

Comparative performance of DeepSeek-R1 and GPT-4o on standardized ophthalmic diagnostic and treatment tasks

Junxiu Zhang

Capital Medical University

Yao Ma

Affiliated Hospital of Shandong University of Traditional Chinese Medicine

Rong Zhang

Shanxi Medical University

Rina Su

Hulunbuir People's Hospital

Ke Ma

cdmake@163.com

Capital Medical University

Research Article

Keywords: Artificial Intelligence, Large Language Models, Diagnostic Accuracy, Clinical Decision Support, Medical Education

Posted Date: October 11th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10798645/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Purpose

To compare the diagnostic and treatment performance of DeepSeek-R1 and GPT-4o large language models (LLMs) in ophthalmology using standardized residency examination cases.

Design:

Cross-sectional comparative study.

Participants:

Thirty-three representative cases drawn from the Chinese Ophthalmology Residency Examination Database, covering 8 subspecialties.

Methods

Each case was processed by DeepSeek-R1 and GPT-4o with identical prompts to act as senior ophthalmologists. Three independent ophthalmologists conducted double-blind evaluations of each model's outputs. Accuracy was scored on a 10-point Likert scale and completeness on a 6-point Likert scale for diagnosis, differential diagnosis, and treatment. Mean scores were compared using paired statistical tests and two-way ANOVA.

Main Outcome Measures:

Accuracy and completeness scores across diagnosis, differential diagnosis, and treatment tasks.

Results

Across all cases, DeepSeek-R1 achieved significantly higher accuracy for diagnosis (8.04 vs 6.46, $P < 0.0001$), differential diagnosis (7.52 vs 5.50, $P < 0.0001$), and treatment (7.62 vs 6.65, $P = 0.002$) compared with GPT-4o. Completeness scores were also superior for DeepSeek-R1 in diagnosis (4.86 vs 3.69, $P < 0.0001$), differential diagnosis (4.44 vs 3.24, $P < 0.0001$), and treatment (4.61 vs 3.90, $P = 0.0001$). Subspecialty analyses revealed the largest advantage for DeepSeek-R1 in retinal diseases, glaucoma, strabismus & amblyopia, and optic nerve disorders.

Conclusions

In this time-stamped, text-only evaluation of standardized ophthalmology cases conducted between February and March 2025, DeepSeek-R1 achieved higher overall mean accuracy and completeness scores than GPT-4o. However, not all subgroup comparisons reached statistical significance, and the findings should be interpreted as preliminary given the limited sample size and text-only design.

Introduction

Large language models (LLMs) have demonstrated rapid progress in natural language understanding and generation, enabling their application in various professional domains, including healthcare¹. In medicine, recent studies have explored the potential of LLMs to assist in clinical decision-making, education, and patient communication^{2–4}. However, the reliability and accuracy of these models in specialized fields such as ophthalmology remain insufficiently characterized.

Ophthalmology presents unique diagnostic challenges due to the diversity of diseases^{5,6}, the reliance on multimodal information (e.g., fundus photography, optical coherence tomography, visual field testing), and the need for precise subspecialty knowledge. Accurate diagnosis and management planning require the integration of patient history, clinical signs, imaging results, and evidence-based guidelines—skills that typically develop through years of clinical training and experience. For this reason, assessing whether LLMs can approach or augment human expert performance is of both academic and practical significance⁷.

In China, the standardized ophthalmology residency training program incorporates comprehensive theoretical and practical assessments, making it a robust platform for evaluating AI-assisted diagnostic reasoning^{8,9}. The Chinese Ophthalmology Residency Examination Database provides well-structured case materials with standardized answer keys across multiple subspecialties, offering an opportunity for objective benchmarking.

Among contemporary LLMs, DeepSeek-R1, created by DeepSeek Artificial Intelligence Co., is a large-scale, open-source model containing approximately 671 billion parameters and launched in early 2025. The model was designed with a strong focus on advanced reasoning capability, achieved through extensive reinforcement-based post-training. In contrast, GPT-4o (OpenAI, May 2024) is a closed-source, general-purpose multimodal system intended for broad comprehension across text, images, and other inputs. Its general-purpose architecture reflects its wide training scope across diverse domains, which may lead to differences in performance on highly specialized medical tasks.

The present study aimed to compare DeepSeek-R1 and GPT-4o in their ability to generate accurate and complete diagnostic and treatment recommendations for standardized ophthalmic cases, using a double-blind evaluation by experienced ophthalmologists. We hypothesized that differences in model training and architecture would result in variable performance across subspecialties, with potential implications for clinical education and decision support in ophthalmology.

Methods

Study Design and Data Source

This was a cross-sectional comparative study conducted using cases from the Chinese Ophthalmology Residency Examination Database. The database contains standardized case materials designed for the national residency training assessment, each including a complete diagnostic description, standardized answer key, and coverage across multiple ophthalmic subspecialties. This study was approved by the Ethics Committee of Beijing Tongren Hospital, Capital Medical University. All methods were performed in accordance with relevant guidelines and regulations. The requirement for informed consent was waived by the ethics committee as all data were fully de-identified and derived from an educational examination database.

