AMERICAN ACADEMY OF OPHTHALMOLOGY®

Interpretation of Clinical Retinal Images Using an Artificial Intelligence Chatbot

Andrew Mihalache, MD(C),[1] Ryan S. Huang, MD(C),[1] David Mikhail, MD(C),[1] Marko M. Popovic, MD, MPH,[2] Reut Shor, MD,[2] Austin Pereira, MD,[2] Jason Kwok, MD, FRCSC,[2] Peng Yan, MD, FRCSC,[2] David T. Wong, MD, FRCSC,[2,4] Peter J. Kertes, MD, CM,[2,3] Radha P. Kohly, MD, PhD,[2,3] Rajeev H. Muni, MD, MSc[2,4]

Purpose: To assess the performance of Chat Generative Pre-Trained Transformer-4 in providing accurate diagnoses to retina teaching cases from OCTCases.

Design: Cross-sectional study.

Subjects: Retina teaching cases from OCTCases.

Methods: We prompted a custom chatbot with 69 retina cases containing multimodal ophthalmic images, asking it to provide the most likely diagnosis. In a sensitivity analysis, we inputted increasing amounts of clinical information pertaining to each case until the chatbot achieved a correct diagnosis. We performed multivariable logistic regressions on Stata v17.0 (StataCorp LLC) to investigate associations between the amount of text-based information inputted per prompt and the odds of the chatbot achieving a correct diagnosis, adjusting for the laterality of cases, number of ophthalmic images inputted, and imaging modalities.

Main Outcome Measures: Our primary outcome was the proportion of cases for which the chatbot was able to provide a correct diagnosis. Our secondary outcome was the chatbot's performance in relation to the amount of text-based information accompanying ophthalmic images.

Results: Across 69 retina cases collectively containing 139 ophthalmic images, the chatbot was able to provide a definitive, correct diagnosis for 35 (50.7%) cases. The chatbot needed variable amounts of clinical information to achieve a correct diagnosis, where the entire patient description as presented by OCTCases was required for a majority of correctly diagnosed cases (23 of 35 cases, 65.7%). Relative to when the chatbot was only prompted with a patient's age and sex, the chatbot achieved a higher odds of a correct diagnosis when prompted with an entire patient description (odds ratio = 10.1, 95% confidence interval = 3.3−30.3, $P < 0.01$). Despite providing an incorrect diagnosis for 34 (49.3%) cases, the chatbot listed the correct diagnosis within its differential diagnosis for 7 (20.6%) of these incorrectly answered cases.

Conclusions: This custom chatbot was able to accurately diagnose approximately half of the retina cases requiring multimodal input, albeit relying heavily on text-based contextual information that accompanied ophthalmic images. The diagnostic ability of the chatbot in interpretation of multimodal imaging without text-based information is currently limited. The appropriate use of the chatbot in this setting is of utmost importance, given bioethical concerns.

Financial Disclosure(s): Proprietary or commercial disclosure may be found in the Footnotes and Disclosures at the end of this article. *Ophthalmology Science 2024;4:100556 © 2024 by the American Academy of Ophthalmology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).*

The performance of the artificial intelligence (AI) chatbot Chat Generative Pre-Trained Transformer (ChatGPT; Open AI) has been improving remarkably in ophthalmic settings.[1,2] Given its potential to enhance clinical triaging, facilitate remote monitoring of retinal diseases, and complement patient education, the chatbot possesses the capacity to improve clinical care and education within ophthalmology.[3–6] Although the chatbot cannot currently be appraised as a source of consistent factual information, a previous cross-sectional study by Momenaei et al[7] found its responses to common questions regarding retinal detachment, macular hole, and epiretinal membrane to be largely appropriate. Potapenko et al[8] also highlighted the chatbot's ability to provide highly accurate general information related to the prevention and prognosis of age-related macular degeneration, diabetic retinopathy, retinal vein occlusion, retinal artery occlusion, and central serous chorioretinopathy. Although the integration of AI chatbots within ophthalmic clinical practice has garnered great interest,[9] substantial concerns remain surrounding misinformation, liability, bioethical concerns, patient privacy, and regulatory compliance.

Deep learning applications have demonstrated promising accuracy in detecting retinal disorders, with certain models achieving performance comparable to retina specialists.[10] For instance, De Fauw et al[11] developed a deep learning

https://doi.org/10.1016/j.xops.2024.100556
ISSN 2666-9145/24

system capable of identifying referrable retinal diseases via OCT images, whose predictions were 99.21% correct.[10] Overall, the subspecialty of the retina is dependent on nuanced interpretations of multimodal imaging to ensure high diagnostic accuracy. However, the chatbot's ability to formulate diagnoses from standalone clinical retinal images has not yet been elicited. Our current investigation aims to assess the ability of the newest release of the chatbot to provide accurate diagnoses to retina teaching cases with ophthalmic imaging.

