Comparative performance analysis of ChatGPT, Scholar GPT, and DeepSeek of the Conselho Brasileiro de Oftalmologia 2022 theoretical exams

Análise comparativa de desempenho entre ChatGPT, Scholar GPT e DeepSeek em provas teóricas do Conselho Brasileiro de Oftalmologia 2022

Diogo Gonçalves dos Santos Martins[1], Thiago Gonçalves dos Santos Martins[2], Eduardo Damasceno[3], Thomaz Gonçalves dos Santos Martins[4], Paulo Schor[1]

[1] Department of Ophthalmology, Universidade Federal de São Paulo, São Paulo, SP, Brazil.
[2] Department of Ophthalmology, Universidade Federal do Rio de Janeiro/Macaé, Macaé, RJ, Brazil.
[3] Department of Ophthalmology, Universidade Federal Fluminense, Niteroi, SP, Brazil.
[4] Department of Ophthalmology, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brazil.

How to cite:
Martins DG, Martins TG, Damasceno E, Martins TG, Schor P. Comparative performance analysis of ChatGPT, Scholar GPT, and DeepSeek of the Conselho Brasileiro de Oftalmologia 2022 theoretical exams. Rev Bras Oftalmol. 2026;85:e0009.

doi:
https://doi.org/10.37039/1982.8551.20260009

Keywords:
Education, medical; Ophthalmology; Artificial intelligence; Generative artificial intelligence; ChatGPT; Scholar GPT; DeepSeek

Descritores:
Educação médica; Oftalmologia; Inteligência artificial; Inteligência artificial generativa; ChatGPT; Scholar GPT; DeepSeek

Received on:
April 15 , 2025

Accepted on:
December 6, 2025

Corresponding author:
Thiago Gonçalves dos Santos Martins
Rua Botucatu, 821 – Vila Clementino
Zip code: 04023-062 – São Paulo, SP, Brasil
E-mail: thiagogsmartins@yahoo.com.br

Institution:
Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brazil.

Conflict of interest:
no conflict of interest.

Financial support:
no financial support for this work.

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

Associate editor:
Bruno Leonardo Barranco Esporcatte
Universidade Federal de São Paulo, São Paulo, SP, Brazil
https://orcid.org/0000-0002-4509-7075

ABSTRACT

Objective: This study presents a comparative analysis of the performance of three artificial intelligence models (ChatGPT, Scholar GPT, and DeepSeek) in solving the CBO 2022 theoretical exams I and II to obtain the Specialist Certification Exam in Ophthalmology.

Methods: A total of 46 valid questions from theoretical exam I and 122 from theoretical exam II were analyzed after the exclusion of 7 cancelled questions. The models were set to operate under standardized conditions, and their responses were compared with the official answer keys.

Results: DeepSeek achieved the highest performance, with an accuracy rate of 97.8% in theoretical exam I and 81.9% in theoretical exam II, consistently outperforming the other models across all knowledge areas. ChatGPT demonstrated intermediate performance, while Scholar GPT had the lowest accuracy rate.

Conclusion: The findings highlight DeepSeek's superiority in highly specialized tasks, emphasizing the relevance of Artificial Intelligence models trained for specific technical domains. However, all models showed limitations in more complex questions, underscoring the need for human supervision in critical applications.

RESUMO

Objetivo: Realizar uma análise comparativa do desempenho de três modelos de inteligência artificial (ChatGPT, Scholar GPT e DeepSeek) na resolução das provas teóricas I e II do Título de Especialista em Oftalmologia de 2022.

Métodos: Foram analisadas 46 questões válidas da prova teórica I e 122 da prova teórica II, após a exclusão de 7 questões anuladas. Os modelos foram configurados para operar em condições padronizadas, e suas respostas foram comparadas aos gabaritos oficiais.

Resultados: O DeepSeek apresentou o melhor desempenho, com taxa de acertos de 97,8% na prova teórica I e 81,9% na prova teórica II, superando consistentemente os demais modelos em todas as áreas de conhecimento. O ChatGPT obteve desempenho intermediário, enquanto o Scholar GPT teve a menor taxa de acertos.