Case Selection

An initial case pool was screened according to the following inclusion criteria: (1) complete diagnostic elements, (2) availability of a standardized answer key, and (3) coverage across 8 major disease categories (retinal diseases, corneal diseases, glaucoma, strabismus & amblyopia, ocular tumors, trauma, optic nerve disorders, and orbital diseases), supplemented with single cases from uveitis, lacrimal disorders, dry eye syndrome, and eyelid disorders. A total of 33 cases met these criteria and were included for analysis (Fig. 1, Table 1). The selection emphasized common and clinically important entities such as retinal disorders and glaucoma, reflecting real-world prevalence. Although subspecialty representation was imbalanced, this mirrors actual case distributions in clinical practice and training curricula.

Model Processing

All model evaluations were conducted between February and March 2025. DeepSeek-R1 (version: January 2025 release) and GPT-4o (version: May 2024 release) were accessed through their respective official application programming interfaces (APIs), which enabled standardized text-based input and output retrieval. Model parameters were set to temperature = 0.1 to ensure deterministic outputs, with all safety filters maintained at default settings.

Each case was independently processed by DeepSeek-R1 and GPT-4o using a uniform prompt: *“Act as a senior ophthalmologist and provide the diagnosis, differential diagnoses, and treatment plan for the following case.”* The exact prompt template used for both models is provided in the Supplementary Materials (*Supplementary Appendix 1*) for reproducibility. Outputs were generated solely in text form, without multimodal content, thus reflecting reasoning ability rather than image interpretation.

Evaluation Procedure

To ensure consistent and objective scoring, all three board-certified ophthalmologist raters received a comprehensive package containing the 33 clinical vignettes alongside their corresponding standardized answer keys. These keys, derived from the Chinese Ophthalmology Residency Examination Database, served as the definitive reference for adjudicating both accuracy and completeness across all tasks. Raters were instructed to evaluate model outputs strictly against these pre-specified standards.

The evaluators scored each model's responses across three tasks—diagnosis, differential diagnosis, and treatment—using two dimensions: accuracy and completeness. All expert raters followed a standardized scoring guideline and conducted ratings independently to reduce subjective bias.

Accuracy (10-point scale) assessed the factual correctness and clinical validity of the response:

1 (Completely Inaccurate): Fundamentally incorrect or clinically unsafe.

10 (Perfectly Accurate): Fully correct, clinically sound, and aligned with the reference standard.

Completeness (6-point scale) evaluated the breadth and thoroughness of the response:

1 (Very Incomplete): Key elements of the task are missing.

6 (Fully Comprehensive): All required components are addressed without omission.

For each of the 33 cases, all three evaluators independently assigned accuracy and completeness scores for the three clinical tasks—diagnosis, differential diagnosis, and treatment. The scores from the three evaluators were averaged to generate a case-level mean for each task and dimension. These case-level means were then aggregated to derive the overall mean scores for each model. For Tables 2 and 3, the values shown in the “Total” row correspond to these overall mean scores across all 33 cases, rather than the arithmetic average of the subspecialty-specific means. The data collection system retained all individual evaluator scores, enabling both case-level and model-level performance analyses.

Statistical Analysis

Consistent with prior research, Likert-scale data were analyzed parametrically as continuous variables. Comparisons between DeepSeek-R1 and GPT-4o were performed using paired t-tests for each task. To assess potential differences among the three ophthalmologist raters in their scoring of accuracy and completeness, a two-way ANOVA was conducted with rater and model type as fixed factors. All statistical analyses were performed using GraphPad Prism (version 10.0.3, GraphPad Software, San Diego, CA, USA). Statistical significance was defined as $p < 0.05$.

Results

Case Selection and Characteristics

From the Chinese Ophthalmology Residency Examination Database, 33 representative cases were selected after applying the predefined inclusion criteria (Fig. 1). These covered 8 major subspecialties with balanced distribution across disease types: retinal diseases (n = 7), corneal diseases (n = 3), glaucoma (n = 4), strabismus & amblyopia (n = 2), ocular tumors (n = 2), trauma (n = 5), optic nerve disorders (n = 3), and orbital diseases (n = 3), in addition to one case each of uveitis, lacrimal disorders, dry eye syndrome, and eyelid disorders (Table 1). Subcategory details are provided in Table 1.

