

Performance of Generative Artificial Intelligence and Medical Students on Undergraduate Ophthalmology Examinations: A Cross-Sectional Comparative Study

Sofia Machado

u020023@ulssjoao.min-saude.pt

Unidade de Saúde Local de São João

Ricardo Roleiro

Unidade de Saúde Local de São João

Amândio Rocha-Sousa

Unidade de Saúde Local de São João

Manuel Falcão

Unidade de Saúde Local de São João

Sónia Torres-Costa

Unidade de Saúde Local de São João

Research Article

Keywords: Artificial intelligence, Large language models, Ophthalmology, Medical education, Medical students

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10923838/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

Background

Generative artificial intelligence (AI) systems are increasingly used in medical education, but their performance in undergraduate ophthalmology remains incompletely characterized. This study aimed to compare five systems with fourth-year medical students on the same ophthalmology examination questions.

Methods

This cross-sectional study included 89 multiple-choice questions from two ophthalmology examinations administered at the University of Porto in 2025, including 12 image-based and 77 were text-only items. The item-level results were available for 229 students. GPT-4.1 Auto, GPT-5.2 Auto, Gemini 3 Flash, Grok 4.1 Auto, and Claude Sonnet 4.5 were tested via a standardized prompt. Performance by question format was compared using Fisher's exact and Mann–Whitney U tests; agreement was assessed using kappa statistics and the McNemar test, and predictors of correct AI responses using mixed-effects logistic regression.

Results

The mean item-level student accuracy was 73.1%. The overall accuracy was 95.5% for Gemini 3 Flash, 93.3% for Claude Sonnet 4.5, 92.1% for GPT-5.2 Auto, 84.3% for Grok 4.1 Auto, and 79.8% for GPT-4.1 Auto. Four of the five systems performed significantly worse on image-based than on text-only questions, whereas student performance was similar across formats. GPT-4.1 Auto had the greatest agreement with the student-majority pattern despite having the lowest AI accuracy. Questions answered correctly by a larger proportion of students were also more likely to be answered correctly by AI systems ($p = 0.002$).

Conclusions

Generative AI systems achieved high accuracy on undergraduate ophthalmology questions, particularly text-based items, but most performed worse on image-based questions. These systems may support learning, but examination performance should not be interpreted as evidence of clinical competence.

1. BACKGROUND

Generative artificial intelligence (AI), particularly large language models (LLMs), has rapidly been incorporated into education to explain complex concepts, create educational materials and support examination preparation (1–3). Although more recent LLMs usually perform at least as well as the

students do on text-based exam questions, they do not consistently perform better than students do across all settings (4–6). Particularly in medical education, the evidence is still heterogeneous, as the results vary considerably according to the model and version evaluated, medical specialty, language, prompting strategy, question format and level of reasoning required (4, 5, 7–11). Studies evaluating these platforms in ophthalmology and comparing them with residents and board-examination candidates have also reported variable results (8, 12, 13).

Far fewer studies have examined how these tools compare with those of medical students. The aim of undergraduate ophthalmology education is to test foundational knowledge and clinical reasoning under a defined curriculum rather than a specialty board examination (14–16). A recent evaluation of an undergraduate-level United Kingdom ophthalmology prize examination suggested that contemporary LLMs could rank within the top performance of human candidates (14). Nevertheless, little is known about the types of questions that these models answer incorrectly and whether their pattern of errors resembles that observed among students, particularly as question difficulty increases.

In our study, we evaluated the performance of four widely used generative AI platforms (ChatGPT, Gemini, Claude, and Grok) on the same ophthalmology examinations completed by fourth-year medical students (MS4) in an undergraduate medical curriculum. We aimed to compare AI accuracy with student performance and to determine whether performance differed according to question format and human-derived item difficulty. We also examined the extent to which AI systems and students shared similar patterns of correct and incorrect responses.

2. METHODS

2.1 Study design

This cross-sectional study compared the performance of five commercially available generative AI systems with that of MS4 on ophthalmology multiple-choice questions. The study protocol was approved by the Ethics Committee of the Faculty of Medicine of the University of Porto/RISE Health (approval no. 374/CEFMUP-RISEHealth/2025; December 16, 2025). Only anonymized student performance data were analyzed, and no individual student could be identified. This study was conducted and is reported in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines for cross-sectional studies (17).

All questions were obtained from two complete ophthalmology examinations (first and second semesters in 2025, Faculty of Medicine of the University of Porto). Each item consisted of a single-best-answer multiple-choice question with five response options. After excluding one duplicate item, 89 unique questions remained (44 from the first semester, 45 from the second). Each item was classified by semester, subspecialty, and format: questions requiring interpretation of a clinical image or diagnostic exam (slit-lamp photography, perimetry, retinography, or fluorescein angiography) were classified as image-based; all others were classified as text-only.

