BRIEF REPORT

What can GPT-4 do for Diagnosing Rare Eye Diseases? A Pilot Study

Xiaoyan Hu · An Ran Ran · Truong X. Nguyen · Simon Szeto ·

Jason C. Yam · Carmen K. M. Chan · Carol Y. Cheung

ABSTRACT

Introduction: Generative pretrained transformer-4 (GPT-4) has gained widespread attention from society, and its potential has been extensively evaluated in many areas. However, investigation of GPT-4's use in medicine, especially in the ophthalmology field, is still limited. This study aims to evaluate GPT-4's capability to identify rare ophthalmic diseases in three simulated scenarios for different end-users, including patients, family physicians, and junior ophthalmologists.
Methods: We selected ten treatable rare ophthalmic disease cases from the publicly available EyeRounds service. We gradually increased the amount of information fed into GPT-4 to simulate the scenarios of patient, family physician, and junior ophthalmologist using GPT-4. GPT-4's responses were evaluated from two aspects: suitability (appropriate or inappropriate) and accuracy (right or wrong) by senior ophthalmologists (> 10 years' experiences).

Results: Among the 30 responses, 83.3% were considered "appropriate" by senior ophthalmologists. In the scenarios of simulated patient, family physician, and junior ophthalmologist, seven (70%), ten (100%), and eight (80%) responses were graded as "appropriate" by senior ophthalmologists. However, compared to the ground truth, GPT-4 could only output several possible diseases generally without "right" responses in the simulated patient scenarios. In contrast, in the simulated family physician scenario, 50% of GPT-4's responses were "right," and in the simulated junior ophthalmologist scenario, the model achieved a higher "right" rate of 90%.
Conclusion: To our knowledge, this is the first proof-of-concept study that evaluates GPT-4's capacity to identify rare eye diseases in simulated scenarios involving patients, family physicians, and junior ophthalmologists. The results indicate that GPT-4 has the potential to serve as a consultation assisting tool for patients and family physicians to receive referral suggestions and an assisting tool for junior ophthalmologists to diagnose rare eye diseases. However, it is important to approach GPT-4 with caution and acknowledge the need for verification and careful referrals in clinical settings.

Keywords: Rare eye disease; Generative pretrained transformer; Large language models; Artificial intelligence

X. Hu · A. R. Ran · T. X. Nguyen · S. Szeto ·
J. C. Yam · C. Y. Cheung (✉)
Department of Ophthalmology and Visual Sciences,
The Chinese University of Hong Kong, Hong Kong
SAR, China
e-mail: carolcheung@cuhk.edu.hk

C. K. M. Chan
Hong Kong Eye Hospital, Hong Kong SAR, China

Key Summary Points

Why carry out this study?

Rare eye diseases are the leading cause of visual impairment and blindness in children and young adults, which can adversely decrease the life quality of patients and their families. Therefore, there is an urgent need to develop automated and accurate tools to quickly and accurately diagnose rare eye diseases to support patients.

Recently, large language models (LLMs), especially GPT (Chat Generative Pre-training Transformer), have motivated numerous researchers to evaluate their ability in various tasks. Nevertheless, the capability of GPT-4 in the ophthalmology field of identifying rare eye diseases is still largely unknown.

This study aims to evaluate the capability and explore the potential implementation of GPT-4 in identifying rare ophthalmic diseases in simulated scenarios of patient, family physician, and junior ophthalmologist.

What was learned from the study?

Most responses (83.3%) output by GPT-4 were graded as "appropriate" by senior ophthalmologists from the perspective of suitability. GPT-4 could provide considerably "right" diagnoses when chief complaints, history of present illness, and descriptions of ophthalmic and other necessary examinations focusing on ocular imaging were provided.

In the future, GPT-4 may serve as a consultation assisting tool for patients and family physicians to obtain referral suggestions and an assisting tool for junior ophthalmologists to diagnose rare eye diseases. However, it is important to approach GPT-4 with caution and acknowledge the need for verification and careful referrals in clinical settings.