Table 1. Distribution of Cases by Disease Category

Subspecialty	Number of Cases	Subcategories and Case Details
Retinal Diseases	7	Diabetic Retinopathy
		DR Stage V (with vitreous hemorrhage)
		Central Retinal Vein Occlusion (left)
		Retinal Detachment
		Rhegmatogenous RD (with choroidal detachment)
		Pediatric Retinopathy
		Retinoblastoma (glaucoma stage)
Corneal Diseases	3	Retinopathy of Prematurity (Stage 4a+/2)
		Macular Disorders
		Central Serous Chorioretinopathy
Glaucoma	4	Age-Related Macular Degeneration (right wet/left dry)
		Infectious Keratitis
		Herpes Simplex Keratitis (dendritic)
		Acanthamoeba Keratitis (ring infiltrate)
Strabismus & Amblyopia	2	Traumatic Injury
		Corneal Perforation (suspected IOFB)
Ocular Tumors	2	Primary Angle-Closure
		Acute PACG (right chronic stage)
Trauma	5	Secondary Glaucoma
		Phacomorphic Glaucoma
		Lens Dislocation-induced Glaucoma
		Ischemic Optic Neuropathy-related Glaucoma
		Comitant Esotropia
Optic Nerve Disorders	3	Partially Accommodative ET (with V-pattern)
		Paralytic Strabismus
		Paralytic Esotropia (right lateral rectus palsy)
Orbital Diseases	3	Intraocular Tumor
		Retinoblastoma
		Orbital Tumor
		Cavernous Hemangioma (right orbit)
Uveitis	1	Orbital Fracture
		Blowout Fracture (left inferior wall)
		Optic Nerve Injury
		Traumatic Optic Neuropathy
		Intraocular Foreign Body
Lacrimal Disorders	1	IOFB with Endophthalmitis
		Lens Dislocation
Dry Eye Syndrome	1	Traumatic Lens Subluxation
		Penetrating Injury
Eyelid Disorders	1	Metallic Foreign Body Penetration
		Inflammatory
Total	33	Optic Neuritis (MS-associated)
		Ischemic
		Non-arteritic Anterior Ischemic Optic Neuropathy
		Traumatic
		Indirect Optic Nerve Injury
		Thyroid Eye Disease
		Thyroid-Associated Orbitopathy (restrictive strabismus)
		Tumor
		Orbital Cavernous Hemangioma
		Fracture
		Orbital Floor Fracture
		Posterior Uveitis
		Vogt-Koyanagi-Harada Syndrome
		Obstructive
		Nasolacrimal Duct Obstruction (right dacryocystitis)
		Age-Related
		Aqueous-Deficient Dry Eye
		Congenital
		Blepharophimosis Syndrome

Comparative Accuracy Across Tasks

Across all evaluated cases, DeepSeek-R1 showed higher mean accuracy scores than GPT-4o in diagnosis, differential diagnosis, and treatment. The most notable differences were observed in retinal diseases, glaucoma, optic-nerve disorders, and ocular tumors. Detailed task- and subspecialty-specific results are presented in Table 2.

Comparative Completeness Across Tasks

DeepSeek-R1 also showed higher completeness scores overall across diagnosis, differential diagnosis, and treatment tasks. The clearest advantages were observed in retinal diseases, glaucoma, optic-nerve disorders, and corneal diseases. Detailed results are presented in Table 3.

A graphical summary of model performance across subspecialties is provided in *Supplementary Fig. 1*, demonstrating DeepSeek-R1’s consistent superiority in both accuracy and completeness measures.

Inter-rater Evaluation of Accuracy and Completeness

To evaluate inter-rater consistency, three ophthalmologist evaluators independently scored the diagnosis, differential diagnosis, and treatment outputs generated by DeepSeek-R1 and GPT-4o. As shown in Fig. 2, the distributions of accuracy and completeness scores were generally consistent across evaluators, with DeepSeek-R1 receiving higher mean values in most task categories.

For diagnosis accuracy, all three evaluators assigned higher average scores to DeepSeek-R1 than to GPT-4o, with narrower variability across raters. A similar pattern was observed for differential-diagnosis accuracy, where DeepSeek-R1 achieved higher mean ratings from each evaluator. In terms of treatment accuracy, DeepSeek-R1’s scores were also numerically higher across evaluators, though the magnitude of difference varied among raters.

Regarding completeness, DeepSeek-R1 demonstrated higher mean values in diagnosis, differential diagnosis, and treatment tasks for each evaluator. The interquartile ranges of DeepSeek-R1’s scores were consistently smaller than those of GPT-4o, indicating more concentrated expert assessments. Overall, the trends across all three evaluators supported consistent scoring patterns favoring DeepSeek-R1 in both accuracy and completeness dimensions.

