Arquivos Brasileiros de
Oftalmologia

ORIGINAL ARTICLE

Performance of generative large language models in answering questions from the Brazilian Retina and Vitreous Society certification exam

Adriano Cypriano Faneli[1], Ricardo Danilo Chagas Oliveira[2], Luis Filipe Nakayama[1], Rodrigo Amaral Torres[2], Cristina Muccioli[1], Caio Vinicius Saito Regatieri[1]

1. Universidade Federal de São Paulo, São Paulo, SP, Brazil.
2. Universidade Federal da Bahia, Salvador, BA, Brazil.

ABSTRACT | Purpose: Natural language models and chatbots, particularly OpenAI's Generative Pre-Trained Transformer architecture, have transformed human interaction with digital interfaces. The latest versions, including ChatGPT-4o, offer enhanced functionalities compared to their predecessors. This study evaluates the accuracy of ChatGPT-4, ChatGPT-4o, and Claude 3.5 Sonnet in answering questions from the Brazilian Retina and Vitreous Society certification exam. **Methods:** We compiled 200 multiple-choice questions from the Brazilian Retina and Vitreous Society 2018 and 2019 exams. Questions were categorized into three domains: Anatomy and Physiology of the Retina, Retinal Pathology, and Diagnosis and Treatment. Using a standardized prompt developed according to prompt design guidelines, we tested ChatGPT-4, ChatGPT-4o, and Claude 3.5 Sonnet, recording their first responses as final. Three retina specialists performed a qualitative analysis of the answers. Accuracy was determined by comparing responses to the official correct answers. Statistical analysis was conducted using chi-square tests and Cohen's Kappa. **Results:** Claude 3.5 Sonnet achieved the highest overall accuracy (72.5%), followed by ChatGPT-4o (66.0%) and ChatGPT-4 (55.5%). Claude 3.5 Sonnet and ChatGPT-4o significantly outperformed ChatGPT-4 ($p<0.01$ and $p=0.03$, respectively), while no significant difference was observed between Claude 3.5 Sonnet and ChatGPT-4o ($p=0.16$). Model responses agreed 74.5% of the time, with a Cohen's κ of 0.47. Retinal Pathology was the best-performing domain for all models, whereas Anatomy and Physiology of the Retina and Diagnosis and Treatment were the weakest domains for Claude 3.5 Sonnet and ChatGPT-4, respectively. **Conclusions:** This study is the first to assess Claude 3.5 Sonnet, ChatGPT-4, and ChatGPT-4o in retina specialist certification exams. Claude 3.5 Sonnet and ChatGPT-4o significantly outperformed ChatGPT-4, highlighting their potential as effective tools for studying retina specialist board exams. These findings suggest that the enhanced functionalities of Claude 3.5 Sonnet and ChatGPT-4o offer substantial improvements in medical education contexts.

Keywords: Artificial intelligence; ChatGPT; Retina; Medical education; Ophthalmology, Large language model; Natural language processing

INTRODUCTION

Large language models (LLMs) and chatbots have become powerful tools, transforming human interaction with digital interfaces in both professional and personal contexts[1,2]. LLMs, including OpenAI's GPT, Google's Gemini, Claude, and Microsoft Copilot, enable advanced chatbots to understand language context, generate rational responses, and engage in lifelike conversations[3]. ChatGPT has garnered particular attention due to its accessibility and ability to enhance user experiences[4].

In healthcare, LLMs have broad potential applications that may revolutionize patient care, research, and medical education[5]. ChatGPT has been applied in scientific and medical contexts, including writing abstracts, conducting literature reviews, overcoming language barriers, simplifying reports, supporting decision-making, and assisting with discharge summaries[6-12]. ChatGPT has rapidly advanced from version 3.5 to 4.0 and now to 4o. The newer version, ChatGPT-4o, integrates text and images, accesses real-time information, maintains

Submitted for publication: April 8, 2025
Accepted for publication: October 28, 2025

Funding: This study received no specific financial support.

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

Corresponding author: Adriano Cypriano Faneli.
E-mail: adrianofaneli@gmail.com

Data Availability Statement: The datasets generated and/or analyzed during this study are available from the corresponding author upon reasonable request.

