AMERICAN ACADEMY OF OPHTHALMOLOGY®

Evaluating the Performance of ChatGPT in Ophthalmology

An Analysis of Its Successes and Shortcomings

Fares Antaki, MD, CM,[1,2,3,4] *Samir Touma, MD, CM,*[1,2,3] *Daniel Milad, MD,*[1,2,3] *Jonathan El-Khoury, MD,*[1,2,3] *Renaud Duval, MD*[1,2]

Purpose: Foundation models are a novel type of artificial intelligence algorithms, in which models are pretrained at scale on unannotated data and fine-tuned for a myriad of downstream tasks, such as generating text. This study assessed the accuracy of ChatGPT, a large language model (LLM), in the ophthalmology question-answering space.

Design: Evaluation of diagnostic test or technology.

Participants: ChatGPT is a publicly available LLM.

Methods: We tested 2 versions of ChatGPT (January 9 "legacy" and ChatGPT Plus) on 2 popular multiple choice question banks commonly used to prepare for the high-stakes Ophthalmic Knowledge Assessment Program (OKAP) examination. We generated two 260-question simulated exams from the Basic and Clinical Science Course (BCSC) Self-Assessment Program and the OphthoQuestions online question bank. We carried out logistic regression to determine the effect of the examination section, cognitive level, and difficulty index on answer accuracy. We also performed a post hoc analysis using Tukey's test to decide if there were meaningful differences between the tested subspecialties.

Main Outcome Measures: We reported the accuracy of ChatGPT for each examination section in percentage correct by comparing ChatGPT's outputs with the answer key provided by the question banks. We presented logistic regression results with a likelihood ratio (LR) chi-square. We considered differences between examination sections statistically significant at a P value of < 0.05.

Results: The legacy model achieved 55.8% accuracy on the BCSC set and 42.7% on the OphthoQuestions set. With ChatGPT Plus, accuracy increased to 59.4% ± 0.6% and 49.2% ± 1.0%, respectively. Accuracy improved with easier questions when controlling for the examination section and cognitive level. Logistic regression analysis of the legacy model showed that the examination section (LR, 27.57; $P = 0.006$) followed by question difficulty (LR, 24.05; $P < 0.001$) were most predictive of ChatGPT's answer accuracy. Although the legacy model performed best in general medicine and worst in neuro-ophthalmology ($P < 0.001$) and ocular pathology ($P = 0.029$), similar post hoc findings were not seen with ChatGPT Plus, suggesting more consistent results across examination sections.

Conclusion: ChatGPT has encouraging performance on a simulated OKAP examination. Specializing LLMs through domain-specific pretraining may be necessary to improve their performance in ophthalmic subspecialties.

Financial Disclosure(s): Proprietary or commercial disclosure may be found in the Footnotes and Disclosures at the end of this article. *Ophthalmology Science 2023;3:100324 © 2023 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/).*

Supplemental material available at www.ophthalmologyscience.org.

Since 2015, significant progress has been made in the application of artificial intelligence (AI) and deep learning (DL) in medicine, particularly in ophthalmology.[1] Deep learning has been widely used for image recognition using various types of ophthalmic data, such as fundus photographs and OCT, and has shown strong results in detecting a wide range of diseases.[2,3] More recently, there has been growing interest in using DL for natural language processing in ophthalmology, which involves using AI to understand and interact with human language.[4]

Natural language processing has received considerable media attention in the past months due to the release of large DL models called foundation models.[5] Foundation models represent a novel paradigm for building AI systems, whereby models are pretrained at *scale* on vast amounts of unannotated multimodal data in a self-supervised

https://doi.org/10.1016/j.xops.2023.100324
ISSN 2666-9145/23

manner. They are subsequently fine-tuned for a myriad of downstream tasks through a process called *transfer learning*.[6,7] The incredible *scale* of foundation models, which can now contain billions of parameters, has been made possible by advances in computer hardware coupled with the *transformer* model architecture and the availability of vast amounts of training data.[6] A prominent example of such models is Generative Pretrained Transformer 3 (GPT-3), a large language model (LLM) that generates human-like text. It was trained on a massive data set of text (> 400 billion words) from the internet, including books, articles, and websites.[8]

There has been recent interest in evaluating the capabilities of LLMs for understanding and generating natural language in medicine.[9,10] The medical domain can pose a significant challenge for LLMs because clinical reasoning often requires years of training and hands-on experience to master. In 2022, Singhal et al[9] demonstrated the capabilities of Pathways Language Model, a 540-billion parameter LLM, by testing it on multiple choice questions from the United States Medical Licensing Examination (USMLE) with an impressive 67.6% accuracy. More recently, Kung et al[11] evaluated the performance of ChatGPT, a generic LLM developed by OpenAI that is based on the GPT-3 series and optimized for dialogue, using multiple choice questions also from the USMLE. They found that ChatGPT achieved overall accuracy of > 50% in most of their experiments and also provided insightful explanations to support its answer choices.