Anonymized item-level student results were obtained from the original examinations. The first- and second-semester examinations were completed by 119 and 110 students, respectively. The official answer key was established before model testing and was independently reviewed by five ophthalmologists, with disagreements resolved by consensus. Item-level student performance was defined as the proportion of students selecting the correct answer; a question was considered "correct" for the student group when this percentage exceeded 50% (sensitivity analyses used 75% and 90% thresholds). Item difficulty was entered into the statistical models as a continuous facility index (0–1).

2.2 Testing procedure

Five commercially available generative AI systems from four platforms were evaluated: GPT-4.1 Auto and GPT-5.2 Auto (OpenAI, San Francisco, CA, USA), Gemini 3 Flash (Google, Mountain View, CA, USA), Grok 4.1 Auto (xAI, Palo Alto, CA, USA), and Claude Sonnet 4.5 (Anthropic, San Francisco, CA, USA). All were accessed in private-browsing sessions using newly created accounts, with no previous conversation history, saved memory or custom instructions on December 22, 2025.

Each conversation was opened with the same standardized prompt (Portuguese, with English translation): "Please answer the following questions from the ophthalmology examination of MS4. For each question, select the single best answer." Questions and options were entered manually without paraphrasing; image-based items included the original image. Questions were submitted sequentially in two consecutive boxes because of platform character limits, with no follow-up prompts, regeneration, or accuracy feedback, and no web search, plug-ins, or external references enabled. Responses selecting a nonexistent option or with no identifiable answer, were classified as "incorrect". There were no missing data: all 89 items were answered by all five AI systems, and item-level response data were available for all students who completed each examination.

2.3 Outcomes

The primary outcome was overall accuracy (proportion of correct answers per AI system). The secondary outcomes include accuracy by question format and subspecialty, agreement between each AI's item-level correctness pattern and the student group, inter-AI agreement, and the association between AI accuracy and the student item facility.

2.4 Statistical analysis

Accuracy was defined as the proportion of correct responses for each AI system. Ninety-five percent confidence intervals for AI accuracy were calculated via the Wilson method; for students whose performance was expressed as the mean of item-level proportions, confidence intervals were estimated via percentile bootstrapping (2 000 resamples). Each AI system's total score was ranked against individual student scores to determine its percentile position within the student distribution for each examination.

Performance on image-based versus text-only questions was compared via Fisher's exact test for each AI system and the Mann–Whitney U test for students. Pairwise item-level comparisons between each AI

system and the student-majority response used the exact binomial McNemar test, which evaluates whether the number of items answered correctly by only the AI system differs from the number answered correctly by only the student majority.

Agreement between each AI system's item-level correctness pattern and the student-majority response (binarized at a threshold of more than 50% correct) was assessed with Cohen's kappa; a sensitivity analysis repeated the calculation at thresholds of 75% and 90%. Agreement among the five AI systems was assessed with Fleiss's kappa. Both coefficients were interpreted per Landis and Koch (1977): less than 0.00, no agreement; 0.00–0.20, slight; 0.21–0.40, fair; 0.41–0.60, moderate; 0.61–0.80, substantial; and 0.81–1.00, almost perfect. Bootstrap confidence intervals (2 000 resamples, percentile method) were calculated for kappa estimates.

To examine predictors of correct AI responses while accounting for the clustering of responses within questions, a mixed-effects logistic regression (GLMM) was fitted with item-level correctness (0/1) as the outcome and the question as a random intercept. Fixed effects included AI system, question format (image-based vs. text-only), their interaction, item facility (proportion of students answering correctly, entered as a continuous covariate), and semester. Three nested models of increasing complexity were compared via the Akaike information criterion and the likelihood ratio test. The BOBYQA algorithm was used for optimization. Cohen's kappa was also calculated separately for each ophthalmology subspecialty with at least five questions.

All analyses were stratified by question format (overall, image-based, text-only). Statistical significance was set at $\alpha = 0.05$. Analyses were performed in R version 4.5.3 (R Foundation for Statistical Computing), using the irr package for Cohen's and Fleiss' kappa, boot for bootstrap confidence intervals, and lme4 for mixed-effects regression.

3. RESULTS

3.1 Question characteristics and student performance

A total of 89 unique multiple-choice questions were included: 44 from the first-semester and 45 from the second semester examinations. Twelve questions (13.5%) were image-based and 77 (86.5%) were text-only. The most frequently represented subspecialties were general ophthalmology ($n = 18$), medical retina ($n = 15$), glaucoma ($n = 13$), neuro-ophthalmology ($n = 9$), strabismus ($n = 7$), cataract and refractive surgery ($n = 6$), and pediatric ophthalmology ($n = 5$).

The mean item-level student accuracy was 73.1% (median, 79.8%; range, 8.4%–100%). Student performance was similar between the first and second semesters (72.2% vs 73.9%; $p = 0.739$) and between image-based and text-only questions (72.8% vs 73.1%; $p = 0.904$). Using the prespecified student-majority threshold, more than half of students answered 73 (82.0%) questions correctly (Table 1).

