ORIGINAL ARTICLE

DOI: 10.4274/hamidiyemedj.galenos.2026.25238

Hamidiye Med J

Performance of GPT-4.1 and Gemini-3.0 on Ocular Oncology Board-Style Questions: A Comparative Study

GPT-4.1 ve Gemini-3.0'ün Oküler Onkoloji Kurul Tarzı Sorulardaki Performansı: Karşılaştırmalı Bir Çalışma

Dilan Yıldız, Emine Betül Akbaş Özyürek

University of Health Sciences Türkiye, Prof. Dr. Cemil Taşcıoğlu City Hospital, Clinic of Ophthalmology, İstanbul, Türkiye

ABSTRACT

Background: Large language models have shown promising performance in general medical education, but evidence regarding their accuracy in ocular oncology is limited. This study compared the performance of GPT 4.1 and Gemini-3.0 on ocular oncology board-style questions.

Materials and Methods: Fifty-eight board-style questions on ocular tumours were obtained from an established ophthalmology question bank. Each question was independently entered into GPT 4.1 and Gemini-3.0 using identical prompts. Accuracy was assessed by ophthalmologists with ocular oncology expertise. Response length and response time were also recorded.

Results: GPT 4.1 answered 63.8% of questions correctly, while Gemini-3.0 answered 65.5% of questions correctly; there was no statistically significant difference between the models ($p = 0.84$). Gemini-3.0 generated significantly longer responses than GPT 4.1 ($p < 0.001$), whereas response times were comparable ($p = 0.29$). No correlation was observed between response length and accuracy.

Conclusion: GPT 4.1 and Gemini-3.0 demonstrated comparable, moderate accuracy on ocular oncology board-style questions. Increased response verbosity did not improve accuracy, highlighting the need for expert oversight when using language models for subspecialty ophthalmology education.

Keywords: Artificial intelligence, ocular tumours, medical education, ophthalmology, question banks

öz

Amaç: Büyük dil modelleri genel tıp eğitimi alanında umut verici performans göstermiştir; ancak oküler onkoloji alanındaki doğruluklarına ilişkin kanıtlar sınırlıdır. Bu çalışma, GPT 4.1 ile Gemini-3.0'ün oküler onkolojiye yönelik kurul (board) tarzı sorulardaki performansını karşılaştırmayı amaçlamıştır.

Gereç ve Yöntemler: Oküler tümörlere ilişkin 58 adet kurul tarzı soru, yerleşik bir oftalmoloji soru bankasından elde edilmiştir. Her bir soru, aynı istemler kullanılarak bağımsız şekilde GPT 4.1 ve Gemini-3.0'e girilmiştir. Doğruluk, oküler onkoloji konusunda uzman oftalmologlar tarafından değerlendirilmiştir. Ayrıca yanıt uzunluğu ve yanıt süresi de kaydedilmiştir.

Bulgular: GPT 4.1 soruların %63,8'ini doğru yanıtlarken, Gemini-3.0 %65,5 doğruluk oranına ulaşmıştır ve modeller arasında istatistiksel olarak anlamlı bir fark bulunmamıştır ($p = 0,84$). Gemini-3.0, GPT 4.1'e kıyasla anlamlı derecede daha uzun yanıtlar üretmiştir ($p < 0,001$),ancak yanıt süresi benzer bulunmuştur ($p = 0,29$). Yanıt uzunluğu ile doğruluk arasında herhangi bir korelasyon saptanmamıştır.

Sonuç: GPT 4.1 ve Gemini-3.0, oküler onkoloji kurul tarzı sorularda benzer ve orta düzeyde doğruluk göstermiştir. Yanıtların daha uzun olması doğruluğu artırmamıştır; bu durum, dil modellerinin yandal oftalmoloji eğitiminde kullanımında uzman denetiminin gerekliliğini vurgulamaktadır.