To our knowledge, the performance of LLMs has not yet been examined in the ophthalmology question-answering space. In this study, we evaluated the performance of ChatGPT in ophthalmology by using 2 popular board preparation question banks: the American Academy of Ophthalmology's Basic and Clinical Science Course (BCSC) Self-Assessment Program and the OphthoQuestions online question bank. These resources have been shown to be effective in studying for board examinations and have been linked to improved performance on the standardized Ophthalmic Knowledge Assessment Program (OKAP) examination, which is taken annually by ophthalmology residents in the United States and Canada.[12,13]

Methods

ChatGPT Is an LLM

ChatGPT (OpenAI) is a fine-tuned LLM based on a model from the GPT-3.5 series called "gpt-3.5-turbo."[14] Generative Pretrained Transformer 3 has a transformer architecture and was trained using billions of text data obtained from writings on the internet. This process is done by training the model to minimize the difference between the predicted word and the actual word in the training data set. Once the model is trained, it can be used to generate new text by providing it with a prompt and allowing it to predict the next word. The model then uses this predicted word as the context for the next prediction, and this process is repeated until a complete sentence or paragraph is generated.[8] ChatGPT goes beyond just predicting the next word because it is optimized for dialogue and was trained using human feedback. This allows it to understand and respond to human expectations when answering questions.[15]

ChatGPT January 9 (Legacy Model) and ChatGPT Plus

We tested 2 versions of ChatGPT. In the initial experiment, we used the free research preview that was released on January 9, 2023, which we will subsequently refer to as the "legacy model." While conducting further experiments, OpenAI unveiled a newly upgraded model on January 30, 2023, that boasted "enhanced factuality and mathematical capabilities." Shortly thereafter, ChatGPT Plus was introduced as a subscription-based service, offering faster responses and priority access.[16] Because previous models of ChatGPT were made inaccessible as new ones were released, we used ChatGPT Plus for subsequent experiments to ensure the stability of results.

Repeatability of ChatGPT Performance

We were only able to run a single experiment with the legacy model of ChatGPT before it was made unavailable by OpenAI. Using the reliable ChatGPT Plus, we conducted multiple experiments to measure the variability and establish the reproducibility of our results. We anticipated that the responses provided by ChatGPT would exhibit some variability across different runs because of the probabilistic nature of LLMs. We repeated the experiments 3 times for each of the BCSC and OphthoQuestions sets by manually composing the prompts and extracting responses from the ChatGPT website.

BCSC and OphthoQuestions

In January 2023, we generated a test set of 260 questions from the BCSC Self-Assessment Program and 260 questions from OphthoQuestions through personal subscription accounts. Permission was obtained from the American Academy of Ophthalmology for use of the underlying BCSC Self-Assessment program materials. Those questions are not publicly accessible, thereby excluding the possibility of prior indexing in any search engine (like Google) or in the ChatGPT training data set. For the BCSC and OphthoQuestions test sets, we randomly generated 260 questions out of a pool of 4458 and 4539 potential questions, respectively. During the process, any questions that included visual information, such as clinical, radiologic, or graphical images, were removed and replaced because ChatGPT does not currently support such data. We generated 20 random questions based on the 13 sections of the OKAP examination: update on general medicine, fundamentals and principles of ophthalmology, clinical optics and vision rehabilitation, ophthalmic pathology and intraocular tumors, neuro-ophthalmology, pediatric ophthalmology and strabismus, oculofacial plastic and orbital surgery, external disease and cornea, uveitis and ocular inflammation, glaucoma, lens and cataract, retina and vitreous, and refractive surgery.

