthalmic image interpretation, while also highlighting the need for task-specific validation [34]. For ChatGPT o4-mini, the HealthBench framework released by OpenAI in 2025 provides a relevant reference for evaluating clinical reasoning and safety in medical AI systems [35,

36]. OpenAI documentation indicates that ChatGPT o4-mini supports text and image inputs and is optimized for fast and effective reasoning, including visual tasks [37,38]. This design is aligned with the technical challenge of simultaneously interpreting clinical records and wide-field fundus images in ROP diagnosis and providing treatment suggestions [21].

Therefore, our study selected Gemini 2.5 Pro and ChatGPT o4-mini for evaluation, as these AI chatbots meet the rigorous requirements of clinical tasks for multimodal data integration. Gemini 2.5 Pro was publicly available through Google’s official platform during the study period [33]. ChatGPT o4-mini used in this study refers to the reasoning-focused o4-mini model released by OpenAI in April 2025 [37], distinct from the earlier ChatGPT-4o mini omni model released in 2024. We used its web-exclusive high-inference mode, an enhanced reasoning variant optimized for complex multimodal and clinical reasoning tasks. All chatbot interactions in this study were conducted exclusively via the official web interfaces of the respective providers, without reliance on API access. These standard web interfaces do not provide end-users with adjustable parameters such as temperature. The temperature value is fixed at a default setting determined by the platform providers and cannot be modified by users. Consequently, all AI chatbot outputs were generated under the default, fixed configuration of the web platforms to ensure consistency and reproducibility.

AI Chatbot-Aided Diagnosis and Treatment Suggestions

We conducted standardized AI chatbot interactions with Gemini 2.5 Pro and ChatGPT o4-mini between June 1 and June 30, 2025. To reduce evasive responses that AI chatbots may generate when directly asked for medical advice (eg, stating “I am not a medical professional and cannot provide treatment suggestions”) [22], this study adopted a 2-stage, specialist-simulation prompting strategy. The initial prompt framed the task as a “clinical simulation analysis scenario” and assigned the AI chatbots the role of an “experienced pediatric fundus disease specialist” with additional context: “This is a clinical simulation analysis scenario. You will act as a pediatric ophthalmologist with rich clinical experience, and based on the basic medical history and a varying number of Retcam fundus images provided, briefly analyze the ROP zone, stage, and whether there is plus disease in this infant (report separately for left and right eyes, omitting the analysis process and directly providing the results).”

In the diagnosis task, the 2 AI chatbots were provided with standardized clinical records and deidentified bilateral fundus images (consistent with those obtained by experts) for each infant. The left and right fundus images were uploaded as separate files (1600×1200 resolution) within this conversation, rather than as composite images with white dividing bars, thus minimizing potential interpretation errors related to image boundaries. Simultaneously, the two AI chatbots were asked: “As a pediatric ophthalmologist with rich clinical experience, based on the provided patient history and Retcam

fundus images provided above, please directly report the ROP zone, stage, and presence of plus disease in this infant”

In the treatment suggestion task, we manually provided the gold-standard diagnoses (not the chatbot’s own outputs) to the AI chatbots, with the following instruction: “Based on my revised diagnoses of this infant, does the child require surgical intervention? If so, what is your recommended optimal surgical plan? Please briefly explain the rationale behind your suggestion”

For each infant, the diagnosis and treatment suggestion tasks were conducted within the same continuous chat session. For different infants, separate independent chat sessions were used with no information carryover between cases. A single query was submitted per case, and only

the initial response was included for analysis. All model hyperparameters (eg, temperature and top-p sampling) were retained at the default values of the official web interface, with no custom parameter adjustments. Screenshots of the response interfaces of the two AI chatbots are shown in [Figures 2](#) and [3](#). All prompts and interactions were conducted in Chinese and that the English prompts shown in the figures are translations. This design separated the AI chatbots’ performance in providing evidence-based treatment suggestions from diagnostic errors, enabling an unbiased evaluation based on the gold-standard diagnoses. At the end of each AI chatbot interaction, 2 specialists (XZ and ZW) independently graded the quality of the AI chatbots’ outputs using an ROP-specific Global Quality Score (GQS) adapted from Bernard et al [39].

Figure 2. Screenshot of Gemini 2.5 Pro’s response interface for an infant with retinopathy of prematurity.

Gemini 2.5 Pro

This is a clinical simulation analysis scenario. You will act as a pediatric ophthalmologist with rich clinical experience, and based on the basic medical history and a varying number of Retcam fundus images I provide, briefly analyze the ROP zoning, staging, and whether there is Plus lesion in this infant (report separately for left and right eyes, omitting the analysis process and directly providing the results)

Of course, I am ready to assist. Please provide the infant's basic medical history (gestational age, birth weight, and postmenstrual age at the time of examination) along with the Retcam fundus images.

I will provide a concise analysis for each eye as requested.

As a pediatric ophthalmologist with rich clinical experience, based on the provided patient history (female infant, born at 27 weeks 2 days, weighing 950g, singleton, delivered by C-section) and two Retcam fundus images provided above, please directly tell me the ROP zone, stage, and presence of plus disease in this infant.

Based on the information and images provided, here is my assessment.

Right Eye (OD)

- **Zone:** Zone II
- **Stage:** Stage 3
- **Plus Disease:** Present

Left Eye (OS)

- **Zone:** Zone II
- **Stage:** Stage 3
- **Plus Disease:** Present

Based on that diagnosis, the infant requires immediate surgical intervention. This condition is classified as **Type 1 ROP**, which carries a high risk of progressing to retinal detachment and permanent blindness if left untreated. Treatment should be administered promptly, typically within 48 to 72 hours.

Recommended Surgical Plan

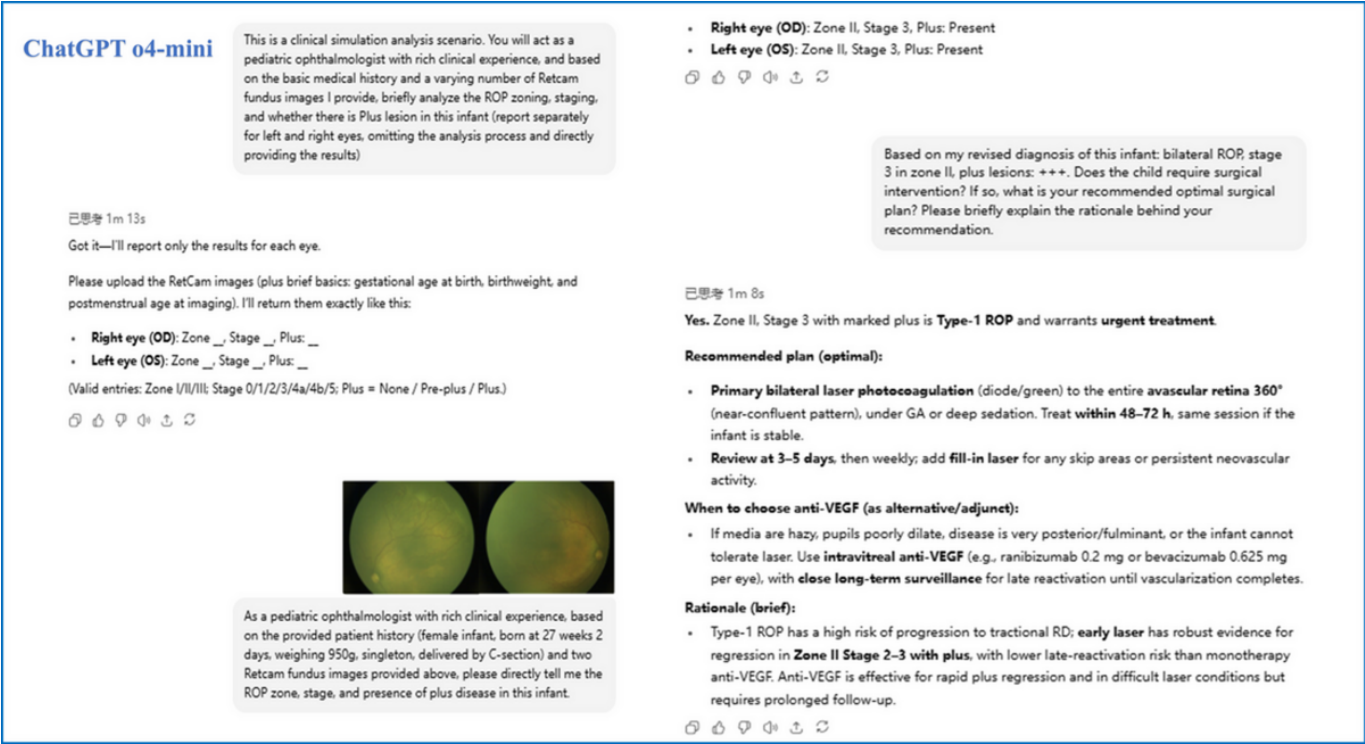
The optimal and most common first-line treatment for this specific presentation is an **intravitreal injection of an anti-VEGF (anti-vascular endothelial growth factor) agent**, such as bevacizumab or ranibizumab, into both eyes.