Edited by

Editor-in-Chief: Newton Kara-Júnior
Associate Editor: Dácio C. Costa

longer context windows, generates precise summaries, and understands language more deeply. However, comparative studies measuring its improvement over ChatGPT-4 remain limited. Similarly, research on Claude 3.5 Sonnet's performance in board-style questions is scarce, with existing studies focusing mainly on image analysis[13,14].

ChatGPT has also demonstrated value in education. Wang et al. and Dai et al. reported its effectiveness as a coaching tool, classroom analyzer, and provider of detailed, coherent feedback[15,16]. In ophthalmology, multiple studies have reported satisfactory results using ChatGPT to answer board-style questions[17-20]. However, studies analyzing subspecialist-level questions remain limited. This study evaluates the accuracy of Claude 3.5 Sonnet, ChatGPT-4, and ChatGPT-4o in answering questions from the Brazilian Retina and Vitreous Society certification exam.

METHODS

This study did not require ethics committee approval, as no human subjects were directly involved. All data were obtained from publicly available sources and contained no identifiable patient information. We compiled 200 multiple-choice questions from the sbrv.org website from the 2018 and 2019 exams, the only exams publicly available from June to July 2024. Each question had four options and no images.

Questions were categorized into three domains based on the knowledge required: Anatomy and Physiology of the Retina, Retinal Pathology, and Diagnosis and Treatment.

Natural language models

GPT and Claude are generative AI models designed to interpret and generate text, enabling human-like dialogue. They were trained on diverse text corpora from books, articles, and online content. By minimizing the difference between predicted and actual words, these models generate coherent text according to instructions[21,22]. ChatGPT-4o offers enhanced functionality, including multimodal capabilities and improved personalization[23]. This study compared ChatGPT-4, ChatGPT-4o, and Claude 3.5 Sonnet to assess their accuracy in answering questions from the Brazilian Retina and Vitreous Society exam.

Prompt design

A standardized prompt was used for all models: "You are a retina specialist taking a certification exam in retinal specialization. Only one option is selected as the best choice. Provide the letter corresponding to the correct option, followed by explaining why this is the correct answer." Questions were presented in Brazilian Portuguese. The prompt was refined according to established prompt design guidelines to ensure clarity and minimize ambiguity[24]. Pilot tests confirmed prompt clarity.

Qualitative analysis

Three Brazilian Retina and Vitreous Society-certified ophthalmologists (R.C., L.N., and R.T.) analyzed all correctly answered questions. Responses were categorized into five qualitative levels:

1. **Extremely incorrect:** Completely wrong, with no correct parts.
2. **Partially incorrect:** Mostly wrong, with a few correct parts.
3. **Neutral/ambiguous:** Balanced correct and incorrect elements, or unclear.
4. **Partially correct:** Mostly correct, with minor errors or omissions.
5. **Extremely correct:** Entirely correct, covering all relevant aspects.

Data collection

The first response from each model was recorded as its answer. The interaction was reset after each question to avoid bias from previous responses. Answers were classified as correct or incorrect based on the official exam answers.

Brazilian Retina and Vitreous Society exam requirements

The exam consists of 100 multiple-choice questions. Candidates have five hours to complete it and must achieve a minimum score of 50% to pass. Eligibility requires completion of an ophthalmology residency and a two-year fellowship in Retina and Vitreous.

Statistical analysis

The primary outcome was model accuracy, calculated by comparing each model's selected answer with the correct answer provided by the Brazilian Retina and

Vitreous Society. Chi-square tests were used to assess differences in accuracy between ChatGPT-4, ChatGPT-4o, and Claude 3.5 Sonnet across all questions and within each domain. Fleiss' Kappa measured interrater agreement between the three models and among the three graders. The Kruskal-Wallis H-test, with Dunn-Bonferroni post-hoc analysis, compared the distribution of qualitative classification scores across the models. All statistical analyses were performed using Stata Statistical Software Release 18 (StataCorp LLC, 2023).