Question Format and Encoding

We aimed to replicate an OKAP examination and therefore maintained the standard multiple choice format with 1 correct answer and 3 incorrect options (distractors). We employed a zero-shot approach for the lead-in prompt, using the prompt "Please select the correct answer and provide an explanation" followed by the question and answer options, without providing any examples.[9] Although more challenging for ChatGPT,[8] we chose this technique because it is the closest to human test-taking. A new session was started in ChatGPT for each question to reduce memory retention bias.

Level of Cognition and Question Difficulty

Because the BCSC and OphthoQuestions questions were not labeled for difficulty, we labeled them according to the cognitive level and calculated a difficulty index.[17] We did this to analyze ChatGPT's performance based on not only the subject but also the type of question and level of difficulty. Despite having no control over the distribution of cognitive level and question difficulty in each of the randomly generated test sets, we elected not to balance them manually to prevent cherry-picking, thereby avoiding bias in the experiment results.

We used a simplified scoring system of low and high cognitive level, instead of the 3-tier system proposed in the OKAP User's Guide.[18] This was done because we found it difficult to distinguish between level 2 and level 3 questions, and we wanted to avoid making assumptions about the intended goal of the questions. Low-cognitive-level questions tested recall of facts and concepts, such as identifying the gene implicated in a known condition. High-cognitive-level questions tested the ability to interpret data, make calculations and manage patients, like in common clinical optics exercises (e.g., cylinder transpositions) or to select the best treatment for specific cancers in unique clinical contexts (e.g., the optimal treatment for metastatic sebaceous cell carcinoma of the eyelid). The difficulty index represented the percentage of individuals who correctly answered a question, as reported by BCSC and OphthoQuestions platforms for each question. Questions with a higher difficulty index are considered easier. The questions were categorized into 3 levels of difficulty: difficult (< 30%), moderate (≥ 30% and < 70%), and easy (≥ 70%).[19]

Statistical Analysis

Accuracy was determined by comparing ChatGPT's answer with the answer key provided by the question banks. The legacy model's accuracy was determined from a single run, whereas the means and standard deviations of the ChatGPT Plus model were derived from data collected over 3 runs. The degree of repeatability for those runs was assessed within each examination section using a κ measure for M raters.[20] We used logistic regression (all the input variables were entered simultaneously) to examine the effect of the examination section, cognitive level, and difficulty index on ChatGPT's answer accuracy. We then performed a post hoc analysis using Tukey's test to determine if there were significant differences in accuracy between examination sections while controlling for question difficulty and cognitive level. By controlling for those factors, we were able to isolate the effect of the examination section on accuracy and determine if there were any meaningful differences between the tested topics. For the ChatGPT Plus model analyses, we combined the results of the 3 runs. Therefore, to account for correlated values, we used a generalized estimating equation model with an exchangeable correlation matrix and a binomial distribution using a logit link.

Results

The Testing Sets Demonstrated Similar Difficulty and Cognitive Levels

The BCSC and OphthoQuestions training sets had a similar level of difficulty ($P = 0.154$) and mostly included easy and moderate questions in a very similar distribution ($P = 0.102$), as illustrated in Figure 1 and Table 1. Likewise, the questions' cognitive levels were comparable between the 2 test sets ($P = 0.425$). Those similarities allowed us to combine the testing sets during further analyses.

ChatGPT Had a Modest Overall Performance Initially but It Improved After a Model Update

In the initial experiment, the legacy model achieved an accuracy of 55.8% on the BCSC set and 42.7% on the OphthoQuestions set. However, with the improved ChatGPT Plus, the accuracy increased to 59.4% ± 0.6% on the BCSC set and 49.2% ± 1.0% on the OphthoQuestions set. Table S2 (available at www.ophthalmologyscience.org) and Figure 2 show the variations in performance between the ChatGPT models and the BCSC and OphthoQuestions sets for the same examination section. Taking the data sets together, the legacy model performed well in general medicine (75%), fundamentals (60%), and cornea (60%), but not as well in neuro-ophthalmology (25%), glaucoma (37.5%), and pediatrics and strabismus (42.5%). The updated ChatGPT Plus model consistently excelled in its strongest subjects: fundamentals (68.3% ± 6.1%), general medicine (65.8% ± 7.4%), and cornea (65.0% ± 3.2%). However, its weakest subject remained neuro-ophthalmology (40.0% ± 9.5%) in addition to oculoplastics (40.8% ± 9.2%) and clinical optics (45.8% ± 10.2%).