3.2 Overall AI performance and ranking within the student distributions

Grok 4.1 Auto and GPT-5.2 Auto each selected once an answer option that was not provided in different questions once. These responses were classified as incorrect.

The item-level accuracy of all five AI systems was numerically higher than the mean student item facility. Gemini 3 Flash achieved the highest accuracy of 95.5%, followed by Claude Sonnet 4.5, GPT-5.2 Auto, Grok 4.1 Auto and GPT-4.1 Auto with accuracies of 93.3%, 92.1%, 84.3% and 79.8%, respectively (Table 2).

Three systems consistently ranked near the top of the student distributions: Gemini 3 Flash (100th and 99.1st percentiles), Claude Sonnet 4.5 (99.6th and 97.3rd), and GPT-5.2 Auto (93.3rd and 99.1st) in the first and second examinations, respectively. Grok 4.1 Auto ranked at the 83.2nd and 85.9th percentiles, while GPT-4.1 Auto showed the greatest variation, dropping from the 83.2nd to the 49.5th percentile.

3.3 Performance according to question format

Performance was lower for image-based than for text-only questions: Gemini 3 Flash (75.0% vs 98.7%; $p = 0.007$), GPT-5.2 Auto (66.7% vs 96.1%; $p = 0.006$), Grok 4.1 Auto (50.0% vs 89.6%; $p = 0.001$), and Claude Sonnet 4.5 (75.0% vs 96.1%; $p = 0.030$). No significant difference according to question format was observed for GPT-4.1 Auto (75.0% vs 80.5%; $p = 0.702$). (Table 2)

3.4 Comparison and agreement with student performance

In paired item-level comparisons, Gemini 3 Flash and Claude Sonnet 4.5 answered correctly more frequently items missed by the student majority than the opposite pattern, showing significant asymmetry in favor of the AI system ($p = 0.012$ and $p = 0.031$, respectively). No significant asymmetry was observed for the remaining systems, although GPT-5.2 Auto showed a similar nonsignificant trend.

GPT-4.1 Auto showed the highest agreement with the student-majority correctness pattern. Agreement was moderate for the complete set of questions ($\kappa = 0.42$; $p < 0.001$) and for text-only questions ($\kappa = 0.45$; $p < 0.001$), but was lower for image-based questions ($\kappa = 0.25$; $p = 0.371$).

Agreement was minimal for the remaining systems: GPT-5.2 Auto, $\kappa = 0.07$; Claude Sonnet 4.5, $\kappa = 0.09$; Grok 4.1 Auto, $\kappa = 0.04$; and Gemini 3 Flash, $\kappa = -0.08$. For text-only questions, GPT-5.2 Auto showed slight agreement with the student-majority response ($\kappa = 0.18$; $p = 0.026$). No statistically significant agreement was found for the other systems.

A sensitivity analysis using alternative student-performance thresholds showed similar findings. GPT-4.1 Auto consistently showed the highest agreement at student-performance thresholds of more than 50%, 75%, and 90%, although the magnitude of agreement decreased as the threshold increased ($\kappa = 0.42$,

0.36, and 0.14, respectively). The kappa values for the other systems remained close to zero across the three thresholds (Table 3).

3.5 Agreement among AI systems

Overall agreement in correctness status among the five AI systems was fair (Fleiss $\kappa = 0.38$; $p < 0.001$). Agreement was moderate for image-based questions ($\kappa = 0.50$; $p < 0.001$) and fair for text-only questions ($\kappa = 0.25$; $p < 0.001$).

3.6 Mixed-effects analysis

In the mixed-effects analysis, a greater proportion of students answering an item correctly was associated with greater odds of a correct AI response ($p = 0.002$), indicating that AI systems were more likely to answer correctly the same questions that were answered correctly by a larger proportion of students. Examination semester was not associated with AI performance ($p = 0.673$). Using Claude Sonnet 4.5 and text-only questions as the reference categories, GPT-4.1 Auto had lower odds of a correct response than Claude Sonnet 4.5 (OR, 0.04; $p < 0.001$), while image-based questions were associated with lower odds of a correct Claude response (OR, 0.01; $p = 0.019$). No model-by-question-format interaction reached statistical significance; however, the limited number of image-based items resulted in imprecise interaction estimates.

3.7 Performance according to ophthalmology subspecialty

Performance varied across subspecialties (Fig. 1). Gemini 3 Flash and Claude Sonnet 4.5 achieved 100% accuracy in several categories, including medical retina, glaucoma, neuro-ophthalmology, strabismus, and cataract and refractive surgery. General ophthalmology was among the lowest-performing categories for several systems, with accuracies of 61.1% for Grok 4.1 Auto and 72.2% for GPT-4.1 Auto. In pediatric ophthalmology, student accuracy was 82.0%, compared with 80.0% for four of the five systems and 100% for Gemini 3 Flash. Detailed subspecialty-level results and agreement analyses are provided in Supplementary Table 1.