RESULTS

As of June 2024, the Brazilian Retina and Vitreous Society had made 200 questions available online. The overall accuracy was 55.5% for ChatGPT-4, 66.0% for ChatGPT-4o, and 72.5% for Claude 3.5 Sonnet. Both ChatGPT-4o and Claude 3.5 Sonnet significantly outperformed ChatGPT-4 ($p=0.03$ and $p<0.01$, respectively), while no significant difference was observed between ChatGPT-4o and Claude 3.5 Sonnet ($p=0.16$; Table 1).

In the 2018 exam, accuracies were 53.0% for ChatGPT-4, 67.0% for ChatGPT-4o, and 67.0% for Claude 3.5 Sonnet. ChatGPT-4o and Claude 3.5 Sonnet significantly outperformed ChatGPT-4 ($p=0.04$; Table 1).

For the 2019 exam, ChatGPT-4 achieved 58.0%, ChatGPT-4o 66.0%, and Claude 3.5 Sonnet 78.0%. Claude 3.5 Sonnet significantly outperformed ChatGPT-4 ($p<0.01$) but not ChatGPT-4o ($p=0.06$). No significant difference was observed between ChatGPT-4o and ChatGPT-4 ($p=0.24$; Table 1).

Fleiss' κ among the three models was 0.44, indicating moderate agreement. Figure 1 presents a Venn diagram of correctly answered questions. Of 200 questions, 90 were correctly answered by all three models. Claude 3.5 Sonnet uniquely answered 20 questions correctly, while ChatGPT-4o and ChatGPT-4 uniquely answered 11 and five questions, respectively. In pairwise overlaps, Claude 3.5 Sonnet and ChatGPT-4o shared 25 correct answers not captured by ChatGPT-4; Claude 3.5 Sonnet and ChatGPT-4 shared 10; ChatGPT-4o and ChatGPT-4 shared 6. Thirty-three questions were missed by all models.

Accuracy by domain was also assessed: Anatomy and Physiology of the Retina, Retinal Pathology, and Diagnosis and Treatment. In the 2018 exam (Figure 2), Retinal Pathology was the best-performing domain for Claude 3.5 Sonnet (74.3%) and ChatGPT-4 (65.7%), while Anatomy and Physiology of the Retina was top for ChatGPT-4o (77.8%). No statistically significant differences were observed across domains ($p>0.05$).

In the 2019 exam (Figure 3), Diagnosis and Treatment was the most accurate domain for Claude 3.5 Sonnet (78.6%), while Retinal Pathology remained highest for ChatGPT-4 (63.4%) and ChatGPT-4o (73.2%). Claude 3.5 Sonnet significantly outperformed ChatGPT-4 in Diagnosis and Treatment ($p=0.01$), while other domain comparisons showed no significant differences ($p>0.05$).

Overall, Retinal Pathology was the best-performing domain for all three models (Figure 4). Claude 3.5 Sonnet (70.4%) significantly outperformed ChatGPT-4 (49.0%) in Diagnosis and Treatment ($p<0.01$). No other statistically significant differences were observed.

The qualitative analysis included 388 correct responses (111 from ChatGPT-4, 132 from ChatGPT-4o, and 145 from Claude 3.5 Sonnet), resulting in 1,164 evaluations by three specialists. Evaluators classified two as "Neutral or ambiguous," 137 as "Partially correct," and 1,025 as "Extremely correct" (Table 2). Fleiss' Kappa among graders was 0.054, indicating slight agreement, mainly due to clustering in two categories.

Table 1. Overall Accuracy per Domain

Total Number of Correctly Answered Questions by Each Model (Percentage Shown in Parentheses)				p-value	p-value	p-value
Exam	Claude 3.5 n (%)	GPT-4o n (%)	GPT-4 *n* (%)	GPT-4o vs. GPT-4)[a]	(GPT-4 vs. Claude 3.5)[a]	(GPT-4o vs. Claude 3.5)[a]
2018 (100 Questions)	67 (67.0)	67 (67.0)	53 (53.0)	0.04	0.04	1
2019 (100 Questions)	78 (78.0)	66 (66.0)	58 (58.0)	0.24	<0.01	0.06
Overall	145 (72.5)	132 (66.0)	111 (55.5)	0.03	<0.001	0.16

[a]Chi-square test.