Table 2. Comparison of accuracy scores between DeepSeek-R1 and GPT-4o across diagnostic, differential-diagnostic, and treatment tasks

Subspecialty	Diagnosis Accuracy			Differential Diagnosis Accuracy			Treatment Accuracy		
	GPT-4o	Deepseek-R1	P Value	GPT-4o	Deepseek-R1	P Value	GPT-4o	Deepseek-R1	P Value
Retinal Diseases	6.83(4.83,9.17)	8.67(7.50,9.83)	0.03	6.17(4.50,6.83)	7.50(6.83,9.17)	0.01	7.83(4.83,8.83)	8.33(6.83,9.17)	0.16
Corneal Diseases	5.22(4.50,5.99)	6.67(4.17,9.17)	0.57	6.83(5.17,8.50)	8.67(8.17,9.17)	0.55	4.17(3.83,4.50)	4.45(4.07,4.83)	0.11
Glaucoma	6.50(5.17,7.50)	8.67(7.17,9.83)	0.08	4.67(3.83,5.88)	6.50(5.83,7.50)	0.03	6.17(4.84,8.50)	7.33(6.83,7.83)	0.32
Strabismus & Amblyopia	6.38(4.50,8.27)	9(8.50,9.50)	0.31	4.33(4.17,4.50)	7.67(6.50,8.83)	0.24	6.17(4.50,7.83)	9.17(8.17,9.17)	0.28
Ocular Tumors	7.33(6.83,7.83)	8.50(8.17,8.83)	0.09	5.50(4.50,6.50)	9.00(8.83,9.17)	0.20	8(7.83,8.17)	8.67(8.50,8.83)	<0.0001
Trauma	6.83(5.17,8.83)	6.50(4.50,9.50)	0.86	5.17(4.83,6.17)	5.83(5.50,8.17)	0.07	6.50(6.27,8.83)	6.83(4.17,9.50)	0.62
Optic Nerve Disorders	5.17(4.50,8.17)	9.50(9.17,9.50)	0.08	4.50(4.17,5.17)	7.83(7.17,8.50)	0.02	4.83(4.17,8.50)	8.83(7.17,9.50)	0.13
Orbital Diseases	7.17(5.83,7.83)	8.83(4.50,8.83)	>0.9999	6.50(5.50,6.50)	7.50(5.50,8.83)	0.24	8.17(6.27,8.17)	7.50(4.17,8.83)	0.47
Others	4.83(4.17,9.17)	7.33(4.83,9.83)	0.15	5.00(4.50,8.17)	7.33(5.50,9.17)	0.17	6.17(4.83,7.83)	8.17(6.17,9.50)	0.15
Total	6.17(5.17,7.83)	8.50(7.50,9.17)	0.00	5.17(4.83,6.17)	7.50(6.83,8.17)	<0.0001	6.50(5.17,7.83)	7.83(7.17,8.83)	0.01

Note: The “Total” row represents the overall mean score across all 33 cases after averaging the three evaluators’ scores for each case; it is not the simple arithmetic mean of the subspecialty-specific rows.

Table 3. Comparison of completeness scores between DeepSeek-R1 and GPT-4o across diagnostic, differential-diagnostic, and treatment tasks

Subspecialty	Diagnosis Completeness			Differential Diagnosis Completeness			Treatment Completeness		
	GPT-4o	Deepseek-R1	P Value	GPT-4o	Deepseek-R1	P Value	GPT-4o	Deepseek-R1	P Value
Retinal Diseases	3.50(2.83,5.50)	5.17(3.83,5.83)	0.01	3.17(2.50,3.83)	4.00(3.17,5.50)	0.01	4.50(3.17,5.17)	4.83(4.17,5.50)	0.16
Corneal Diseases	3.17(2.83,3.50)	4.33(3.17,5.50)	0.39	3.83(2.50,5.17)	5.67(5.50,5.88)	0.44	2.83(2.50,3.17)	3.83(3.50,4.17)	<0.0001
Glaucoma	3.00(2.83,3.50)	5.33(3.83,5.83)	0.04	2.83(2.17,3.17)	3.83(2.83,4.88)	0.05	3.17(2.50,4.17)	4.33(3.83,4.50)	0.13
Strabismus & Amblyopia	3.83(2.83,4.83)	5.50(5.50,5.50)	0.34	2.50(2.50,2.50)	4.83(4.17,5.50)	0.18	3.50(2.83,4.17)	4.83(4.50,5.17)	0.16
Ocular Tumors	4(3.50,4.50)	5.50(5.17,5.83)	0.07	3.17(2.50,3.83)	5.33(4.83,5.83)	0.05	4.67(4.17,5.17)	5.33(4.83,5.83)	<0.0001
Trauma	3.83(3.17,5.17)	4.17(3.50,5.17)	0.07	3.17(2.83,3.17)	4.50(3.50,5.17)	0.02	3.83(3.17,5.17)	3.83(2.83,5.50)	0.73
Optic Nerve Disorders	2.83(2.17,5.17)	5.50(5.17,5.83)	0.18	2.50(2.17,2.83)	4.83(4.17,5.17)	0.02	2.50(2.50,5.17)	5.50(4.17,5.50)	0.16
Orbital Diseases	3.50(3.17,4.50)	3.83(3.50,5.83)	0.18	2.83(2.83,3.83)	3.83(3.50,5.83)	0.09	5.17(3.17,5.17)	4.50(2.83,5.83)	0.81
Others	3.50(2.83,5.83)	4.33(3.50,5.83)	0.19	3.17(2.50,5.17)	4.50(3.17,5.17)	0.19	3.83(3.17,5.17)	5.00(3.17,5.83)	0.19
Total	3.50(2.83,3.83)	5.17(4.17,5.50)	<0.0001	3.17(2.50,3.17)	4.17(3.50,5.17)	<0.0001	4.17(3.17,4.50)	4.50(4.17,5.17)	<0.0001

Note: The “Total” row represents the overall mean score across all 33 cases after averaging the three evaluators’ scores for each case; it is not the simple arithmetic mean of the subspecialty-specific rows.