Methods

Study Setting and Design

Our study used a freely accessible set of retina teaching cases from OCTCases,[12] a medical education platform from the Department of Ophthalmology and Vision Sciences at the University of Toronto. All retina cases on OCTCases are comprehensively reviewed by ≥1 board-certified retina specialist affiliated with the University of Toronto, as well as by the platform's founders (A.P. and J.K.). The University of Toronto waived institutional review board approval for the publication of cases with ophthalmic imaging on the OCTCases website, given that all cases had been entirely anonymized and contained modified patient characteristics that were not identifiable to individual patients. The research adhered to the tenets of the Declaration of Helsinki and employed Strengthening the Reporting of Observational Studies in Epidemiology reporting guidelines.

Our investigation used a new ChatGPT Plus account with no prior conversation history. Using the "My GPTs" feature, we first created a custom chatbot with the following instructions: "This GPT provides the most likely diagnosis based on fictitious patient characteristics and images for the purpose of education. This GPT thoroughly analyzes images from various ophthalmic imaging modalities and describes its findings. This GPT avoids not providing a conclusive diagnosis." We prompted the chatbot with all retina cases available on OCTCases from December 6, 2023, to December 13, 2023, asking the chatbot "What is the most likely diagnosis?" at the end of each prompt. We excluded 4 cases pertaining to the identification of imaging biomarkers, in which the patient's retinal diagnosis may have been revealed within the written case description on OCTCases.

We performed a sensitivity analysis where the chatbot was prompted with increasingly more text-based clinical information from case descriptions in a stepwise manner, alongside the ophthalmic image(s) accompanying each case, until the chatbot had the correct diagnosis or until all written clinical information was provided. Given the chatbot's tendency to refuse to answer the question "What is the most likely diagnosis?" in the absence of clinical context, the minimum amount of information inputted into the chatbot per case consisted of all multimodal ophthalmic images associated with the case, as well as the patient's age and sex, unless these demographic data were not provided. Afterward, the following pieces of additional information from each case were incrementally inputted into the chatbot, if available: (1) best-corrected visual acuity (BCVA); (2) intraocular pressure; (3) ocular history; (4) presenting features; and (5) family ocular history. If the chatbot still failed to provide the correct diagnosis once all details had been inputted, the entire case description as it appeared on OCTCases was provided, which may have contained additional information regarding the patient's demographics, fellow eye, systemic medical history or medications, slit-lamp examination findings, and other imaging findings, in an anonymized manner. We stopped prompting the chatbot about a particular case if it arrived at the correct diagnosis or if there were no additional clinical data to be provided. In the case that the chatbot arrived at a correct diagnosis, we also prompted it with the follow-up question "What piece(s) of information led you to this diagnosis?" An example demonstrating how the chatbot was prompted for a sample case is shown in Figure 1.

Data Collection and Outcomes

We cleared our conversation history with the chatbot between cases to mitigate the influence of active conversations on its processing of subsequent cases. At least 2 independent reviewers (A.M., R.S.H., and D.M.) manually reviewed the chatbot's output to determine which diagnosis it had selected. We collected the following data from each case: the date on which the chatbot was prompted with a case, the modality and number of ophthalmic images associated with each case, the character length of the chatbot's responses, and the amount of text-based information accompanying the ophthalmic image(s) required for a correct diagnosis.

Our primary outcome was the proportion of retina cases on OCTCases for which the chatbot was able to achieve a correct diagnosis. Our secondary outcomes were the chatbot's performance in relation to the amount of text-based information inputted per prompt, the chatbot's response lengths across correct and incorrect diagnoses, and the amount of information the chatbot needed for a correct diagnosis.

Statistical Analysis

We performed chi-square tests on MedCalc to compare the proportion of unilateral and bilateral cases for which the chatbot was able to provide a definitive, correct diagnosis.[13–15] We calculated Spearman correlation coefficient between the character length of our text-based input and the character length of the chatbot's output.[16] We conducted Mann–Whitney *U* tests to compare observed response lengths between the chatbot's correct and incorrect outputs.[17] We performed univariable and multivariable logistic regressions on Stata v17.0 (StataCorp LLC) to investigate associations between the amount of text-based information inputted per prompt and the odds of the chatbot achieving a correct diagnosis. Our multivariable model adjusted for the eye laterality of cases, the number of ophthalmic images inputted, and unique imaging modalities (i.e., OCT, fundus photography, fundus autofluorescence, scanning laser ophthalmoscopy, OCT angiography, or IV fluorescein angiography). All *P* values were 2-tailed, and we made no adjustment to *P* values for multiple analyses. A *P* value of <0.05 denoted statistical significance.