ChatGPT Plus Answers Were Consistent across Runs with Substantial to Almost Perfect Repeatability

The κ values were 0.769 (95% confidence interval [CI], 0.699−0.839) for the BCSC set and 0.798 (95% CI, 0.728−0.868) for the OphthoQuestions set. When these sets were combined, the resulting κ value was 0.786 (95% CI, 0.736−0.835). These findings indicate that the CIs fell within the range of substantial to almost perfect repeatability.[21] The mean overall accuracies for each of the runs are shown in Table S2, and they had a maximum difference of 1.9%.

ChatGPT's Accuracy Depends on the Examination Section, Cognitive Level, and Question Difficulty

Our initial experiments on the legacy model showed that the examination section (likelihood ratio [LR], 27.57; $P = 0.006$) followed by question difficulty (LR, 24.05; $P < 0.001$) were most predictive of answer accuracy (Tables 3, S4, available at www.ophthalmologyscience.org). The initial experiments also showed that while controlling for question difficulty and cognitive level, there were significant differences in the legacy model's performance between general medicine and each of glaucoma ($P = 0.002$), neuro-ophthalmology ($P < 0.001$), and ophthalmic pathology and intraocular tumors ($P = 0.029$) (Fig S3, available at www.ophthalmologyscience.org). Similarly, we found that accuracy improved with increased difficulty index (easier questions) even when controlling for the examination section and cognitive level. Figures S4 and S5 (available at www.ophthalmologyscience.org) provide the results of the post hoc analysis for each of the testing sets.

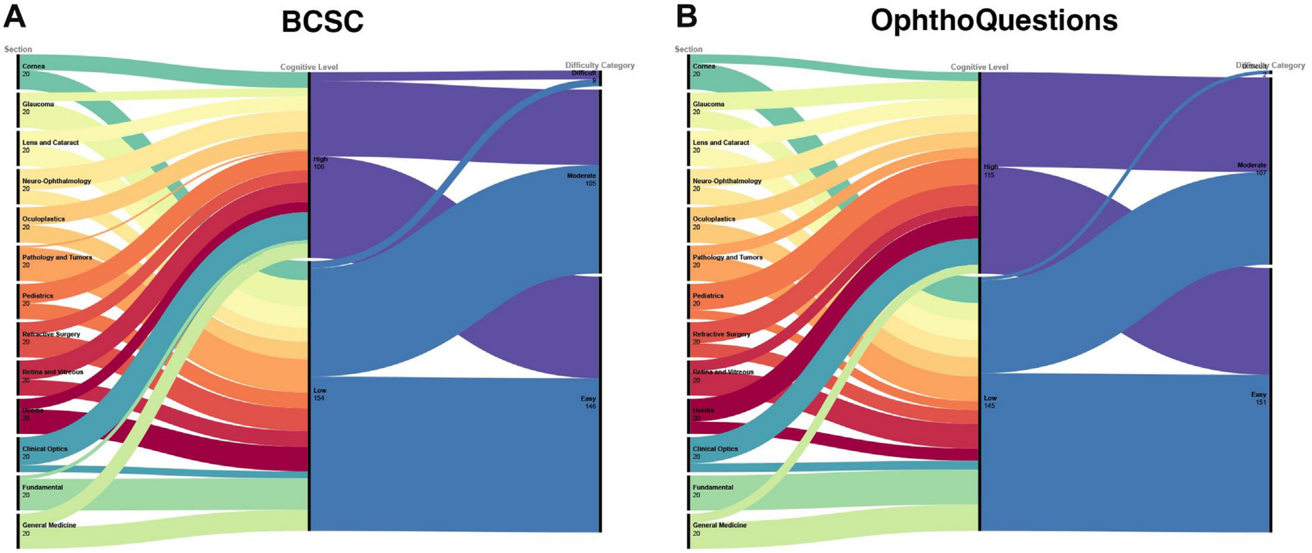

Figure 1. Alluvial diagram illustrating the distribution of questions across examination sections, cognitive level, and question difficulty. Despite having been generated at random, the Basic and Clinical Science Course (BCSC) and OphthoQuestions test sets have a similar distribution of questions with high and low cognitive levels and similar difficulty.