Discussion

This study compared the reasoning performance of two large language models, DeepSeek-R1 and GPT-4o, using standardized ophthalmology case materials evaluated by experienced clinicians. Across diagnosis, differential diagnosis, and therapeutic tasks, DeepSeek-R1 achieved higher mean scores for both accuracy and completeness than GPT-4o, indicating more precise and comprehensive outputs under structured clinical assessment conditions. Subspecialty-level analyses further demonstrated that the differences were most evident in retinal diseases, glaucoma, and neuro-ophthalmic disorders, where multi-step reasoning is often required.

The superior performance of DeepSeek-R1 in diagnosis accuracy was most evident in retinal diseases, glaucoma, strabismus & amblyopia, and optic nerve disorders. These conditions often require the integration of multimodal clinical data—such as fundus imaging, perimetry, and neuroimaging findings—into a coherent diagnostic synthesis^{15,16}. The higher scores for DeepSeek-R1 suggest a possible advantage in integrating and prioritizing relevant diagnosis cues, which may reflect differences in model training data or reasoning strategies. In contrast, for conditions such as corneal diseases, trauma, and orbital diseases^{17,18}, the performance gap between DeepSeek-R1 and GPT-4o was less pronounced. These entities often present with more distinct clinical signs or imaging findings, which may be easier for both models to interpret with similar accuracy.

The advantage of DeepSeek-R1 in completeness metrics, especially in diagnosis and differential diagnosis, indicates that it may be more consistent in providing comprehensive explanations and enumerating relevant differential possibilities. However, in treatment completeness, no statistically significant difference was detected at the subspecialty level, despite DeepSeek-R1 trending higher. This may reflect that evidence-based treatment pathways for many ophthalmic diseases are well-established, and both models have comparable access to such information.

Our findings align with and extend the growing body of evidence highlighting both the capabilities and limitations of large language models in clinical reasoning. Zhao *et al*¹. found that both GPT-4o and DeepSeek-R1 provided reasonable treatment guidance for osteoporotic vertebral compression fractures, although occasional deviations from established clinical guidelines underscored the need for cautious implementation. Chan *et al*². observed comparable diagnostic reasoning between DeepSeek-R1 and GPT-4 in complex clinicopathologic scenarios, suggesting that open-source models can achieve parity with proprietary systems when appropriately optimized. Menezes *et al*³. further demonstrated GPT-4o's robustness in processing multilingual medical documentation, highlighting its adaptability across healthcare contexts. Collectively, these studies reinforce that LLM performance depends strongly on task structure, data modality, and domain alignment—factors that similarly influenced model behavior in the present ophthalmic evaluation.

Nevertheless, the observed differences should be interpreted cautiously. First, the evaluation was based on text-only model outputs rather than multimodal input, and therefore did not capture image-based interpretation, which is critical in ophthalmology. Second, although the scoring framework employed standardized Likert anchors and blinded expert review, the number of evaluators and test cases remained limited, potentially constraining generalizability. The small sample size limits statistical power and increases the risk of both Type I and Type II errors. Future studies with larger samples are required to validate the observed trends. Inter-rater reliability was not formally quantified, which may introduce variability in subjective scoring. Future studies should formally assess inter-rater reliability using established metrics such as intraclass correlation coefficients. Third, both models occasionally produced partially correct but incomplete reasoning chains, highlighting the ongoing challenge of ensuring factual reliability in complex diagnostic contexts. Additionally, this study represents a time-

stamped comparison of specific model versions evaluated between February and March 2025. Given the rapid evolution of LLMs and vendor updates, these findings may not generalize to newer model generations or updated deployments. Finally, DeepSeek-R1's higher mean scores in this text-based evaluation do not imply clinical applicability.

Further validation in real-world decision-making scenarios, particularly those involving multimodal data such as fundus photographs, OCT scans, and visual field tests, is needed to better approximate actual diagnostic workflows. Prospective studies conducted in live clinical environments could clarify the role of LLMs in supporting real-time decision-making, including in resource-limited settings. Additional work exploring optimized prompting strategies aligned with local clinical guidelines and incorporating multilingual medical terminology may further improve performance. Comparative evaluations involving ophthalmology trainees and LLMs could also provide insight into how these models might complement clinical education and help standardize diagnostic quality across healthcare systems.