Rationale for Recommendation

1. **Directly Targets the Disease Driver:** ROP is primarily driven by abnormally high levels of VEGF, which causes the irregular blood vessel growth and leakage (the "Plus disease"). Anti-VEGF medication works by directly neutralizing this factor, leading to a rapid and often dramatic regression of the Plus disease and the abnormal vessels.
2. **Efficacy in Zone II with Severe Plus:** For ROP in Zone II with severe Plus disease, anti-VEGF therapy has been shown to be highly effective. It is less destructive than laser therapy, preserving the peripheral retina and allowing for the potential of more normal vessel development toward the front of the eye.
3. **Lower Risk of Myopia:** Traditional laser photocoagulation, while also an effective treatment, involves ablating a large area of the peripheral retina. This is known to induce a significant degree of high myopia (nearsightedness) later in life. Anti-VEGF treatment is associated with better refractive outcomes.

It is crucial to note that while anti-VEGF is the recommended initial step, the infant will require **diligent, long-term follow-up**. The disease can recur weeks or months after an injection, sometimes requiring additional treatment with either another injection or supplemental laser therapy.

Figure 3. Screenshot of ChatGPT o4-mini’s response interface for an infant with retinopathy of prematurity.



The primary outcome of this study was to explore the performance of Gemini 2.5 Pro and ChatGPT o4-mini in analyzing infants with ROP and generating coherent, accurate diagnoses and treatment suggestions by comparing their outputs to the specialists’ consensus gold standard (including consistency rates of diagnoses and treatment suggestions). For treatment modality selection, a response that selects an effective treatment modality consistent with gold-standard

decision criteria was rated as correct. As a secondary outcome, we used the mean GQS to quantify the subjective quality and clinical utility of the 2 AI chatbots’ output results.

ROP-Specific GQS

As shown in Table 1, the scoring system adopted a 5-point Likert scale (1=unusable, 5=excellent).

Table 1. Retinopathy of prematurity specific Global Quality Score criteria.

Score	Diagnosis	Treatment suggestion
5 (Excellent)	All bilateral diagnoses fully consistent with the gold standard	The treatment suggestions consistent with the gold standard, integrate individual characteristics and ICROP3 ^a guidelines, mention posttreatment monitoring, no hallucinations
4 (Good)	5 diagnoses consistent with the gold standard	The treatment suggestions consistent with the gold standard, integrate ICROP3 guidelines, lack of individualized details or posttreatment monitoring, no hallucinations
3 (Moderate)	3-4 diagnoses consistent with the gold standard	The treatment suggestions inconsistent with the gold standard, integrate ICROP3 guidelines, no hallucinations
2 (Poor)	1-2 diagnoses consistent with the gold standard	The treatment suggestions inconsistent with the gold standard, not integrate ICROP3 guidelines, hallucinations exist
1 (Unusable)	No diagnosis consistent with the gold standard or no response provided	The treatment suggestions contain fundamental errors (eg, no treatment required) or no suggestion is provided, hallucinations exist

^aICROP3: third edition of the International Classification of retinopathy of prematurity.

The ROP-specific GQS was adapted from the tool developed by Bernard et al [39], originally designed for web-site usability assessment. Given the lack of standardized quantitative instruments for evaluating clinical AI chatbots’ outputs, the GQS has been adopted in ophthalmic AI studies [40] and digital health research [41], supporting its use in our investigation. To better assess the quality of AI chatbot outputs in real ROP case analyses, we modified the GQS according to the ICROP3 guidelines [11] to

generate ROP-specific GQS. We also incorporated hallucination assessment for treatment suggestion outputs [42]. Prior to scoring, 2 ROP specialist raters (XZ and ZW) underwent standardized training covering ICROP3 criteria and ROP-specific GQS scoring rules, with pilot calibration and discussion of borderline cases. To minimize bias, AI chatbots’ outputs were deidentified and provided as plain text, ensuring raters were blinded to the source of AI chatbots. Interrater reliability

for assessing AI chatbots was evaluated using the intraclass correlation coefficient (ICC).

Statistical Analysis

All statistical analyses were performed using R (version 4.4.1; R Foundation for Statistical Computing). The “gold standard” was established by the consensus of 3 pediatric retinal specialists (XZ, ZW, and GZ). Interrater reliability regarding ROP zone, stage, plus disease, treatment requirement, and treatment modality were assessed using the Fleiss kappa coefficient. Any disagreements among the 3 pediatric retinal specialists were resolved through discussion until a consensus was reached. Generalized estimating equations (GEE) were used to compare the consistency rates and GQS of the 2 models while accounting for intereye correlation. As all enrolled eyes in this study were treatment-requiring ROP, the metric for treatment requirement was equivalent to sensitivity. Subgroup analyses of the 2 AI chatbots for diagnosis and providing treatment suggestions were conducted between Zone I and Zone II (as shown in Table 2), because Zone

III cases were insufficient for statistical analysis. For the 3 Zone III eyes, both ChatGPT o4-mini and Gemini 2.5 Pro misclassified all 3 eyes as Zone II. For the Zone I subgroup, complete separation was observed for Gemini 2.5 Pro in both the treatment requirement and treatment modality comparisons, as it achieved 100% correct classification. This separation precludes the use of conventional GEE Wald-type estimates due to their unreliability. Consequently, for these 2 comparisons, we applied a penalized GEE approach incorporating Firth likelihood penalty, which provides reliable estimates of the regression coefficients while properly accounting for the inherent correlation between eyes. For the GQS, each AI chatbot had 140 individual rater scores for each task. The mean GQS for each response was calculated by averaging the scores from the two raters, followed by the overall mean score. Score category frequencies and hallucination rates were calculated using the original integer scores provided by both raters. $P<.05$ was considered statistically significant.

Table 2. Subgroup analysis of two AI chatbots for diagnosis and providing treatment suggestions in Zones I and II.

Outcome	Zone I ^a (n=15)				Zone II ^b (n=122)			
	Gemini 2.5 Pro, n (%)	ChatGPT o4-mini, n (%)	OR ^c (95% CI)	P value	Gemini 2.5 Pro, n (%)	ChatGPT o4-mini, n (%)	OR (95% CI)	P value
Zone identification	10 (66.7)	2 (13.3)	0.08 (0.01-0.63)	.02	100 (82.0)	118 (96.7)	6.49 (1.53-27.55)	.01
Stage	9 (60.0)	12 (80.0)	2.70 (0.69-10.64)	.16	80 (65.6)	84 (68.9)	1.16 (0.60-2.23)	.66
Plus disease	13 (86.7)	13 (86.7)	1.00 (0.04-26.23)	1.0	94 (77.0)	65 (53.3)	0.34 (0.17-0.68)	.003
Treatment requirement	15 (100)	14 (93.3)	0.18 ^d (0.02-1.54)	.12 ^d	113 (92.6)	110 (90.2)	0.73 (0.45-1.17)	.19
Treatment modality	15 (100)	5 (33.3)	0.01 ^d (0.002-0.05)	<.001 ^d	40 (32.8)	83 (68.0)	4.36 (1.75-10.88)	.002

^aFor Zone I, no significant differences were observed in most outcomes, with the exception of ROP zone identification and treatment modality.

^bFor Zone II, ChatGPT o4-mini significantly outperformed Gemini 2.5 Pro in ROP zone identification (OR 6.49; $P=.011$) and treatment modality selection (OR 4.36; $P=.002$). Conversely, ChatGPT o4-mini showed significantly lower consistency in plus disease identification compared with Gemini 2.5 Pro (OR 0.34; $P=.003$). $P<.05$ was considered statistically significant.

^cOR: odds ratio.

^dFor the Zone I subgroup, complete separation occurred for Gemini 2.5 Pro in the treatment requirement and treatment modality comparisons (100% correct). Accordingly, a Firth-penalized GEE was applied to address this separation. Other statistical comparisons were performed using GEE. All ORs reflect ChatGPT o4-mini versus Gemini 2.5 Pro.

Ethical Considerations

The study was approved by the Institutional Review Board (IRB) of SZEH (IRB No. 2025KYPJ120) prior to its initiation. Informed consent was waived by the IRB due to the retrospective study design and the use of fully deidentified data for secondary analysis. To ensure compliance with ethical standards and data privacy regulations, all sensitive personal information was removed from case descriptions, only ROP-related key factors such as GA and BW were retained to balance research accuracy and privacy protection. All fundus images were anonymized, free of any potential biometric identifiers or embedded textual metadata, and subjected to quality control in adherence to data privacy

protocols. This study did not recruit any additional subjects and no compensation was provided to participants.