With the improved ChatGPT Plus model, question difficulty (LR, 32.30; $P < 0.001$), followed by examination section (LR, 23.40; $P = 0.024$), and cognitive level (LR, 5.60; $P = 0.018$) were all predictive of ChatGPT Plus's answer accuracy (Tables 3, S5 available at www.ophthalmologyscience.org). Although there was a significant global effect of examination section on ChatGPT Plus's performance, our analysis of pairwise differences, which accounted for multiple comparisons, did not reveal any statistically significant differences in performance across the examination sections (Fig S6, available at www.ophthalmologyscience.org). This contrasts with the results obtained from the legacy model, making it a noteworthy finding. Similar to the legacy model, accuracy improved with easier questions even when controlling the examination section and cognitive level. The results of the post hoc analysis for each testing set can be found in Figures S7 and S8 (available at www.ophthalmologyscience.org).

Discussion

In the past months, there has been significant interest in examining the utility of LLMs in medicine.[5] Despite having encouraging impacts in various industries, it is important to thoroughly evaluate their performance and biases before determining their clinical usefulness.[4,22] In this study, we

Table 1. Baseline Characteristics of the Testing Sets

	BCSC (n = 260)	OphthoQuestions (n = 260)	*P* Value*
Difficulty index			
Mean	0.69	0.719	0.154
Median	0.73	0.750	
Interquartile range	0.28	0.260	
Range	0.00–0.94	0.21–0.96	
Difficulty category			0.102
Easy	146 (56.2%)	151 (58.1%)	
Moderate	105 (40.4%)	107 (41.2%)	
Difficult	9 (3.5%)	2 (0.8%)	
Cognitive level			0.425
High	106 (40.7%)	115 (44.2%)	
Low	154 (59.3%)	145 (55.8%)	

BCSC = Basic and Clinical Science Course.
*Mann–Whitney *U* nonparametric test (difficulty index) and chi-square test (difficulty category and cognitive level)

