Assessing the Medical Reasoning Skills of GPT-4 in Complex Ophthalmology Cases

Daniel Milad, MD[1,2], Fares Antaki, MDCM[1,3,4], Jason Milad[5], Andrew Farah[6], Thomas Khairy[6], David Mikhail[7], Charles-Édouard Giguère[8], Samir Touma, MDCM[1,2], Allison Bernstein, MD[1,2], Andrei Szigiato[9], Taylor Nayman[1,2], Guillaume Mullie[10], Renaud Duval, MD[1,2]

1. Department of Ophthalmology, University of Montreal, Montreal, Quebec, Canada
2. Department of Ophthalmology, Hôpital Maisonneuve-Rosemont, Montreal, Quebec, Canada
3. Institute of Ophthalmology, University College London, London, United Kingdom
4. CHUM School of Artificial Intelligence in Healthcare (SAIH), Centre Hospitalier de l'Université de Montréal (CHUM), Montreal, Quebec, Canada
5. Department of Software Engineering, University of Waterloo, Waterloo, Ontario, Canada
6. Department of Medicine, McGill University, Montreal, Quebec, Canada
7. Department of Medicine, University of Toronto, Toronto, Ontario, Canada
8. Institut universitaire en santé mentale de Montréal (IUSMM), Montreal, Quebec, Canada
9. Department of Ophthalmology, Hôpital du Sacré-Coeur de Montréal, Montreal, Quebec, Canada
10. Cité-de-la-Santé Hospital, Laval, Quebec, Canada

ORCID of the authors: Daniel Milad (0000-0002-0693-3421), Fares Antaki (0000-0001-6679-7276), David Mikhail (0009-0009-0831-1915), Samir Touma (0000-0002-6365-0946), Renaud Duval (0000-0002-3845-3318)

Corresponding authors:
Renaud Duval, MD
Department of Ophthalmology, Université de Montréal, Montreal, Quebec, Canada.
renaud.duval@gmail.com

Funding: None.
Competing interests: None
Ethics Approval: Ethics approval was not required for this project.
Patient Consent: Patient consent was not required as this work did not involve patients.

Word count: 3005; abstract 249, **Tables**: 1, **Figures**: 4, **References**: 27, **Supplemental Material**: Tables/Figures 1/1

Keywords: artificial intelligence; foundation models; medical education; GPT-4; ChatGPT; ophthalmology; Generative Pretrained Transformer; clinical accuracy

SYNOPSIS/PRECIS

GPT-4 demonstrates strong diagnostic and decision-making accuracy in complex ophthalmology cases.

ABSTRACT

Background/Aims

This study assesses the proficiency of Generative Pre-trained Transformer (GPT)-4 in answering questions about complex clinical ophthalmology cases.

Methods

We tested GPT-4 on 422 *JAMA* Ophthalmology Clinical Challenges, and prompted the model to determine the diagnosis (open-ended question) and identify the next-step (multiple-choice question). We generated responses using two zero-shot prompting strategies, including Zero-Shot Plan-and-Solve+ (PS+), to improve the reasoning of the model. We compared the best performing model to human graders in a benchmarking effort.

Results

Using PS+ prompting, GPT-4 achieved mean accuracies of 48.0% (95% CI [43.1%, 52.9%]) and 63.0% (95% CI [58.2%, 67.6%]) in diagnosis and next step, respectively. Next-step accuracy did not significantly differ by subspecialty (p=0.44). However, diagnostic accuracy in Pathology and Tumors was significantly higher than in Uveitis (p=0.027). When the diagnosis was accurate, 75.2% (95% CI [68.6%, 80.9%]) of the next steps were correct. Conversely, when the diagnosis was incorrect, 50.2% (95% CI [43.8%, 56.6%]) of the next steps were accurate. The next step was three times more likely to be accurate when the initial diagnosis was correct (p<0.001). No significant differences were observed in diagnostic accuracy and decision-making between board-certified ophthalmologists and GPT-4. Amongst trainees, senior residents outperformed GPT-4 in diagnostic accuracy (p = <0.001 and 0.049) and in accuracy of next step (p = 0.002 and 0.020).