Results

Overview

Demographic characteristics of the 70 infants are shown in Table 3. Specifically, the mean GA was 26.79 (SD 2.14) weeks, and the mean BW was 888.60 (SD 277.82) g. Among all infants, 57.14% (40/70) were male; 64.29% (45/70) were singleton pregnancies; and the proportion of vaginal deliveries was comparable to that of cesarean sections.

Table 3. Demographic characteristics of infants with retinopathy of prematurity (N=70).

Characteristics	Values
Sex, n (%)	
Male	40 (57.14)
Female	30 (42.86)
Gestational age (weeks), mean (SD)	26.79 (2.14)
Birth weight (g), mean (SD)	888.60 (277.82)
Pregnancy type, n (%)	
Singleton	45 (64.29)
Multiple pregnancy	25 (35.71)
Delivery type, n (%)	
Vaginal delivery	37 (52.86)
Cesarean section	33 (47.14)

Diagnosis and treatment suggestions for the 140 eyes (70 infants) are shown in Table 4. The most common diagnosis was Zone II, Stage 3 ROP. Specifically, Zone II involvement was observed in 62 of 70 left eyes (88.6%) and 60 of 70 right eyes (85.7%). Stage 3 was noted in 63 of 70 left eyes (90.0%) and 64 of 70 right eyes (91.4%). Furthermore, plus

disease was highly prevalent, affecting 64 of 70 eyes (91.4%) bilaterally. All eyes required surgical treatment. Laser therapy was the primary modality, administered to 50 of 70 left eyes (71.4%) and 48 of 70 right eyes (68.6%). The remaining eyes were treated with anti-VEGF injections.

Table 4. Baseline characteristics of infants with retinopathy of prematurity (ROP; N=70).

Characteristic and category	Left eye, n (%)	Right eye, n (%)
ROP Zone		
Zone I	7 (10.0)	8 (11.4)
Zone II	62 (88.6)	60 (85.7)
Zone III	1 (1.4)	2 (2.9)
ROP Stage		
Stage 1	1 (1.4)	1 (1.4)
Stage 2	6 (8.6)	5 (7.1)
Stage 3	63 (90.0)	64 (91.4)
Plus disease		
Absent (-)	6 (8.6)	6 (8.6)
Present (+)	64 (91.4)	64 (91.4)
Treatment requirement		
Yes	70 (100.0)	70 (100.0)
Treatment modality		
Laser therapy	50 (71.4)	48 (68.6)
IVI ^a	20 (28.6)	22 (31.4)

^aIVI: intravitreal injection.

Interrater Reliability

Interrater reliability among the 3 specialists for the gold standard was substantial to excellent across all 5 tasks: Fleiss kappa for ROP zone=0.65 (95% CI 0.36-0.82), for ROP stage=0.72 (95% CI 0.43-0.90), for plus disease=0.66 (95% CI 0.34-0.86), and for treatment modality=0.82 (95% CI 0.68-0.92). For treatment requirement, all 3 specialists (XZ, ZW, and GZ) reached unanimous consensus, resulting in no calculated kappa value for this endpoint. Initial agreement was achieved in 84.3% (59/70) of cases; the remaining 11 cases with disagreement were resolved through discussion until full consensus was reached for all tasks. Subsequently, the consistency of GQS scoring for two AI

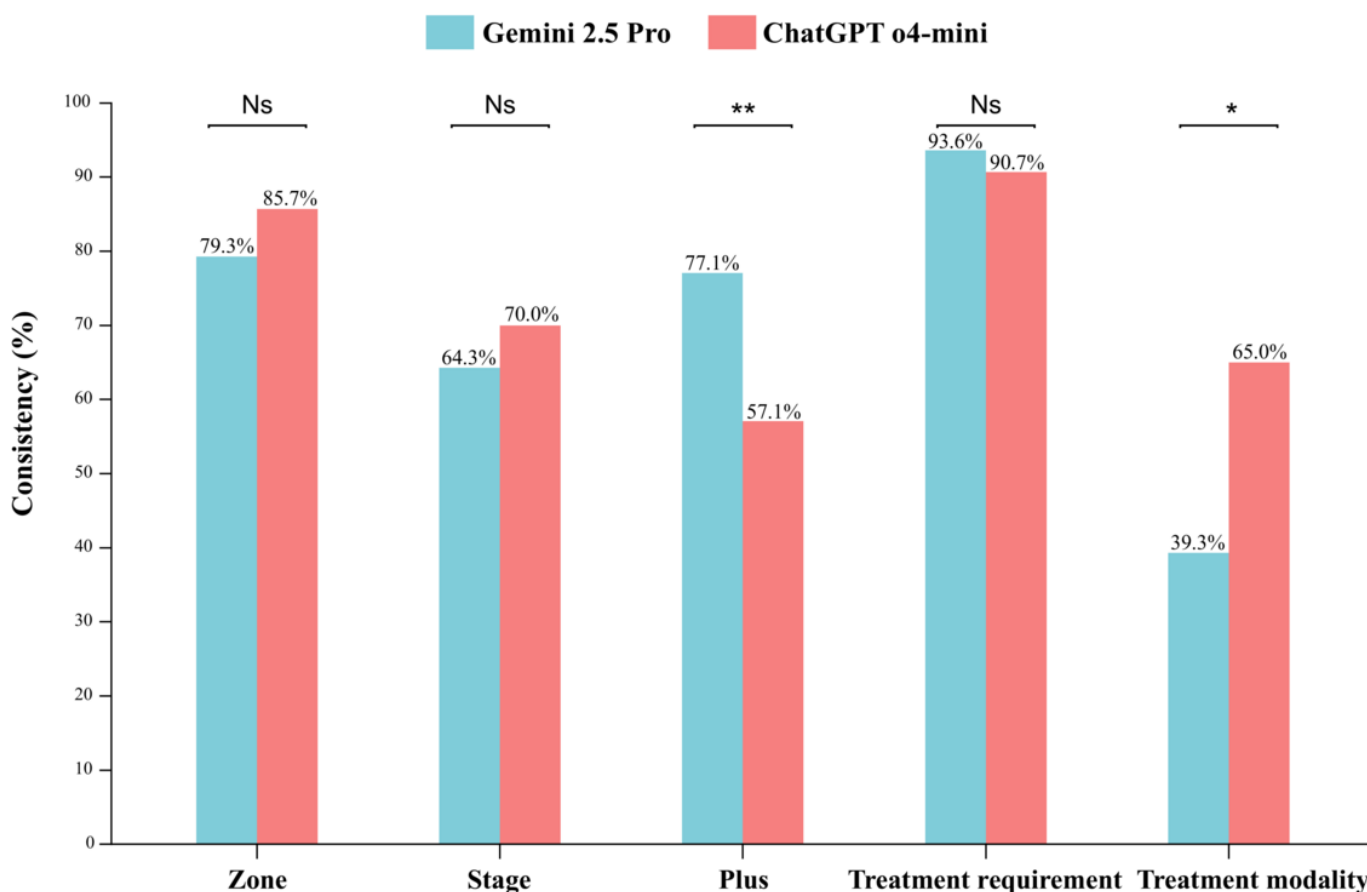
chatbots' outputs was evaluated by two independent ROP specialists (XZ and ZW) raters to ensure reliability of the secondary outcome measures. For Gemini 2.5 Pro, excellent interrater reliability was observed for both diagnostic GQS consistency (ICC=0.85; 95% CI 0.77-0.90) and treatment suggestion GQS consistency (ICC=0.81; 95% CI 0.71-0.88). Similarly, ChatGPT o4-mini demonstrated strong interrater reliability for diagnostic GQS consistency (ICC=0.86; 95% CI 0.78-0.91) and treatment suggestion GQS consistency (ICC=0.83; 95% CI 0.66-0.91).