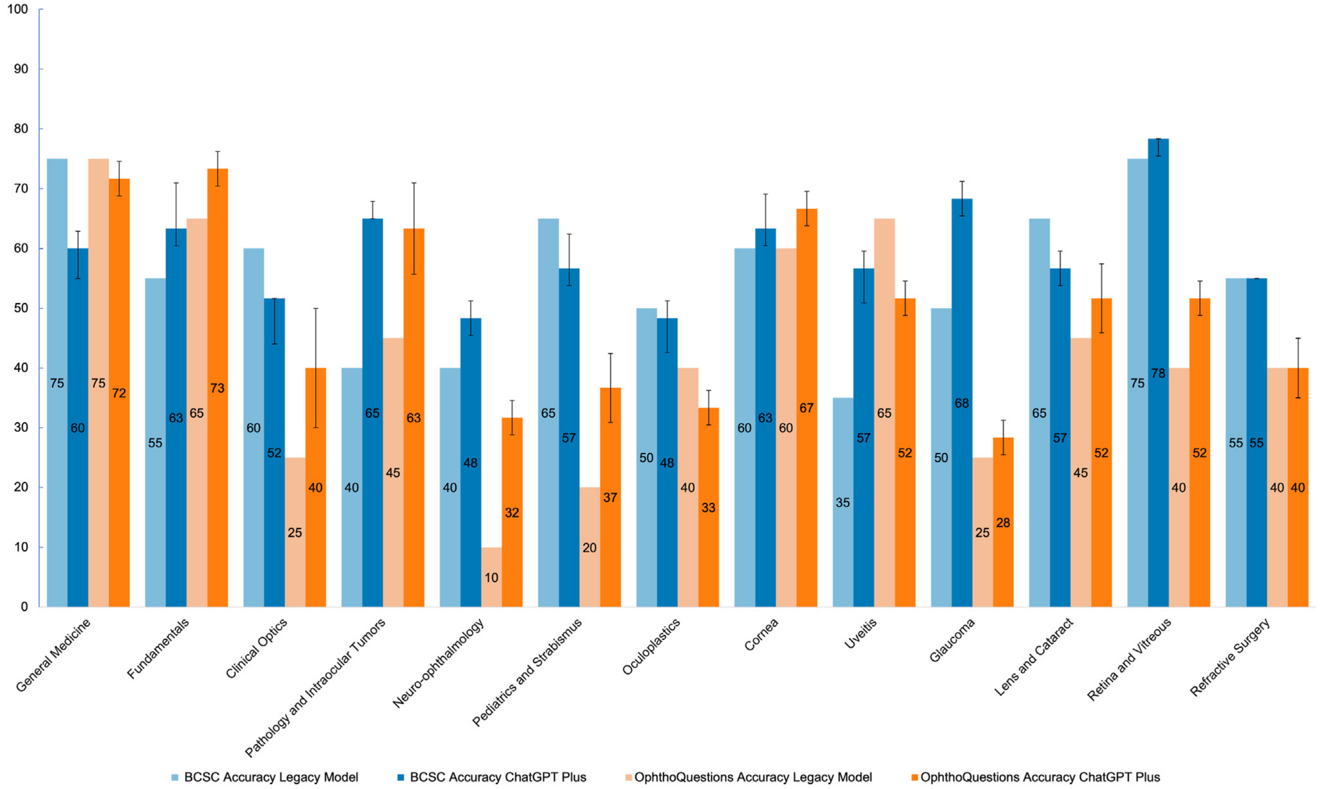

Figure 2. Bar plot of the accuracy of ChatGPT across examination sections and ChatGPT models for the Basic and Clinical Science Course (BCSC) and OphthoQuestions testing sets. The ChatGPT Plus model accuracy is shown with error bars representing the standard deviation from the 3 experimental runs.

provide evidence on the performance of ChatGPT, a non—domain-specific LLM, in responding to questions similar to those found on the OKAP examination.

During experimentation, we observed notable improvement in ChatGPT's performance as the model was updated. The most powerful model (ChatGPT Plus) achieved an accuracy of 59.4% on the simulated OKAP examination using the BCSC testing set and 49.2% on the OphthoQuestions testing set. To put the results into perspective, we aggregated historical human performance data in Table S6 (available at www.ophthalmologyscience.org). The data indicate that, on average, humans score 74% on the BCSC question bank; and in the last 3 years, the group of ophthalmology residents who completed their training in 2022 obtained an average score of 63% on OphthoQuestions. Despite being slightly out of range of human performance, we believe that this outcome is noteworthy and promising within ophthalmology, as our results approach ChatGPT's performance on the USMLE despite being a much more specialized examination.[11] Our findings are also encouraging as ChatGPT's accuracy in ophthalmology is similar to the typical accuracy seen in general medical question answering by state-of-the-art LLMs, typically ~ 40% to 50%, as reported in publications from 2022.[9]

We found that the accuracy of the legacy model mostly depended on the examination section, even when controlling for question difficulty and cognitive level. This effect was less pronounced in the updated version of ChatGPT. ChatGPT consistently excelled in general medicine, fundamentals, and cornea. The model's high performance in these areas might be attributed to the vast amount of training data and resources available on the internet for those topics. In contrast, the legacy model performed poorest in neuro-ophthalmology as well as ophthalmic pathology and intraocular tumors. Those are highly specialized domains that are

Table 3. Comparing the Likelihood Ratios for Examination Section, Cognitive Level, and Difficulty Index for the Legacy and ChatGPT Plus Models (Testing Sets Combined)

Effects	LR Chisq	Df	Pr(>Chisq)
Model: legacy			
Section	27.57	12	0.006*
Cognitive level	3.54	1	0.06
Difficulty index	24.05	1	< 0.001*
Model: ChatGPT Plus			
Section	23.40	12	0.024*
Cognitive level	5.60	1	0.018*
Difficulty index	32.30	1	< 0.001*

BCSC = Basic and Clinical Science Course; LR = likelihood ratio.
*Statistically significant at the 0.05 level

considered challenging even within the ophthalmology community. For example, up to 40% of patients referred to a neuro-ophthalmology subspecialty service are misdiagnosed,[23] and similar referral patterns are observed in ocular oncology.[24] The updated ChatGPT Plus model continued to perform poorly in neuro-ophthalmology, but its performance in pathology and intraocular tumors improved.

Understanding why ChatGPT makes mistakes is important. We found that question difficulty was predictive of ChatGPT's accuracy, even when controlling for the examination section and cognitive level. ChatGPT was more accurate when a higher percentage of human peers obtained the right answer for a specific question. This discovery is comforting as it suggests that ChatGPT's responses align, to a certain degree, with the collective understanding of ophthalmology trainees. In parallel, Kung et al[11] showed that the accuracy of ChatGPT is heavily influenced by concordance and insight, indicating that inaccurate responses are caused by a lack of training information for the USMLE. We plan to perform a similar qualitative analysis to identify areas for improvement in the ophthalmology space. Incorporating ChatGPT with other specialized foundation models that are trained using domain-specific sources (such as EyeWiki) might be required to improve its accuracy.

