Arquivos Brasileiros de
Oftalmologia
PUBLICAÇÃO OFICIAL DO
CONSELHO BRASILEIRO
DE OFTALMOLOGIA

ORIGINAL ARTICLE

Assessing a large language model for glaucoma knowledge: ChatGPT-5 versus residents

Mauro Gobira[1,2], Rodrigo Moreira[1], Flavio J. L. Galhardo Carvalho Filho[1], Kevin Waquim Pessoa Carvalho[1], Francisco N. Murta[1], Lucas Antônio Avelar Carvalho[1], Rubens Belfort Jr.[1,2], Ivan M. Tavares[2]

1.Ophthalmology Department, Instituto da Visão, São Paulo, SP, Brazil.

2. Department of Ophthalmology and Visual Sciences, Escola Paulista de Medicina, Universidade Federal de São Paulo, São Paulo, SP, Brazil.

ABSTRACT

Purpose: To assess the performance of a contemporary large language model (ChatGPT-5) against ophthalmology residents on a standardized set of glaucoma multiple-choice questions. **Methods:** We conducted a cross-sectional comparative study with 189 text-only glaucoma multiple-choice questions from the Cybersight question bank. ChatGPT-5 was tested under standardized conditions, with each item placed in a new chat and limited to letter-only outputs. Six ophthalmology residents from a Brazilian training program (two Postgraduate Year 1, two Postgraduate Year 2, and two Postgraduate Year 3) answered the same questions under supervision. Accuracy was calculated using the official key. McNemar's exact test was used to compare items between ChatGPT-5 and residents, and matched odds ratios and 95% confidence intervals (95% CIs) were calculated using the Haldane–Anscombe correction. **Results:** ChatGPT-5 received 164 of 189 correct responses (86.8%; 95% CI, 81.2–90.9). Residents' overall accuracy was 62.9% (713/1,134; 95% CI, 60.0–65.6). The top-performing resident earned 76.7%. ChatGPT-5 outperformed all residents in head-to-head comparisons, with odds ratios ranging from 1.84 (95% CI, 1.10–3.08) to 13.15 (95% CI, 5.93–29.20), all p≤0.023. ChatGPT-5 correctly answered 17/189 items (9.0%), but fewer than half of residents were correct ("large language model-only wins"), whereas residents were more successful on items that ChatGPT-5 overlooked. **Conclusions:** ChatGPT-5 outperformed ophthalmology residents on text-based glaucoma multiple-choice questions, indicating its potential as a subspecialty education and assessment tool. Generalizability is limited by the single question bank, text-only items, a small resident cohort, and the evaluation of one large language model version at a single time point. Before incorporating these findings into clinical decision-making, larger, multimodal, and longitudinal studies are required.

KEYWORDS: Glaucoma; Artificial intelligence; Large language models; Education, medical; Medical staff, hospital

http://dx.doi.org/10.5935/0004-2749.2025-0283

Submitted for publication:
October 6, 2025

Accepted for publication:
February 10, 2026

Funding:
This study received no specific financial support.

Disclosure of potential conflicts of interest:
The authors declare no potential conflicts of interest.

Corresponding author:
Mauro Gobira
E-mail: maurogobira19@gmail.com

Approved by the following research ethics committee:
Centro Universitário FMABC (CAAE: 85336324.7.0000.0082).

Data Availability Statement:
The datasets generated and/or analyzed during the current study are included in the manuscript.

Edited by
Editor-in-Chief: Newton Kara-Júnior
Associate Editor: Rodrigo P. C. Lira

INTRODUCTION

Glaucoma is a progressive optic neuropathy characterized by retinal ganglion cell loss and optic nerve damage, which is frequently associated with increased intraocular pressure (IOP)[1]. It is the world's leading cause of irreversible blindness, with many cases undiagnosed[2]. Early detection and intervention are critical to avoiding permanent vision loss[3]. Despite advances in diagnostics and treatment, there are still gaps in timely diagnosis and management[4]. Integrating advanced artificial intelligence (AI) tools, such as large language models (LLMs), has the potential to assist clinicians in knowledge synthesis and decision support for glaucoma treatment.

LLMs are AI systems that use extensive text corpora to predict and generate human-like language[5]. They use transformer architectures and self-attention to encode contextual meaning[6]. ChatGPT is a well-known LLM developed by OpenAI that progresses from GPT-3 to GPT-4 and beyond, gradually improving fluency, contextual understanding, and reasoning abilities. Its evolution reflects its increasing ability to assist in drafting medical documents, simulating clinical scenarios, and answering biomedical questions[7].