Conclusão: Os resultados evidenciam a superioridade do DeepSeek em tarefas altamente especializadas, reforçando a relevância de modelos de Inteligência Artificial treinados para domínios técnicos específicos. No entanto, todos os modelos demonstraram limitações em questões mais complexas, ressaltando a necessidade de complementação com supervisão humana para aplicações críticas.

INTRODUCTION

Artificial Intelligence (AI) has emerged as a transformative tool across various fields, including education, professional assessment, and decision-making support. Advanced language models such as ChatGPT, Scholar GPT, and DeepSeek represent significant milestones in this domain, demonstrating remarkable capabilities ranging from text generation to solving complex problems. ChatGPT and ScholarGPT represent general-purpose and academically fine-tuned language models, respectively, while DeepSeek stands out for its open-weight architecture and focus on advanced reasoning capabilities. This comparative analysis offers an opportunity to evaluate how distinct design paradigms impact model performance within a highly specialized domain. By contrasting a generalist model, an academically optimized tool, and a reasoning-oriented system, we aim to explore how each approach responds to the demands of ophthalmology education. To strengthen this focus, we restructured the transition from the broader discussion on AI in education to the specific challenges posed by the Specialist Certification Exam in Ophthalmology. These exams, which require high precision and domain-specific expertise, provide a rigorous context for testing the real-world applicability of large language models in complex, technical assessments. However, their performance can vary considerably depending on the context and the specificity of the tasks, highlighting the need to evaluate their effectiveness in highly specialized scenarios, such as technical exams and professional evaluations.[1,2]

METHODS

In this study, we conducted a comparative analysis of the performance of three AIs models (ChatGPT, Scholar GPT, and DeepSeek) on the theoretical exams I and II of the 2022 Ophthalmology Specialist Title theoretical exams I and II. These exams, designed to assess in-depth technical knowledge in Ophthalmology, offer an ideal framework for evaluating the models' capacity to address complex, context-dependent questions. The selection of the 2022 edition was further justified by the broader availability of candidate performance data for that year, enabling more robust benchmarking and contextual interpretation of the results.

A total of 50 questions from theoretical exam I and 125 from theoretical exam II were analyzed, with 4 and 3 questions cancelled, respectively. This resulted in 46 and 122 valid questions. Questions containing figures were included in the analysis. The cancelled questions were excluded to ensure that only valid and officially recognized questions were considered, preventing distortions in the results and ensuring the integrity of the evaluation.

To ensure a fair and consistent comparison, the three AI models were evaluated under standardized conditions. Uniform prompt structures were applied to each question to maintain input consistency. Response times were controlled by imposing a maximum limit of 60 seconds per query, preventing discrepancies arising from variations in processing speed. When applicable, all interactions were conducted using the same user account and interface to minimize variability related to different access points or user profiles.

The evaluation environment consisted of a workstation equipped with an Intel Core i7-9700K processor running at 3.6 GHz, 16 GB of DDR4 RAM, and a stable high-speed internet connection with an average bandwidth of 100 Mbps. ChatGPT (version GPT-4.0) and ScholarGPT (version 1.2) were accessed via their official web platforms, operating remotely on the providers' cloud infrastructure, with average response times ranging from 10 to 30 seconds per query. DeepSeek (V3 version) was deployed on a dedicated local server, allowing for efficient use of its open-weight architecture, with average response times of approximately 15 seconds per query.

All responses generated by the models were independently evaluated by two board-certified ophthalmologists using a predefined scoring rubric that assessed accuracy, completeness, and relevance. Any scoring discrepancies were resolved through consensus discussions to ensure objectivity and reliability in the evaluation process.

- ChatGPT: GPT-4-Turbo version.[3]
- Scholar GPT: GPT-4-Turbo version.[3]
- DeepSeek: DeepSeek-V3 version.[4]

All models were instructed to provide direct and objective responses using the uploaded PDF file of the 2022 Ophthalmology Specialist Title theoretical exams I and II. The questions were answered on the same day (March 2, 2025) to ensure that the models were using their most updated versions.

