Clinical Ophthalmology Dovepress

Open Access Full Text Article ORIGINAL RESEARCH

Comparison of Large Language Models in Diagnosis and Management of Challenging Clinical Cases

Sujeeth Krishna Shanmugam, David J Browning

Department of Ophthalmology, Wake Forest University School of Medicine, Winston-Salem, NC, USA

Correspondence: David J Browning, Department of Ophthalmology, Wake Forest University School of Medicine, 1 Medical Center Boulevard, Winston-Salem, NC, 27157, USA, Email djbrowni@Wakehealth.edu

Clinical Ophthalmology downloaded from https://www.dovepress.com/
For personal use only.

Purpose: Compare large language models (LLMs) in analyzing and responding to a difficult series of ophthalmic cases.

Design: A comparative case series involving LLMs that met inclusion criteria tested on twenty difficult case studies posed in open-text format.

Methods: Fifteen LLMs accessible to ophthalmologists were tested against twenty case studies published in JAMA Ophthalmology. Each case was presented in identical, open-ended text fashion to each LLM and open-ended responses regarding differential diagnosis, next diagnostic tests and recommended treatments were requested. Responses were recorded and assessed for accuracy against published correct answers. The main outcome was accuracy of LLMs against the correct answers. Secondary outcomes included comparative performance on the differential diagnosis, ancillary testing, and treatment subtests; and readability of responses.

Results: Scores were normally distributed and ranged from 0–35 (with a maximum score of 60) with a mean ± standard deviation of 19 ± 9. Scores for three of the LLMs (ChatGPT 3.5, Claude Pro, and Copilot Pro) were statistically significantly higher than the mean. Two of the high-performing LLMs were paid subscription (Claude Pro and Copilot Pro) and one was free (ChatGPT 3.5). While there were no clinical or statistical differences between ChatGPT 3.5 and Claude Pro, a separation of +5 points, or 0.56 standard deviations, between Copilot Pro and the other highly ranked LLMs was present. Readability of all tested programs were above the AMA (American Medical Association) reading level recommendations to public consumers of eight grade.

Conclusion: Subscription LLMs were more prevalent among highly ranked LLMs suggesting that these perform better as ophthalmic assistants. While readability was poor for the average person, the content was understood by a board-certified ophthalmologist. The accuracy of LLMs is not high enough to recommend patient care in standalone mode, but aiding clinicians in patient care and prevent oversights is promising.

Keywords: large language model, LLM, chatbot, artificial intelligence, AI

Introduction

Chatbots, a colloquial term for large language models (LLMs) that use artificial intelligence (AI) to answer user's questions, were conceived in the late 1960s. However, advancements in technology, processing power, and software led to breakthroughs in the past 5 years. OpenAI released ChatGPT3.5 in 2022 which inspired many other LLMs.[1]

LLMs have wide applicability, including medicine. Among the uses published in medicine so far include predicting diseases from symptoms and signs, promoting mental health and weight loss, predicting protein folding, helping with drug development, student preparation for examinations and education, helping patients understand informed consent, educating patients on myopia, extracting information from electronic health records, and helping clinicians in analyzing and managing real-world clinical situations.[2–8] LLMs are already used in some practice settings including clinical documentation of the patient encounter in electronic health records.[9]

ChatGPT3.5 has had an accuracy greater than 50% in answering samples of medical questions with diminishing rates according to difficulty.[10,11] In ophthalmology, Chat GPT Plus correctly answered 59.4% of the questions in the AAO BCSC Self-Assessment Program,[12] ChatGPT3.5 correctly answered more than 80% of complex open-ended vitreoretinal clinical scenarios devised by clinicians.[2] ChatGPT 4 produced more accurate answers than glaucoma specialists in answering a panel of glaucoma questions.[13]

Received: 23 July 2024
Accepted: 6 November 2024
Published: 12 November 2024

© 2024 Shanmugam and Browning. This work is published and licensed by Dove Medical Press Limited. The full terms of this license are available at https://www.dovepress.com/terms.php and incorporate the Creative Commons Attribution – Non Commercial (unported, v3.0) License (http://creativecommons.org/licenses/by-nc/3.0/). By accessing the work you hereby accept the Terms. Non-commercial uses of the work are permitted without any further permission from Dove Medical Press Limited, provided the work is properly attributed. For permission for commercial use of this work, please see paragraphs 4.2 and 5 of our Terms (https://www.dovepress.com/terms.php).

Few studies have tested LLMs besides ChatGPT. Our purpose in this study was to fill this gap and compare all available LLMs as assistants to ophthalmologists facing challenging clinical scenarios.