LLMs have shown promise in educational and clinical settings, but they are still limited by well-documented challenges such as the generation of inaccurate or fabricated information, variable performance across medical domains, and bias in the absence of domain-specific fine-tuning[8-10]. Furthermore, most evaluation studies report methodological heterogeneity and often only moderate accuracy in specialized medical tasks, raising concerns about their clinical applicability[8-10].

There is a scarcity of empirical data comparing LLMs to human trainees in ophthalmology, particularly in glaucoma. To close this gap, the current study compares the performance of an advanced LLM (ChatGPT-5) to ophthalmology residents on a standardized set of glaucoma multiple-choice questions (MCQs). This method allows for a direct assessment of accuracy and error patterns, yielding objective evidence about the potential and limitations of LLMs in subspecialty ophthalmic education and assessment.

METHODS

Ethics statement

The study followed the Declaration of Helsinki guidelines and was approved by the Centro Universitário FMABC Research Ethics Committee (CAAE: 33842220.7.2001.0082; Approval No. 7.677.884). All procedures followed the Brazilian National Health Council's Resolution 466/2012. All resident participants gave informed consent to the use of de-identified performance data.

Study design and setting

We conducted a cross-sectional comparative performance study of a contemporary LLM ("ChatGPT-5," OpenAI) versus ophthalmology residents on glaucoma MCQs. The corpus included 189 publicly available glaucoma MCQs from the Cybersight question bank (Orbis International) at https://cybersight.org/. Items were written by glaucoma subspecialists, were text-only (no images), and included a single best answer key. On August 8–9, 2025, all LLM runs were carried out using the official OpenAI web interface (OpenAI, San Francisco, California) and Google Chrome (Google LLC, Mountain View, California) on a 2020 MacBook Air (Apple Inc., Cupertino, California).

LLM evaluation protocol

To prevent context carryover, each MCQ was entered into a separate, empty chat. Every item was preceded by a fixed instruction.

"You're an ophthalmology specialist. Read the MCQs below and select the single best answer. Please provide only the letter that corresponds to your choice, with no further explanation."

The model produced only letters, which were recorded in real time in a scoring spreadsheet. No follow-up questions, chain-of-thought requests, or clarification exchanges were used.

Human comparators

The same 189 MCQs were administered under supervision to six ophthalmology residents from a single Brazilian training program (two Postgraduate Year 1 [PGY-1], two Postgraduate Year 2 [PGY-2], and two Postgraduate Year 3 [PGY-3] equivalents). Testing was done individually and proctored, with no access to external resources. Residents did not have prior access to the specific subset of items or the answer key. From August 8 to 20, all residents answered the questions in a single sitting. Responses were recorded electronically as letters and compared to the key. Resident identifiers were changed to Resident 01–06 for all analyses and reporting.

Outcomes

The primary outcome was ChatGPT-5's accuracy on the same 189 items for all residents. Secondary outcomes included paired head-to-head performance (LLM vs each resident) using item-matched discordant pairs as well as an item-level comparison between ChatGPT-5 and the per-item mean resident accuracy (defined as the proportion of residents correctly answering that item). We summarized "LLM-only wins," which were defined as items answered correctly by ChatGPT-5 while <50% of residents were correct.

Statistical analysis

Resident performance was summarized as an aggregated proportion of all resident responses (correct/total across participants), with item-level descriptive summaries based on the per-item mean proportion of residents answering correctly. Accuracy proportions are reported using 95% Wilson confidence intervals. For item-matched comparisons between ChatGPT-5 and each resident, we used McNemar's test with two-sided exact binomial p-values calculated on discordant pairs. The matched odds ratio (OR=*b*/*c*) was used to summarize the effect size, with *b* representing items correct by the LLM and incorrect by the resident, and *c* representing items incorrect by the LLM and correct by the resident, respectively. To stabilize estimates when *b* or *c* equaled (or approached) zero, we used the Haldane–Anscombe correction (adding 0.5 to each discordant cell) and calculated 95% confidence intervals (95% CIs) on the log scale. We also described and visualized the per-item difference (LLM correctness indicator minus mean resident correctness). All tests were two-sided, with α=0.05. Analyses and figures were created using Python 3.11 (pandas, NumPy, SciPy for exact binomial tests, and Matplotlib).