Overall Performance Comparison of Two AI Chatbots

We compared the consistency between Gemini 2.5 Pro and ChatGPT o4-mini in 5 core clinical tasks (ROP zone, stage, plus disease identification, treatment requirement, and treatment modality). As shown in Figure 4, Gemini 2.5 Pro and ChatGPT o4-mini demonstrated comparable consistency in ROP zoning (79.3% vs 85.7%), staging (64.3% vs 70.0%), and treatment requirement (93.6% vs 90.7%), with no statistically significant differences (all $P>.05$). However, Gemini 2.5 Pro demonstrated significantly better consistency in diagnosing plus disease (77.1% vs 57.1%; $P=.006$), suggesting its advantage in visual pattern recognition tasks. In contrast, ChatGPT o4-mini showed significantly better consistency in treatment modality (65.0% vs 39.3%; $P=.01$),

indicating strengths in providing treatment suggestions based on the gold standard diagnoses. As shown in Table 2, subgroup analyses revealed distinct strengths and limitations between the two AI chatbots across ROP zones. All odds ratios (ORs) reflect ChatGPT o4-mini versus Gemini 2.5 Pro. In Zone I, Gemini 2.5 Pro maintained superior consistency, particularly in treatment modality task (100.0% vs 33.3%; OR=0.01; $P<.001$). Conversely, in Zone II, ChatGPT o4-mini significantly outperformed Gemini 2.5 Pro in ROP zone identification (96.7% vs 82.0%; OR=6.49; $P=.01$) and treatment modality selection (68.0% vs 32.8%; OR=4.36; $P=.002$). In addition, ChatGPT o4-mini showed significantly lower consistency in plus disease detection within Zone II compared with Gemini 2.5 Pro (53.3% vs 77.0%; OR=0.34; $P=.003$).

Figure 4. Comparison of consistency between Gemini 2.5 Pro and ChatGPT o4-mini in diagnosis and providing treatment suggestions for retinopathy of prematurity. The Y-axis represents consistency (%). The X-axis denotes specific tasks (ROP zoning, staging, plus disease identification, treatment requirement, treatment modality). Ns: not significant; *: $P<.05$, significant difference; **: $P<.01$, significant difference. Gemini 2.5 Pro outperformed ChatGPT o4-mini in plus disease identification ($P=.006$), whereas ChatGPT o4-mini was superior in treatment modality selection ($P=.01$).



Subgroup Analyses of Two AI Chatbots According to ROP Zone

We conducted subgroup analyses to explore the two AI chatbots' performance across distinct ROP risk zones, including Zone I (high-risk) and Zone II (intermediate-risk), to identify potential strengths and limitations within specific populations.

As shown in Table 2, in Zone I (n=15), ChatGPT o4-mini achieved numerically lower consistency rates than Gemini

2.5 Pro in zone identification (13.3% vs 66.7%; OR=0.08; $P=.02$), and treatment modality (33.3% vs 100%; OR=0.01; $P<.001$). No significant differences were observed for ROP staging ($P=.16$), plus disease identification ($P>.99$), and treatment requirement ($P=.12$) in Zone I.

In Zone II (n=122), ChatGPT o4-mini exhibited clear advantages in clinical decision-making. It significantly outperformed Gemini 2.5 Pro in both ROP zone identification (96.7% vs 82.0%; OR=6.49; $P=.01$) and treatment modality

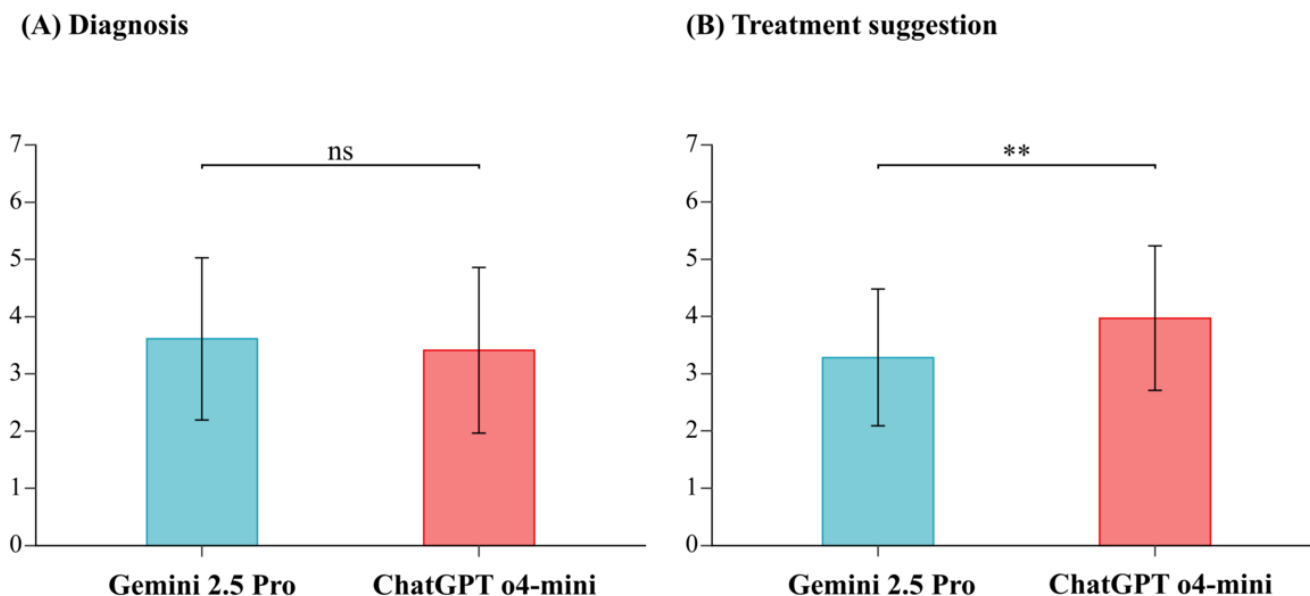
selection (68.0% vs 32.8%; OR=4.36; $P=.002$). Conversely, ChatGPT o4-mini showed significantly lower consistency in plus disease identification compared with Gemini 2.5 Pro (53.3% vs 77.0%; OR=0.34; $P=.003$). For ROP staging and treatment requirement, no statistically significant differences were observed between the two AI chatbots in Zone II.

GQS Results Analysis

The GQS results are presented in Figure 5. Based on the content evaluated by the GQS, hallucinations in treatment suggestion were identified in ratings of 1 or 2. For the diagnosis task, Gemini 2.5 Pro obtained a higher mean GQS score (3.61; SD 1.42) than ChatGPT o4-mini (3.41; SD 1.45), with no statistically significant difference ($P=.39$). Notably, the similar SDs across both AI chatbots suggest comparable variability in diagnostic output quality. Gemini 2.5 Pro had score 1 ratings in 7.14% (10/140) and score 2 ratings in 12.14% (17/140), while ChatGPT o4-mini yielded score 1 ratings in 7.86% (11/140) and score 2 ratings in 15.71% (22/140). In terms of treatment suggestion task,

ChatGPT o4-mini significantly outperformed Gemini 2.5 Pro (mean GQS: 3.97; SD 1.26 vs 3.29; SD 1.19; $P=.001$). The small difference in SDs suggests similar variability in treatment suggestion output quality. Gemini 2.5 Pro had score 1 ratings in 7.86% (11/140) and score 2 ratings in 14.29% (20/140), corresponding to a rating-level hallucination rate of 22.14% (31/140), while ChatGPT o4-mini yielded score 1 ratings in 7.14% (10/140) and score 2 ratings in 7.14% (10/140), corresponding to a rating-level hallucination rate of 14.29% (20/140). Hallucinations may manifest as proposing ineffective treatment modalities not endorsed by the ICROP3 guidelines, exaggerating the indications for a particular intervention, or fabricating unproven reasons for a treatment modality preference. These results showed that ChatGPT o4-mini achieved a significantly higher mean GQS score than Gemini 2.5 Pro for treatment suggestions and received fewer hallucination-related ratings, supporting its superior performance in providing clinical ROP treatment suggestions in this study.

Figure 5. Comparison of global quality scores (GQS) for diagnosis and providing treatment suggestion in retinopathy of prematurity between Gemini 2.5 Pro and ChatGPT o4-mini. The bar chart displays the mean GQS for both (A) Diagnosis and (B) Treatment suggestion. For diagnostic task, the average GQS was 3.61 for Gemini 2.5 Pro vs 3.41 for ChatGPT o4-mini ($P=.39$). For treatment suggestion task, the average GQS was 3.29 for Gemini 2.5 Pro vs 3.97 for ChatGPT o4-mini ($P=.001$). Error bars represent SD. ns: not statistically significant difference; **($P<.01$ (significant difference)).



Discussion

Principal Findings and Comparison to Prior Work

This study compared the performance of Gemini 2.5 Pro and ChatGPT o4-mini for diagnosing and providing treatment suggestions in ROP. The results provided key empirical evidence for the performance of AI chatbots to bridge the gap between the complex diagnostic requirements of ROP and the shortage of pediatric ophthalmologists. Furthermore, our findings may offer directions for improvement, with potential implications for the wider application and evaluation of AI chatbots in other imaging-driven specialties, consistent with the broader translational framework of AI in smart health care [43].