The evaluation of the responses was based on the official answer keys of the exams, with the following criteria: correct if the response provided by the model exactly matched the correct answer in the answer key; incorrect if the response provided by the model did not match the correct answer or was left blank; cancelled questions, if these were excluded from the analysis, as mentioned earlier.

RESULTS

Quantitative metrics, organized by area of knowledge, were used to provide a detailed evaluation of the models' performance in different categories. Below are the analysis methods, the results obtained, and interpretations of the data.

Quantitative metrics

Accuracy rate

The accuracy rate was calculated as the percentage of correctly answered questions relative to the total number of valid questions. The results obtained were as follows:

Theoretical exam I (46 valid questions; Figure 1):

- DeepSeek: 45 correct answers (97.8% accuracy).
- ChatGPT: 39 correct answers (84.8% accuracy).
- Scholar GPT: 38 correct answers (82.6% accuracy).

Theoretical exam II (122 valid questions; Figure 2)

- DeepSeek: 100 correct answers (81.9% accuracy)
- ChatGPT: 92 correct answers (75.4% accuracy)
- Scholar GPT: 90 correct answers (73.8% accuracy)

Relative performance

To compare the performance of the models in each exam, we calculated the percentage difference in the accuracy rates between them:

Theoretical exam I:

- DeepSeek outperformed ChatGPT by 13 percentage points (97.8% *versus* 84.8%).
- DeepSeek outperformed Scholar GPT by 15.2 percentage points (97.8% *versus* 82.6%).

Theoretical exam II:

- DeepSeek outperformed ChatGPT by 6.5 percentage points (81.9% *versus* 75.4%).
- DeepSeek outperformed Scholar GPT by 8.1 percentage points (81.9% *versus* 73.8%).

Comparative analysis by area of knowledge

To conduct a more detailed analysis, the questions were organized by area of knowledge (e.g., histology, physiology, embryology) based on the content of each CBO theoretical exam I and II. This categorization allowed for identifying whether the models' performance varied depending on the subject matter. The percentage of correct answers by area of knowledge was calculated to assess the differences in performance across the various ophthalmology topics (figure 1 and figure 2).

DISCUSSION

The results of this comparative study between ChatGPT, Scholar GPT, and DeepSeek on the theoretical exams I and II of the 2022 Ophthalmology Specialist Title Exams reveal important insights into the performance of AI models in specialized contexts. The analysis not only highlights DeepSeek's superiority but also raises questions about the limitations and potential applications of these models in professional and educational environments.[5,6]

DeepSeek demonstrated superior performance compared to ChatGPT and ScholarGPT, which may be attributed to its specialized training architecture and optimized reasoning strategies tailored for multiple-choice question formats. Unlike the more generalist design of ChatGPT and ScholarGPT, DeepSeek appears to have been trained on datasets more closely aligned with the specific content and structure of the Specialist Certification Exam in Ophthalmology. This specialization likely enhances its ability to interpret and accurately respond to complex,