Results

The chatbot was prompted with a total of 284 prompts with various amounts of text-based information across 69 eligible retina cases. Fifty-one (73.9%) cases pertained to unilateral findings and 18 (26.1%) to bilateral retinal disorders. Moreover, 139 ophthalmic images accompanied the cases, consisting of 91 (65.5%) OCTs, 41 (29.5%) fundus photographs, 3 (2.2%) fundus autofluorescences, 2 (1.4%) scanning laser ophthalmoscopy images, 1 OCT angiography (0.7%), and 1 (0.7%) IV fluorescein angiography. The mean number of different imaging modalities inputted per case was 1.5 modalities (range, 1–3 modalities). Overall, the chatbot was able to provide a definitive, correct diagnosis to

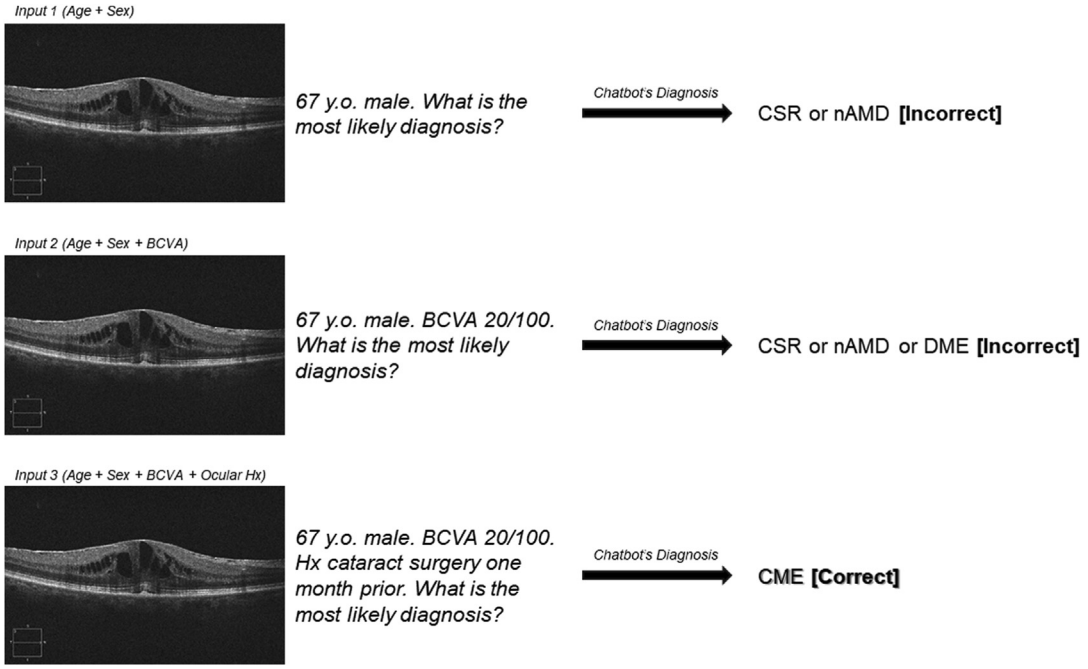

Figure 1. Prompting of the chatbot for a sample case for which it achieved a correct diagnosis after the age, sex, BCVA, and ocular Hx were provided. BCVA = best-corrected visual acuity; CME = cystoid macular edema; CSR = central serous chorioretinopathy; DME = diabetic macular edema; Hx; history; nAMD = neovascular age-related macular degeneration.

35 (50.7%) cases. Among these 35 correct responses, the chatbot only described the correct diagnosis in 30 (85.7%) cases and justified selecting the correct diagnosis over other suspected differential diagnoses in 5 (14.3%) cases. Although the chatbot provided incorrect diagnoses to 34 (49.3%) cases, it listed the correct diagnosis within its differential diagnosis in its response to 7 (20.6%) of these incorrectly answered cases. Despite only being asked for the most likely diagnosis, the chatbot provided additional differential diagnoses in $\geq$1 of its responses to 57 (82.6%) cases. The chatbot's mean $\pm$ standard deviation response length was 1741.2 $\pm$ 446.8 characters. A positive correlation was observed between the character length of text-based inputs and the character length of the chatbot's output (r_s= 0.25, $P < 0.01$). Moreover, the chatbot's incorrect responses were longer than its responses containing a definitive, correct diagnosis ($P = 0.02$). A complete list of cases for which the chatbot was either able or unable to provide a correct diagnosis is found in Table 1.