RESULTS

ChatGPT-5 answered 164 of 189 items (86.8%; 95% CI, 81.2–90.9). The residents' overall accuracy was 62.9% (713/1.134; 95% CI, 60.0–65.6), resulting in an absolute difference of +23.9 percentage points in favor of ChatGPT-5. Compared with the best individual resident (76.7%), the difference was +10.1 percentage points (Table 1 and Figure 1).

McNemar tests revealed that ChatGPT-5 outperformed each resident, with discordant pairs consistently favoring the LLM; matched odds ratios ranged from 1.84 (95% CI, 1.10–3.08) versus Resident 04 to 13.15 (5.93–29.20) versus Resident 06 with all two-sided exact p-values ≤0.023 (Table 2).

Table 1. Participant accuracy on 189 glaucoma MCQs

Participant	Correct	Total	Accuracy (%)	95% CI (Wilson)
ChatGPT-5	164	189	86.77	81.20–90.90
Resident 01	132	189	69.84	63.00–75.90
Resident 02	134	189	70.90	64.10–76.90
Resident 03	123	189	65.08	58.00–71.50
Resident 04	145	189	76.72	70.20–82.20
Resident 05	94	189	49.74	42.70–56.80
Resident 06	85	189	44.97	38.10–52.10
Mean of Residents	713	1134	62.87	60.00–65.60

95% CI= 95% confidence interval; MCQs= multiple-choice questions.

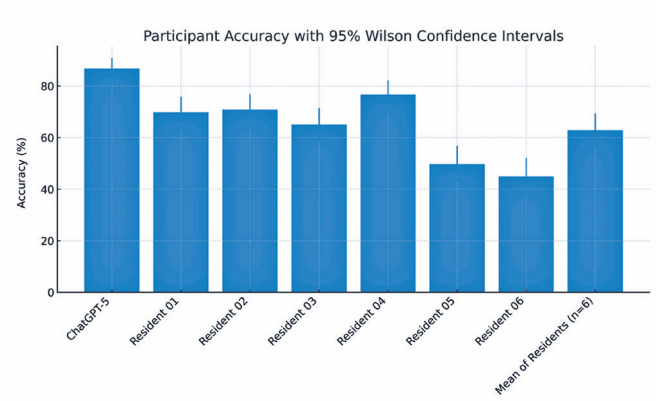

Figure 1. Overall accuracy of ChatGPT-5 and residents on 189 glaucoma MCQs.

Item-level agreement patterns also revealed a systematic advantage for ChatGPT-5 over the mean resident performance, with superiority on 164/189 (86.8%) items and inferiority on 25/189 (13.2%); ties were absent with six resident responses per-item. Notably, on 17/189 (9.0%) questions, ChatGPT-5 was correct, while fewer than half of residents answered correctly (mean resident accuracy <50%), indicating "LLM-only wins" on items that were relatively difficult for residents. However, among the 25 items missed by ChatGPT-5, residents achieved ≥50% correctness on 23 of them, suggesting that LLM errors tended to occur on items most residents solved (Figure 2). Collectively, these results, summarized in table 1 (participant accuracies with Wilson 95% CIs), table 2 (pairwise McNemar comparisons with Haldane–Anscombe corrected odds ratios and exact two-sided p-values), and visualized in figure 1 (accuracy with 95% CIs) and Figure 2 (distribution of per-item performance differences), demonstrate consistent superiority of ChatGPT-5 over residents on this set of glaucoma MCQs.

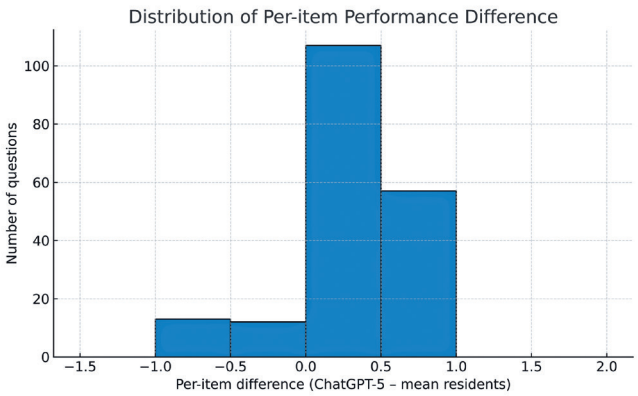

Figure 2. Histogram of per-item performance difference (ChatGPT-5 minus mean resident accuracy).