4. DISCUSSION

In this study, generative AI models performed well on a MS4 ophthalmology exam. Studies using ophthalmology residency or board examinations have reported performance approaching that of human candidates (8, 18–21), but these findings are not equivalent to comparisons with undergraduate students, where examinations primarily evaluate foundational knowledge and the application of core concepts within a defined curriculum. Therefore, performance relative to residents or board candidates cannot be directly extrapolated to undergraduate education. Also, direct comparisons between AI systems and undergraduate medical students remain relatively limited, as summarized in Table 4. To our knowledge, this is one of the first studies to compare these systems with undergraduate medical students in the context of ophthalmology education in Europe. Importantly, the models and students

were evaluated on the same questions from actual university examinations rather than on a publicly available question bank.

In our study, Gemini 3 Flash, Claude Sonnet 4.5, and GPT-5.2 Auto achieved overall accuracies above 90%, while GPT-4.1 Auto and Grok 4.1 Auto had lower accuracy than the other systems, although their accuracies remained numerically above mean student performance (84.3% and 79.8%, respectively, vs 73.1%). These results are greater than those reported in many earlier ophthalmology examination studies (Table 4). Part of this difference may reflect improvements in more recent model generations. Nevertheless, comparisons across studies remain difficult because the source and difficulty of the questions, examination language, model version, prompt structure, presence of images, and testing interface differ substantially between published evaluations (16, 22) (Table 4).

The question format was an important determinant of AI performance in our study. Four of the five systems performed significantly worse on image-based questions, as already reported in other studies, whereas student performance remained similar across formats (5, 10, 23–26). The performance of image-based questions is particularly relevant in ophthalmology, where diagnosis and management often depend on the interpretation of visual findings, rather than on textual knowledge alone. Nonetheless, the magnitude of the difference observed in our study should be interpreted cautiously, because only 12 questions included images.

An important consideration when interpreting the high AI accuracy observed in our study is the difference in preparation between students and AI systems. The questions in the present study were written by faculty members involved in teaching the course, and students had attended the corresponding lectures and received the associated materials to get prepared. By contrast, the AI systems were given only the question stem, response options, and associated image (when applicable), without access to course materials or targeted preparation. Despite this difference, the AI systems achieved high accuracy on locally developed, Portuguese-language ophthalmology questions.

Our mixed-effects analysis further showed that AI accuracy was associated with student item-level performance: questions answered correctly by a larger proportion of students were also more likely to be answered correctly by the AI systems, consistent with previous reports that question difficulty influences LLM performance (27). This does not imply that students and AI systems used the same reasoning processes or exhibited identical error patterns (18, 28). Indeed, accuracy and agreement captured different aspects of performance: while GPT-4.1 Auto had lower overall accuracy but showed the greatest similarity to the student-majority pattern of responses, Gemini 3 Flash achieved the highest accuracy and correctly answered several questions missed by most students, resulting in a low chance-corrected agreement.

Our results support the potential use of these systems as supplementary learning tools from an educational perspective. They may assist students in reviewing content, discussing alternative answer choices or identifying areas requiring further study (1). However, their performance also raises concerns about the continued use of unsupervised, text-only multiple-choice assessments. Exams that

incorporate supervised testing, image interpretation, oral justification, simulated patients and practical clinical skills may provide a more meaningful evaluation of ophthalmology training, ensuring that it reflects the visual nature of ophthalmology.

Several limitations should also be acknowledged. The study was conducted at a single medical school and included a relatively small number of questions, particularly in terms of image-based items and questions within individual subspecialties. Each system answered each question only once, preventing assessment of response variability or reproducibility. Although newly created accounts, and private-browsing sessions minimized contextual carryover and user-specific personalization, the systems were accessed through proprietary consumer interfaces whose underlying configurations may change over time. The findings therefore reflect the configurations available on December 22, 2025 and may not be fully reproducible following subsequent platform updates. Since then, newer versions of Gemini and ChatGPT, with enhanced reasoning capabilities, have been released and may offer performance improvements (29, 30).

5. CONCLUSIONS

Contemporary generative AI systems performed well on real undergraduate ophthalmology examinations, with three of the five systems achieving overall accuracies above 90%. Questions that were easier for students were also more likely to be answered correctly by the AI systems, although the pattern of correct and incorrect responses differed between models. Performance was lower on image-based questions for most systems, but this finding should be interpreted cautiously given the limited number of image-based items. These results suggest that generative AI may have a useful place in ophthalmology education, while also reinforcing that success on written examinations is not equivalent to clinical competence.

Abbreviations

AI

Artificial intelligence

CI

Confidence interval

GLMM

Generalized linear mixed-effects model

GPT

Generative pre-trained transformer

IQR

Interquartile range

LLM

Large language model

MCQ

Multiple-choice question

MS4

Fourth-year medical student

OR

Odds ratio

STROBE

Strengthening the Reporting of Observational Studies in Epidemiology

Declarations

Ethics approval and consent to participate:

The study protocol was approved by the Ethics Committee of the Faculty of Medicine of the University of Porto/RISE Health (approval no. 374/CEFMUP-RISEHealth/2025; December 16, 2025). The requirement for informed consent was waived by the Ethics Committee because only anonymized examination data were analyzed.