Combined accuracy of Claude 3.5 Sonnet, ChatGPT-4o, and ChatGPT-4 across the three domains—Anatomy and Physiology of the Retina, Retinal Pathology, and Diagnosis and Treatment—using results from the 2018 and 2019 exams. Bars represent the percentage of correctly answered questions for each model within each domain. Retinal Pathology was the best-performing domain for all three models. Claude 3.5 Sonnet (70.4%) significantly outperformed ChatGPT-4 (49.0%) in the Diagnosis and Treatment domain ($p<0.01$). No other statistically significant differences were observed between models across the domains ($p>0.05$).

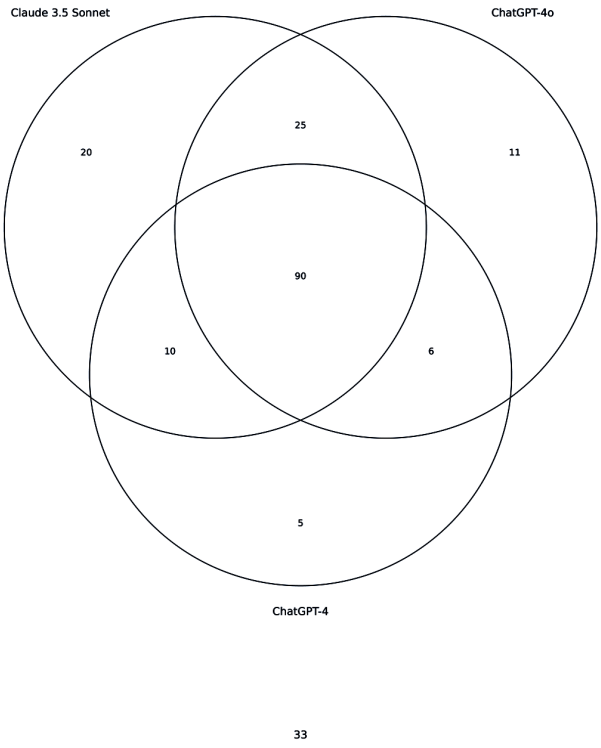

Figure 1. Venn diagram of correct answers by Claude 3.5 Sonnet, ChatGPT-4o, and ChatGPT-4. This Venn diagram illustrates the correct answers provided by Claude 3.5 Sonnet, ChatGPT-4o, and ChatGPT-4. The left circle represents Claude 3.5 Sonnet, the top circle represents ChatGPT-4o, and the bottom circle represents ChatGPT-4. Overlapping areas indicate questions correctly answered by more than one model, with the central overlap showing questions correctly answered by all three models. Numbers outside the overlapping areas represent questions correctly answered by only one model. The number outside all circles represents questions not correctly answered by any model.

The Kruskal-Wallis H-test revealed a significant difference in qualitative classification across models ($p<0.001$). Post-hoc Dunn-Bonferroni analysis showed Claude 3.5 Sonnet received significantly higher ratings than ChatGPT-4o ($p<0.001$) and ChatGPT-4 ($p<0.001$), due to a greater proportion of "Extremely Correct" responses. No significant difference was observed between ChatGPT-4o and ChatGPT-4 ($p=0.57$).

DISCUSSION

In this study, we evaluated the accuracy of Claude 3.5 Sonnet, ChatGPT-4, and ChatGPT-4o in answering 200 questions from the Brazilian Retina and Vitreous Society Retina Specialist Certification Exam. To our knowledge, this is the first study to assess language model performance on a retina specialist certification exam and the first to directly compare ChatGPT-4, ChatGPT-4o, and Claude 3.5 Sonnet on ophthalmology board-style questions.

Our results indicate that ChatGPT-4o and Claude 3.5 Sonnet achieved significantly higher accuracy than ChatGPT-4. The models demonstrated moderate agreement, reflecting consistency in predictions. However, a substantial number of questions were not correctly answered by any model, highlighting opportunities for refinement and alternative approaches. Importantly, all models met the minimum 50% accuracy threshold required for human candidates to pass the Brazilian Retina and Vitreous Society exam.