Table 2. Paired head-to-head comparisons versus ChatGPT-5

Comparison	Discordant: LLM correct/ Resident wrong	Discordant: LLM wrong/ Resident correct	Total discordant pairs	Odds ratio (LLM vs. resident)	95% CI for OR
ChatGPT-5 vs. Resident 01	45	13	58	3.37	1.84–6.19
ChatGPT-5 vs. Resident 02	48	18	66	2.62	1.53–4.48
ChatGPT-5 vs. Resident 03	58	17	75	3.34	1.96–5.70
ChatGPT-5 vs. Resident 04	41	21	63	1.84	1.10–3.08
ChatGPT-5 vs. Resident 05	87	17	104	5.0	2.99–8.35
ChatGPT-5 vs. Resident 06	85	6	91	13.15	5.93–29.20

95% CI= 95% confidence interval; LLM= large language model; OR= odds ratio.

DISCUSSION

To the best of our knowledge, this is the first medical study to evaluate ChatGPT-5 in ophthalmology and glaucoma, with results indicating a clear accuracy advantage over residents on validated glaucoma MCQs. Benchmarking emerging LLMs is timely given their rapid adoption in educational and clinical workflows; recent peer-reviewed surveys show that many health professionals already use these tools for tasks such as document drafting, patient education, and decision support[11]. Such early adoption emphasizes the importance of rigorous, domain-specific performance assessments. In this context, ChatGPT-5's strong results on Cybersight's subspecialty glaucoma items, which were written by glaucoma specialists and cover pathophysiology, epidemiology, diagnosis, and treatment, indicate potential relevance for subspecialty education and assessment.

Prior research found significant variability in medical QA performance across models and settings. GPT-4 outperformed GPT-3.5 with an accuracy of approximately 81% in US Medical Licensing Examination-style evaluations[7]. In Brazil's National Examination for Medical Degree Revalidation, GPT-4 scored 87.7%, outperforming the majority of human candidates[12]. As previously stated, measured accuracy varies depending on the model (e.g., GPT-3.5 vs. GPT-4), item difficulty, targeted audience (generalist vs. subspecialist), and whether questions are patient- or clinician-facing[7].

In ophthalmology, the highest LLM accuracies have typically been observed in specialist MCQ contexts. On the American Academy of Ophthalmology Basic and Clinical Science Course self-assessment (>1,000 items), GPT-4 scored 82.4%, outperforming human examinees (75.7%) and GPT-3.5 (65.9%)[13]. In the Ophthalmic Knowledge Assessment Program (OKAP) practice questions, GPT-4 achieved 81% versus 57% for GPT-3.5 across 180 text-only items[14]. Earlier evaluations on BCSC and OphthoQuestions reported 59.4% and 49.2% with ChatGPT Plus (legacy: 55.8% and 42.7%), indicating a reliance dependence on both the question bank and model version[15]. A broader analysis encompassing USMLE items and the Ophthalmology Board Written Qualifying Examination (WQE) discovered 70% overall for GPT-4, including 62.9% on OB-WQE when images were excluded, highlighting the influence of exam level and visual content[16]. More recently, on Taiwan's National Medical Licensing Examination ophthalmology set, GPT-4o scored 92.9% compared to GPT-4's 69.2%, indicating that newer models will continue to improve[17].

Focusing on glaucoma, the reported LLM performance ranges from patient FAQs to case-based diagnostics. Using guideline-anchored expert scoring, GPT-4 produced 88.7% completely correct responses to commonly asked glaucoma questions; only 3.8% were partially misleading, and none were entirely incorrect[18]. In a clinic-facing FAQ study, ChatGPT scored 97% accuracy versus 77% for Google on 30 curated questions graded by three glaucoma specialists[19]. With 24 curated patient questions to GPT-3.5, 70.8% of responses were rated appropriate, and allowing self-correction increased full-score responses from 30.6% to 57.1%[20]. In case-based tasks, ChatGPT provided the correct provisional diagnosis in 72.7% (8/11) of glaucoma cases, matching or outperforming senior residents[21]. Using textual case reports from OHTS (1,585 subjects; 3,170 eyes), GPT-4 achieved 87% accuracy (AUC, 0.76; specificity 90%; sensitivity 61%) for glaucoma classification[22]. In addition to these findings, a study comparing glaucoma and retina scenarios discovered that a GPT-4 chatbot outperformed glaucoma specialists and matched retina specialists in terms of diagnostic accuracy and completeness[23].