Conclusion

Improved prompting enhances GPT-4's performance in complex clinical situations, although it does not surpass ophthalmology trainees in our context. Specialized LLMs hold promise for future assistance in medical decision-making and diagnosis.

KEY MESSAGES

What is already known on this topic

Clinicians are exploring the use of large language models (LLMs) like Generative Pre-trained Transformer (GPT) to improve diagnostic accuracy and clinical decision-making in medicine, notably in ophthalmology. Studies show that GPT-4 outperforms previous models in ophthalmology question banks, but its text generation method reveals limitations in critical thinking. Early research using ophthalmology case reports suggests a high agreement between LLMs and experts, yet the application of LLMs in a large set of ophthalmology clinical challenges remains unexplored.

What this study adds

This study assesses GPT-4's performance on ophthalmological cases featured in the Journal of the American Medical Association (*JAMA*) Ophthalmology Clinical Challenges section, showcasing its diagnostic and decision-making capabilities. It also evaluates the efficacy of various prompting strategies and positions GPT-4's performance in relation to ophthalmology trainees.

How this study might affect research, practice, or policy

This study underscores the potential of LLMs within ophthalmology, suggesting a future where AI complements clinical expertise. By demonstrating that GPT-4 can achieve commendable performance in complex ophthalmology cases, this study may catalyze the discussion on integrating AI in clinical decision support systems and encourage policy frameworks that facilitate the responsible deployment of LLMs in patient care.

INTRODUCTION

Globally, clinicians and scientists alike are contemplating the potential uses of large language models (LLMs) in improving diagnostic precision and supporting clinical decision-making processes. (1) LLMs, which represent fine-tuned foundation models trained on large datasets, can produce coherent text and demonstrate complex reasoning capabilities. (2–5) Generative Pre-trained Transformer (GPT)-4 currently sets the industry standard in the LLM domain, showing considerable improvements over its predecessors in the medical domain. (4) Notably, GPT-4's diagnostic and clinical decision-making abilities seem to be enhanced as it continues to learn. (5)

In ophthalmology, our group has previously studied the performance of GPT in medical question-answering. We have shown that GPT-4 can achieve an accuracy of 72.9% on the large ophthalmology question banks, outperforming GPT-3.5 by 18.3%. (6,7) Since our original work, numerous subsequent studies have corroborated our findings.(6–11) We have also shown that GPT-4 performs best in recall questions compared to ones involving clinical decision-making. (7) Thus, the ability of LLMs to engage in true critical thinking, beyond simply generating text by predicting the next most probable word, or “token”, remains to be determined. (5)

Evaluating the performance of LLMs on diagnosing case reports from the literature may be useful to determine how well they can handle complex, real-world medical cases. To date, only a handful of studies have studied that in ophthalmology, with sample sizes between 11 and 22 cases covering neuro-ophthalmology, glaucoma, and cornea. (12–14) These initial findings indicate a high level of agreement between LLMs and experts, highlighting a potential role for LLMs in clinical decision-making.

In this work, we explore the performance of GPT-4 in answering questions about complex ophthalmological cases published in the Journal of the American Medical Association (*JAMA*) Ophthalmology Clinical Challenges section. These reports represent challenging ophthalmological cases, where clinicians attempt to determine the diagnosis (open-ended) and the best next diagnostic or treatment step (multiple-choice question). We explore multiple prompting strategies to enhance the performance model. We then compare this performance to the accuracy of ophthalmology trainees as a benchmark.

MATERIALS AND METHODS

JAMA Ophthalmology’s Clinical Challenges

In July 2023, GPT-4 was prompted using 422 case studies from *JAMA* Ophthalmology’s Clinical Challenges section. These case studies were designed to assess both diagnostic prediction and identification of the best next step using a multiple-choice question. The challenges were classified in one of the 13 ophthalmology subspecialties, as categorized by the American Academy of Ophthalmology in their Basic and Clinical Science Course. (15) The study title, case, and figure descriptions were provided to GPT-4 and human graders. Figures were excluded as GPT-4 could not process images at the time of writing (August

2023). Discussions were also excluded to avoid data leakage, as the answers were often revealed in this section.