Anahtar Kelimeler: Oftalmoloji, oküler tümörler, soru bankaları, tıp eğitimi, yapay zekâ

Address for Correspondence: Dilan Yıldız, MD, University of Health Sciences Türkiye, Prof. Dr. Cemil Taşcıoğlu City Hospital, Clinic of Ophthalmology, İstanbul, Türkiye

E-mail: dilanicar@gmail.com **ORCID ID:** orcid.org/0000-0001-6575-9825

Received: 13.04.2026 **Accepted:** 31.07.2026 **Epub:** 30.10.2026

Cite this article as: Yıldız D, Akbaş Özyürek EB. Performance of GPT-4.1 and Gemini-3.0 on ocular oncology board-style questions: a comparative study. Hamidiye Med J. [Epub Ahead of Print]

Copyright© 2026 The Author(s). Published by Galenos Publishing House on behalf of University of Health Sciences Türkiye, Hamidiye Faculty of Medicine. This is an open access article under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND) International License.

Introduction

Artificial intelligence (AI) has rapidly transformed multiple aspects of medicine, including diagnostics, prognostication, and medical education. Among recent advances, large language models (LLMs) have attracted significant attention due to their ability to generate human-like text, synthesize complex information, and respond to domain-specific questions with minimal prompting. Models such as ChatGPT (OpenAI) and Gemini (Google) are increasingly explored as potential tools for clinical decision support, examination preparation, and continuing medical education across medical specialties (1-3).

In ophthalmology, AI applications have traditionally focused on image-based tasks, particularly in retinal diseases, glaucoma, and diabetic retinopathy screening (4-6). More recently, interest has expanded toward natural language processing-based tools, including LLMs, which may assist ophthalmologists in interpreting clinical information, summarizing literature, and preparing for board examinations (7). However, the reliability and limitations of these models remain areas of active investigation, especially in subspecialties that require nuanced clinical reasoning.

Ocular oncology represents a particularly challenging domain for AI-based language models. The spectrum of ocular tumors—including uveal melanoma, retinoblastoma, conjunctival melanoma, choroidal metastases, and orbital tumors demands precise knowledge of epidemiology, clinical presentation, imaging characteristics, histopathology, staging systems, and management strategies (8-10). Board-style examination questions in ocular oncology often integrate multimodal information, such as clinical photographs, fundus images, and scenario-based decision-making, making them rigorous tests of both factual knowledge and clinical reasoning.

Medical board examinations play a critical role in ophthalmology training by assessing competence and readiness for independent practice. Educational resources such as EyeQuiz have become widely used by trainees and specialists preparing for ophthalmology board and subspecialty examinations. These question banks reflect real-world examination standards and emphasize high-yield clinical concepts (11). Evaluating LLM performance using such board-style questions provides a pragmatic framework for assessing their potential utility and risks in ophthalmic education.

Several recent studies have evaluated the performance of ChatGPT and other LLMs on medical examinations, including the United States Medical Licensing Examination, specialty board questions, and ophthalmology knowledge assessments (12-15). While many reports demonstrate that LLMs can achieve moderate to high accuracy, results vary widely depending on question format, specialty, and evaluation criteria. Importantly, high accuracy in general medical knowledge does not necessarily translate to subspecialty competence, particularly in fields where visual interpretation and contextual reasoning are essential.

Comparative evaluations of different LLMs are especially relevant, as models differ in architecture, training data, and optimization strategies. ChatGPT has been reported to provide more structured and concise responses in some medical contexts, whereas Gemini has been described as producing more verbose and explanatory outputs (16,17). However, direct comparisons of these models in ophthalmology, particularly in ocular oncology, remain scarce.

Beyond accuracy, additional performance metrics are increasingly recognized as important. Response length may influence educational usability, as excessively verbose answers can obscure key learning points, while overly brief responses may omit critical details. Similarly, response time is relevant for real-time educational or clinical support applications. Despite this, few studies have systematically compared LLMs using a multidimensional evaluation framework that includes accuracy, verbosity, and response latency.

Furthermore, the presence of image-based content introduces additional complexity. Although current LLMs have limited direct image interpretation capabilities compared with vision-specific AI models, image-referenced questions may indirectly affect reasoning complexity and response behavior. Understanding how LLMs perform on questions with and without image components is therefore relevant for ophthalmology-specific applications.