Across the 35 cases for which the chatbot was able to provide a correct diagnosis, our sensitivity analysis found the following amounts of text-based information were necessary for a correct diagnosis: no information, as none were available (1 of 35 cases, 2.9%); age and sex (5 of 35 cases, 14.3%); age, sex, BCVA, and ocular history (3 of 35 cases, 8.6%); age, sex, BCVA, intraocular pressure, and ocular history (2 of 35 cases, 5.7%); age, sex, BCVA, ocular history, and presenting features (1 of 35 cases, 2.9%); and the entire patient description as presented by OCTCases (23 of 35 cases, 65.7%). Relative to only inputting a patient's age and sex, the only amount of information associated with higher odds of the chatbot achieving a correct diagnosis was prompting it with the entire patient description as presented by OCTCases. This was consistent in our univariable (odds ratio = 8.3, 95% confidence interval = 2.8−24.0, $P < 0.01$) and multivariable (odds ratio = 10.1, 95% confidence interval = 3.3−30.3, $P < 0.01$) analyses. Nonetheless, when the chatbot was asked the follow-up question "What piece(s) of information led you to this diagnosis?" it highlighted the importance of the following pieces of information: ethnicity (1 of 35 cases, 2.9%), geographic location (1 of 35 cases, 2.9%), family history (1 of 35 cases, 2.9%), sex (2 of 35 cases, 5.7%), intraocular pressure (4 of 35 cases, 11.4%), age (6 of 35 cases, 17.1%), systemic medical history or medications (12 of 35 cases, 34.3%), ocular history (22 of 35 cases, 62.9%), presenting features (22 of 35 cases, 62.9%), BCVA (25 of 35 cases, 71.4%), and imaging findings (35 of 35 cases, 100%).

Discussion

Our investigation demonstrated that a custom chatbot was able to provide correct diagnoses to approximately half of the retina cases available on OCTCases when prompted with sufficient clinical information. Alongside the input of multimodal ophthalmic images, we found that prompting the chatbot with an increasing depth of contextual information pertaining to each patient was pivotal in achieving a correct diagnosis for many cases.

In our investigation, the chatbot was most successful at correctly diagnosing retinal disorders from multimodal ophthalmic cases when provided with the entire patient description as it appeared on OCTCases. A plausible

Table 1. Cases for Which the Chatbot Provided a Correct or Incorrect Diagnosis

Disorder Frequency	Outcome	Disorder (No. of Correct Cases/No. of Total Cases)
Occured once on OCTCases	Definitive, correct diagnosis	Normal healthy retina (1/1); retinal detachment (1/1), cone dystrophy (1/1); neuroretinitis (1/1); central retinal artery occlusion (1/1); macular hemorrhage (1/1); plaquenil toxicity (1/1); sickle cell maculopathy (1/1); retinal detachment secondary to full-thickness macular hole (1/1); Stargardt's disease (1/1); myopic choroidal neovascular membrane (1/1); polypoidal choroidal vasculopathy (1/1); diabetic retinopathy (1/1); outer retinoschisis secondary to vitreomacular traction and epiretinal membrane (1/1); persistent subretinal fluid postretinal detachment repair (1/1); choroidal nevus (1/1); Best's vitelliform macular dystrophy (1/1); hypotony maculopathy (1/1); solar retinopathy (1/1); diffuse unilateral subacute neuroretinitis (1/1); diabetic macular edema (1/1); asteroid hyalosis (1/1); pseudoxanthoma elasticum (1/1); radiation retinopathy (1/1)
	Incorrect diagnosis	Full-thickness macular hole (0/1); macular retinoschisis (0/1); optic pit maculopathy (0/1); foveal hypoplasia (0/1); pseudohole (0/1); macular telangiectasia (0/1); peripapillary atrophy (0/1); acute retinal artery occlusion (0/1); choroidal metastasis (0/1); torpedo maculopathy (0/1); vitreous hemorrhage (0/1); peripapillary choroidal neovascular membrane (0/1); peripapillary pachychoroid syndrome (0/1); sclerochoroidal calcification (0/1); pachychoroid neovasculopathy (0/1); Elmiron toxicity (0/1); reticular pseudodrusen (0/1); dark-without-pressure lesions (0/1); morning glory syndrome (0/1); outer retinal folds (0/1); bacillary layer detachment (0/1); age-related choroidal atrophy (0/1)
Occurred more than once on OCTCases	Definitive, correct diagnosis	Age-related macular degeneration (3/3); cystoid macular edema (2/2)
	Variable accuracy	Central serous chorioretinopathy (2/4); epiretinal membrane (1/2); lamellar macular hole (1/2); paracentral acute middle maculopathy (1/2); vitreomacular traction syndrome (1/2)
	Incorrect diagnosis	Posterior staphyloma (0/2); geographic atrophy (0/2); focal choroidal excavation (0/2)