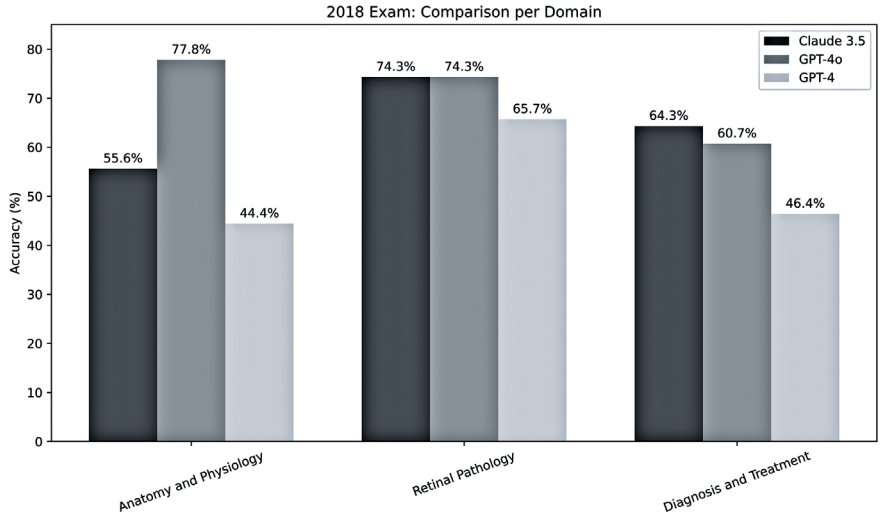

Figure 2. Accuracy per domain in the 2018 exam. Accuracy of Claude 3.5 Sonnet, ChatGPT-4o, and ChatGPT-4 across the three domains—Anatomy and Physiology of the Retina, Retinal Pathology, and Diagnosis and Treatment—in the 2018 exam. Bars represent the percentage of correctly answered questions for each model within each domain. No statistically significant differences were observed between the models across the domains ($p>0.05$).

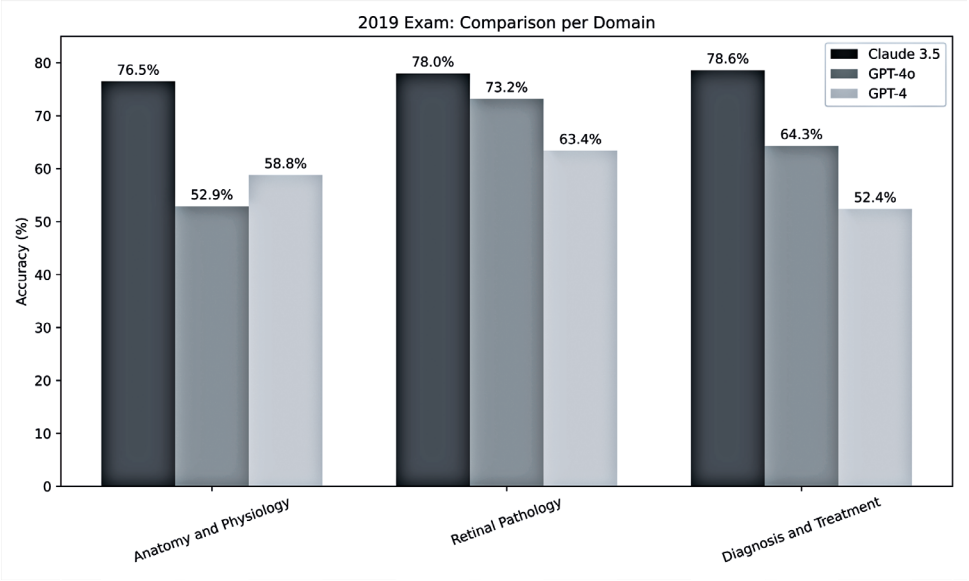

Figure 3. Accuracy per domain in the 2019 exam. Accuracy of Claude 3.5 Sonnet, ChatGPT-4o, and ChatGPT-4 across the three domains—Anatomy and Physiology of the Retina, Retinal Pathology, and Diagnosis and Treatment—in the 2019 exam. Bars represent the percentage of correctly answered questions for each model within each domain. Claude 3.5 Sonnet significantly outperformed ChatGPT-4 in the Diagnosis and Treatment domain (p=0.01), while other differences were not statistically significant (p>0.05).