GPT-4 Access and Parameters

We accessed GPT-4, OpenAI's latest LLM, using the Application Programming Interface (API). (4) This allowed us to design customized automated mass prompting techniques using Google Sheets. The API, unlike the ChatGPT web application, guarantees data privacy by not using user data to enhance the GPT model. Furthermore, GPT-4's "temperature", referring to the degree of randomness in its responses when given identical prompts, was set to 0.3. The temperature scale goes from 0 to 1, with 0 yielding the most conservative responses, and 1 yielding highly creative responses. Although the ideal temperature has not yet been defined for this use case, our most recent paper determined that a temperature of 0.3 achieved the highest accuracy. (4,7)

Prompt Engineering

The "What to Do Next?" questions from *JAMA* Ophthalmology's Clinical Challenges follow a standardized multiple-choice question format, with one correct option and three incorrect options (distractors). The exact same information (case report, multiple-choice question, and answer options) was provided to GPT-4 and human graders.

Recent studies have shown that different strategies of zero-shot prompting lead to different results. (16) Thus, we compared the use of two zero-shot prompting strategies : the first consisted of what our team collectively agreed would be most logical, whilst the second consisted of a Zero-Shot Plan-and-Solve + (PS+) prompt (**Figure 1**). Proposed by Wang et al., Zero-Shot-PS+ prompting consists of asking GPT to build a plan to divide the task into smaller subtasks, to then be able to carry out the subtasks with detailed instructions. Although the original Plan-and-Solve (PS) prompting strategy described uses similar methodology, it suffers from calculation errors and low quality reasoning steps. (16) In PS+, more detailed instructions address these weaknesses. PS+ demonstrates superiority over PS and basic Zero-Shot Chain-of-Thought (CoT) strategies, such as the "Let's think step by step" prompt. (16)

Human Benchmarking

Since historical data on human performance is not publicly available on *JAMA* Ophthalmology, three practicing board-certified ophthalmologists and three ophthalmology trainees were recruited to answer five randomly selected clinical challenges from each of the 13 ophthalmology subspecialties. The ophthalmologists specialised in comprehensive ophthalmology, glaucoma and medical retina. The trainees had various levels of training: postgraduate years two, three and four. We compared the results of human graders to GPT-4 on the same subset of clinical challenges to contextualize our findings.

Statistical Analysis

We compared GPT-4 answers to those provided by *JAMA* Ophthalmology. When grading the open-ended diagnosis questions, we prioritized specificity in evaluating correct answers. Initially, three junior trainees jointly assessed the answers. Answers were deemed correct if both the general primary diagnosis and the specific etiology of subtype were correct. For example, if the specific etiology was "acute posterior multifocal placoid pigment epitheliopathy", mentioning only the general primary diagnosis like "posterior uveitis" was deemed insufficient and marked incorrect. In another example from our dataset, if the

specific etiology was “UL97- and UL54-resistant Cytomegalovirus retinitis”, mentioning “Cytomegalovirus retinitis” was marked as correct. When junior trainees were unsure, further adjudication was performed by a senior clinician. The efficacy of both zero-shot prompting strategies was evaluated using Generalized Estimating Equations (GEE), considering the overlap in question sets. The GEE, facilitated by the *geepack* package, accommodated for data correlation, with significant findings further examined via post-hoc analysis and Dunnet’s method for p-value adjustment. Logistic regression allowed us to study the influence of subspecialty on accuracy. All analyses were conducted with R version 4.3.1 using a 5% significance threshold.

The same approach was employed to compare the performance of both GPT-4 prompting strategies to human graders. Human grader concordance was quantified using kappa statistics, with kappa values interpreting agreement levels. Kappa can be interpreted as 0-0.2 none to slight agreement, 0.21-0.4 fair agreement, 0.41-0.6 moderate agreement, 0.61-0.8 substantial agreement, and 0.81-1.00 near perfect agreement. In this section, GPT-4 was tested on the subset of clinical challenges that underwent human grading; thus, the accuracy reported may differ slightly from the ones reported in the previous section.