The tasks of ROP zoning, staging, and treatment requirement are relatively structured because they are based on standardized disease classification and treatment decision criteria provided by the ICROP3 guidelines [11]. This guideline helps AI chatbots transform complex clinical observation tasks into highly standardized and repeatable operational procedures, thus demonstrating consistently reliable performance in these tasks [11,44]. Both AI chatbots were provided with the gold-standard diagnoses (zone, stage, and presence of plus disease) before answering the question “Is surgical treatment required?,” as this task was highly objective and did not require complex clinical reasoning. Based on these standardized features, the 2 AI chatbots in this study showed relatively consistent performance on these tasks by leveraging their multimodal processing capabilities and the standardized criteria provided by the ICROP3 guidelines [11, 45].

The identification of plus disease in ROP requires precise visual discrimination of morphological changes in retinal vasculature [46]. Gemini 2.5 Pro exhibited superior consistency in identifying plus disease in this study. This advantage may be related to the broader ability of modern vision-language models to integrate visual features through multimodal representation learning and adaptive feature integration mechanisms [47,48]. In contrast, although ChatGPT o4-mini supports text-image reasoning and visual tasks according to OpenAI documentation [49], its lower consistency in plus disease identification in this study suggests that general-purpose multimodal reasoning may not fully translate into optimal recognition of subtle ROP vascular abnormalities. This performance difference appears to be task-dependent. Prior ophthalmology studies have shown that the relative performance of AI chatbots varies across examination-based and clinical ophthalmology tasks [31,50,51]. This finding does not contradict our results but rather highlights the “performance specialization” of AI chatbots [50,51]. These differences may be associated with model-specific optimization strategies, training data composition, input format, and task types [45,52]. Specifically, the training data varied: our study used wide-field fundus images, whereas the exam study used typical test-related images. Furthermore, these tasks impose different cognitive demands: our study focused

on high-fidelity visual discrimination, whereas exam-based studies require textual integration of visual materials of diverse complexity. Therefore, vision-optimized AI chatbots such as Gemini 2.5 Pro are more adaptable for clinical tasks that rely on fine visual pattern recognition, as their performance aligns closely with the visual dependency of the task.

In tasks of determining treatment modality, which demand the integration of multidimensional clinical information (including clinical text records and wide-field fundus images), AI chatbots need to simulate specialists’ multifactor trade-off logic. Consistent with prior ophthalmic multimodal AI chatbot research [21], ChatGPT o4-mini may have benefited from integrating structured text records and fundus images when deducing the optimal treatment modality, yielding superior performance. Previous medical large language model studies also suggest that language models can perform well on structured medical knowledge and reasoning tasks [53, 54]. For questions based on established medical knowledge with clear and objective answers, these AI chatbots achieved strong performance. This indicates that when tasks are governed by explicit rules and objective criteria, the performance gap among different AI chatbot architectures narrows, and their capabilities converge.

Subgroup analysis showed that in high-risk Zone I, characterized by severe posterior retinal lesions and vascular abnormalities [11], Gemini 2.5 Pro outperformed ChatGPT o4-mini in ROP zone identification and treatment modality selection. This finding may again reflect Gemini 2.5 Pro’s stronger task-specific visual discrimination in cases requiring fine vascular pattern recognition, suggesting its potential value in the identification of complex retinal vascular diseases. Conversely, in the intermediate-risk Zone II with relatively mild vascular lesions, ChatGPT o4-mini significantly outperformed Gemini 2.5 Pro in both zone identification and treatment modality selection. Since treatment decisions in Zone II rely more heavily on the comprehensive clinical context rather than isolated visual examination results [11], ChatGPT o4-mini may have benefited from integrating structured clinical records and fundus images when generating treatment suggestions, consistent with prior ophthalmic multimodal AI chatbot research [21]. It is worth noting that ChatGPT o4-mini’s ability to identify plus disease is not as good as Gemini 2.5 Pro, which is consistent with the overall comparison results. The above results confirm that the performance of the AI chatbot is closely related to the risk of the ROP zone. Gemini 2.5 Pro is more suitable for high-risk Zone I cases that require precise visual identification, while ChatGPT o4-mini is more suitable for medium-risk Zone II cases. This also verifies the overall performance of the two AI chatbots in different clinical scenarios and their robustness in handling different retinal complexities.

The difference in GQS between ChatGPT o4-mini and Gemini 2.5 Pro varies across task dimensions, reflecting the interplay between specialist scoring logic and AI chatbot performance. According to our ROP-specific GQS criteria, specialists mainly evaluated completeness and clinical practicality. When scoring, specialists particularly focus on whether the AI chatbots’ output fully covers the core

diagnostic elements of ROP cases and whether it conforms to the follow-up or treatment suggestions. In the diagnosis dimension, no significant difference in GQS was observed between the two AI chatbots, indicating comparable performance in capturing ROP diagnostic details. In contrast, ChatGPT o4-mini achieved a significantly higher GQS in the treatment suggestion dimension. This advantage may be partly explained by its structured output format, which is compatible with clinical workflows and may help specialists extract key information efficiently [55]. However, because the training data and optimization details of commercial AI chatbots are not fully disclosed, this interpretation should be regarded as a task-specific observation rather than a confirmed model-level mechanism. Although Gemini 2.5 Pro performs well in isolated image analysis, it often experiences logical discontinuities when integrating multidimensional textual information of ROP cases, resulting in a decreased matching degree between the output and actual clinical requirements.

Notably, the SD patterns differed between the 2 AI chatbots. For Gemini 2.5 Pro, the SDs for diagnostic GQS were higher than that for treatment suggestion GQS, while ChatGPT o4-mini exhibited higher SDs in treatment suggestion GQS. This aligns with the common observation of insufficient output consistency in AI chatbots for complex clinical decisions, particularly evident in multimodal information integration tasks [56]. The output quality and stability of AI chatbots are positively associated with their ability to structure clinical knowledge and are influenced by task complexity, further supporting the emphasis on transparency, usability, and validity in explainable AI systems for clinical decision support [57].

Limitations

This study has several limitations. First, we evaluated AI chatbots performance on moderately severe and well-documented ROP cases (excluding stage 4 and above) rather than the full clinical spectrum encountered in routine ROP screening scenarios, failing to reflect real-world data quality challenges and limiting the clinical applicability of our findings. Furthermore, the single-center convenience sampling design and potential selection bias restrict geographic and demographic generalizability. Notably, although some findings from the Zone I subgroup analysis reached statistical significance, these results should be interpreted with caution due to the small sample size, and validation in larger cohorts is still needed. Second, given the exploratory nature of this study, formal adjustments for multiple comparisons (eg, Bonferroni correction) were not applied to avoid an inflated risk of Type II error (false-negative conclusions). While this represents a potential limitation, the robust effect sizes observed for the primary endpoints largely mitigated this concern. Third, the lack of cross-platform validation across different software and hardware environments limits the generalizability of the AI chatbots and their reliability in diverse clinical settings. Fourth, the AI

chatbots have different training data cutoffs, resulting in asynchronous knowledge updates that may be inconsistent with contemporary clinical practice standards, with Gemini 2.5 Pro trained up to January 2025 and ChatGPT o4-mini up to June 2024. This discrepancy in training data cutoff dates between the two AI chatbots may also contribute to the observed differences in their diagnostic performance. Fifth, there is still a potential model hierarchy mismatch between Gemini 2.5 Pro (flagship high-computation model) and ChatGPT o4-mini (efficiency-focused model), despite our use of the enhanced reasoning mode of ChatGPT o4-mini. Sixth, our prompt instructions forced the 2 AI chatbots to omit analytical processes and directly output results, which may have artificially lowered their performance. The prompts did not explicitly request eye-specific separate treatment suggestions, though the models spontaneously generated treatment suggestions for each eye separately, which may affect the accuracy of our analysis results. Furthermore, in the treatment suggestion task, we manually input gold-standard diagnoses rather than using the chatbots' own outputs, and all diagnostic and therapeutic queries were completed within a single continuous chat session, which precluded evaluation of their end-to-end clinical management performance. All interactions were continuous, so the treatment suggestions generated by chatbots were not produced within an isolated conversational context. Moreover, all interactions were conducted in Chinese, and the AI chatbot performance may vary across prompting languages. Finally, the absence of standardized tools to quantify "hallucinations in treatment suggestion" prevents objective assessment of potential risks in chatbot outputs, weakening their credibility as clinical decision support tools.

Future Directions

While our results, validated against specialist consensus as the gold standard, appeared promising, rigorous external validation—including multicenter studies, prospective trials, and cross-platform verification—is essential before clinical implementation to confirm generalizability and ensure patient safety. Future studies can validate the end-to-end clinical management performance of AI chatbots based on their own independent diagnoses, rather than manual input of gold-standard diagnoses. In addition, further balanced cross-model comparisons are also needed to explore how model size, parameter count, and architecture shape the performance of AI chatbots in multimodal ROP tasks.