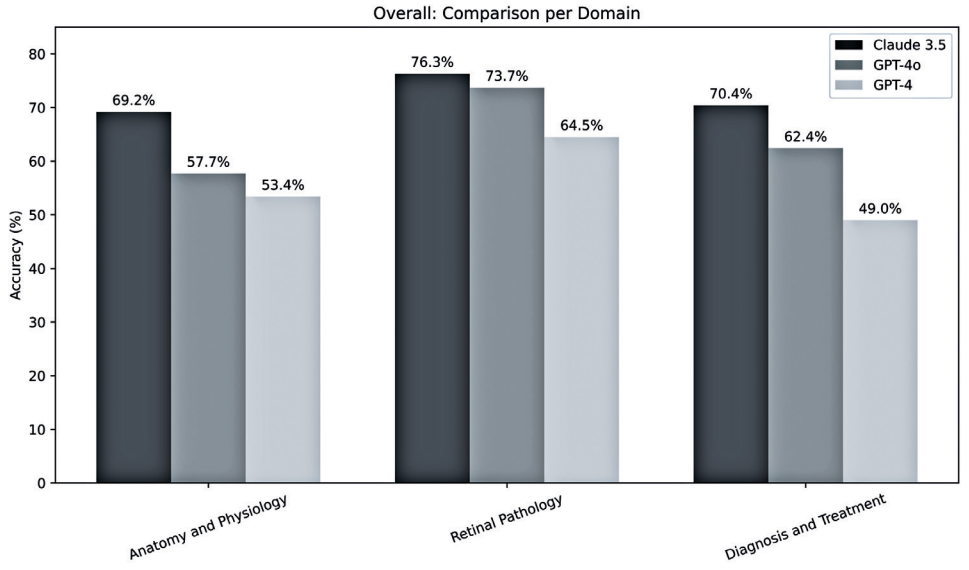

Figure 4. Overall accuracy per domain. Combined accuracy of Claude 3.5 Sonnet, ChatGPT-4o, and ChatGPT-4 across the three domains—Anatomy and Physiology of the Retina, Retinal Pathology, and Diagnosis and Treatment—using results from the 2018 and 2019 exams. Bars represent the percentage of correctly answered questions for each model within each domain. Retinal Pathology was the best-performing domain for all three models. Claude 3.5 Sonnet (70.4%) significantly outperformed ChatGPT-4 (49.0%) in the Diagnosis and Treatment domain (p<0.01). No other statistically significant differences were observed between models across the domains (p>0.05).

Table 2. Distribution of Qualitative Response Ratings for Each Model

	Qualitative Grading Distribution of Responses Across Large Language Models		
Response Ratings	Claude 3.5	GPT-4º	GPT-4
Extremely incorrect	0	0	0
Partially incorrect	0	0	0
Neutral or ambiguous	0	1	1
Partially correct	18	59	60
Extremely correct	417	336	272

The table summarizes the distribution of qualitative ratings for responses generated by Claude 3.5 Sonnet, ChatGPT-4o, and ChatGPT-4. Responses were classified into five categories: extremely incorrect, partially incorrect, neutral or ambiguous, partially correct, and extremely correct. The table presents the total number of responses in each category, highlighting the models' qualitative performance in providing accurate and well-explained answers.

The qualitative analysis provided additional insights into performance differences. Claude 3.5 Sonnet consistently produced higher-quality responses. The Kruskal-Wallis *H*-test revealed significant differences in response ratings among the models ($p<0.001$), and post-hoc Dunn-Bonferroni tests confirmed that Claude 3.5 Sonnet received significantly more "Extremely Correct" ratings compared with ChatGPT-4o ($p<0.001$) and ChatGPT-4 ($p<0.001$). This suggests that Claude 3.5 Sonnet is particularly effective for tasks requiring a deeper understanding of medical concepts, making it well-suited as a study tool for retina specialist certification exams.