Conclusions

In this time-stamped, text-only evaluation conducted between February and March 2025, differences in performance were observed between DeepSeek-R1 and GPT-4o across standardized ophthalmic diagnostic, differential diagnostic, and treatment tasks. In this limited set of 33 cases, DeepSeek-R1 tended to achieve higher mean scores for accuracy and completeness, particularly in subspecialties requiring multi-step clinical reasoning.

Importantly, this study was designed as an exploratory, hypothesis-generating comparison rather than a definitive benchmark. Given the modest sample size, reliance on subjective expert scoring, and text-only evaluation framework, the present findings should be interpreted as preliminary and do not establish the superiority of any AI platform for clinical use. Further validation in larger, multimodal, and real-world clinical settings is required to determine whether observed performance differences are robust, generalizable, and clinically meaningful.

Declarations

Competing interests

The authors declare no competing interests.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethical Approval and accordance

Not applicable. This study utilizes publicly available, fully de-identified educational materials (provided in the supplementary files) and does not involve human research participants or live vertebrates, thereby exempting it from institutional ethical review.

Consent to participate

Not applicable. This study is based on publicly available educational datasets provided as supplementary material and does not involve human subjects.

Data Availability

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

Author Contribution

J.Z. design the study, wrote the main manuscript text and analysis the data. Y.M., R.Z., and S.R. answered all the case and provide their primary diagnosis and differential diagnosis of the 33 cases. K.M. supervise the study.

References

1. Fahrner LJ, Chen E, Topol E, Rajpurkar P. The generative era of medical AI. *Cell* Jul. 2025;10(14):3648–60. 10.1016/j.cell.2025.05.018.
2. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. *Jama* Oct. 2024;15. 10.1001/jama.2024.21700.
3. Johri S, Jeong J, Tran BA, et al. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nat Med* Jan. 2025;31(1):77–86. 10.1038/s41591-024-03328-5.
4. Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat Med* Feb. 2024;27. 10.1038/s41591-024-02855-5.
5. Waisberg E, Ong J, Kamran SA, et al. Generative artificial intelligence in ophthalmology. *Surv Ophthalmol* May. 2024;16. 10.1016/j.survophthal.2024.04.009.
6. Morse AR. Telemedicine in ophthalmology: promise and pitfalls. *Ophthalmology* Apr. 2014;121(4):809–11. 10.1016/j.opthta.2013.10.033.
7. Ting DSW, Peng L, Varadarajan AV, et al. Deep learning in ophthalmology: The technical and clinical considerations. *Prog Retin Eye Res*. 2019;09:72:100759. 10.1016/j.preteyeres.2019.04.003.
8. Wang YE, Zhang C, Chen AC, Pineles S, Hou J. Current Status of Ophthalmology Residency Training in China: The Experience From Well-Structured Centers Around the Country. *Asia Pac J Ophthalmol (Phila)*. Jul-Aug. 2020;9(4):369–373. 10.1097/apo.0000000000000284

9. Ming S, Yao X, Guo X, et al. Performance of ChatGPT in Ophthalmic Registration and Clinical Diagnosis: Cross-Sectional Study. *J Med Internet Res* Nov. 2024;14:26:e60226. 10.2196/60226.
10. Wang L, Mao Y, Wang L, Sun Y, Song J, Zhang Y. Suitability of GPT-4o as an Evaluator of Cardiopulmonary Resuscitation Skills Examinations. *Resuscitation* Sep. 2024;27:110404. 10.1016/j.resuscitation.2024.110404.
11. Sun Z, Yang J, Zhang N, Wu H, Li C. GPT-4o is more like a real person: potentials in surgical oncology. *Int J Surg* Jul. 2024;2. 10.1097/js9.0000000000001898.
12. Sandmann S, Hegselmann S, Fujarski M et al. Benchmark evaluation of DeepSeek-R1 large language models in clinical decision-making. *Nat Med*. Apr 23. 2025; 10.1038/s41591-025-03727-2
13. Fink MA, Bischoff A, Fink CA, et al. Potential of ChatGPT and GPT-4 for Data Mining of Free-Text CT Reports on Lung Cancer. *Radiology*. 2023;09(3):e231362. 10.1148/radiol.231362.
14. Zhang J, Ma Y, Zhang R, et al. A comparative study of GPT-4o and human ophthalmologists in glaucoma diagnosis. *Sci Rep* Dec. 2024;5(1):30385. 10.1038/s41598-024-80917-x.
15. Cen LP, Ji J, Lin JW, et al. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nat Commun*. 2021;08 10(1):4828. 10.1038/s41467-021-25138-w.
16. Koh JEW, Ng EYK, Bhandary SV, Hagiwara Y, Laude A, Acharya UR. Automated retinal health diagnosis using pyramid histogram of visual words and Fisher vector techniques. *Comput Biol Med*. Jan 2018;1:92:204–9. 10.1016/j.compbimed.2017.11.019.
17. Alzubaidi R, Sharif MS, Qahwaji R, Ipson S, Brahma A. In vivo confocal microscopic corneal images in health and disease with an emphasis on extracting features and visual signatures for corneal diseases: a review study. *Br J Ophthalmol* Jan. 2016;100(1):41–55. 10.1136/bjophthalmol-2015-306934.
18. Guillon-Rolf R, Hau S, Larkin DFP. Clinical and confocal imaging findings in congenital corneal anaesthesia. *Br J Ophthalmol*. 2021;11(11):1491–6. 10.1136/bjophthalmol-2020-316672.
19. Yao JJ, Aggarwal M, Lopez RD, Namdari S. Large Language Models in Orthopaedics: Definitions, Uses, and Limitations. *J Bone Joint Surg Am* Aug. 2024;7(15):1411–8. 10.2106/jbjs.23.01417.
20. Marshall R, Xu H, Dalvin LA et al. Accuracy and Completeness of Large Language Models About Antibody-Drug Conjugates and Associated Ocular Adverse Effects. *Cornea*. Aug 7. 2024; 10.1097/ico.0000000000003664
21. Shi B, Chen L, Pang S, et al. Large Language Models and Artificial Neural Networks for Assessing 1-Year Mortality in Patients With Myocardial Infarction: Analysis From the Medical Information Mart for Intensive Care IV (MIMIC-IV) Database. *J Med Internet Res* May. 2025;12:27:e67253. 10.2196/67253.
22. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. *Med Educ* Apr. 2024;19. 10.1111/medu.15402.
23. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions. *JMIR Med Educ* Jun. 2023;1:9:e48291.