Consent for publication:

Not applicable

Funding:

None

Author Contribution

S.M., M.F., and S.T.C. conceived the study. All authors contributed to the study design. S.M., R.R., M.F. and A.R.S. contributed to data acquisition and curation. A.R.S., M.F. and S.T.C. supervised the study. S.M. and R.R. drafted the original manuscript. All authors contributed to the interpretation of the results, critically revised the manuscript, reviewed and approved the final version and agreed to be accountable for their contributions and the integrity of the work.

Acknowledgements:

Not applicable

Availability of data and materials:

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests: The authors declare that they have no competing interests

References

1. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions. *JMIR Med Educ*. 2023;9:e48291.
2. Xu X, Chen Y, Miao J. Opportunities, challenges, and future directions of large language models, including ChatGPT in medical education: a systematic scoping review. *J Educ Eval Health Prof*. 2024;21:6.
3. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. *Med Educ*. 2024;58(11):1276–85.
4. Qiu Y, Liu C. Capable exam-taker and question-generator: the dual role of generative AI in medical education assessment. *Global Med Educ*. 2025;2(1):135–48.
5. Al-Thani SN, Anjum S, Bhutta ZA, Bashir S, Majeed MA, Khan AS, et al. Comparative performance of ChatGPT, Gemini, and final-year emergency medicine clerkship students in answering multiple-choice questions: implications for the use of AI in medical education. *Int J Emerg Med*. 2025;18(1):146.
6. Hersh W, Fultz Hollis K. Results and implications for generative AI in a large introductory biomedical and health informatics course. *NPJ Digit Med*. 2024;7(1):247.
7. Moshirfar M, Altaf AW, Stoakes IM, Tuttle JJ, Hoopes PC. Artificial Intelligence in Ophthalmology: A Comparative Analysis of GPT-3.5, GPT-4, and Human Expertise in Answering StatPearls Questions. *Cureus*. 2023;15(6):e40822.
8. Balci AS, Yazar Z, Ozturk BT, Altan C. Performance of Chatgpt in ophthalmology exam; human versus AI. *Int Ophthalmol*. 2024;44(1):413.
9. Rodrigues Alessi M, Gomes HA, Oliveira G, Lopes de Castro M, Grenteski F, Miyashiro L, et al. Comparative Performance of Medical Students, ChatGPT-3.5 and ChatGPT-4.0 in Answering Questions From a Brazilian National Medical Exam: Cross-Sectional Questionnaire Study. *JMIR AI*. 2025;4:e66552.
10. Sarantopoulos A, Pana ZD, Larentzakis A, Kondylis S, Maina A, Ziogas N, et al. ChatGPT-4.0 and Medical Students: A Recognition-Gated Comparative Evaluation on Image-Based Medical Examinations. *J Med Educ Curric Dev*. 2026;13:23821205261449387.
11. Friederichs H, Friederichs WJ, Marz M. ChatGPT in medical school: how successful is AI in progress testing? *Med Educ Online*. 2023;28(1):2220920.
12. 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* (2017).

2023;15(2):e184–e7.