RESULTS

Within the collection of 422 Clinical Challenges, the sections on Retina and Vitreous, Uveitis, and Neuro-Ophthalmology were notably popular, comprising 23% (96/422), 16% (67/422), and 16% (67/422) of the total, respectively. No challenges were published on the topics of Refractive Surgery, Clinical Optics and Fundamentals (**Supplemental Figure 1**).

Traditional Zero-Shot GPT-4 Prompting

Using traditional Zero-Shot prompting strategies, GPT-4 achieved mean accuracies of 41.5% (95% confidence interval (CI) [36.8%, 46.3%]) and 60.4% (95% CI [55.6%, 65.1%]) in diagnosis and next step, respectively. Diagnostic and next-step accuracy did not significantly differ by subspecialty ($p=0.13$ and $p=0.41$, respectively).

We observed the following patterns: when the diagnosis was accurate, 74.9% (95% CI [67.6%, 81.0%]) of the next steps were correct. Conversely, when the diagnosis was incorrect, 50.2% (95% CI [43.8%, 56.6%]) of the next steps were accurate. The next step was three times more likely to be accurate when the initial diagnosis is correct ($p<0.001$). This was seen amongst all subspecialty cases, with no significant differences found between subspecialties ($p=0.41$).

GPT-4 Zero-Shot Plan-and-Solve + Prompting outperforms Traditional Zero-Shot Prompting

Using Zero-Shot PS+ prompting, GPT-4 achieved mean accuracies of 48.0% (95% CI [43.1%, 52.9%]) and 63.0% (95% CI [58.2%, 67.6%]) in diagnosis and next step, respectively (**Figure 2**). Next-step accuracy did not significantly differ by subspecialty ($p=0.44$). However, diagnostic accuracy in Pathology and Tumors was significantly higher than in Uveitis ($p=0.027$).

When the diagnosis was accurate, 75.2% (95% CI [68.6%, 80.9%]) of the next steps were correct. Conversely, when the diagnosis was incorrect, 50.2% (95% CI [43.8%, 56.6%]) of the next steps were accurate (**Figure 3**). The next step remained approximately three times more likely to be accurate when the initial diagnosis was correct ($p<0.001$).

Across all subspecialty challenges, Zero-Shot PS+ prompting outperformed traditional Zero-Shot prompting in diagnostic accuracy ($p = 0.006$), but did not show a statistically significant difference in accuracy for determining the next step ($p = 0.18$) (**Table 1**). There was no observed subspecialty effect in the relationship for diagnosis ($p = 0.13$) or the next step ($p = 0.89$) (**Supplemental Table 1**).

GPT-4 Versus Ophthalmologists and Ophthalmology Trainees

We then compared the performance of GPT-4 to the six human graders. Since the overall agreement amongst the board-certified ophthalmologists was moderate to substantial (kappa = 0.66, 95% CI [0.44, 0.86] for diagnostic accuracy and kappa = 0.63, 95% CI [0.39, 0.85] for next step), the comparison with GPT-4 was done with each ophthalmologist separately. There were no statistically significant differences in diagnostic performance when comparing ophthalmologists to GPT-4 Zero-Shot PS+, with respective accuracies of 48.9% ($p = 0.562$), 59.6% ($p = 0.649$), and 68.1% ($p = 0.477$). Similarly, there was no statistically significant differences in performance for next step determination, with respective scores of 59.6% ($p = 0.998$), 59.6% ($p = 0.998$), and 72.3% ($p = 0.416$). (**Figure 4**)

The agreement amongst trainee graders was low (kappa = 0.45, 95% CI [0.29, 0.62] for next step and kappa = 0.27, 95% CI [0.10, 0.43] for diagnostic accuracy), and as such the comparison with GPT-4 was also done with each trainee separately. Both senior residents significantly outperformed GPT-4 Zero-Shot PS+ in diagnostic performance, with respective accuracies of 78.7% ($p = 0.049$) and 85.1% ($p < 0.001$). Similarly, both senior residents outperformed in next step determination, with respective accuracies of 78.7% ($p = 0.020$) and 85.1% ($p = 0.002$). There were no significant differences in diagnostic performance and next step determination when compared with the junior resident, with respective accuracies of 51.1% ($p = 0.75$) and 57.4% ($p = 1.00$). (**Figure 4**)