Given these gaps in the literature, a focused evaluation of LLM performance in ocular oncology education is warranted. This study aims to compare GPT-4.1 and Gemini 3.0 using board-style ocular tumor questions sourced from an established ophthalmology question bank. We assess model performance across three key dimensions: answer accuracy, response length, and response time, and perform additional subgroup analyses stratified by tumor type and image availability. By providing a detailed comparative analysis, this study seeks to inform educators, trainees, and clinicians about the strengths and limitations of contemporary LLMs in the context of ophthalmic oncology education.

Materials and Methods

Study Design

This comparative observational study evaluated the performance of two LLMs, GPT-4.1 (OpenAI, San Francisco,

CA, USA) and Gemini-3.0 (Google DeepMind, Mountain View, CA, USA), on board-style ocular oncology questions. Model performance was assessed across three predefined domains: answer accuracy, response length, and response time.

Question Selection

A total of 58 board-style multiple-choice questions related to ocular tumors were included (Supplementary Material). Questions were obtained from EyeQuiz.com, a widely used ophthalmology board preparation platform. The topics encompassed tumors of the eyelid and orbit, the ocular surface, choroid, retina, and iris, reflecting the breadth of ocular oncology content encountered in ophthalmology examinations.

Questions were selected to represent a range of difficulty levels and clinical scenarios. Items containing image-based content (e.g., fundus photographs or clinical images referenced in the question stem) were recorded in a separate variable. To ensure ethical compliance, questions were used strictly for educational research purposes, and no proprietary content was reproduced verbatim in the manuscript.

Each question was entered independently into both models using the same standardized prompt: "Select the single best answer and provide a brief explanation for your choice."

No follow-up prompting, clarification requests, or prompt-engineering techniques were used. All queries were performed between March and April 2025 using the publicly available versions of GPT-4.1 and Gemini-3.0.

Response Collection

For each question, the first complete response generated by each model was recorded. Responses were not edited or truncated. The following parameters were documented:

- Response accuracy
- Response length, measured as word count
- Response time, measured in seconds from submission to completion of the response

Response time was defined as the interval between the submission of the question and the completion of the generated response, measured in seconds using a digital stopwatch under stable internet conditions.

Accuracy Assessment

Answer accuracy was evaluated independently by two fellowship-trained ophthalmologists with experience in ocular oncology independently evaluated all responses. Responses were classified as:

- Correct (fully concordant with accepted reference standards)
- Incorrect (partially correct or incorrect)

Standard ophthalmology textbooks and established guidelines were used as reference benchmarks. Disagreements were resolved through consensus discussion. Inter-observer agreement was evaluated using Cohen's kappa coefficient.

Subgroup Variables

Additional variables were analyzed to explore potential factors influencing model performance:

- Tumor location (eyelid and orbit, ocular surface, choroid, retina, iris)
- Image-based content (present vs. absent)

These subgroup analyses were performed for both accuracy and response-behavior metrics.

Statistical Analysis

Statistical analyses were conducted using IBM SPSS Statistics for Windows, version 27.0 (IBM Corp., Armonk, NY, USA) software. Continuous variables were reported as mean ± standard deviation, and categorical variables as counts and percentages.

Comparisons between GPT-4.1 and Gemini 3.0 were performed using appropriate parametric or non-parametric tests based on the data distribution. Correlations between variables were assessed using Pearson or Spearman correlation coefficients, as appropriate.

Analysis of variance was used to evaluate differences across tumor subtypes. A p-value < 0.05 was considered statistically significant.

Ethical Considerations

This study did not involve human subjects or patient data, and therefore did not require institutional review board approval. The study adhered to ethical standards for educational research and responsible use of artificial intelligence tools.

Results

Study Sample

A total of 58 ocular oncology board-style questions were included in the analysis. Questions covered tumors of the eyelid and orbit (27.6%), the ocular surface (8.6%), the choroid (37.9%), the retina (22.4%), and the iris (3.4%) (Figure 1). Image-based content was present in 43.1% of questions.