Across the three domains—Anatomy and Physiology of the Retina, Retinal Pathology, and Diagnosis and Treatment—the models performed comparably. Retinal Pathology was the best-performing domain for all models, likely due to the factual nature of the questions and the extensive coverage of retinal pathology in the models' training data. Claude 3.5 Sonnet and ChatGPT-4o were comparatively weaker in Anatomy and Physiology of the Retina, while Diagnosis and Treatment was the weakest domain for ChatGPT-4. Notably, Claude 3.5 Sonnet outperformed ChatGPT-4 in the Diagnosis and Treatment domain, likely reflecting its advanced architecture and ability to synthesize information from medical texts.

Previous studies on LLM performance in retina-related questions are limited. Mihalache et al. reported poor performance of ChatGPT-3.5 on 166 retina-related questions from the OphthoQuestions database, with the model failing to answer any correctly[8]. In contrast, Taloni et al. reported higher accuracies for ChatGPT-3.5 (75.8%) and ChatGPT-4 (78.9%) on American Academy of Ophthalmology self-assessment questions[25]. Our study extends these findings by focusing on subspecialty-level questions specifically designed for board certification, which are more challenging than those intended for recent ophthalmology residency graduates.

Despite the promising results, LLM limitations must be acknowledged, particularly in healthcare contexts. These models operate as "black boxes", often providing answers without transparent reasoning or references, and are prone to generating fabricated information (hallucinations)[26-29]. Future research should explore mechanisms to mitigate these limitations.

This study has additional limitations. Only two exam years (2018 and 2019) were publicly available, restricting the sample size. Model performance might vary with a larger or more diverse question set, and domain comparisons may reach statistical significance with additional data. Rapidly evolving LLM versions suggest that future accuracy could surpass current results. Some exam questions may have been included in model training data, limiting generalizability, though the consistency of our results with other studies suggests minimal impact. Lastly, LLM responses cannot be directly compared to human performance, affecting interpretation.

In conclusion, this study provides the first comparative assessment of generative language models on retina specialist certification exams. Claude 3.5 Sonnet and ChatGPT-4o demonstrated superior accuracy and response quality, highlighting their potential as valuable tools in medical education and specialist training.

AUTHORS' CONTRIBUTIONS

Significant Contribution to Conception and Design: Adriano Cypriano Faneli, Luis Filipe Nakayama. **Data Acquisition:** Adriano Cypriano Faneli, Ricardo Danilo Chagas Oliveira, Luis Filipe Nakayama, Rodrigo Amaral Torres. **Data Analysis and Interpretation:** Adriano Cypriano Faneli, Luis Filipe Nakayama. **Manuscript Drafting:** Adriano Cypriano Faneli, Luis Filipe Nakayama, Cristina Muccioli, Caio Vinicius Saito Regatieri. **Significant Intellectual Content Revision of the Manuscript:** Cristina Muccioli, Caio Vinicius Saito Regatieri. **Final Approval of the Submitted Manuscript:** Adriano Cypriano Faneli, Ricardo Danilo Chagas Oliveira, Luis Filipe Nakayama, Rodrigo Amaral Torres, Cristina Muccioli, Caio Vinicius Saito Regatieri.

Statistical Analysis: Adriano Cypriano Faneli, Luis Filipe Nakayama. **Obtaining Funding:** not applicable. **Supervision of Administrative, Technical, or Material Support:** Cristina Muccioli, Caio Vinicius Saito Regatieri. **Research Group Leadership:** Cristina Muccioli, Caio Vinicius Saito Regatieri.