DISCUSSION

In this study, we demonstrate that enhanced prompting strategies can improve GPT-4's performance, that GPT-4 performs well in complex clinical scenarios and that GPT-4 does not currently outperform ophthalmology trainees. We selected GPT-4 as the state-of-the-art LLM for this study since it has been shown to outperform its predecessors and other publicly available LLMs such as Google Bard and Claude-2. (17,18)

Enhanced prompting techniques have demonstrated their potential to augment the performance of GPT-4. (16,19) While there are infinite prompting strategies—like few-shot chain-of-thought prompting, which provides several exemplary chains of thought to the model through multiple prompts sent by the user—such strategies were incompatible with the constraints of ChatGPT's API, which is currently only designed for Zero-Shot prompting. However, Zero-Shot prompting can be refined to improve GPT's accuracy. A novel advancement in this area, Zero-Shot PS+, entails directing the LLM to formulate a strategy

by breaking down the main task into simpler subtasks, executing these with meticulous logical instructions. (16) The enhancement of GPT's performance by various prompting strategies points out a major limitation in our current methods of evaluating LLMs. Since testing all possible prompting strategies is unrealistic, there is a pressing need for standard frameworks and guidelines to evaluate LLMs in medicine.

With the implementation of Zero-Shot PS+, GPT-4 achieved a diagnostic accuracy of 48% and was 63% accurate in identifying the most appropriate next step. Diagnostic accuracy was significantly higher in Pathology and Tumors than in Uveitis ($p=0.027$), possibly due to the complex and often difficult diagnoses in uveitis. The likelihood of subsequent step accuracy was tripled when the initial diagnosis was correct, a trend that held across various subspecialty cases. This was likely the case due to GPT's method of generating content by predicting the next most probable "token". (5) While conjectural, this may indicate that GPT-4's proficiency lies in its ability to recall information and draw rapid inferences, rather than in iterative reasoning or reevaluation of decisions as new information, such as multiple-choice answer options, is presented. Consequently, GPT-4 appears predisposed to determining the optimal next step based on its initial diagnosis, without reconsidering this decision in light of subsequent information.

Prior research has explored GPT's utility in analyzing ophthalmology case reports on a limited basis. Madadi et al. detailed GPT's concordance with neuro-ophthalmologists in 22 case reports, highlighting strong alignment with experts. (12) Delsoz et al. evaluated GPT's performance on 11 glaucoma cases, with findings indicative of a diagnostic precision comparable to that of senior ophthalmology residents. (13) Lastly, Delsoz et al. explored GPT-4's application to 20 cornea case reports, showcasing once again a robust performance. (14) Collectively, these studies signal a growing interest and recognition of GPT's potential in ophthalmological evaluations. The diagnostic accuracy of GPT-4 in these studies ranged from 72.7%-85%, higher than our achieved combined diagnostic accuracy of 48%. The limited number of cases in these studies makes further comparison challenging. Beyond ophthalmology, interest persists: a study testing GPT-4 Vision (GPT-4V) on general medical cases found it outperformed physicians in 934 cases, but its performance declined when images were introduced. (20)

Since no historical human performance metrics were published on these clinical challenges, we created a human benchmark for performance comparison. We used the performances of ophthalmology trainees at varying educational stages—first through third years—and of practicing, board-certified ophthalmologists as a benchmark. This approach was chosen to capture a snapshot of the progression in clinical proficiency and to contextualize GPT-4's performance within the current landscape of clinical learning. Within our limited comparative framework, we observed poor agreement among trainees and higher consensus among ophthalmologists. This suggests more variability in performance among trainees, reflecting the nature of continuous learning during residency. With each year of residency representing a significant jump in knowledge, senior residents performed better as they approached board examinations. Surprisingly, GPT-4 performed similarly to the ophthalmologists, a noteworthy finding that should be interpreted with caution due to the limited sample size. Also, both senior trainees outperformed GPT-4 and the consultants included in our study. This discrepancy could be attributed to a potential sampling bias where the trainees may have been preparing for upcoming examinations, making them more familiar with the specific