Accuracy Comparison

GPT-4.1 answered 63.8% of questions correctly (37/58), whereas Gemini 3.0 achieved a slightly higher accuracy of 65.5% (38/58). The difference in overall accuracy between

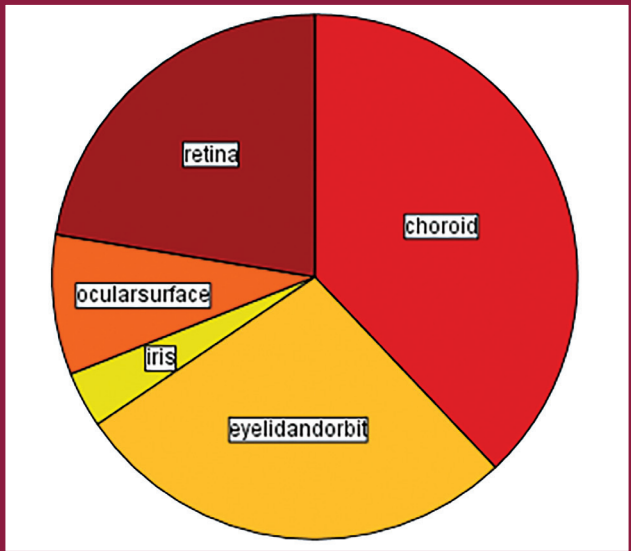


Figure 1. Distribution of ocular oncology board-style questions according to anatomical tumor location.

the two models was not statistically significant (p = 0.84). Accuracy did not significantly vary across tumor locations for either model (p = 0.145).

A strong positive correlation was observed between the accuracy scores of GPT-4.1 and Gemini-3.0 (Pearson r = 0.51, p < 0.001), indicating that both models tended to perform similarly on the same questions.

Response Length

Gemini-3.0 generated significantly longer responses, with a mean word count of 265.5 ± 56.4, compared to 75.8 ± 29.8 words for GPT-4.1. Response lengths showed a strong positive correlation between the two models (Pearson r = 0.66, p < 0.001), suggesting consistent verbosity patterns across questions.

Image-based questions were associated with increased response length for both models, particularly for Gemini-3.0, although the increase did not reach statistical significance (p = 0.371).

Response Time

Mean response time was 14.7 ± 4.9 seconds for GPT-4.1 and 13.8 ± 4.5 seconds for Gemini-3.0. The difference in response time was not statistically significant (p=0.29). However, image-based questions were moderately correlated with longer response times for both GPT-4.1 (r = 0.30, p = 0.021) and Gemini-3 (r = 0.36, p = 0.005).

The accuracy, response length, and response time are summarized in Table 1.

Table 1. Comparison of accuracy, response length, and response time between GPT-4.1 and Gemini-3.0.

Metric	GPT-4.1	Gemini-3.0	p-value
Accuracy (%)	63.8	65.5	0.84
Mean response length (words)	75.8 ± 29.8	265.5 ± 56.4	<0.001
Mean response time (seconds)	14.7 ± 4.9	13.8 ± 4.5	0.29

Response times between the two models were positively correlated (Pearson r = 0.45, p < 0.001), indicating similar computational complexity across questions.

Subgroup Analysis

Neither model demonstrated statistically significant differences in accuracy by tumor type or image availability (p = 0.724 and p = 0.105, respectively). Both GPT-4.1 and Gemini 3.0 showed slightly lower performance on retina-related questions; however, this trend did not reach statistical significance (p = 0.562) (Figure 2).

Discussion

In this study, we performed a comparative evaluation of two contemporary LLMs, GPT-4.1 and Gemini-3.0, using board-style questions focused on ocular oncology. Our findings demonstrate that both models achieved moderate and comparable accuracy, with no statistically significant difference in overall performance. However, notable differences were observed in response length and response behavior, highlighting important considerations for the educational use of LLMs in ophthalmology.

The overall accuracy rates of GPT-4.1 (63.8%) and Gemini-3.0 (65.5%) are consistent with previously reported performances of LLMs on specialty-level medical examinations (3,12-14). Prior studies evaluating ChatGPT in ophthalmology have shown accuracy ranging from 55% to 70%, depending on question complexity and subspecialty focus (7,13). Our results extend these observations specifically to ocular oncology, a field characterized by complex clinical reasoning and relatively low exposure during general ophthalmology training.