Our study builds on previous research by conducting paired, item-matched comparisons of ChatGPT-5 and six ophthalmology residents on 189 text-only subspecialty glaucoma MCQs. ChatGPT-5 outperformed all residents in overall accuracy and in all head-to-head McNemar comparisons, indicating a consistent advantage on this standardized corpus. These findings support the use of LLMs as adjuncts for study, teaching, and item review in educational setting, rather than as a substitute for supervised clinical training.

Beyond educational assessment, artificial intelligence has shown significant promise in glaucoma screening and increasing access to ophthalmologic care[24,25]. Deep learning–based systems have demonstrated high diagnostic accuracy for detecting glaucomatous optic neuropathy using fundus photographs, indicating their potential use in large-scale screening and referral prioritization[24]. When integrated into teleophthalmology workflows, such systems may allow for earlier identification of at-risk individuals, optimize specialist referral pathways, and help reduce barriers to eye care in underserved populations[25]. Although LLMs are not intended for image interpretation or autonomous diagnosis, their complementary use in guideline-based risk stratification, clinical decision support, and patient-facing education has the potential to increase the efficiency and reach of glaucoma screening programs, especially when combined with validated imaging-based artificial intelligence tools and existing screening recommendations[3].

This work has limitations. The evaluation was limited to text-only items and did not consider image-dependent tasks (such as fundus photographs, OCT/OCTA, visual fields, gonioscopy, and slit-lamp videos). The resident cohort, which includes six trainees at various levels within a single institution, is still relatively small and may not reflect broader training environments. We used a single question bank, tested only one model/version, and did not assess longitudinal variability, calibration, or downstream clinical outcomes. As a result, findings should not be extrapolated to diagnostic or clinical decision-making scenarios.

Future research should use multimodal designs (fundus photos, OCT/OCTA, perimetry, gonioscopy, and slit-lamp video), include multi-center human comparators across training levels (students, residents by PGY, fellows, and subspecialists), evaluate multiple model versions with repeated runs/seeds, and test cross-bank generalization (Cybersight, BCSC, OphthoQuestions, and OKAP). Prospective, preregistered protocols with transparent reporting would enhance comparability and aid in defining educational impact, error modes, and the conditions under which LLM assistance is reliable and safe for ophthalmic applications.

In conclusion, ChatGPT-5 outperformed ophthalmology residents on text-based glaucoma MCQs, demonstrating that modern LLMs can match or exceed trainee-level knowledge in this subspecialty domain. These findings support LLMs' potential role as adjunct tools for study, teaching, and educational assessment, while emphasizing that their use should supplement, not replace, structured training and clinical supervision.

AUTHORS' CONTRIBUTIONS:

Significant contribution to conception and design: Mauro Gobira, Rodrigo Moreira, Flavio J. L. Galhardo Carvalho Filho, Kevin Waquim Pessoa Carvalho, Francisco N. Murta, Lucas Antônio Avelar Carvalho.

Data Acquisition: Mauro Gobira, Rodrigo Moreira, Flavio J. L. Galhardo Carvalho Filho, Kevin Waquim Pessoa Carvalho, Francisco N. Murta, Lucas Antônio Avelar Carvalho.

Data Analysis and interpretation: Mauro Gobira, Rodrigo Moreira, Flavio J. L. Galhardo Carvalho Filho, Kevin Waquim Pessoa Carvalho, Francisco N. Murta, Lucas Antônio Avelar Carvalho, Rubens Belfort Jr.

Manuscript Drafting: Mauro Gobira, Rodrigo Moreira, Flavio J. L. Galhardo Carvalho Filho, Kevin Waquim Pessoa Carvalho, Francisco N. Murta, Lucas Antônio Avelar Carvalho.

Significant intellectual content revision of the manuscript: Mauro Gobira, Rodrigo Moreira, Flavio J. L. Galhardo Carvalho Filho, Kevin Waquim Pessoa Carvalho, Francisco N. Murta, Lucas Antônio Avelar Carvalho, Rubens Belfort Jr., Ivan M. Tavares.

Final approval of the submitted manuscript: Mauro Gobira, Rodrigo Moreira, Flavio J. L. Galhardo Carvalho Filho, Kevin Waquim Pessoa Carvalho, Francisco N. Murta, Lucas Antônio Avelar Carvalho, Rubens Belfort Jr., Ivan M. Tavares.

Statistical analysis: Mauro Gobira.

Obtaining funding: not applicable.