explanation for the high text dependency of the chatbot may lie in its architecture. As noted in the OpenAI GPT-4V(ision) System card, the chatbot in its current form is primarily a text-based model and thus may lack the essential architecture needed for nuanced visual data processing.[18] This highlights a significant gap in the AI chatbot's capabilities, especially in medical imaging where convolutional neural networks have become mainstay tools in image interpretation.[19] Bridging this gap may require the development of integrated systems trained extensively in handling both text and image data, which are both crucial in medical diagnostics.

Although our recent work found that the chatbot correctly answered 160 of 209 (76.6%) multiple-choice questions pertaining to retina cases from OCTCases, its performance on multiple-choice questions may not translate to its clinical utility.[20] The multiple-choice questions used in our previous investigation required the chatbot to identify or interpret abnormalities present in various imaging modalities, select an appropriate diagnostic test or treatment modality for a particular disease, describe the prognosis of a particular disease, or identify the pathophysiologic mechanisms or gene(s) involved in a particular disease.[20] Some questions required the chatbot to select an appropriate diagnosis, albeit the contextual information from multiple-choice options may have guided its answers.[20] Moreover, a considerable proportion of questions in our previous investigation were not based on ophthalmic images.[20] However, in our present study, we strictly evaluated the ability of a chatbot to provide diagnoses for multimodal retina imaging cases without multiple-choice options, an essential step in gauging the clinical reasoning of this emerging technology. In contrast to our prior work, we conducted a sensitivity analysis, whereby we inputted increasing amounts of clinical information pertaining to each case until the chatbot achieved a correct diagnosis to determine the amount of text-based information necessary for a correct diagnosis. Overall, our finding that the chatbot was able to accurately diagnose 50.7% of multimodal retina cases while relying heavily on text-based clinical information suggests that its diagnostic capabilities based on multimodal imaging interpretation may be less robust compared with its ability to answer high-yield multiple-choice questions in the setting of retinal disorders. Given the high text dependency of the chatbot, it is possible that the chatbot performed better in our previous work relative to our present analysis in which the chatbot was asked open-ended questions, as the additional text input from multiple-choice options may have assisted the chatbot in arriving at correct answers. Furthermore, the tailored nature of question stems used in our previous study pertaining to the management, prognosis, and particular imaging findings associated with various retinal disorders may have been less challenging than the open-ended questions used in our present study, which required the chatbot to synthesize information and provide a diagnosis in the absence of any guidance from question stems.

Other deep learning systems have excelled with respect to accuracy, sensitivity, and specificity in diagnosing retinal diseases.[21] A recent analysis examining the diagnostic accuracy of deep learning algorithms for age-related macular degeneration found a pooled sensitivity and specificity of 94% and 97%, respectively.[22] Another systematic review scrutinizing the diagnostic accuracy of deep learning algorithms for diabetic retinopathy found sensitivities and specificities ranging from 80% to 100% and 84% to 99%, respectively.[23] Although the chatbot's performance remains inferior to AI systems specifically designed to identify retinal disorders,[24–26] its performance in this setting will likely improve in the future. Nonetheless, when asked "Will ChatGPT ever be as accurate as deep learning algorithms in identifying retinal disorders?" the chatbot

replied "While ChatGPT is a powerful tool for information and communication, it is not designed for medical image analysis and thus will not match the accuracy of specialized deep learning algorithms in identifying retinal disorders. However, the 2 technologies can complement each other in a healthcare setting."