13. Mihalache A, Popovic MM, Muni RH. Performance of an Artificial Intelligence Chatbot in Ophthalmic Knowledge Assessment. *JAMA Ophthalmol.* 2023;141(6):589–97.
14. Akinniranye O, Akinniranye O. Performance of Large Language Models and Top-Decile Doctors on an Undergraduate Ophthalmology Examination. *Cureus.* 2025;17(11):e97687.
15. Arora A, Parmar UPS, Anjali, Mathur P, Kishore K, Singh U, et al. When Large Language Models (LLMs) walk into a Bachelor's in optometry examination: Comparing the performances of LLMs and bachelor of optometry students. *Indian J Ophthalmol.* 2026;74(4):516–22.
16. Zhang Z, Zhang H, Pan Z, Bi Z, Wan Y, Song X, et al. Evaluating Large Language Models in Ophthalmology: Systematic Review. *J Med Internet Res.* 2025;27:e76947.
17. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *PLoS Med.* 2007;4(10):e296.
18. Thirunavukarasu AJ, Mahmood S, Malem A, Foster WP, Sanghera R, Hassan R, et al. Large language models approach expert-level clinical knowledge and reasoning in ophthalmology: A head-to-head cross-sectional study. *PLOS Digit Health.* 2024;3(4):e0000341.
19. Bahir D, Zur O, Attal L, Nujeidat Z, Knaanie A, Pikkell J, et al. Gemini AI vs. ChatGPT: A comprehensive examination alongside ophthalmology residents in medical knowledge. *Graefes Arch Clin Exp Ophthalmol.* 2025;263(2):527–36.
20. Ming S, Yao X, Guo X, Guo Q, Xie K, Chen D, et al. Performance of ChatGPT in Ophthalmic Registration and Clinical Diagnosis: Cross-Sectional Study. *J Med Internet Res.* 2024;26:e60226.
21. Rocha H, Chong YJ, Thirunavukarasu AJ, Wong YL, Wong SW, Chang YH, et al. Performance of Foundation Models vs Physicians in Textual and Multimodal Ophthalmological Questions. *JAMA Ophthalmol.* 2026;144(1):5–13.
22. Wickramaratna S, Pirwani M-A. Evaluation of Large Language Model performance on a high-stakes undergraduate Ophthalmology examination: a comparative study. 2026.
23. Wei J, Wang X, Huang M, Xu Y, Yang W. Evaluating the Performance of ChatGPT on Board-Style Examination Questions in Ophthalmology: A Meta-Analysis. *J Med Syst.* 2025;49(1):94.
24. Xu P, Chen X, Zhao Z, Shi D. Unveiling the clinical incapacities: a benchmarking study of GPT-4V(ision) for ophthalmic multimodal image analysis. *Br J Ophthalmol.* 2024;108(10):1384–9.
25. Mikhail D, Milad D, Antaki F, Milad J, Farah A, Khairy T et al. Multimodal Performance of GPT-4 in Complex Ophthalmology Cases. *J Pers Med.* 2025;15(4).
26. Cai LZ, Shaheen A, Jin A, Fukui R, Yi JS, Yannuzzi N, et al. Performance of Generative Large Language Models on Ophthalmology Board-Style Questions. *Am J Ophthalmol.* 2023;254:141–9.
27. 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.

28. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. 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.
29. OpenAI, Introducing OAI. 2026 [updated 5 March 2026. Available from: <https://openai.com/index/introducing-gpt-5-4/>
30. Google, DeepMind. Gemini 3.1 Pro: A smarter model for your most complex tasks [Internet]. Google DeepMind; 2026 [updated 19 February 2026. Available from: <https://deepmind.google/blog/gemini-3-1-pro-a-smarter-model-for-your-most-complex-tasks/>
31. Patel J, Robinson P, Illing E, Anthony B. Is ChatGPT 3.5 smarter than Otolaryngology trainees? A comparison study of board style exam questions. *PLoS ONE.* 2024;19(9):e0306233.
32. Strong E, DiGiammarino A, Weng Y, Kumar A, Hosamani P, Hom J, et al. Chatbot vs Medical Student Performance on Free-Response Clinical Reasoning Examinations. *JAMA Intern Med.* 2023;183(9):1028–30.
33. Meyer A, Riese J, Streichert T. Comparison of the Performance of GPT-3.5 and GPT-4 With That of Medical Students on the Written German Medical Licensing Examination: Observational Study. *JMIR Med Educ.* 2024;10:e50965.
34. Bharatha A, Ojeh N, Fazle Rabbi AM, Campbell MH, Krishnamurthy K, Layne-Yarde RNA, et al. Comparing the Performance of ChatGPT-4 and Medical Students on MCQs at Varied Levels of Bloom's Taxonomy. *Adv Med Educ Pract.* 2024;15:393–400.
35. Ros-Arlanzon P, Gutarra-Avila R, Arrarte-Esteban V, Bertomeu-Gonzalez V, Hernandez-Blasco L, Masia M, et al. When AI models take the exam: large language models vs medical students on multiple-choice course exams. *Med Educ Online.* 2025;30(1):2592430.
36. Salvador R, Vas D, Oleaga L, Matute-Gonzalez M, Castillo-Fortuno A, Setoain X, et al. Comparing ChatGPT and medical student performance in a real image-based Radiology and Applied Physics in Medicine exam. *Radiologia (Engl Ed).* 2025;67(4):101638.

Tables

Table 1
Characteristics of the examination questions and student performance

Characteristic	Semester 1	Semester 2	Overall
Students, n	119	110	229
Examination items, n	44	45	89
Text-only items, n (%)	38 (86.3)	39 (86.7)	77 (86.5)
Image-based items, n (%)	6 (13.7)	6 (13.3)	12 (13.5)
Student examination score, median (IQR), %	72.73 (65.91–79.55)	77.78 (71.11–82.22)	-
Score range, %	45.45–90.91	44.44–97.78	-
Students achieving ≥ 90%, n (%)	1 (0.8)	9 (8.2)	-
Student item facility, median (range), %	79.0 (8.4–100)	80 (9.1–99.1)	79.8 (8.4–100.0)
Items correctly answered by > 50% of students, n (%)	37 (84.1)	36 (80.0)	73 (82.0)
Items correctly answered by ≤ 50% of students, n (%)	7 (15.9)	9 (20.0)	16 (18.0)

Table 2
Overall and format-specific performance of generative AI systems and medical students