Methods

A web search was conducted using the search terms "artificial intelligence (AI) chatbot" and "large language models" to gather a list of available LLMs. This list was analyzed to exclude any LLMs that were not designed to answer user's questions such as those programs that generated images in response to user's input. Eight LLMs met the inclusion criteria and seven of them also had a paid "pro" version with advanced features. Twenty Clinical Challenges published in JAMA Ophthalmology were randomly chosen and presented to each free AI and their responses were recorded. Each question was inserted individually and after an answer was produced and copied, the associated history tab was deleted. The pro version of each LLM was purchased, and the same steps were conducted. Each LLM had customizable options such as "temperature" and they were all set to default numbers and remained unchanged throughout testing. All fifteen AIs were tested in a five-day window from 6/17/2024-6/23/2024 with the free and paid versions of each AI being tested on the same day in that respective order.

We used only the text part of the Clinical Challenges as none of the LLMs that were tested accept image inputs. The descriptions of the images were included in the posed material.

Each case presentation was followed by a three-part question:

1. What was the differential diagnosis?
2. What was the best next step or steps for narrowing the differential diagnosis?
3. What treatment was recommended?

The answers were prespecified and derived from the published answer in the Clinical Challenge, but we did not use a multiple-choice format. Rather we asked the LLM for a conversational response to simulate a consultation with a colleague in real-life.

The primary outcome was the score of each LLM when tested on the bank of questions. Table 1 lists the LLMs tested. Table 2 lists the 20 Clinical Challenges used. Twelve concerned problems in the subspecialty of Retina, four in Cornea and External Disease, four in Orbit and Oculoplastics, and one in Glaucoma. We used Clinical Challenges from 2023–2024 to reduce the possibility that the challenges were used as part of the LLM's training data. The use of published clinical challenges to test LLMs has been piloted in other specialties.[14]

Table 1 Clinical Challenges Submitted to Large Language Models

Clinical Challenge Number	Title of Clinical Challenge	Month of Publication
1	Multifocal Unilateral Orange Fundus Tumors in a Young Man (doi: 10.1001/jamaophthalmol.2024.0207)	3/2024
2	Sudden Vision Loss (doi: 10.1001/jamaophthalmol.2024.0212)	3/2024
3	An Atypical Optic Nerve Head Mass (doi: 10.1001/jamaophthalmol.2023.6349)	3/2024
4	Acute Corneal Edema in a Middle-Aged Patient (doi: 10.1001/jamaophthalmol.2024.1094)	5/2024
5	A Case of Unilateral Intraocular Pressure Elevation in a Patient with Glaucoma (doi: 10.1001/jamaophthalmol.2022.4735)	1/2023
6	Acute Unilateral Proptosis after Blunt Orbital Trauma in an Adolescent Patient (doi: 10.1001/jamaophthalmol.2022.4748)	1/2023
7	An Amelanotic Choroidal Lesion in a 68-Year-Old Man (doi: 10.1001/jamaophthalmol.2022.5336)	2/2023

(*Continued*)

Table 1 (Continued).

Clinical Challenge Number	Title of Clinical Challenge	Month of Publication
8	Ocular Tumor in a Woman with Breast and Kidney Carcinomas (doi: 10.1001/jamaophthalmol.2022.5544)	2/2023
9	Immunotherapy for a Choroidal Pigmented Lesion (doi: 10.1001/jamaophthalmol.2024.1518)	5/2024
10	Episodic Upper Eyelid Edema in an African American Patient (doi: 10.1001/jamaophthalmol.2024.1517)	5/2024
11	Postoperative Inflammation After Anterior Segment Surgery (doi: 10.1001/jamaophthalmol.2024.0580)	4/2024
12	A Case of a Bumpy Retina (doi: 10.1001/jamaophthalmol.2024.0047)	3/2024
13	A Suspicious Pigmented Lesion of the Conjunctiva (doi: 10.1001/jamaophthalmol.2024.0004)	2/2024
14	Maculopathy in a 52-Year-Old Patient with Diabetes (doi: 10.1001/jamaophthalmol.2023.6938)	2/2024
15	Bilateral Blurry Vision After a Liver Transplant (doi: 10.1001/jamaophthalmol.2023.6710)	2/2024
16	Cystic-Appearing Eyelid Lesion in a 62-Year-Old Man (doi: 10.1001/jamaophthalmol.2023.6637)	2/2024
17	Older Woman with Proptosis, Ptosis, and Blurred Vision (doi: 10.1001/jamaophthalmol.2023.6035)	1/2024
18	Bilateral Hypopyon in a Young Woman (doi: 10.1001/jamaophthalmol.2023.0939)	6/2023
19	Red Eye and Choroidal Detachment in an Older Woman (doi: 10.1001/jamaophthalmol.2023.1513)	7/2023
20	An Unusual Spontaneous Acute Deepening of the Anterior Chamber (doi: 10.1001/jamaophthalmol.2023.1824)	7/2023

Table 2 Large Language Models Tested

Number	Large Language Model
1	ChatGPT 3.5
2	ChatGPT 4.0
3	Chatsonic
4	Chatsonic Pro
5	Claude
6	Claude Pro
7	Copilot
8	Copilot Pro
9	Gemini
10	Gemini Pro
11	Meta

(*Continued*)

Table 2 (Continued).