Importantly, the absence of a significant accuracy difference between GPT-4.1 and Gemini-3.0 suggests that model generation alone does not guarantee superior performance in subspecialty domains. The strong positive correlation between the accuracy scores of both models indicates that they tended to succeed or fail on the same questions. This finding implies that question-specific complexity—rather than model architecture—may be a dominant factor influencing performance, particularly for topics requiring nuanced interpretation or integration of clinical details.

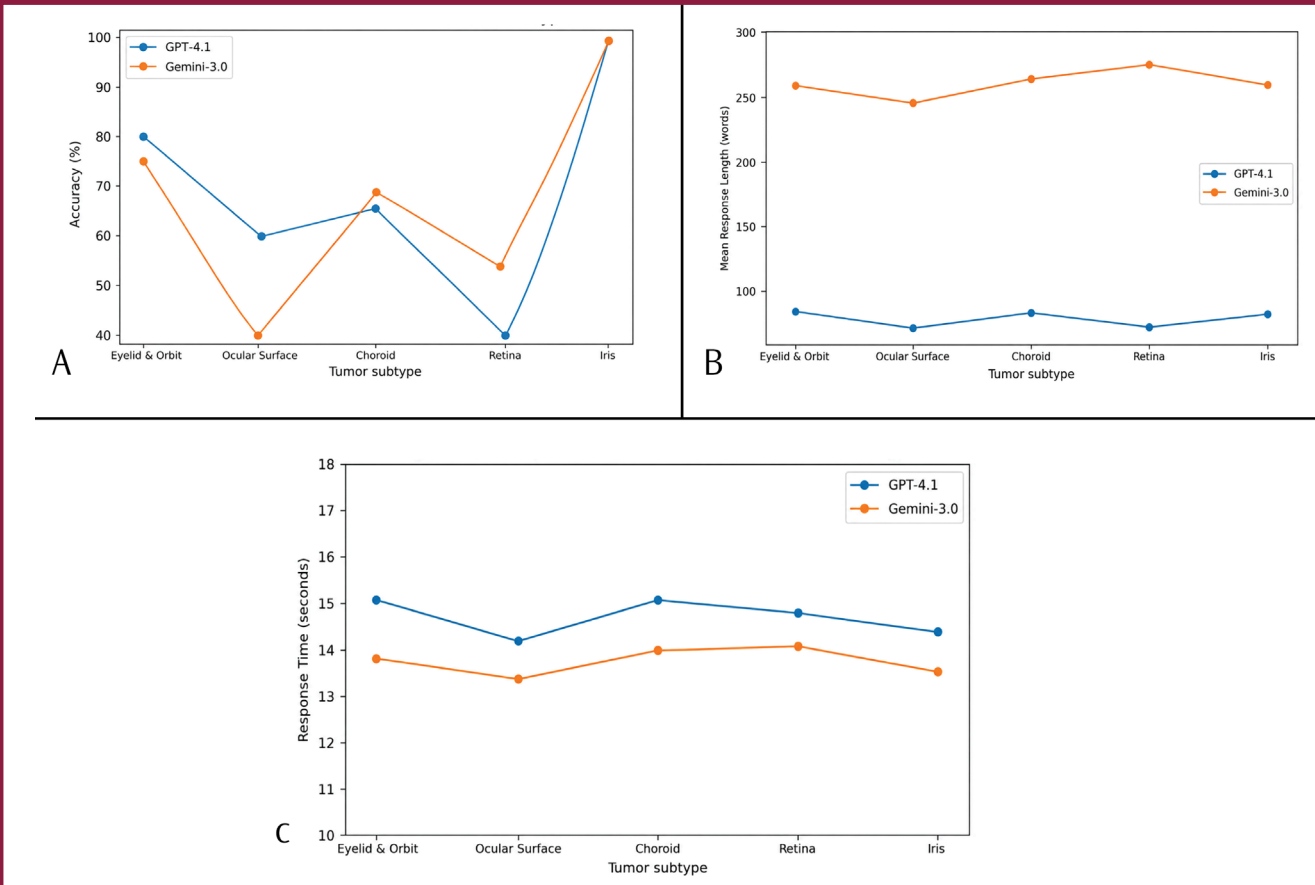


Figure 2. (A-C): Comparison of accuracy, mean response length, and mean response time for GPT-4.1 and Gemini 3.0 across ocular tumor subtypes, including eyelid and orbit, ocular surface, choroid, retina, and iris.