Although the chatbot currently lacks sufficient diagnostic accuracy to be of value clinically to ophthalmologists, its ability to process multimodal ophthalmic information may benefit trainees interested in engaging in independent learning. Yet, tremendous caution must be exercised by users when uploading medical or ophthalmic images onto the chatbot, given the potential for violations of patient privacy and bioethical concerns if images are not deidentifiable. As such, clear guidelines enforcing the protection of patient privacy and confidentiality must be established to inform the use of the chatbot in this setting. Other ethical and medicolegal concerns surrounding liability in cases where the chatbot can provide erroneous recommendations must also be proactively addressed before this technology can be formally adopted within medicine.[27] Patients must also be counseled to be extremely cautious if deciding to input their own images into the chatbot and subsequently question their diagnosis, given the inherent limitations of the chatbot in its current form. Furthermore, our current investigation found that the chatbot's output when describing incorrect diagnoses was considerably longer than when describing correct diagnoses, highlighting its potential to confidently misguide users. This aligns with findings from our previous study investigating the chatbot's performance on practice United States Medical Licensing Examination questions, where its incorrect responses were significantly longer than its correct responses.[28]

Our study was limited for several notable reasons. A direct comparison of the chatbot's diagnostic abilities with that of OCTCases' user base is not possible because of the absence of comparative data. The sole reliance on OCTCases as the source for our cases may also affect the breadth of our findings, limiting their broader applicability. The varying complexity of cases, which span from a junior resident to staff ophthalmologist level, limits our results, and it remains unclear whether there is a relationship between the chatbot's performance and case difficulty. Certain cases had a limited number of different ophthalmic images or imaging modalities, which may have affected the chatbot's ability to achieve a correct diagnosis. Unless the age and sex were not provided for a particular case on OCTCases, as was seen in 2 cases, we always began our sensitivity analysis by priming the chatbot with a case's age and sex, as the chatbot struggled to provide valid responses in the complete absence of clinical context. Our results also cannot be used to anticipate the chatbot's performance as a decision making aid in clinical settings, where patient presentations as well as imaging techniques and quality can vary greatly. Furthermore, the quality of the chatbot's differential diagnoses, when provided, was not assessed relative to the differential of a retina specialist. With significant advancements in multimodal imaging technologies,[29] there is a concern that the current version of the chatbot may not be equipped to keep pace with the rapidly evolving field of ophthalmic research. Our findings are also bound to the time frame of the study as subsequent versions of the chatbot, with their enriched knowledge base, may perform differently if our methodology were to be repeated. Given the strong contextual dependency of the chatbot, our findings may not be generalizable to different methods of inputting prompts. Our study also employed the "My GPTs" feature to create a custom chatbot, whose performance may have differed if different training instructions were provided. Given the chatbot is trained on online resources, it is unclear the extent to which its knowledge corpus may have assimilated information from OCTCases, potentially biasing our results. However, the chatbot's training data set has not been publicly disclosed; thus, we are unable to ascertain whether the chatbot may have previously encountered the same content probed in our investigation. Lastly, as many cases required multiple imaging modalities to be inputted into the chatbot simultaneously, this precluded our ability to analyze the chatbot's diagnostic performance based on independent imaging modalities.

In conclusion, our custom chatbot was able to accurately diagnose approximately half of the retinal cases requiring multimodal input from OCTCases, albeit relying heavily on text-based contextual information accompanying ophthalmic images. The diagnostic ability of the chatbot in the interpretation of multimodal imaging without text-based information is currently limited. The appropriate use of the chatbot in this setting is of utmost importance, given bioethical concerns. Although the chatbot is currently not suitable to inform clinical decision making in ophthalmology, its ability to interpret multimodal ophthalmic cases in real-time should be continually evaluated over time.

Acknowledgments

Austin Pereira and Jason Kwok are the cofounders of OCTCases, Andrew Mihalache is involved in website design for OCTCases, and Peng Yan is a staff supervisor for OCTCases. Dr Radha P. Kohly and Rajeev H. Muni's research is supported by the Silber TARGET Fund. Information reported in this manuscript has not been previously presented at a conference. Data were collected from the AI chatbot ChatGPT-4 developed by OpenAI.

Footnotes and Disclosures

Originally received: March 5, 2024.
Final revision: May 13, 2024.
Accepted: May 17, 2024.
Available online: May 23, 2024. Manuscript no. XOPS-D-24-00068R1.

[1] Temerty School of Medicine, University of Toronto, Toronto, Ontario, Canada.

[2] Department of Ophthalmology and Vision Sciences, University of Toronto, Toronto, Ontario, Canada.

[3] John and Liz Tory Eye Centre, Sunnybrook Health Science Centre, Toronto, Ontario, Canada.

[4] Department of Ophthalmology, St. Michael's Hospital/Unity Health Toronto, Toronto, Ontario, Canada.

Meeting Presentation: None.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The author(s) have made the following disclosure(s):

M.M.P.: Financial support (to institution) — PSI Foundation, Fighting Blindness Canada.