Despite its encouraging performance, the imminent implementation of ChatGPT in ophthalmology may be limited because it does not have the capability to process images. This is a significant limitation because ophthalmology is a field that heavily relies on visual examination and imaging to diagnose, treat, and monitor patients. Large language model, such as ChatGPT, may need to incorporate other transformer models that can handle multiple types of data, such as the Contrastive Language-Image Pretraining model,[25] which can classify images and generate a text description that ChatGPT can then use to respond to a question. Although this approach shows potential, it is limited by its reliance on a large amount of image−text pairs from the internet (in the case of the Contrastive Language-Image Pretraining model) that are not specific to our domain. These data may not be sufficient to accurately distinguish subtle and specific differences relevant to medicine and ophthalmology.[26] For instance, the Contrastive Language-Image Pretraining model may not be able to accurately caption a "superior" retinal detachment that would need a pneumatic retinopexy, as opposed to an "inferior" retinal detachment that might require a scleral buckle.

Although we could not obtain repeat experiments on the legacy model, we conducted the ChatGPT Plus experiments thrice to ensure the consistency of our findings. This process proved to be extremely labor-intensive. We believe that the availability of an application programming interface for ChatGPT may facilitate more thorough validation of this technology in the future and potentially alleviate the labor-intensive nature of the process. Generally, we found that ChatGPT Plus provided highly consistent and repeatable results, but some variations occurred. We expected that because ChatGPT is a probabilistic model that works by predicting the likelihood of a particular sequence of words appearing in a language. The model calculates the probability of each possible next word given the previous words in the sequence, and the probability distribution of each next word is based on statistical patterns learned from its training data. Until recently, ChatGPT could not be made more deterministic, but the latest application programming interface release now permits such modifications. By adjusting the "temperature" setting, you can either decrease it to maintain the model's emphasis on the prompt's intention (more deterministic) or increase it to allow it to digress (more probabilistic). Determining the appropriate temperature for each case use and each clinical context may be necessary as we experiment further with those models.

As the performance of ChatGPT improves (perhaps through prompting strategies and through updates by OpenAI), it will be important to work collectively toward building safeguards for our patients.[4] Those will include protecting vulnerable populations from biases and evaluating the potential harm or risk of acting on the answers provided by LLMs, such as ChatGPT. This will be particularly important for high-level decision-making questions that may be challenging to train for because of inconclusive training data on the internet, reflecting the variability in research data as well as global practice patterns. We are excited about the potential of ChatGPT in ophthalmology, but we remain cautious when considering the potential clinical applications of this technology.

Acknowledgments

The authors thank Mr. Charles-Édouard Giguère, statistician at the Institut Universitaire en Santé Mentale de Montréal, for his assistance in the statistical analysis. We thank the American Academy of Ophthalmology for generously granting us permission to use the underlying BCSC Self-Assessment Program materials.

Footnotes and Disclosures

Originally received: February 3, 2023.
Final revision: April 21, 2023.
Accepted: April 25, 2023.
Available online: May 5, 2023. Manuscript no. XOPS-D-23-00026R2.

[1] Department of Ophthalmology, Université de Montréal, Montréal, Quebec, Canada.

[2] Centre Universitaire d'Ophtalmologie (CUO), Hôpital Maisonneuve-Rosemont, CIUSSS de l'Est-de-l'Île-de-Montréal, Montréal, Quebec, Canada.

[3] Department of Ophthalmology, Centre Hospitalier de l'Université de Montréal (CHUM), Montréal, Quebec, Canada.

[4] The CHUM School of Artificial Intelligence in Healthcare (SAIH), Centre Hospitalier de l'Université de Montréal (CHUM), Montréal, Quebec, Canada.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The authors have made the following disclosures: F.A.: Grants − Bayer.

S.T.: Grants − Bayer.