Conclusion

In this study of ROP, ChatGPT o4-mini appears to be more promising for evidence-based treatment suggestions based on the gold standard diagnoses, while Gemini 2.5 Pro's visual acuity highlights its potential for targeted ROP diagnostic screening, particularly in identifying plus disease. As these AI chatbots continue to evolve, further validation in larger and more diverse cohorts is warranted to establish their clinical utility and generalizability.

Acknowledgments

We used Gemini 2.5 Pro (Google) [33] and ChatGPT o4-mini (OpenAI) [49] to generate response interfaces for infants with retinopathy of prematurity (ROP). These responses were subsequently reviewed and revised by the research team. Figures 2 and 3 were generated using Gemini 2.5 Pro (Google) [33] and ChatGPT o4-mini (OpenAI) [49].

Funding

This work was supported by Shenzhen Medical Research Fund (C2301005, C2501034, A2403020), National Natural Science Foundation of China (82301269, 82271103, 82401315, 82401272, 82301226), Shenzhen Science and Technology R&D Fund Program (JCYJ20240813152703005, JCYJ20250604184009011), Guangdong Basic and Applied Basic Research Foundation (2026A1515012558), Sanming Project of Medicine in Shenzhen (SZSM202311018).

Data Availability

Core data are provided in the manuscript. The raw AI outputs generated during this study, including diagnostic and treatment suggestions from both models, are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: SP, XZ, ZW, and GZ.

Methodology: SP, XZ, ZW, and GZ.

Investigation: DY and ND.

Resources: DY and ND.

Validation: DY and ND.

Software: SP and XZ.

Formal analysis: SP and XZ.

Data curation: SP and XZ.

Visualization: SP and XZ.

Writing – original draft: SP and XZ.

Writing – review and editing: SP, XZ, ZW, and GZ.

Funding acquisition: KC, ZY, WY, WW, and WC.

Project administration: KC, ZY, WY, WW, and WC.

Supervision: KC, ZY, WY, WW, and WC.

GZ, WC, WY contributed equally as the corresponding authors.

Conflicts of Interest

None declared.

References

1. Gilbert C. Retinopathy of prematurity: a global perspective of the epidemics, population of babies at risk and implications for control. *Early Hum Dev*. Feb 2008;84(2):77-82. [doi: [10.1016/j.earlhumdev.2007.11.009](https://doi.org/10.1016/j.earlhumdev.2007.11.009)] [Medline: [18234457](https://pubmed.ncbi.nlm.nih.gov/18234457/)]
2. Blencowe H, Lawn JE, Vazquez T, Fielder A, Gilbert C. Preterm-associated visual impairment and estimates of retinopathy of prematurity at regional and global levels for 2010. *Pediatr Res*. Dec 2013;74 Suppl 1(Suppl 1):35-49. [doi: [10.1038/pr.2013.205](https://doi.org/10.1038/pr.2013.205)] [Medline: [24366462](https://pubmed.ncbi.nlm.nih.gov/24366462/)]
3. García H, Villasis-Keever MA, Zavala-Vargas G, Bravo-Ortiz JC, Pérez-Méndez A, Escamilla-Núñez A. Global prevalence and severity of retinopathy of prematurity over the last four decades (1985-2021): a systematic review and meta-analysis. *Arch Med Res*. Feb 2024;55(2):102967. [doi: [10.1016/j.arcmed.2024.102967](https://doi.org/10.1016/j.arcmed.2024.102967)] [Medline: [38364488](https://pubmed.ncbi.nlm.nih.gov/38364488/)]
4. Wood EH, Chang EY, Beck K, Hadfield BR, Quinn AR, Harper CA 3rd. 80 Years of vision: preventing blindness from retinopathy of prematurity. *J Perinatol*. Jun 2021;41(6):1216-1224. [doi: [10.1038/s41372-021-01015-8](https://doi.org/10.1038/s41372-021-01015-8)] [Medline: [33674712](https://pubmed.ncbi.nlm.nih.gov/33674712/)]
5. Quinn GE, Ying G shuang, Daniel E, et al. Validity of a telemedicine system for the evaluation of acute-phase retinopathy of prematurity. *JAMA Ophthalmol*. Oct 2014;132(10):1178-1184. [doi: [10.1001/jamaophthalmol.2014.1604](https://doi.org/10.1001/jamaophthalmol.2014.1604)] [Medline: [24970095](https://pubmed.ncbi.nlm.nih.gov/24970095/)]
6. Vinekar A, Jayadev C, Mangalesh S, Shetty B, Vidyasagar D. Role of tele-medicine in retinopathy of prematurity screening in rural outreach centers in India - a report of 20,214 imaging sessions in the KIDROP program. *Semin Fetal Neonatal Med*. Oct 2015;20(5):335-345. [doi: [10.1016/j.siny.2015.05.002](https://doi.org/10.1016/j.siny.2015.05.002)] [Medline: [26092301](https://pubmed.ncbi.nlm.nih.gov/26092301/)]
7. Hennein L, Jastrzebski B, Shah AS. Use of telemedicine in pediatric ophthalmology in the underserved population. *Semin Ophthalmol*. Feb 2023;38(2):116-123. [doi: [10.1080/08820538.2022.2152703](https://doi.org/10.1080/08820538.2022.2152703)] [Medline: [36529958](https://pubmed.ncbi.nlm.nih.gov/36529958/)]