One of the most striking differences between the two models was response length. Gemini-3.0 produced substantially longer responses than GPT-4.1, despite similar accuracy rates. This observation aligns with prior reports suggesting that some LLMs prioritize explanatory completeness over conciseness (16,17). From an educational perspective, verbosity may be a double-edged sword: while detailed explanations can enhance conceptual understanding, excessive length may obscure key learning points or introduce irrelevant information. Our findings reinforce the observation that longer responses do not necessarily translate into higher accuracy, which is an important consideration for trainees relying on AI-generated explanations.

Response time was comparable between the two models, with both demonstrating modest delays when addressing image-referenced questions. Although, in this study, current LLMs do not directly interpret images, image-associated questions may increase cognitive complexity by requiring indirect inference or contextual reasoning. The observed correlation between image presence and response time suggests that such questions impose greater processing demands on language models, a phenomenon also noted in previous AI evaluations within ophthalmology (4,6).

Subgroup analyses revealed no significant differences in accuracy based on tumor location or image availability. However, both models demonstrated slightly lower performance on retina-related questions. This trend, although not statistically significant, may reflect the increased diagnostic complexity and reliance on imaging interpretation commonly associated with retinal tumors and retinoblastoma (9,10). These findings underscore the current limitations of language-only AI systems in domains where visual reasoning plays a critical role.

The use of board-style questions from an established ophthalmology question bank strengthens the real-world relevance of this study. Unlike examinations designed specifically to test AI performance, board-style questions reflect the expectations and standards that ophthalmology

trainees face. Our results, therefore, provide practical insight into how LLMs may function as adjunctive tools for examination preparation rather than definitive sources of truth.

Despite promising performance, the clinical and educational use of LLMs warrants caution. Incorrect or partially correct responses—observed in over one-third of cases—carry the risk of reinforcing misconceptions if used without expert supervision. This concern has been emphasized in prior discussions regarding AI-assisted medical education (1,15). Consequently, LLMs should be viewed as supplementary educational tools, ideally integrated with validated resources and expert oversight.

Recent studies have increasingly examined the performance of LLMs in ophthalmology, focusing on accuracy, readability, and clinical applicability across different subspecialties (18,19). However, most published studies have evaluated either patient-oriented or general ophthalmology examination questions, rather than subspecialty-specific, board-style ocular oncology questions. Our study expands this growing body of literature by assessing the performance of GPT-4.1 and Gemini-3.0 on ocular tumor board-style questions, using accuracy, response length, and response time as comparative metrics.

Biçer and Şahlı (20) evaluated ChatGPT-4 and Gemini-2.0 by having them answer frequently asked questions about retinitis pigmentosa, and found no significant difference in accuracy between the two models, despite measurable differences in readability indices. Their findings suggest that both LLMs are capable of producing accurate disease-related information, particularly when responses are simplified for patient comprehension (20). Similarly, in our study, GPT-4.1 and Gemini-3.0 demonstrated comparable accuracy across ocular tumor subtypes, reinforcing the notion that newer-generation LLMs reach a performance plateau for ophthalmic knowledge-based questions. However, unlike the RP study, our analysis targeted expert-level, board-style questions rather than patient FAQs, highlighting that comparable accuracy persists even under more specialized and cognitively demanding conditions.

In contrast to readability-focused studies, our results demonstrate that increased verbosity does not necessarily translate into higher accuracy. Gemini-3.0 consistently generated significantly longer responses than GPT-4.1, yet this verbosity did not confer an advantage in correctness. This observation partially aligns with prior work by Karataş and Karataş in pediatric ophthalmology, who compared ChatGPT-4.0 with DeepSeek-R1 using American Academy of Ophthalmology BCSC-derived multiple-choice questions. While DeepSeek-R1 achieved numerically higher accuracy, the difference was not statistically significant, underscoring that response quality depends more on model reasoning capabilities than response length or stylistic detail (21).