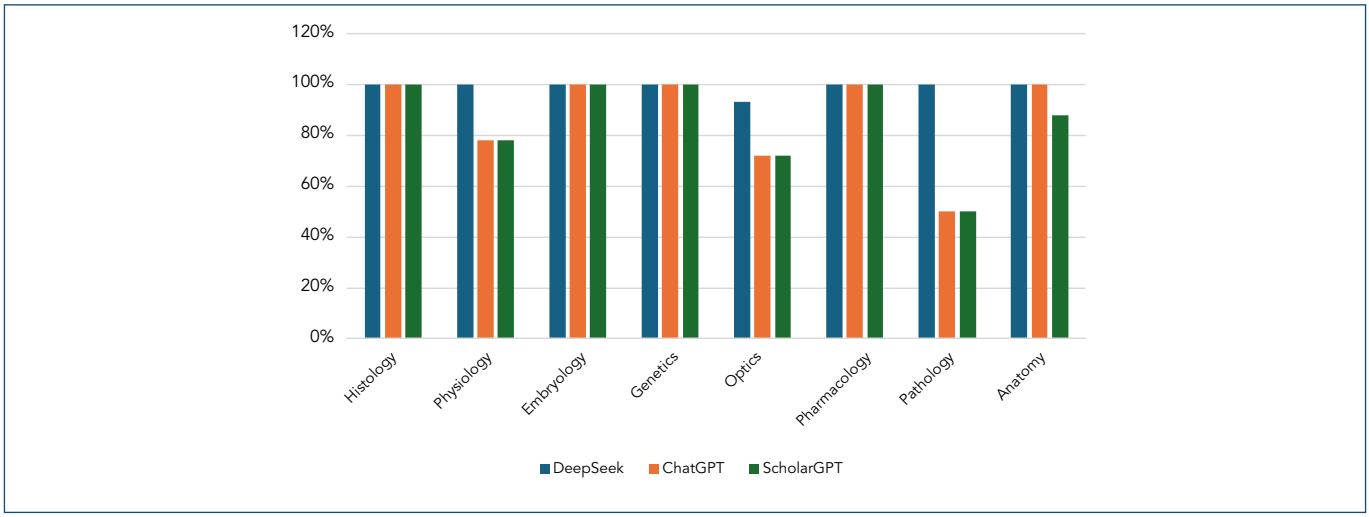

Figure 1. Accuracy of algorithms in solving theoretical test 1 of the *Conselho Brasileiro de Oftalmologia* for the year 2022.

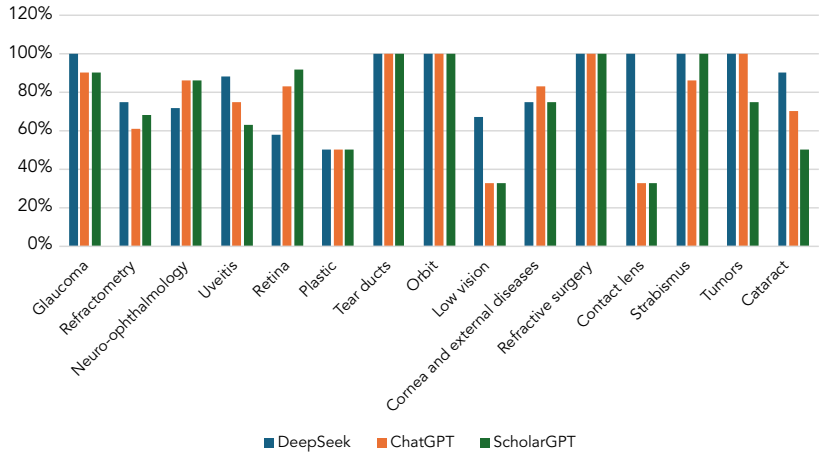

Figure 2. Accuracy of algorithms in solving theoretical test 2 of the *Conselho Brasileiro de Oftalmologia* for the year 2022.

domain-specific questions. These findings underscore the importance of task-specific model training when applying AI tools in professional and technical contexts that demand deep subject-matter expertise.[7]

Although both ChatGPT and Scholar GPT are language models developed by OpenAI, their performances were distinct. ChatGPT relies on a broad and diverse dataset from the internet, including books, websites, articles, and other public sources, while Scholar GPT is a specialized version of GPT focused on academic and scientific contexts. However, despite its access to academic databases such as peer-reviewed articles and scientific publications, Scholar GPT had the worst comparative performance among the models tested.[8]

The quality and recency of the data used in training the models are determining factors for performance. DeepSeek may have access to more recent and relevant databases for the CBO context, while ChatGPT and Scholar GPT may rely on more generic or outdated information. This difference may explain DeepSeek's higher accuracy in technical and specific questions.[9]