8. Takeda Y, Kaneko Y, Sugimoto M, Yamashita H, Sasaki A, Mitsui T. Prediction models for retinopathy of prematurity using nonimaging machine learning approaches: a regional multicenter study. *Ophthalmol Sci*. 2025;5(4):100715. [doi: [10.1016/j.xops.2025.100715](https://doi.org/10.1016/j.xops.2025.100715)] [Medline: [40182984](https://pubmed.ncbi.nlm.nih.gov/40182984/)]
9. Barrero-Castillero A, Corwin BK, VanderVeen DK, Wang JC. Workforce shortage for retinopathy of prematurity care and emerging role of telehealth and artificial intelligence. *Pediatr Clin North Am*. Aug 2020;67(4):725-733. [doi: [10.1016/j.pcl.2020.04.012](https://doi.org/10.1016/j.pcl.2020.04.012)] [Medline: [32650869](https://pubmed.ncbi.nlm.nih.gov/32650869/)]
10. Fouzdar Jain S, Song HH, Al-Holou SN, Morgan LA, Suh DW. Retinopathy of prematurity: preferred practice patterns among pediatric ophthalmologists. *Clin Ophthalmol*. 2018;12:1003-1009. [doi: [10.2147/OPTH.S161504](https://doi.org/10.2147/OPTH.S161504)] [Medline: [29881255](https://pubmed.ncbi.nlm.nih.gov/29881255/)]
11. Chiang MF, Quinn GE, Fielder AR, et al. International Classification of Retinopathy of Prematurity, Third Edition. *Ophthalmology*. Oct 2021;128(10):e51-e68. [doi: [10.1016/j.ophtha.2021.05.031](https://doi.org/10.1016/j.ophtha.2021.05.031)] [Medline: [34247850](https://pubmed.ncbi.nlm.nih.gov/34247850/)]
12. Fiererson WM, American Academy of Pediatrics Section on Ophthalmology, American Academy of Ophthalmology, American Association for Pediatric Ophthalmology and Strabismus, American Association of Certified Orthoptists. Screening examination of premature infants for retinopathy of prematurity. *Pediatrics*. Dec 2018;142(6):e20183061. [doi: [10.1542/peds.2018-3061](https://doi.org/10.1542/peds.2018-3061)] [Medline: [30478242](https://pubmed.ncbi.nlm.nih.gov/30478242/)]
13. Abidin ZU, Naqvi RA, Kim HS, Kim HS, Jeong D, Lee SW. Optimizing optic cup and optic disc delineation: Introducing the efficient feature preservation segmentation network. *Eng Appl Artif Intell*. Mar 2025;144:110038. [doi: [10.1016/j.engappai.2025.110038](https://doi.org/10.1016/j.engappai.2025.110038)]
14. Zhang Z, Zhang H, Pan Z, et al. Evaluating large language models in ophthalmology: systematic review. *J Med Internet Res*. Oct 27, 2025;27:e76947. [doi: [10.2196/76947](https://doi.org/10.2196/76947)] [Medline: [41144954](https://pubmed.ncbi.nlm.nih.gov/41144954/)]
15. Jin K, Yu T, Grzybowski A. Multimodal artificial intelligence in ophthalmology: applications, challenges, and future directions. *Surv Ophthalmol*. 2026;71(1):158-167. [doi: [10.1016/j.survophthal.2025.07.003](https://doi.org/10.1016/j.survophthal.2025.07.003)] [Medline: [40683606](https://pubmed.ncbi.nlm.nih.gov/40683606/)]
16. Liu W, Kan H, Jiang Y, Geng Y, Nie Y, Yang M. MED-ChatGPT CoPilot: a ChatGPT medical assistant for case mining and adjunctive therapy. *Front Med (Lausanne)*. 2024;11:1460553. [doi: [10.3389/fmed.2024.1460553](https://doi.org/10.3389/fmed.2024.1460553)] [Medline: [39478827](https://pubmed.ncbi.nlm.nih.gov/39478827/)]
17. Huang KA, Choudhary HK, Hardin WM, Prakash N. Comparative analysis of ChatGPT-4o and Gemini Advanced performance on diagnostic radiology in-training exams. *Cureus*. Mar 2025;17(3):e80874. [doi: [10.7759/cureus.80874](https://doi.org/10.7759/cureus.80874)] [Medline: [40255788](https://pubmed.ncbi.nlm.nih.gov/40255788/)]
18. Sandmann S, Hegselmann S, Fujarski M, et al. Benchmark evaluation of DeepSeek large language models in clinical decision-making. *Nat Med*. Aug 2025;31(8):2546-2549. [doi: [10.1038/s41591-025-03727-2](https://doi.org/10.1038/s41591-025-03727-2)] [Medline: [40267970](https://pubmed.ncbi.nlm.nih.gov/40267970/)]
19. David D, Zloto O, Katz G, et al. The use of artificial intelligence based chat bots in ophthalmology triage. *Eye (Lond)*. Mar 2025;39(4):785-789. [doi: [10.1038/s41433-024-03488-1](https://doi.org/10.1038/s41433-024-03488-1)] [Medline: [39592814](https://pubmed.ncbi.nlm.nih.gov/39592814/)]
20. Tomita K, Nishida T, Kitaguchi Y, Kitazawa K, Miyake M. Image recognition performance of GPT-4V(ision) and GPT-4o in ophthalmology: use of images in clinical questions. *Clin Ophthalmol*. 2025;19:1557-1564. [doi: [10.2147/OPTH.S494480](https://doi.org/10.2147/OPTH.S494480)] [Medline: [40357454](https://pubmed.ncbi.nlm.nih.gov/40357454/)]
21. Ma R, Cheng Q, Yao J, et al. Multimodal machine learning enables AI chatbot to diagnose ophthalmic diseases and provide high-quality medical responses. *NPJ Digit Med*. Jan 27, 2025;8(1):64. [doi: [10.1038/s41746-025-01461-0](https://doi.org/10.1038/s41746-025-01461-0)] [Medline: [39870855](https://pubmed.ncbi.nlm.nih.gov/39870855/)]
22. Carlà MM, Gambini G, Baldascino A, et al. Exploring AI-chatbots' capability to suggest surgical planning in ophthalmology: ChatGPT versus Google Gemini analysis of retinal detachment cases. *Br J Ophthalmol*. Sep 20, 2024;108(10):1457-1469. [doi: [10.1136/bjo-2023-325143](https://doi.org/10.1136/bjo-2023-325143)] [Medline: [38448201](https://pubmed.ncbi.nlm.nih.gov/38448201/)]
23. Duan N, Zhao X, Wu Z, et al. Exploring the capabilities of three artificial intelligence chatbots in diagnosis and decision-making of age-related macular degeneration. *Br J Ophthalmol*. Jun 10, 2026;bjo-2025-328860. [doi: [10.1136/bjo-2025-328860](https://doi.org/10.1136/bjo-2025-328860)] [Medline: [42270292](https://pubmed.ncbi.nlm.nih.gov/42270292/)]
24. Gupta A, Al-Kazwini H. Evaluating ChatGPT's diagnostic accuracy in detecting fundus images. *Cureus*. 2024;16:e73660. [doi: [10.7759/cureus.73660](https://doi.org/10.7759/cureus.73660)]
25. Zhao X, Wu Z, Liu Y, et al. Eyecare-cloud: an innovative electronic medical record cloud platform for pediatric research and clinical care. *EPMA J*. Sep 2024;15(3):501-510. [doi: [10.1007/s13167-024-00372-6](https://doi.org/10.1007/s13167-024-00372-6)] [Medline: [39239111](https://pubmed.ncbi.nlm.nih.gov/39239111/)]
26. Xu P, Wu Y, Jin K, Chen X, He M, Shi D. DeepSeek-R1 outperforms Gemini 2.0 Pro, OpenAI o1, and o3-mini in bilingual complex ophthalmology reasoning. *Adv Ophthalmol Pract Res*. 2025;5(3):189-195. [doi: [10.1016/j.aopr.2025.05.001](https://doi.org/10.1016/j.aopr.2025.05.001)] [Medline: [40678192](https://pubmed.ncbi.nlm.nih.gov/40678192/)]
27. Bellanda VCF, Santos MLD, Ferraz DA, Jorge R, Melo GB. Applications of ChatGPT in the diagnosis, management, education, and research of retinal diseases: a scoping review. *Int J Retina Vitreous*. Oct 17, 2024;10(1):79. [doi: [10.1186/s40942-024-00595-9](https://doi.org/10.1186/s40942-024-00595-9)] [Medline: [39420407](https://pubmed.ncbi.nlm.nih.gov/39420407/)]
28. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? *Patterns (NY)*. Mar 8, 2024;5(3):100943. [doi: [10.1016/j.patter.2024.100943](https://doi.org/10.1016/j.patter.2024.100943)] [Medline: [38487804](https://pubmed.ncbi.nlm.nih.gov/38487804/)]