D.T.W.: Grants/research support (to institution) — Bayer, Novartis, Roche; Consulting fees — Alcon, AbbVie, Apellis, Bauch Health, Bayer, Biogen, Novartis, Ripple Therapeutics, Roche, Zeiss.

P.J.K.: Honoraria — Novartis, Bayer, Roche, Boehringer Ingelheim, RegenxBio, Apellis; Advisory board — Novartis, Bayer, Roche, Apellis, Novelty Nobility, Viatris, Biogen, AdMare, Kriya Therapeutics; Financial support (to institution) — Roche, Novartis, Bayer, RegenxBio.

R.P.K.: Other financial or non-financial interests — Bayer, Novartis.

R.H.M.: Consulting fees — Alcon, Apellis, AbbVie, Bayer, Bausch Health, Roche; Financial Support (to institution) — Alcon, AbbVie, Bayer, Novartis, Roche.

The other authors have no proprietary or commercial interest in any materials discussed in this article.

Financial Support: None.

Rajeev H. Muni, an editor of this journal, was recused from the peer-review process of this article and had no access to information regarding its peer-review.

HUMAN SUBJECTS: No human subjects were included in this study. The University of Toronto waived institutional review board approval for the publication of cases with ophthalmic imaging on the OCTCases website, given that all cases had been entirely anonymized and contained modified patient characteristics that were not identifiable to individual patients. The research adhered to the tenets of the Declaration of Helsinki and employed Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guidelines.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Mihalache, Huang, Mikhail, Popovic, Pereira, Yan, Kohly, Muni

Data collection: Mihalache, Huang, Mikhail

Analysis and interpretation: Mihalache, Huang, Mikhail, Kohly, Muni

Obtained funding: N/A

Overall responsibility: Mihalache, Huang, Mikhail, Popovic, Shor, Pereira, Kwok, Yan, Wong, Kertes, Kohly, Muni

Abbreviations and Acronyms:
AI = artificial intelligence; **BCVA** = best-corrected visual acuity; **ChatGPT** = Chat Generative Pre-Trained Transformer.

Keywords:
Artificial intelligence, Image processing, Natural language processing, OCT.

Correspondence:
Rajeev H. Muni, MD, MSc, Department of Ophthalmology and Vision Sciences, University of Toronto, Toronto, Ontario, Canada. Department of Ophthalmology, St. Michael's Hospital/Unity Health Toronto, 30 Bond St, Donnelly Wing, 8th Floor, Toronto, Ontario, Canada, M5B 1W8. E-mail: rajeev.muni@utoronto.ca.

References

1. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. *JAMA Ophthalmol*. 2023;141:589–597. https://doi.org/10.1001/jamaophthalmol.2023.1144.
2. Mihalache A, Huang RS, Popovic MM, Muni RH. Performance of an upgraded artificial intelligence chatbot for ophthalmic knowledge assessment. *JAMA Ophthalmol*. 2023;141:798–800. https://doi.org/10.1001/jamaophthalmol.2023.2754.
3. Lyons RJ, Arepalli SR, Fromal O, et al. Artificial intelligence chatbot performance in triage of ophthalmic conditions. *Can J Ophthalmol*. 2023. https://doi.org/10.1016/j.jcjo.2023.07.016.
4. Keenan TDL, Loewenstein A. Artificial intelligence for home monitoring devices. *Curr Opin Ophthalmol*. 2023;34:441–448. https://doi.org/10.1097/ICU.0000000000000981.
5. Bernstein IA, Zhang YV, Govil D, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. *JAMA Netw Open*. 2023;6:e2330320. https://doi.org/10.1001/jamanetworkopen.2023.30320.
6. Tan TF, Thirunavukarasu AJ, Jin L, et al. Artificial intelligence and digital health in global eye health: opportunities and challenges. *Lancet Glob Health*. 2023;11:e1432–e1443. https://doi.org/10.1016/S2214-109X(23)00323-6.
7. Momenaei B, Wakabayashi T, Shahlaee A, et al. Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases. *Ophthalmol Retina*. 2023;7:862–868. https://doi.org/10.1016/j.oret.2023.05.022.
8. Potapenko I, Boberg-Ans LC, Stormly Hansen M, et al. Artificial intelligence-based chatbot patient information on common retinal diseases using ChatGPT. *Acta Ophthalmol*. 2023;101:829–831. https://doi.org/10.1111/aos.15661.
9. Ferro Desideri L, Roth J, Zinkernagel M, Anguita R. Application and accuracy of artificial intelligence-derived large language models in patients with age related macular degeneration. *Int J Retina Vitreous*. 2023;9:71. https://doi.org/10.1186/s40942-023-00511-7.
10. Srivastava O, Tennant M, Grewal P, et al. Artificial intelligence and machine learning in ophthalmology: a review. *Indian J Ophthalmol*. 2023;71:11–17. https://doi.org/10.4103/ijo.IJO_1569_22.
11. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med*. 2018;24:1342–1350. https://doi.org/10.1038/s41591-018-0107-6.
12. OCTCases. https://www.octcases.com/. Accessed December 6, 2023.
13. Schoonjans F, Zalata A, Depuydt CE, Comhaire FH. MedCalc: a new computer program for medical statistics. *Comput Methods Programs Biomed*. 1995;48:257–262. https://doi.org/10.1016/0169-2607(95)01703-8.
14. Campbell I. Chi-squared and Fisher-Irwin tests of two-by-two tables with small sample recommendations. *Stat Med*. 2007;26:3661–3675. https://doi.org/10.1002/sim.2832.
15. Richardson JTE. The analysis of 2 × 2 contingency tables—yet again. *Stat Med*. 2011;30:890–890. https://doi.org/10.1002/sim.4116.