INTRODUCTION

There are approximately 7000 rare diseases, and patients with rare diseases are estimated to constitute about 10% of the population [1]. Many rare diseases can adversely decrease the life quality of patients and their families. However, timely and accurate diagnoses remain difficult [2]. Rare eye diseases are the leading cause of visual impairment and blindness in children and young adults in Europe. Over 900 eye disorders are included in this heterogeneous group of conditions, ranging from relatively prevalent disorders, such as retinitis pigmentosa, to very rare entities, such as developmental eye anomalies [3]. Therefore, there is an urgent need to develop automated and accurate tools to quickly and accurately diagnose rare eye diseases to support patients.

Deep learning methods have already been approved to achieve good performance in many healthcare tasks, and some works have attempted to utilize deep learning methods to address the challenges of detecting rare eye diseases. Burlina et al. [4] suggested the potential benefits of using low-shot methods for rare ophthalmic disease diagnostics when a limited number of annotated training retinal images is available. Yoo et al. [5] introduced a method that combined the few-shot learning and generative adversarial network to improve the applicability of deep learning in the optical coherence tomography diagnosis of rare retinal diseases. However, these methods only output diagnosis results, do not offer explanations, and cannot interact with end-users. Studies using conversational chatbots that can be used by different end-users by interacting with people to diagnose rare eye diseases with explanations are lacking.

Applying expert knowledge to refine artificial intelligence models' output is often carried out in practice, and there have been various efforts to investigate this field. Recently, large language models (LLMs), especially ChatGPT (Chat Generative Pre-training Transformer), trained by reinforcement learning from human feedback strategy, have attracted public, media, and scientific attention from various fields worldwide

[6] and motivated numerous researchers to evaluate their ability in various tasks, e.g., data analysis [7], software development [8], and education [9]. A few reports have already demonstrated the potential applications of ChatGPT in medicine, even in the field of ophthalmology. In the medical field, Kanjee et al. [10] proposed that GPT-4 could provide a numerically superior mean differential quality score in a complex diagnostic challenge compared with some differential diagnosis generators. Sorin et al. [11] assessed the potential application of ChatGPT in patient management in breast tumor board decisions as a clinical decision support tool. In the ophthalmology field, Mihalache et al. [12] designed a study to evaluate ChatGPT's ability to answer practice questions for board certification in ophthalmology. Balas et al. [13] investigated ChatGPT's accuracy in formulating provisional and differential diagnoses from text case report descriptions. Antaki et al. [14] tested ChatGPT on two popular multiple-choice question banks commonly used to prepare for the high-stakes Ophthalmic Knowledge Assessment Program examination, and ChatGPT showed encouraging performance on the examination. Rasmussen et al. [15] evaluated the performance of ChatGPT's responses to typical patient-related questions on vernal keratoconjunctivitis. Nevertheless, the capability of GPT-4 in the ophthalmology field of identifying rare eye diseases is still largely unknown [16].

In this study, we aim to qualitatively evaluate the ability of GPT-4, the recent successor to ChatGPT, in identifying rare ophthalmic diseases in simulated patient, family physician, and junior ophthalmologist scenarios.

METHODS

We selected ten cases of treatable rare ophthalmic disease [17] with confirmed diagnosis (i.e., the ground truth) from the publicly available EyeRounds service [18]. For each case, we simulated different end-users, including patients, family physicians, and junior ophthalmologists, utilizing GPT-4. Because these end-users have different information available,