minutiae often presented in these cases. Additionally, it is crucial to note that two of the three ophthalmologists are subspecialists, possibly contributing to their exposure being more focused and distant from other subspecialties, unlike trainees who are currently undergoing broader training. Furthermore, the complexity of the cases might have influenced those with more clinical experience to answer based on their real-world experiences rather than adhering strictly to textbook answers, which residents are more exposed to. This adds another layer of complexity to the interpretation of our findings. The future potential for GPT-4 or subsequent language models to equal or surpass the proficiency of senior trainees—or even experienced ophthalmologists—remains a provocative and open question. Given the fast pace of innovation in this domain, it is plausible to conjecture that these models may soon approximate the diagnostic capabilities of human clinicians.

Since *JAMA* Ophthalmology's Clinical Challenges are behind a paywall, it is likely that GPT was not trained on this data. However, due to the opaque nature of GPT's training dataset, we can not know for certain. If the task training examples from *JAMA Ophthalmology* were included in GPT's pre-training data, this would introduce the risk of task contamination, disrupting this study's zero-shot nature. Recent findings demonstrate that for classification tasks with no possibility of task contamination, LLMs rarely exhibit noteworthy enhancements in both zero-shot and few-shot methodologies. (21) This critical limitation, when applied broadly to LLMs, may constrain their overall potential, revealing that they may not evolve and learn as rapidly as initially speculated. While our results are promising, they should not be misconstrued as an indication that GPT-4's operational proficiency is equivalent to that of an ophthalmologist. Performance on online clinical challenges, much like for physicians, does not encompass the full spectrum of the practice of medicine. Soft skills, such as communication, professionalism and bedside manner all represent essential skills which are not accounted for in this evaluation. (22,23) Our study design intentionally focused on using a single, highly vetted dataset with a large sample size. The dataset from *JAMA* Ophthalmology, with their rigorous review process and low acceptance rate, ensures a high level of quality. However, it is important to acknowledge that this single dataset may suffer from publication bias, potentially containing more impressive cases than what ophthalmologists encounter in their daily practice. This highlights the need for the development and availability of new benchmarking datasets for research purposes in ophthalmology.

In ophthalmology, a specialty that heavily relies on imaging, the forthcoming GPT-4 Vision, an extension of the GPT-4 model, aims to add visual information processing, representing a significant step towards creating Large Multimodal Models (LMMs) that can handle the complexities of medical data. (24) This advancement could revolutionize our approach, allowing us to include image data from Clinical Challenges in our evaluations. Specialized foundation models designed for ophthalmology are expected to greatly influence our field, and we are now starting to see their emergence. (25) In the future, accuracy, safety and validity of these specialized LLMs will need to be assessed before considering clinical implementation. (26) Lastly, there is a growing trend towards using LLMs for generating differential diagnoses, employing a few-shot prompting technique characterized by multiple prompts, each adding new clinical information to iteratively refine the final list of potential diagnoses. This approach will likely offer the greatest utility to clinicians and should be prioritized in future projects. (27)

To conclude, GPT-4's performance on complex clinical challenges in ophthalmology is promising, although it does not yet rival the expertise of human trainees. Currently, it will likely play a strong role in educational settings, suggesting a valuable role for specialized LLMs in the future of medical decision assistance.

Data sharing statement: All data produced in the present study is available upon reasonable request to the authors. *JAMA* Ophthalmology's Clinical Challenges are proprietary and access is available to be distributed through their website.

Contributions: Conception and design of the study (DM, FA, RD); data collection (DM, FA, JM, AF, TK, DM2, ST, AB, AS, TN, GM); data analysis (DM, FA, JM, DM2, CEG); writing of the manuscript and preparation of figures (DM, FA, JM, DM2); supervision (RD); review and discussion of the results (all authors); edition and revision of the manuscript (all authors).