Figures

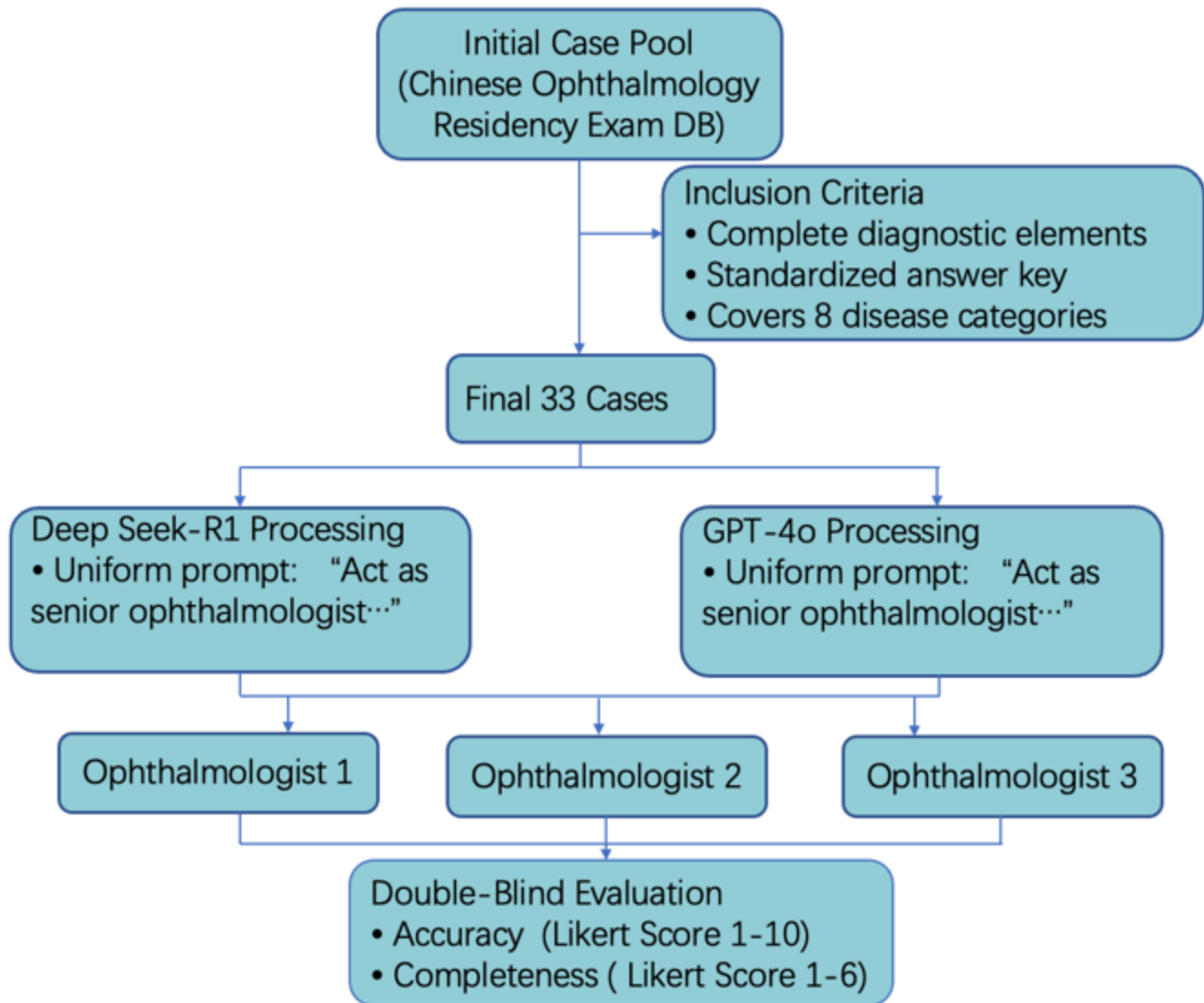


Figure 1

Overview of study design and evaluation workflow

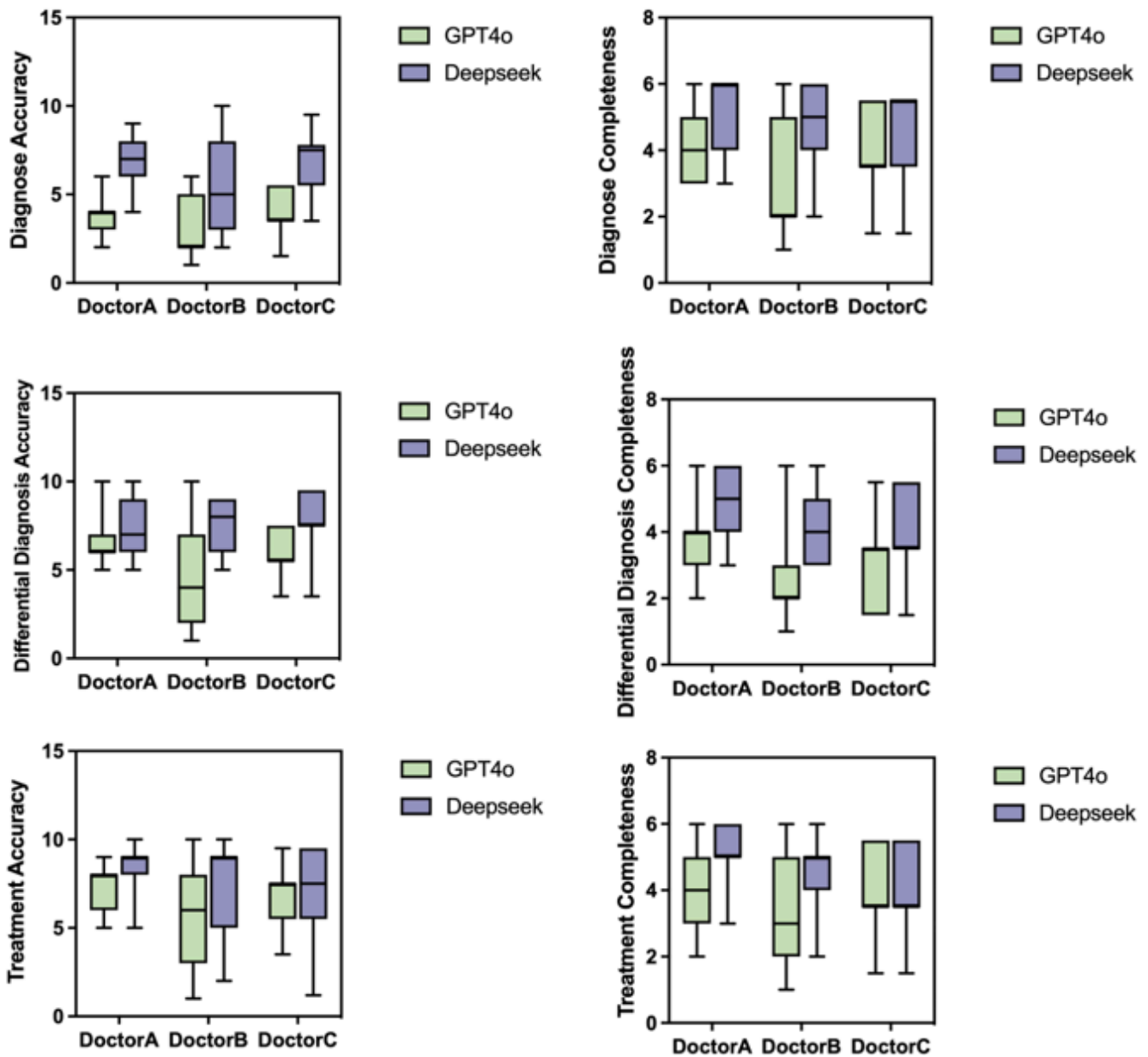


Figure 2

Box Plot Comparison of Model Performance Across Raters

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SM..docx](#)
- [SM127cases.pdf](#)