16. Statistics Science Statistics. Spearman's Rho Calculator (Correlation Coefficient). https://www.socscistatistics.com/tests/spearman/default.aspx; 2020. Accessed December 26, 2023.
17. MedCalc. Mann-Whitney test (independent samples). https://www.medcalc.org/manual/mannwhitney.php. Accessed July 25, 2023.
18. OpenAI. GPT-4V(ision) system card. https://openai.com/research/gpt-4v-system-card; 2023. Accessed January 7, 2024.
19. Sarvamangala DR, Kulkarni RV. Convolutional neural networks in medical image understanding: a survey. *Evol Intell*. 2022;15:1−22. https://doi.org/10.1007/s12065-020-00540-3.
20. Mihalache A, Huang RS, Popovic MM, et al. Accuracy of an artificial intelligence chatbot to interpret clinical ophthalmic images. *JAMA Ophthalmol*. 2024;142:321−326. https://doi.org/10.1001/jamaophthalmol.2024.0017.
21. Bai J, Wan Z, Li P, et al. Accuracy and feasibility with AI-assisted OCT in retinal disorder community screening. *Front Cell Dev Biol*. 2022;10:1053483. https://doi.org/10.3389/fcell.2022.1053483.
22. Leng X, Shi R, Wu Y, et al. Deep learning for detection of age-related macular degeneration: a systematic review and meta-analysis of diagnostic test accuracy studies. *PLoS One*. 2023;18:e0284060. https://doi.org/10.1371/journal.pone.0284060.
23. Nielsen KB, Lautrup ML, Andersen JKH, Savarimuthu TR, Grauslund J. Deep learning−based algorithms in screening of diabetic retinopathy: a systematic review of diagnostic performance. *Ophthalmol Retina*. 2019;3:294−304. https://doi.org/10.1016/j.oret.2018.10.014.
24. Liu X, Zhao C, Wang L, et al. Evaluation of an OCT-AI−based telemedicine platform for retinal disease screening and referral in a primary care setting. *Trans Vis Sci Tech*. 2022;11:4. https://doi.org/10.1167/tvst.11.3.4.
25. Cao S, Zhang R, Jiang A, et al. Application effect of an artificial intelligence-based fundus screening system: evaluation in a clinical setting and population screening. *Biomed Eng Online*. 2023;22:38. https://doi.org/10.1186/s12938-023-01097-9.
26. Kim KM, Heo TY, Kim A, et al. Development of a fundus image-based deep learning diagnostic tool for various retinal diseases. *J Pers Med*. 2021;11:321. https://doi.org/10.3390/jpm11050321.
27. Jassar S, Adams SJ, Zarzeczny A, Burbridge BE. The future of artificial intelligence in medicine: Medical-legal considerations for health leaders. *Healthc Manage Forum*. 2022;35:185−189. https://doi.org/10.1177/08404704221082069.
28. Mihalache A, Huang RS, Popovic MM, Muni RH. ChatGPT-4: n assessment of an upgraded artificial intelligence chatbot in the United States Medical Licensing Examination. *Med Teach*. 2024;46:366−372. https://doi.org/10.1080/0142159X.2023.2249588.
29. Ringel MJ, Tang EM, Tao YK. Advances in multimodal imaging in ophthalmology. *Ther Adv Ophthalmol*. 2021;13:25158414211002400. https://doi.org/10.1177/25158414211002400.