Evaluator	Overall	Text-only	Image-based	p value
Gemini 3 Flash	85/89 (95.5%) [89.0–98.2]	76/77 (98.7%) [93.0–99.8]	9/12 (75.0%) [46.8–91.1]	0.007 ^a
Claude Sonnet 4.5	83/89 (93.3%) [86.1–96.9]	74/77 (96.1%) [89.2–98.7]	9/12 (75.0%) [46.8–91.1]	0.030 ^a
GPT-5.2 Auto	82/89 (92.1%) [84.6–96.1]	74/77 (96.1%) [89.2–98.7]	8/12 (66.7%) [39.1–86.2]	0.006 ^a
Grok 4.1 Auto	75/89 (84.3%) [75.3–90.4]	69/77 (89.6%) [80.8–94.6]	6/12 (50.0%) [25.4–74.6]	0.001 ^a
GPT-4.1 Auto	71/89 (79.8%) [70.3–86.8]	62/77 (80.5%) [70.3–87.8]	9/12 (75.0%) [46.8–91.1]	0.702 ^a
Students	73.1% [68.0–77.9]	73.1% [67.5–78.3]	72.8% [60.3–84.9]	0.904^b

Values for AI systems are presented as n/N (%) [95% CI]. Student performance represents the mean item-level proportion of correct responses and is therefore presented as a percentage [95% CI]. P values compare performance between text-only and image-based questions. ^a Fisher's exact test; ^b Mann–Whitney U test.

Table 3

Agreement between each AI system and the student-majority, overall and according to question format.

AI system	Overall κ	p value	Image-based κ	p value	Text-only κ	p value
GPT-4.1	0.42	< 0.001	0.25	0.371	0.45	< 0.001
GPT-5.2 Auto	0.07	0.447	-0.29	0.273	0.18	0.026
Claude Sonnet 4.5	0.09	0.310	0.25	0.371	0.06	0.488
Grok 4.1 Auto	0.04	0.714	0.00	1.000	0.10	0.336
Gemini 3 Flash	-0.08	0.338	-0.25	0.371	-0.02	0.635
The student-majority classification considered an item correct when more than 50% of students selected the correct answer. Agreement was assessed using Cohen's kappa.						

Table 4

Selected studies directly comparing generative AI systems with undergraduate medical students on examination-style assessments.

Study	Question set	AI models	Human comparator	Findings
<i>Patel et al., 2024</i>	30 board-style Otolaryngology–Head & Neck Surgery multiple-choice questions (text only)	GPT-3.5	Medical students 1, 2, 3 and 4	ChatGPT-3.5 (54.7%) significantly outperformed MS1 (28.8%), MS2 (31.4%), and MS3 (36.0%) students. It performed similarly to MS4 students (37.8%). (31)
<i>Strong et al., 2023</i>	14 open-ended clinical reasoning cases (text only)	GPT-3.5 (prior study), GPT-4 (follow-up study)	First- and second-year medical students and physician assistant students	GPT-4 scored a mean 4.2 percentage points higher than students. Its passing rate was 93% (vs. 85% of students). GPT-4 also substantially outperformed GPT-3.5 (43%). (32)
<i>Meyer, Riese & Streichert, 2024</i>	834 original multiple-choice questions from 3 official written German medical licensing exams (text only)	GPT-3.5, GPT-4	Medical students who took the same exams	GPT-4 averaged 85% correct, ranking in the 92.8th, 99.5th, and 92.6th percentiles. (33)
<i>Bharatha et al., 2024</i>	304 MCQs from basic science and clinical course exams (text only)	ChatGPT-4	Medical students who took the same exams	ChatGPT-4 scored higher overall (73.7%) than students (66.7%). ChatGPT-4's performance depended on course type. (34)
<i>Friederichs et al., 2023</i>	Progress test spanning multiple specialties (incl. dermatology, radiology, legal medicine, and others) (text only)	GPT	Medical students taking the same progress test	ChatGPT answered 65.5% of questions correctly, better than student average across most specialties. Results varied by field (very high in dermatology, very low in radiology and legal medicine).(11)
<i>Alessi et al., 2025</i>	333 multiple-choice questions from the Brazilian Progress Test across different areas (text only)	GPT-3.5, GPT-4.0	Medical students who took the same test	GPT-4 scored higher than GPT-3.5 in every subject, but differences remained statistically significant only in basic sciences and gynecology/obstetrics. Both generally scored above student averages. (9)