they may provide different input when using GPT-4. We assumed that these three end-users would input the following information into GPT-4, respectively: Scenario 1 (patient): chief complaints; Scenario 2 (family physician): chief complaints and history of present illness; Scenario 3 (junior ophthalmologist): chief complaints, history of present illness, and descriptions of ophthalmic and other necessary examinations focusing on ocular imaging. GPT-4 was accessed on May 10, 2023, via https://chat.openai.com/, and all responses were obtained and recorded at that time. The prompts were from EyeRounds, including chief complaints, history of present illness, and descriptions of ophthalmic and other necessary examinations focusing on ocular imaging for different scenarios with the question 'What eye disease may I/he/she have?' We evaluated GPT-4's responses in two different aspects: suitability (appropriate or inappropriate) and accuracy (right or wrong). Senior ophthalmologists, who had > 10 years' experience and were blinded to the ground truth, graded GPT-4's responses as "appropriate" or "inappropriate." We assigned each case to a senior ophthalmologist (> 10 years' experience) specialized in the relevant field for the grading. An "appropriate" GPT-4 response was defined as no misconceptions and had reasonable descriptions of diagnosis differentiation process based on input information in each scenario. Each response was further classified as "right" or "wrong." A "right" response was defined as GPT-4 confirming the diagnosis the same as the ground truth.

This article is based on an online database and does not contain any new studies with human participants performed by any of the authors; therefore, ethics committee approval was not required.

RESULTS

Twenty-five out of 30 (83.3%) responses were graded as "appropriate" by senior ophthalmologists. For the simulated patient, family physician, and junior ophthalmologist scenarios, seven (70%), ten (100%), and eight (80%) responses were graded as "appropriate" by

senior ophthalmologists, respectively. When comparing with the ground truth in the simulated patient scenario, GPT-4 could only output several possible diseases generally, and no responses were "right." In the simulated family physician scenario, five (50%) responses output by GPT-4 were right. In the simulated junior ophthalmologist scenario, most of the responses output by GPT-4, 9 (90%) were "right." Details are summarized in Table 1.

DISCUSSION

Our study found that in the scenario of patient and family physician, most of GPT-4's responses were "appropriate." However, in these two scenarios, GPT-4 could not output "right" responses for most cases. Specifically, in the patient scenario, GPT-4 tended to output several possible but relatively broad and common eye diseases (e.g., refractive errors, retinal diseases, and glaucoma). In the family physician scenario, GPT-4 started to output more specific responses (e.g., case 7 as optic neuritis); however, most of the responses were still "wrong." The reason could be that the prompts for these two simulated scenarios had insufficient information related to eye conditions and GPT-4 could not ask for additional information such as visual acuity or medical and ocular history to

Table 1 Evaluation GPT-4's output for the ten cases in different scenarios

Rare eye diseases	Scenario-1: patient		Scenario-2: family physician		Scenario-3: junior ophthalmologist	
	Grade for GPT-4 output	Compared to ground truth	Grade for GPT-4 output	Compared to ground truth	Grade for GPT-4 output	Compared to ground truth
Case 1: Behçet's disease	Appropriate	Wrong	Appropriate	Right	Appropriate	Right
Case 2: Best vitelliform macular dystrophy	Inappropriate	Wrong	Appropriate	Wrong	Inappropriate	Right
Case 3: Charles Bonnet syndrome	Appropriate	Wrong	Appropriate	Right	Inappropriate	Right
Case 4: Coloboma	Inappropriate	Wrong	Appropriate	Wrong	Appropriate	Right
Case 5: Cystinosis	Appropriate	Wrong	Appropriate	Wrong	Appropriate	Right
Case 6: Idiopathic intracranial hypertension	Inappropriate	Wrong	Appropriate	Right	Appropriate	Right
Case 7: Leber hereditary optic neuropathy	Appropriate	Wrong	Appropriate	Right	Appropriate	Wrong
Case 8: Optic neuritis	Appropriate	Wrong	Appropriate	Right	Appropriate	Right
Case 9: Retinitis pigmentosa	Appropriate	Wrong	Appropriate	Wrong	Appropriate	Right
Case 10: Retinoblastoma	Appropriate	Wrong	Appropriate	Wrong	Appropriate	Right