REFERENCES

1. Betzler BK, Chen H, Cheng CY, Lee CS, Ning G, Song SJ, et al. Large language models and their impact in ophthalmology. Lancet Digit Health. 2023 Dec 1;5(12):e917–24.
2. Nath S, Marie A, Ellershaw S, Korot E, Keane PA. New meaning for NLP: the trials and tribulations of natural language processing with GPT-3 in ophthalmology. Br J Ophthalmol. 2022 Jul;106(7):889–92.
3. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language Models are Few-Shot Learners. In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2020 [cited 2023 Oct 24]. p. 1877–901. Available from: https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html
4. OpenAI. GPT-4 Technical Report [Internet]. arXiv; 2023 [cited 2023 Oct 24]. Available from: http://arxiv.org/abs/2303.08774
5. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, et al. Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study. J Med Internet Res. 2023 Aug 22;25:e48659.
6. 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 Dec 1;3(4):100324.
7. Antaki F, Milad D, Chia MA, Giguère CÉ, Touma S, El-Khoury J, et al. Capabilities of GPT-4 in ophthalmology: an analysis of model entropy and progress towards human-level medical question answering. Br J Ophthalmol [Internet]. 2023 Nov 3 [cited 2023 Nov 4]; Available from: https://bjo.bmj.com/content/early/2023/11/02/bjo-2023-324438
8. Teebagy S, Colwell L, Wood E, Yaghy A, Faustina M. Improved Performance of ChatGPT-4 on the OKAP Exam: A Comparative Study with ChatGPT-3.5 [Internet]. medRxiv; 2023 [cited 2023 Oct 24]. p. 2023.04.03.23287957. Available from: https://www.medrxiv.org/content/10.1101/2023.04.03.23287957v1
9. 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 Oct;254:141–9.
10. Mihalache A, Popovic MM, Muni RH. Performance of an Artificial Intelligence Chatbot in Ophthalmic Knowledge Assessment. JAMA Ophthalmol. 2023 Jun 1;141(6):589–97.
11. Raimondi R, Tzoumas N, Salisbury T, Di Simplicio S, Romano MR. Comparative analysis of large language models in the Royal College of Ophthalmologists fellowship exams. Eye. 2023 May 9;1–4.
12. Madadi Y, Delsoz M, Lao PA, Fong JW, Hollingsworth TJ, Kahook MY, et al. ChatGPT Assisting Diagnosis of Neuro-ophthalmology Diseases Based on Case Reports. MedRxiv Prepr Serv Health Sci. 2023 Sep 14;2023.09.13.23295508.
13. Delsoz M, Raja H, Madadi Y, Tang AA, Wirostko BM, Kahook MY, et al. The Use of ChatGPT to Assist in Diagnosing Glaucoma Based on Clinical Case Reports. Ophthalmol Ther. 2023 Sep 14;
14. Delsoz M, Madadi Y, Munir WM, Tamm B, Mehravaran S, Soleimani M, et al. Performance of ChatGPT in Diagnosis of Corneal Eye Diseases. MedRxiv Prepr Serv Health Sci. 2023 Aug 28;2023.08.25.23294635.
15. McCannel CA, Bhatti MT. The Basic and Clinical Science Course of the American Academy of Ophthalmology: The 50th Anniversary of a Unicorn Among Medical Textbooks. JAMA Ophthalmol. 2022 Mar 1;140(3):225–6.
16. Wang L, Xu W, Lan Y, Hu Z, Lan Y, Lee RKW, et al. Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models [Internet]. arXiv; 2023 [cited 2023 Oct 24]. Available from: http://arxiv.org/abs/2305.04091
17. Espejel JL, Ettifouri EH, Alassan MSY, Chouham EM, Dahhane W. GPT-3.5, GPT-4,