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

HUMAN SUBJECTS: No human subjects were included in this study.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Antaki, Touma, Milad, El-Khoury, Duval

Analysis and interpretation: Antaki, Duval

Data collection: Antaki, Touma, Milad, El-Khoury

Obtained funding: N/A

Overall responsibility: Antaki, Touma, Milad, El-Khoury, Duval

Abbreviations and Acronyms:

AI = artificial intelligence; **BCSC** = Basic and Clinical Science Course; **CI** = confidence interval; **DL** = deep learning; **GPT** = Generative Pretrained Transformer; **LLM** = large language model; **LR** = likelihood ratio; **OKAP** = Ophthalmic Knowledge Assessment Program; **USMLE** = United States Medical Licensing Examination.

Keywords:

Artificial intelligence, ChatGPT, Generative Pretrained Transformer, Medical education, Ophthalmology.

Correspondence:

Renaud Duval, MD, Centre Universitaire d'Ophtalmologie (CUO), Hôpital Maisonneuve-Rosemont, 5415 Boulevard de l'Assomption, Montréal, Québec, Canada, H1T 2M4. E-mail: renaud.duval@gmail.com.

References

1. Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol*. 2019;103:167−175.
2. Schmidt-Erfurth U, Sadeghipour A, Gerendas BS, et al. Artificial intelligence in retina. *Prog Retin Eye Res*. 2018;67:1−29.
3. Antaki F, Coussa RG, Kahwati G, et al. Accuracy of automated machine learning in classifying retinal pathologies from ultra-widefield pseudocolour fundus images. *Br J Ophthalmol*. 2023;107:90−95.
4. Nath S, Marie A, Ellershaw S, et al. New meaning for NLP: the trials and tribulations of natural language processing with GPT-3 in ophthalmology. *Br J Ophthalmol*. 2022;106:889−892.
5. Topol E. When M.D. is a Machine Doctor. Available at: https://erictopol.substack.com/p/when-md-is-a-machine-doctor. Accessed January 20, 2023.
6. Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:210807258*. 2021.
7. Wiggins WF, Tejani AS. On the opportunities and risks of foundation models for natural language processing in radiology. *Radiol Artif Intell*. 2022;4:e220119.
8. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst*. 2020;33:1877−1901.
9. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:221213138*. 2022.
10. Liévin V, Hother CE, Winther O. Can large language models reason about medical questions? *arXiv preprint arXiv:220708143*. 2022.
11. Kung TH, Cheatham M, Medinilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2:e0000198.
12. Zafar S, Wang X, Srikumaran D, et al. Resident and program characteristics that impact performance on the Ophthalmic Knowledge Assessment Program (OKAP). *BMC Med Educ*. 2019;19:190.
13. Lee AG, Oetting TA, Blomquist PH, et al. A multicenter analysis of the ophthalmic knowledge assessment program and American Board of Ophthalmology written qualifying examination performance. *Ophthalmology*. 2012;119:1949−1953.
14. OpenAI. ChatGPT: Optimizing Language Models for Dialogue. Available at: https://openai.com/blog/chatgpt/. Published 2022. Accessed January 20, 2023.
15. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:220302155*. 2022.
16. OpenAI, Introducing ChatGPT Plus. Available at: https://openai.com/blog/chatgpt-plus. Accessed March 2, 2023.
17. Taib F, Yusoff MSB. Difficulty index, discrimination index, sensitivity and specificity of long case and multiple choice questions to predict medical students' examination performance. *J Taibah Univ Med Sci*. 2014;9:110−114.
18. Americal Academy of Ophthalmology, OKAP Exam. Available at: https://www.aao.org/okap-exam. Published 2022. Accessed January 21, 2023.
19. Hingorjo MR, Jaleel F. Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency. *J Pak Med Assoc*. 2012;62:142−147.
20. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull*. 1971;76:378−382.
21. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159−174.
22. Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. *NPJ Digit Med*. 2021;4:93.
23. Stunkel L, Mackay DD, Bruce BB, et al. Referral patterns in neuro-ophthalmology. *J Neuroophthalmol*. 2020;40:485−493.
24. Law C, Krema H, Simpson ER. Referral patterns of intraocular tumour patients to a dedicated Canadian ocular oncology department. *Can J Ophthalmol*. 2012;47:254−261.
25. Radford A, Kim JW, Hallacy C, et al. Learning Transferable Visual Models From Natural Language Supervision. Proceedings of the 38th International Conference on Machine Learning; 2021; Proceedings of Machine Learning Research. *arxiv preprint*. https://doi.org/10.48550/arXiv.2103.00020.
26. Wang Z, Wu Z, Agarwal D, Sun J. Medclip: contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:221010163*. 2022.