Our findings also are consistent with broader ophthalmology assessments of LLM performance. Previous studies evaluating ChatGPT-3.5 and ChatGPT-4.0 on ophthalmology question banks reported incremental improvements in accuracy with newer models but continued variability across subspecialties (19-21). Notably, tumor-related and retina-based questions were among the most challenging categories, consistent with our observation that retinal and choroidal tumor subtypes had lower accuracy than eyelid and iris tumors. This pattern suggests that ocular oncology, particularly posterior segment tumors, remains a complex domain where LLM reasoning may be constrained by limited training data or nuanced diagnostic pathways.

Additionally, the literature emphasizes the importance of standardized, objectively graded questions when evaluating LLM performance (22). In contrast, studies using patient-oriented or open-ended questions tend to emphasize readability and accessibility over precision. Our methodology bridges this gap by using expert-curated board questions with definitive answers, enabling objective assessment of accuracy and capturing response length and response time as secondary performance indicators.

Collectively, these comparisons indicate that while modern LLMs such as GPT-4.1 and Gemini-3.0 demonstrate reliable baseline knowledge across ophthalmology, their performance in subspecialty oncology remains heterogeneous. The lack of correlation between verbosity and accuracy, as observed in our study, challenges the assumption that more detailed responses are inherently superior. Instead, concise and focused answers, as produced by GPT-4.1, may be more suitable for examination-oriented or clinical decision-support contexts.

Although most incorrect responses reflected partially accurate reasoning or selection of plausible distractors, several responses demonstrated clinically misleading explanations and overconfident justifications. In particular, some retina-related and posterior-segment tumor questions contained fabricated diagnostic reasoning or omitted critical differential diagnoses. While these hallucination-like responses were relatively infrequent, they highlight an important safety concern regarding the unsupervised educational or clinical use of LLMs in subspecialty ophthalmology. Users may misinterpret confident but inaccurate responses as reliable clinical guidance, underscoring the need for expert verification.

Importantly, no prior studies have simultaneously evaluated accuracy, verbosity, and response time across ocular tumor subtypes. By incorporating these multidimensional performance metrics, our study provides

a more comprehensive assessment of LLM behavior in subspecialty ophthalmology and highlights critical limitations that must be addressed before clinical or educational deployment.

Study Limitations

This study has several limitations that should be acknowledged. First, the number of board-style questions included was relatively limited, and all questions were obtained from a single ophthalmology question bank. Although this approach enhances standardization, it may limit the generalizability of the findings to other examination formats or educational resources. Second, answer accuracy was assessed using a binary classification system (correct vs incorrect), which may not fully capture partially correct or clinically acceptable responses. A more granular scoring system could provide additional insight into model performance. Third, this study evaluated only text-based responses and did not directly assess image interpretation capabilities. Given the visual nature of ocular oncology, this represents an important limitation. Image-referenced questions may have influenced the complexity of reasoning without enabling true visual analysis. Fourth, response time measurements were influenced by external factors, such as network conditions and platform-related latency. These factors could not be fully controlled despite efforts to maintain consistent testing conditions. Finally, LLMs are continuously evolving. The performance of GPT-4.1 and Gemini 3.0 may change with future updates, and the results presented here should be interpreted as a snapshot of model capabilities at the time of evaluation rather than definitive assessments.

Conclusion

In this comparative study of board-style ocular oncology questions, GPT-4.1 and Gemini 3.0 demonstrated similarly moderate accuracy, with no statistically significant difference in overall performance. While Gemini 3 generated substantially longer responses, response times were comparable between the two models. Importantly, increased response length did not correspond to higher accuracy.

These findings suggest that LLMs may serve as adjunctive educational tools in ophthalmology training, particularly for examination preparation and conceptual review. However, given the observed error rates and limitations in subspecialty reasoning, LLMs should not be used as standalone sources of information in ocular oncology.

The study reflects model performance at a single time point and results may vary with future model updates Future research should focus on larger question sets, multimodal AI systems incorporating image analysis, and longitudinal studies evaluating the impact of LLM-assisted learning on trainee knowledge and clinical decision-making. With appropriate oversight, LLMs may play a complementary role in ophthalmic education while reinforcing, rather than replacing, expert-guided learning.