or BARD? Evaluating LLMs Reasoning Ability in Zero-Shot Setting and Performance Boosting Through Prompts [Internet]. arXiv; 2023 [cited 2024 Jan 22]. Available from: http://arxiv.org/abs/2305.12477
18. Hochmair HH, Juhasz L, Kemp T. Correctness Comparison of ChatGPT-4, Bard, Claude-2, and Copilot for Spatial Tasks [Internet]. arXiv; 2024 [cited 2024 Jan 22]. Available from: http://arxiv.org/abs/2401.02404
19. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models [Internet]. arXiv; 2023 [cited 2023 Nov 4]. Available from: http://arxiv.org/abs/2201.11903
20. Buckley T, Diao JA, Rodman A, Manrai AK. Accuracy of a Vision-Language Model on Challenging Medical Cases [Internet]. arXiv; 2023 [cited 2023 Nov 25]. Available from: http://arxiv.org/abs/2311.05591
21. Li C, Flanigan J. Task Contamination: Language Models May Not Be Few-Shot Anymore [Internet]. arXiv; 2023 [cited 2024 Jan 22]. Available from: http://arxiv.org/abs/2312.16337
22. Hamel P, Boisjoly H, Corriveau C, Fallaha N, Lahoud S, Luneau K, et al. Using the CanMEDS roles when interviewing for an ophthalmology residency program. Can J Ophthalmol J Can Ophtalmol. 2007 Apr;42(2):299–304.
23. Ha JF, Longnecker N. Doctor-Patient Communication: A Review. Ochsner J. 2010;10(1):38–43.
24. Yang Z, Li L, Lin K, Wang J, Lin CC, Liu Z, et al. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision) [Internet]. arXiv; 2023 [cited 2023 Nov 27]. Available from: http://arxiv.org/abs/2309.17421
25. Zhou Y, Chia MA, Wagner SK, Ayhan MS, Williamson DJ, Struyven RR, et al. A foundation model for generalizable disease detection from retinal images. Nature. 2023 Oct;622(7981):156–63.
26. Eriksen AV, Möller S, Ryg J. Use of GPT-4 to Diagnose Complex Clinical Cases. NEJM AI [Internet]. 2023 Nov 9 [cited 2023 Nov 27]; Available from: https://onepub-media.nejmgroup-production.org/ai/media/ec2de32e-9aa9-49f0-8f37-45becf6be3ed.pdf
27. McDuff D, Schaekermann M, Tu T, Palepu A, Wang A, Garrison J, et al. Towards Accurate Differential Diagnosis with Large Language Models [Internet]. arXiv; 2023 [cited 2023 Dec 9]. Available from: http://arxiv.org/abs/2312.00164

TABLES

Table 1. Comparison of GPT-4 prompting strategy accuracy.

Prompting Strategy	Diagnostic Accuracy	Next Step Accuracy
Traditional Zero-Shot	41.5% [36.9 %, 46.2 %]	60.4 % [55.7 %, 65.0 %]
Zero-Shot PS+	47.9% [43.2 %, 52.7 %]	63.0 % [58.3 %, 67.5 %]

Table 1 presents the mean accuracy followed by the 95% confidence interval in brackets. Across all subspecialty challenges, Zero-Shot PS+ prompting outperformed traditional Zero-Shot prompting in diagnostic accuracy ($p = 0.006$), but did not show a statistically significant difference in accuracy for determining the next step ($p = 0.18$).

FIGURE LEGENDS

Figure 1. GPT-4 Zero-Shot Prompting Strategies.

The text in brackets (title, case, figure description and answers) vary per clinical challenge. The lead-in prompt and question remain the same.

Figure 2. GPT-4 Zero-Shot Plan-and-Solve+ Prompting Accuracy by Subspecialty.

Using Zero-Shot PS+ prompting, GPT-4 achieved mean accuracies of 48.0% (95% CI [43.1%, 52.9%]) and 63.0% (95% CI [58.2%, 67.6%]) in diagnosis and next step, respectively.

Figure 3. Accuracy of GPT-4 Zero-Shot Plan-and-Solve+ prompting next step predictions based on correctness of initial diagnosis.

When the diagnosis was accurate, 75.2% (95% CI [68.6%, 80.9%]) of the next steps were correct. Conversely, when the diagnosis was incorrect, 50.2% (95% CI [43.8%, 56.6%]) of the next steps were accurate. The next step remained approximately three times more likely to be accurate when the initial diagnosis was correct ($p<0.001$).

Figure 4. Performance of GPT-4 Zero-Shot Plan-and-Solve + prompting compared to human performance.

There were no significant differences in diagnostic performance nor next step determination when comparing ophthalmologists and junior residents to GPT-4 Zero-Shot PS+. However, senior residents significantly outperformed GPT-4 Zero-Shot PS+ in diagnostic performance and in next step determination.

GPT Generative pre-trained transformer; *PGY* Ophthalmology Residency Postgraduate Year