REFERENCES

1. Bart NK, Pepe S, Gregory AT, Denniss AR. Emerging Roles of Artificial Intelligence (AI) in Cardiology: Benefits and Barriers in a 'Bravfet e New World'. Heart Lung Circ. 2023;32(8):883-8.
2. Liu PR, Lu L, Zhang JY, Huo TT, Liu SX, Ye ZW. Application of Artificial Intelligence in Medicine: an Overview. Curr Med Sci. 2021;41(6):1105-15.
3. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595.
4. Sardana D, Fagan TR, Wright JT. ChatGPT: A disruptive innovation or disrupting innovation in academia? J Am Dent Assoc. 2023;154(5):361-4.
5. Wei H, Qiu J, Yu H, Yuan W, editors. MEDCO: Medical Education Copilots Based on a Multi-agent Framework. In: Del Bue A, Canton C, Pont-Tuset J, Tommasi T, editors. Computer Vision - ECCV 2024 Workshops. ECCV 2024. Lecture Notes in Computer Science, vol 15630. Springer, Berlin, 2025, pp. 119-35.
6. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health. 2023;5(3):e107-8.
7. Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y, et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digit Med. 2023;6(1):75.
8. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence Chatbot in ophthalmic knowledge assessment. JAMA Ophthalmol. 2023;141(6):589-97.
9. Singh S, Djalilian A, Ali MJ. ChatGPT and ophthalmology: exploring its potential with discharge summaries and operative notes. Semin Ophthalmol. 2023;38(5):503-7.
10. Ruksakulpiwat S, Kumar A, Ajibade A. Using ChatGPT in medical research: current status and future directions. J Multidiscip Healthc. 2023;16:1513-20.
11. Teixeira da Silva JA. Can ChatGPT rescue or assist with language barriers in healthcare communication? Patient Educ Couns. 2023;115:107940.
12. Rao A, Kim J, Kamineni M, Pang M, Lie W, Dreyer KJ, et al. Evaluating GPT as an adjunct for radiologic decision making: GPT-4 versus GPT-3.5 in a breast imaging pilot. J Am Coll Radiol. 2023;20(10):990-7.
13. Fan KS, Fan KH. Dermatological Knowledge and Image Analysis Performance of Large Language Models Based on Specialty Certificate Examination in Dermatology. Dermato. 2024;4(4):124-35.
14. Kurokawa R, Ohizumi Y, Kanzawa J, Kurokawa M, Sonoda Y, Nakamura Y, et al. Diagnostic performances of Claude 3 Opus and Claude 3.5 Sonnet from patient history and key images in Radiology's "Diagnosis Please" cases. Jpn J Radiol. 2024;42(12):1399-402.
15. Dai W, Lin J, Jin H, Li T, Tsai YS, Gašević D, et al., editors. Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT. 2023 IEEE International Conference on Advanced Learning Technologies (ICALT); 2023 10-13 July 2023.
16. Wang R, Demszky D. Is ChatGPT a Good Teacher Coach? Measuring Zero-Shot Performance For Scoring and Providing Actionable Insights on Classroom Instruction In: Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educacional Applications (BEA 2023). Association for Computational Linguistics. 2023:626-67.
17. Nakayama LF, Gobira MC, Moreira RC, Regatieri CV, Andrade E, Belfort Jr, R. Performance of chatGPT-3.5 answering questions from the Brazilian Council of Ophthalmology Board Examination. Pan Am J Ophthalmol. 2023;5(1):17.
18. Yaïci R, Cieplucha M, Bock R, Moayed F, Bechrakis NE, Berens P, et al. [ChatGPT and the German board examination for ophthalmology: an evaluation]. Ophthalmologie. 2024;121(7):554-64. German.
19. Haddad F, Saade JS. Performance of ChatGPT on Ophthalmology-related questions across various examination levels: observational study. JMIR Med Educ. 2024;10:e50842.
20. 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.
21. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inform Process Syst. 2020;33:1877-901. https://doi.org/10.48550/arXiv.2005.14165
22. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CI, Mishkin P, et al. Training language models to follow instructions with human feedback. Adv Neural Inform Process Syst. 2022;35:27730-27744. https://doi.org/10.48550/arXiv.2203.02155
23. OpenAI. Hello GPT-4o. 2024. Available from: https://openai.com/index/hello-gpt-4o/
24. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res. 2023;25:e50638.
25. Taloni A, Borselli M, Scarsi V, Rossi C, Coco G, Scorcia V, et al. Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology. Sci Rep. 2023;13(1):18562.
26. Naveed H, Khan AU, Qiu S, Saqib M, Anwar S, Usman M, et al. A comprehensive overview of large language models. ACM Trans Intel Sys Technol. 2024;16(5):1-72.
27. Nazi ZA, Peng W. Large language models in healthcare and medical domain: a review. Informatics (MDPI). 2024;11(3):57.
28. Bhattacharyya M, Miller VM, Bhattacharyya D, Miller LE. High rates of fabricated and inaccurate references in chatgpt-generated medical content. Cureus. 2023;15(5):e39238.
29. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. Sci Rep. 2023;13(1):14045.