29. Joseph A, Joseph K, Joseph A. A pilot evaluation of the diagnostic accuracy of ChatGPT-3.5 for multiple sclerosis from case reports. *Transl Neurosci*. Jan 1, 2024;15(1):20220361. [doi: [10.1515/tnsci-2022-0361](https://doi.org/10.1515/tnsci-2022-0361)] [Medline: [39726894](https://pubmed.ncbi.nlm.nih.gov/39726894/)]
30. Chen CH, Hsieh KY, Huang KE, Lai HY. Comparing vision-capable models, GPT-4 and Gemini, with GPT-3.5 on Taiwan's pulmonologist exam. *Cureus*. Aug 2024;16(8):e67641. [doi: [10.7759/cureus.67641](https://doi.org/10.7759/cureus.67641)] [Medline: [39185287](https://pubmed.ncbi.nlm.nih.gov/39185287/)]
31. Silhadi M, Nassrallah WB, Mikhail D, Milad D, Harissi-Dagher M. Assessing the performance of Microsoft Copilot, GPT-4 and Google Gemini in ophthalmology. *Can J Ophthalmol*. Aug 2025;60(4):e507-e514. [doi: [10.1016/j.jcjo.2025.01.001](https://doi.org/10.1016/j.jcjo.2025.01.001)] [Medline: [39863285](https://pubmed.ncbi.nlm.nih.gov/39863285/)]
32. Antaki F, Chopra R, Keane PA. Vision-language models for feature detection of macular diseases on optical coherence tomography. *JAMA Ophthalmol*. Jun 1, 2024;142(6):573-576. [doi: [10.1001/jamaophthalmol.2024.1165](https://doi.org/10.1001/jamaophthalmol.2024.1165)] [Medline: [38696177](https://pubmed.ncbi.nlm.nih.gov/38696177/)]
33. Google Cloud. Gemini 25 Pro. URL: <https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro> [Accessed 2026-09-07]
34. Chia MA, Antaki F, Zhou Y, Turner AW, Lee AY, Keane PA. Foundation models in ophthalmology. *Br J Ophthalmol*. Sep 20, 2024;108(10):1341-1348. [doi: [10.1136/bjo-2024-325459](https://doi.org/10.1136/bjo-2024-325459)] [Medline: [38834291](https://pubmed.ncbi.nlm.nih.gov/38834291/)]
35. Liu J, Liu S. Dissecting HealthBench: disease spectrum, clinical diversity, and data insights from multi-turn clinical AI evaluation benchmark. *J Med Syst*. Jul 28, 2025;49(1):100. [doi: [10.1007/s10916-025-02232-w](https://doi.org/10.1007/s10916-025-02232-w)] [Medline: [40719790](https://pubmed.ncbi.nlm.nih.gov/40719790/)]
36. OpenAI. Introducing HealthBench. 2025. URL: <https://openai.com/index/healthbench> [Accessed 2026-09-07]
37. OpenAI o3 and o4-mini System Card. OpenAI. 2025. URL: <https://openai.com/index/o3-o4-mini-system-card> [Accessed 2026-09-07]
38. O4-mini: fast, cost-efficient reasoning model. OpenAI. 2025. URL: <https://developers.openai.com/api/docs/models/o4-mini> [Accessed 2026-09-07]
39. Bernard A, Langille M, Hughes S, Rose C, Leddin D, Veldhuyzen van Zanten S. A systematic review of patient inflammatory bowel disease information resources on the world wide web. *Am J Gastroenterol*. Sep 2007;102(9):2070-2077. [doi: [10.1111/j.1572-0241.2007.01325.x](https://doi.org/10.1111/j.1572-0241.2007.01325.x)] [Medline: [17511753](https://pubmed.ncbi.nlm.nih.gov/17511753/)]
40. Reyhan AH, Mutaf Ç, Uzun İ, Yüksekayla F. A performance evaluation of large language models in keratoconus: a comparative study of ChatGPT-3.5, ChatGPT-4.0, Gemini, Copilot, Chatsonic, and Perplexity. *J Clin Med*. Oct 30, 2024;13(21):6512. [doi: [10.3390/jcm13216512](https://doi.org/10.3390/jcm13216512)] [Medline: [39518652](https://pubmed.ncbi.nlm.nih.gov/39518652/)]
41. Rao M, Xiujun T, Haoyu W. Evaluating GPT-4 responses on scars or keloids for patient education: large language model evaluation study. *JMIR Med Inform*. Feb 27, 2026;14:e78838. [doi: [10.2196/78838](https://doi.org/10.2196/78838)] [Medline: [41773665](https://pubmed.ncbi.nlm.nih.gov/41773665/)]
42. Maywood MJ, Parikh R, Deobhakta A, Begaj T. Performance assessment of an artificial intelligence chatbot in clinical vitreoretinal scenarios. *Retina*. Jun 1, 2024;44(6):954-964. [doi: [10.1097/IAE.0000000000004053](https://doi.org/10.1097/IAE.0000000000004053)] [Medline: [38271674](https://pubmed.ncbi.nlm.nih.gov/38271674/)]
43. Abbas SR, Seol H, Abbas Z, Lee SW. Exploring the role of artificial intelligence in smart healthcare: a capability and function-oriented review. *Healthcare (Basel)*. Jul 8, 2025;13(14):1642. [doi: [10.3390/healthcare13141642](https://doi.org/10.3390/healthcare13141642)] [Medline: [40724669](https://pubmed.ncbi.nlm.nih.gov/40724669/)]
44. Campbell JP, Chiang MF, Chen JS, et al. Artificial intelligence for retinopathy of prematurity: validation of a vascular severity scale against international expert diagnosis. *Ophthalmology*. Jul 2022;129(7):e69-e76. [doi: [10.1016/j.ophtha.2022.02.008](https://doi.org/10.1016/j.ophtha.2022.02.008)] [Medline: [35157950](https://pubmed.ncbi.nlm.nih.gov/35157950/)]
45. Hartsock I, Rasool G. Vision-language models for medical report generation and visual question answering: a review. *Front Artif Intell*. 2024;7:1430984. [doi: [10.3389/frai.2024.1430984](https://doi.org/10.3389/frai.2024.1430984)] [Medline: [39628839](https://pubmed.ncbi.nlm.nih.gov/39628839/)]
46. Campbell JP, Ataer-Cansizoglu E, Bolon-Canedo V, et al. Expert diagnosis of plus disease in retinopathy of prematurity from computer-based image analysis. *JAMA Ophthalmol*. Jun 1, 2016;134(6):651-657. [doi: [10.1001/jamaophthalmol.2016.0611](https://doi.org/10.1001/jamaophthalmol.2016.0611)] [Medline: [27077667](https://pubmed.ncbi.nlm.nih.gov/27077667/)]
47. Yu J, Huang Z, Zhuge Y, et al. MoE-Adapters++: toward more efficient continual learning of vision-language models via dynamic mixture-of-experts adapters. *IEEE Trans Pattern Anal Mach Intell*. Dec 2025;47(12):11912-11928. [doi: [10.1109/TPAMI.2025.3597942](https://doi.org/10.1109/TPAMI.2025.3597942)] [Medline: [40788794](https://pubmed.ncbi.nlm.nih.gov/40788794/)]
48. Liu J, Chen S, He X, et al. VALOR: Vision-audio-language omni-perception pretraining model and dataset. *IEEE Trans Pattern Anal Mach Intell*. 2025;47(2):708-724. [doi: [10.1109/TPAMI.2024.3479776](https://doi.org/10.1109/TPAMI.2024.3479776)]
49. Introducing openai o3 and o4-mini. OpenAI. URL: <https://openai.com/index/introducing-o3-and-o4-mini> [Accessed 2026-09-07]
50. Koga S, Du W. Challenges of integrating chatbot use in ophthalmology diagnostics. *JAMA Ophthalmol*. Sep 1, 2024;142(9):883-884. [doi: [10.1001/jamaophthalmol.2024.2303](https://doi.org/10.1001/jamaophthalmol.2024.2303)] [Medline: [38958958](https://pubmed.ncbi.nlm.nih.gov/38958958/)]
51. Cai LZ, Shaheen A, Jin A, et al. Performance of generative large language models on ophthalmology board-style questions. *Am J Ophthalmol*. Oct 2023;254:141-149. [doi: [10.1016/j.ajo.2023.05.024](https://doi.org/10.1016/j.ajo.2023.05.024)] [Medline: [37339728](https://pubmed.ncbi.nlm.nih.gov/37339728/)]
52. Li Z, Song D, Yang Z, et al. VisionUnite: a vision-language foundation model for ophthalmology enhanced with clinical knowledge. *IEEE Trans Pattern Anal Mach Intell*. 2025;47(12):11848-11862. [doi: [10.1109/TPAMI.2025.3598734](https://doi.org/10.1109/TPAMI.2025.3598734)]

53. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. *Nat Med*. Mar 2025;31(3):943-950. [doi: [10.1038/s41591-024-03423-7](https://doi.org/10.1038/s41591-024-03423-7)] [Medline: [39779926](https://pubmed.ncbi.nlm.nih.gov/39779926/)]
54. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. Feb 2023;2(2):e0000198. [doi: [10.1371/journal.pdig.0000198](https://doi.org/10.1371/journal.pdig.0000198)] [Medline: [36812645](https://pubmed.ncbi.nlm.nih.gov/36812645/)]
55. Jiang H, Xia S, Yang Y, et al. Transforming free-text radiology reports into structured reports using ChatGPT: a study on thyroid ultrasonography. *Eur J Radiol*. Jun 2024;175:111458. [doi: [10.1016/j.ejrad.2024.111458](https://doi.org/10.1016/j.ejrad.2024.111458)] [Medline: [38613868](https://pubmed.ncbi.nlm.nih.gov/38613868/)]
56. Huang M, Wang X, Zhou S, et al. Comparative performance of large language models for patient-initiated ophthalmology consultations. *Front Public Health*. 2025;13:1673045. [doi: [10.3389/fpubh.2025.1673045](https://doi.org/10.3389/fpubh.2025.1673045)] [Medline: [41059182](https://pubmed.ncbi.nlm.nih.gov/41059182/)]
57. Abbas Q, Jeong W, Lee SW. Explainable AI in clinical decision support systems: a meta-analysis of methods, applications, and usability challenges. *Healthcare (Basel)*. Aug 29, 2025;13(17):2154. [doi: [10.3390/healthcare13172154](https://doi.org/10.3390/healthcare13172154)] [Medline: [40941506](https://pubmed.ncbi.nlm.nih.gov/40941506/)]

Abbreviations

BW: birth weight
GA: gestational age
GEE: generalized estimating equation
GQS: Global Quality Score
ICC: intraclass correlation coefficient
ICROP3: the third edition of the International Classification of ROP
IRB: Institutional Review Board
OR: odds ratio
ROP: retinopathy of prematurity
SZEH : Shenzhen Eye Hospital
VEGF: vascular endothelial growth factor

Edited by Andrew Coristine; peer-reviewed by Masab Mansoor, Seung Won Lee, Stefan Lang; submitted 29.Oct.2025; final revised version received 16.Aug.2026; accepted 18.Aug.2026; published 07.Oct.2026

Please cite as:

Peng S, Zhao X, Wu Z, Yuan D, Duan N, Cui K, Yu Z, Yang W, Wei W, Chi W, Zhang G
Comparison of Two AI Chatbots for Diagnosis and Providing Treatment Suggestions in Retinopathy of Prematurity: Retrospective Study
J Med Internet Res 2026;28:e86726
URL: <https://www.jmir.org/2026/1/e86726>
doi: [10.2196/86726](https://doi.org/10.2196/86726)

© Shaojuan Peng, Xinyu Zhao, Zhenquan Wu, Duo Yuan, Na Duan, Kaixuan Cui, Zhen Yu, Weihua Yang, Wenbin Wei, Wei Chi, Guoming Zhang. Originally published in the Journal of Medical Internet Research (<https://www.jmir.org>), 07.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on <https://www.jmir.org/>, as well as this copyright and license information must be included.