Supervision of Administrative, technical, or material support: Mauro Gobira.

Research group leadership: Mauro Gobira, Rubens Belfort Jr., Ivan M. Tavares.

REFERENCES

1. Weinreb RN, Aung T, Medeiros FA. The pathophysiology and treatment of glaucoma: a review. JAMA. 2014;311(18):1901–11.
2. Tham YC, Li X, Wong TY, Quigley HA, Aung T, Cheng CY. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. Ophthalmology. 2014;121(11):2081–90.
3. Chou R, Selph S, Blazina I, Bougatsos C, Jungbauer R, Fu R, et al. Screening for glaucoma in adults: updated evidence report and systematic review for the US Preventive Services Task Force. JAMA. 2022;327(20):1998–2012.

4. Jonas JB, Aung T, Bourne RR, Bron AM, Ritch R, Panda-Jonas S. Glaucoma. Lancet. 2017;390(10108):2183–93.
5. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33. Available from:https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html
6. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:5998–6008.
7. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLoS Digit Health. 2023;2(2):e0000198.
8. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv. 2023;55(12):1–38.
9. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health. 2023;5(3):e107–8.
10. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023;9:e45312. Erratum in: JMIR Med Educ. 2024 Feb 27;10:e57594.
11. Kisvarday S, Yan A, Yarahuan J, Kats DJ, Ray M, Kim E, et al. ChatGPT use among pediatric health care providers: cross-sectional survey study. JMIR Form Res. 2024;8:e56797.
12. Gobira M, Nakayama LF, Moreira R, Andrade E, Regatieri CVS, Belfort R Jr. Performance of ChatGPT-4 in answering questions from the Brazilian National Examination for Medical Degree Revalidation. Rev Assoc Med Bras. 2023;69(10):e20230848.
13. Taloni A, Borselli M, Scarsi V, Rossi C, Coco G, Scorcia V, Giannaccare G. Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology. Sci Rep. 2023;13(1):18562.
14. Teebagy S, Colwell L, Wood E, Yaghy A, Faustina M. Improved performance of ChatGPT-4 on the OKAP examination: a comparative study with ChatGPT-3.5. J Acad Ophthalmol. 2023;15(2):e184–7.
15. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of Its Successes and Shortcomings. Ophthalmol Sci. 2023;3(4):100324.
16. Haddad F, Saade JS. Performance of ChatGPT on ophthalmology-related questions across various examination levels: observational study. JMIR Med Educ. 2024;10:e50842.
17. Lin CW, Chen CY, Wu P, Chen CL, Lai CH. Assessing GPT-4o and GPT-4 in answering and explaining ophthalmology examination questions from Taiwan's medical licensing test. Taiwan J Ophthalmol. 2025; 15(4):647–54.
18. Kerci SG, Sahan B. An analysis of ChatGPT-4 to respond to glaucoma-related questions. J Glaucoma. 2024;33(7):486–9.
19. Cohen SA, Fisher AC, Xu BY, Song BJ. Comparing the accuracy and readability of glaucoma-related question responses and educational materials by Google and ChatGPT. J Curr Glaucoma Pract. 2024;18(3):110–6.
20. Tan DN, Tham YC, Koh V, Loon SC, Aquino MC, Lun K, et al. Evaluating Chatbot responses to patient questions in the field of glaucoma. Front Med (Lausanne). 2024;11:1359073.
21. Delsoz M, Raja H, Madadi Y, Tang AA, Wirostko BM, Kahook MY, et al. The use of ChatGPT to assist in diagnosing glaucoma based on clinical case reports. Ophthalmol Ther. 2023;12(6):3121–32.
22. Raja H, Huang X, Delsoz M, Madadi Y, Poursoroush A, Munawar A, et al. Diagnosing glaucoma based on the Ocular Hypertension Treatment Study dataset using Chat Generative Pre-trained Transformer as a large language model. Ophthalmol Sci. 2024;5(1):100599.
23. Huang AS, Hirabayashi KE, Barna L, Parikh D, Pasquale LR. Assessment of a large language model's responses to questions and cases about glaucoma and retina management. JAMA Ophthalmol. 2024;142(4):371–5.
24. Li Z, He Y, Keel S, Meng W, Chang RT, He M. Efficacy of a Deep Learning System for Detecting Glaucomatous Optic Neuropathy Based on Color Fundus Photographs. Ophthalmology. 2018;125(8):1199–206.
25. Ting DS, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol. 2019;103(2):167–75.