Study	Question set	AI models	Human comparator	Findings
<i>Ros-Arlanzón et al., 2025</i>	Final MCQ exams in Infectious Diseases, Neurology, Respiratory Medicine, and Cardiovascular Medicine (text and image questions)	OpenAI o1, GPT-4o, DeepSeek R1, Microsoft Copilot, Gemini 1.5 Flash	442 student scores	Across all four courses, every LLM's mean score exceeded the student median, and in several cases beat the single highest-scoring student (LLM means 7.46–9.88 vs. student means 4.28–7.32).(35)
Al-Thani et al., 2025	160 Emergency Medicine MCQs: 98 text-only and 62 image-based.	GPT-4; Gemini 1.5	125 final-year medical students	Students scored 79.4%, ChatGPT 72.5%, and Gemini 54.4%. On image-based questions, accuracy was 62.9%, 54.8%, and 24.2%, respectively.(5)
<i>Salvador et al., 2025</i>	10-question practical examination based on radiological and nuclear medicine images.	GPT-4o	33 third-year medical students	Students achieved a higher mean score than GPT-4o, 7.78 vs. 6.05/10. GPT-4o ranked at the 9th percentile of the student distribution.(36)
<i>Sarantopoulos et al., 2026</i>	90 image-based MCQs in Anatomy, Pathology, and Pediatrics	GPT-4.0	127 undergraduate medical students	ChatGPT performed below students in Anatomy and above students in Pediatrics. Model failed to recognize all Pathology images.(10)
<i>This table provides a focused narrative overview and is not intended to represent a systematic review of the literature. Studies were selected based on the presence of a direct comparison between generative AI systems and undergraduate or pre-licensure students completing examination-style assessments. Studies involving only residents, specialists, or board-examination candidates were excluded.</i>				

Figures

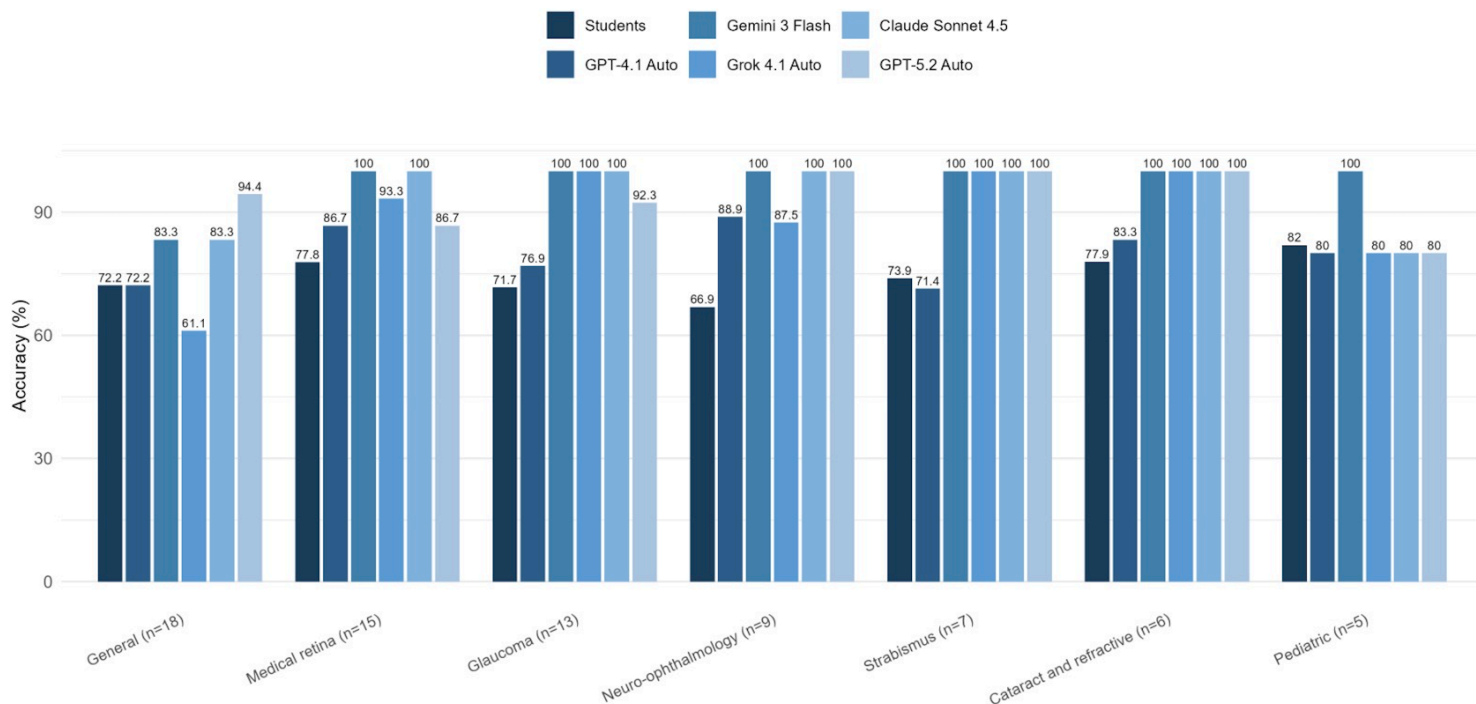


Figure 1

Accuracy (%) by ophthalmology subspecialty.

Legend: Subspecialties with $n \geq 5$ questions

Supplementary Files

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

- [TablesBMC.docx](#)
- [SUPPLEMENTARYTABLE.docx](#)