◄**Fig. 1** Example of case 7 in different scenarios. The input and output of case 7 in different scenarios, which is the only case that GPT-4 got "wrong" in the scenario of junior ophthalmologist. The ground truth was Leber hereditary optic neuropathy (LHON). **a** Input and output of GPT-4 in the scenario of patient. GPT-4 output several possible but relatively broad and common eye diseases. **b** Input and output of GPT-4 in the scenario of family physicians. GPT-4 output more specific diagnostic results. **c** Input and output of GPT-4 in the scenario of junior ophthalmologist. GPT-4 output the most possible diagnosis as optic neuritis and listed reasons why it obtained this result. Although the primary diagnosis output by GPT-4 (optic neuritis) was different from the ground truth (LHON), GPT-4 still output "However, it's a bit unusual for optic neuritis to be painless and bilateral in a patient this age, so other conditions such as Leber's Hereditary Optic Neuropathy (LHON) should also be considered"

further diagnose diseases as ophthalmologists usually do. This indicates that the current GPT-4 is not a suitable diagnostic tool in the scenarios of patient and family physician. Nevertheless, GPT-4 may still serve as a consultation assisting tool for referral suggestions in the future.

In the scenario of junior ophthalmologist, GPT-4 provided a more specific diagnosis, 90% of responses were "right," and it could explain how it obtained the diagnosis in detail. For the only case classified as "wrong," GPT-4's primary diagnosis was optic neuritis, which was different from the ground truth (i.e., case 7, Leber's hereditary optic neuropathy, LHON). Nevertheless, GPT-4 still mentioned that LHON should be considered (Fig. 1), and the output of why GPT-4 gave its diagnosis as optic neuritis was graded as "appropriate" by senior ophthalmologists. Our results indicate that GPT-4 may serve as an assisting tool for junior ophthalmologists to diagnose rare eye diseases quickly and accurately.

There are some inherent limitations of GPT-4. First, it may raise the concern of patient's privacy when enquiry is uploaded to the OpenAI server for computation, especially in the field of healthcare. Second, GPT-4 may output misconceptions as it was originally

designed for general purposes instead of making clinical diagnoses and trained on unverified data. Third, OpenAI has not publicly disclosed the specific information on datasets used for model training, meaning there is a risk of overestimating the capabilities of GPT-4 if Eye-Rounds were used for training the model. In addition, GPT-4 may generate different responses and different primary diagnoses even if end-users feed the same input into GPT-4 multiple times, which means that GPT-4 still has a lack of robustness and cannot provide end-users with consistent suggestions and diagnoses. Lastly, technical details of how GPT-4 generates the responses are not known. This lack of transparency hinders users' ability to have fine-tuned control of the generated responses [19], which may bring adverse effects to end-users for medical purposes. In addition to these concerns, GPT-4 faces several other challenges, including the need for huge computational resources, and can only function effectively in large computational environments; it has difficulty delivering up-to-date information, and "hallucinations" occur [20]. In conclusion, despite GPT-4's impressive capabilities across various domains, we must still acknowledge its limitations.

Future research should compare GPT-4 with other state-of-the-art LLMs, e.g., Bard or LLaMA, using different languages in the ophthalmology field. Artificial intelligence chatbots that are designed and trained specifically for ophthalmic diagnosis purposes and chatbots that can actively ask for information that end-users have not provided, as ophthalmologists usually do, are warranted. Moreover, direct inputting of inputting images into GPT-4 will be available to the public next year. It can be anticipated that if the model can capture information from images and output relevant descriptions, it can potentially be applied in clinical settings to assist junior ophthalmologists to diagnose rare eye diseases.

CONCLUSION

To our knowlege, this is the first proof-of-concept brief report that shows GPT-4 can

potentially identify rare eye diseases in simulated patient, family physician, and junior ophthalmologist scenarios. The results indicate GPT-4's huge potential as a consultation assisting tool for patients and family physicians to obtain referral suggestions. Additionally, GPT-4 may serve as an assisting tool for junior ophthalmologists to diagnose rare eye diseases quickly and accurately in the future, especially when feeding images into GPT-4 becomes available and GPT-4 can capture underlying information from images. However, it is important to approach GPT-4 with caution and acknowledge the need for verification and careful referrals in clinical settings.