Ethics

Ethics Committee Approval: Ethical approval is not required for this study in accordance with local or national guidelines.

Informed Consent: Not applicable, as this study did not involve human participants or patient data.

Footnotes

Authorship Contributions

Concept: D.Y., Design: D.Y., Data Collection or Processing: D.Y., E.B.A.Ö., Analysis or Interpretation: D.Y., E.B.A.Ö., Literature Search: D.Y., Writing: D.Y.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

Supplementary Material Link: https://d2v96fxpocvxx.cloudfront.net/f942abf9-a73b-4057-b48b-c3a4b7d0c2a5/content-images/e528b9f1-f9de-42f8-a50b-8cf537c4fb39.pdf

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25:44-56. [Crossref]
2. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022;28:31-38. [Crossref]
3. 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:e0000198. [Crossref]
4. Ting DSW, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol. 2019;103:167-175. [Crossref]
5. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316:2402-2410. [Crossref]
6. Schmidt-Erfurth U, Sadeghipour A, Gerendas BS, Waldstein SM, Bogunović H. Artificial intelligence in retina. Prog Retin Eye Res. 2018;67:1-29. [Crossref]
7. Heinke A, Radgoudarzi N, Huang BB, Baxter SL. A review of ophthalmology education in the era of generative artificial intelligence. Asia Pac J Ophthalmol (Phila). 2024;13:100089. [Crossref]
8. Shields JA, Shields CL. Intraocular tumors: an atlas and textbook. 3rd ed. Philadelphia (PA): Wolters Kluwer; 2015. [Crossref]
9. Singh AD, Turell ME, Topham AK. Uveal melanoma: trends in incidence, treatment, and survival. Ophthalmology. 2011;118:1881-1885. [Crossref]

10. Kaliki S, Shields CL. Retinoblastoma: achieving new standards with methods of chemotherapy. Indian J Ophthalmol. 2015;63:103-109. [Crossref]
11. EyeQuiz. Ophthalmology board review and question bank [Internet]. Accessed 2025. [Crossref]
12. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. Correction: 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. 2024;10:e57594. [Crossref] Erratum for: JMIR Med Educ. 2023;9:e45312. [Crossref]
13. 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;3:100324. [Crossref]
14. Longwell JB, Hirsch I, Binder F, Gonzalez Conchas GA, Mau D, Jang R, et al. Performance of large language models on medical oncology examination questions. JAMA Netw Open. 2024;7:e2417641. [Crossref]
15. 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. [Crossref]
16. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620:172-180. [Crossref] Erratum in: Nature. 2023;620:E19. [Crossref]
17. Google Research. Gemini: a family of highly capable multimodal models. Technical report. 2024. [Crossref]
18. Mihalache A, Grad J, Patil NS, Huang RS, Popovic MM, Mallipatna A, et al. Google Gemini and Bard artificial intelligence chatbot performance in ophthalmology knowledge assessment. Eye (Lond). 2024;38:2530-2535. [Crossref]
19. Shean R, Shah T, Sobhani S, Tang A, Setayesh A, Bolo K, et al. OpenAI o1 large language model outperforms GPT-4o, Gemini 1.5 Flash, and human test takers on ophthalmology board-style questions. Ophthalmol Sci. 2025;5:100844. [Crossref]
20. Biçer Ö, Şahlı E. Evaluation of the artificial intelligence chatbots in frequently asked questions about retinitis pigmentosa: a comparative analysis between ChatGPT-4 and Gemini-2.0. Int J Retina Vitreous. 2025;12:1. [Crossref]
21. Karataş G, Karataş ME. Artificial intelligence in pediatric ophthalmology: a comparative study of ChatGPT-4.0 and DeepSeek-R1 performance. Strabismus. 2026;34:61-67. [Crossref]
22. Srinivasan S, Ai X, Zou M, Zou K, Kim H, Lo TWS, et al. Ophthalmological question answering and reasoning using OpenAI o1 vs other large language models. JAMA Ophthalmol. 2025;143:740-748. [Crossref]