Context interpretation is an area where AI models still face challenges. Questions requiring inference, abstract reasoning, or practical knowledge can be particularly difficult for models like ChatGPT and Scholar GPT. In contrast, DeepSeek may have more refined algorithms for handling contextual nuances, contributing to its higher accuracy rate.[10]

Although DeepSeek excelled, it is important to acknowledge that none of the models achieved 100% accuracy, especially in theoretical exam II, which had a higher level of complexity. This suggests that in scenarios requiring practical knowledge or subjective interpretation, AI does not yet fully replace human evaluation. The combination of AI with human oversight may be the most effective approach to ensuring accuracy and fairness in assessments.[11]

The presence of biases in training data is a significant concern. If models are trained on data that inadequately represents the diversity of contexts and perspectives, it may lead to inaccurate or unfair evaluations. For instance, questions involving specific cultural or regional contexts may be misinterpreted by models trained predominantly on data from other regions, as the sources used for the response may not reflect the local context in which the question was designed.[12-14]

CONCLUSION

This study demonstrated that DeepSeek outperformed ChatGPT and Scholar GPT in the *Conselho Brasileiro de Oftalmologia* 2022 theoretical exams, highlighting the importance of specialized models for specific tasks. However, all models showed limitations, particularly in more complex or subjective questions. The integration of Artificial Intelligence models into educational platforms emerges as a promising area, with research exploring how these tools can be used to provide customized feedback, identify knowledge gaps, and adapt learning content to meet individual students' needs. Future research should focus on refining these technologies with an emphasis on practical applications in education and professional assessment

AUTHOR'S CONTRIBUTION

Diogo Gonçalves dos Santos Martins: Conceptualization, writing original draft, writing review & editing; Paulo Schor: Conceptualization, writing original draft, writing

review & editing; Eduardo Damasceno: Conceptualization, writing original draft, writing review & editing; Thiago Gonçalves dos Santos Martins: Conceptualization, writing original draft, writing review & editing; Thomaz Gonçalves dos Santos Martins: Conceptualization, writing original draft, writing review & editing.

REFERENCES

1. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan JM, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877-901.
2. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota; 2019. volume 1, p. 4171-86.
3. Zang K. optimizing language models for dialogue. ChatGPT; 2023 [cited Dec 2025 6]. Available from: https://kpzhang.github.io/report/ChatGPT-KZ-Feb2023.pdf
4. DeepSeek. DeepSeek: advanced AI for specialized applications. DeepSeek Technical Report. 2023.
5. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI Blog. 2019 [cited 2025. Dec 6]. Available from: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
6. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman S. GLUE: a multi-task benchmark and analysis platform for natural language understanding. In: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Brussels, Belgium. Association for Computational Linguistics; 2018. p. 353–5.
7. Rajpurkar P, Zhang J, Lopyrev K, Liang P. SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas. Association for Computational Linguistics; 2026. p. 2383–92.
8. Luckin R, Holmes W, Griffiths M, Forcier LB. Intelligence unleashed: an argument for AI in education. Pearson Education; 2016.
9. Chassignol M, Khoroshavin A, Klimova A, Bilyatdinova A. Artificial Intelligence trends in education: a narrative overview. Procedia Comput Sci. 2018;136:16-24.
10. Holmes W, Bialik M, Fadel C. Artificial Intelligence in education: promises and implications for teaching and learning. Boston: Center for Curriculum Redesign; 2019 [cited 2025 Dec 6]. Available from: https://discovery.ucl.ac.uk/id/eprint/10139722/
11. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model cards for model reporting. In: FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency. New York: Association for Coputing Machinery; 2019. p. 220-9.
12. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. In: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Minin. San Francisco; 2016. p.1135-44.
13. Zhang Z, Han X, Liu Z, Jiang X, Sun M, Liu Q. ERNIE: Enhanced Language Representation with Informative Entities. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, July 28- August 2, 2019. c2019 Association for Computational Linguistics: 2019. p.1441-51.
14. Luckin R. Towards artificial intelligence-based assessment systems. Nat Hum Behav. 2017;1(3):1-3.