Declarations

by any of the authors; therefore, ethics committee approval was not required.

Open Access. This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License, which permits any non-commercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc/4.0/.

REFERENCES

1. Haendel M, Vasilevsky N, Unni D, Bologa C, Harris N, Rehm H, et al. How many rare diseases are there? Nat Rev Drug Discov. 2020;19(2):77–8.

2. Ronicke S, Hirsch MC, Türk E, Larionov K, Tientcheu D, Wagner AD. Can a decision support system accelerate rare disease diagnosis? Evaluating the potential impact of Ada DX in a retrospective study. Orphanet J Rare Dis. 2019;14:1–12.

3. Black GC, Sergouniotis P, Sodi A, Leroy BP, Van Cauwenbergh C, Liskova P, et al. The need for widely available genomic testing in rare eye diseases: an ERN-EYE position statement. Orphanet J Rare Dis. 2021;16:1–8.

4. Burlina P, Paul W, Mathew P, Joshi N, Pacheco KD, Bressler NM. Low-shot deep learning of diabetic retinopathy with potential applications to address artificial intelligence bias in retinal diagnostics and rare ophthalmic diseases. JAMA Ophthalmol. 2020;138(10):1070–7.

5. Yoo TK, Choi JY, Kim HK. Feasibility study to improve deep learning in OCT diagnosis of rare retinal diseases with few-shot classification. Med Biol Eng Comput. 2021;59:401–15.

6. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA. 2023;329(10):842–4.

7. Macdonald C, Adeloye D, Sheikh A, Rudan I. Can ChatGPT draft a research article? An example of population-level vaccine effectiveness analysis. J Glob Health. 2023. https://doi.org/10.7189/jogh.13.01003.

8. Surameery NMS, Shakor MY. Use ChatGPT to solve programming bugs. Int J Inf Technol Comput Eng (IJITC). 2023;3(01):17–22.

9. Topsakal O, Topsakal E. Framework for a foreign language teaching software for children utilizing AR, Voicebots and ChatGPT (large language models). J Cognit Syst. 2022;7(2):33–8.

10. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA. 2023;330:1–78.

11. Sorin V, Klang E, Sklair-Levy M, Cohen I, Zippel DB, Balint Lahat N, et al. Large language model (ChatGPT) as a support tool for breast tumor board. NPJ Breast Cancer. 2023;9(1):44.

12. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence Chatbot in ophthalmic knowledge assessment. JAMA Ophthalmol. 2023. https://doi.org/10.1001/jamaophthalmol.2023.2754.

13. Balas MI, Edsel B. Conversational AI models for ophthalmic diagnosis: comparison of ChatGPT and the Isabel Pro differential diagnosis generator. JFO Open Ophthalmol. 2023;1:100005.

14. 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.

15. Rasmussen MLR, Larsen A-C, Subhi Y, Potapenko I. Artificial intelligence-based ChatGPT chatbot responses for patient and parent questions on vernal keratoconjunctivitis. Graefe's Archiv Clin Exp Ophthalmol. 2023. https://doi.org/10.1007/s00417-023-06078-1.

16. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI Chatbot for medicine. N Engl J Med. 2023;388(13):1233–9.

17. Mukamal R. 20 Rare eye conditions that ophthalmologists treat: American Academy of Ophthalmology. 2020. https://www.aao.org/eye-health/

tips-prevention/20-rare-eye-conditions-that-ophthalmologists-treat. Accessed 25 Apr 2023.

18. Ophthalmology Cases: EyeRounds.org. 2014. http://eyerounds.org/cases.htm. Accessed 25 Apr 2023.

19. Zhang C, Zhang C, Li C, Qiao Y, Zheng S, Dam SK, et al. One small step for generative AI, one giant leap for AGI: a complete survey on ChatGPT in AIGC era. 2023. arXiv preprint arXiv:230406488.

20. Choi JY, Yoo TK. New era after ChatGPT in ophthalmology: advances from data-based decision support to patient-centered generative artificial intelligence. Ann Transl Med 2023.