Number	Large Language Model
12	Perplexity
13	Perplexity Pro
14	Vello
15	Vello Pro

The answer documents from the LLMs were then anonymized with a number system only known to the recorder and given to the same board-certified ophthalmologist for grading. Grading consisted of a +1 if the answer was deemed correct by the published answer and, 0 if the answer was incorrect. The grades by LLM were then anonymized again, only known to the grader, and sent back to the recorder for analysis. After analysis of the data, the LLMs were de-anonymized.

Secondary outcomes were readability of the LLM response and skill relative to arriving at a differential diagnosis (DDX), further workup recommendations, and treatment recommendations. The answer documents including all the responses for all twenty questions per assessed LLMs were copied and pasted into an online software program called Readable, using default settings, which used its AI engine to gather readability data for each LLM's response per Flesch Kincaid reading levels.[15]

Statistical analysis was conducted using MedCalc 22.026 to test the data for normality (Shapiro–Wilk test) and Python to apply a one sample *T*- test comparing each test score to the mean for the normally distributed accuracy data and a Wilcoxon signed-rank test comparing each readability score against the median score for the non-normally distributed readability data. An alpha of 0.0033 was chosen by applying a Bonferroni correction factor of 15 to an initial alpha of 0.05.

Although our study did not involve human research subjects, we obtained approval from the Wake Forest University School of Medicine institutional review board (#00113679).

Results

Accuracy and readability data are shown in Tables 3 and 4, respectively. The Meta AI program refused to give answers to questions regarding medical care. The total scores were normally distributed. The scores ranged from 0 to 35 with a mean ± standard deviation of 19 ± 9. The scores for three of the LLMs (ChatGPT 3.5, Claude Pro, and Copilot Pro) were statistically significantly superior to the mean score for the sample. There was no clinically meaningful difference

Table 3 Accuracy of Large Language Models on Clinical Challenges

Program	Differential Diagnosis Score	Ancillary Testing Score	Treatment Score	Total Score
ChatGPT 3.5	8	12	10	30
ChatGPT 4.0	4	8	7	19
Chatsonic	5	3	4	12
Chatsonic Pro	9	8	8	25
Claude	10	7	7	24
Claude Pro	9	10	10	29
Copilot	3	6	3	12
Copilot Pro	10	11	14	35

(*Continued*)

Table 3 (Continued).

Program	Differential Diagnosis Score	Ancillary Testing Score	Treatment Score	Total Score
Gemini	6	3	5	14
Gemini Pro	6	5	11	22
Meta	0	0	0	0
Perplexity	7	5	4	16
Perplexity Pro	2	7	4	13
Vello	7	2	5	14
Vello Pro	6	6	8	20

Table 4 Readability of Clinical Challenges Text and Large Language Model Responses

Program	Readability Scored as Grade Level
Clinical Challenges Text	13.2
ChatGPT 3.5	16.8
ChatGPT 4.0	15.1
Chatsonic	13.6
Chatsonic Pro	13.5
Claude	17.0
Claude Pro	14.0
Copilot	15.1
Copilot Pro	13.9
Gemini	13.3
Gemini Pro	14.7
Meta	10.7
Perplexity	15.6
Perplexity Pro	14.0
Vello	29.0
Vello Pro	17.4

Note: Higher readability scores correspond to text that is more difficult to understand.

between LLMs ChatGPT3.5 and Claude Pro, but there was a separation of 5 points, or 0.56 standard deviations between Copilot Pro and the other highly ranked LLMs. Considering that the maximum score was 60, the average of the LLMs was 32% correct answers with Copilot Pro scoring 58% correct.

One LLM, ChatGPT, had two versions available for testing. Chat GPT 4.0 performed worse than ChatGPT 3.5 on these clinical challenges. The scores were 30 and 19, respectively. Paid versions and free versions were available for 7 of the LLMs. In 5 of 7 cases, the paid versions had higher scores than the free versions. For Perplexity, there was no meaningful performance difference in the scores between paid and free versions.

There were three components to the total score – the sub score for providing a differential diagnosis, the sub score for suggested further ancillary testing, and the sub score for treatment recommendation. None of the LLMs were statistically different from the mean sub score in providing a differential diagnosis or suggesting ancillary testing. The treatment recommendation score of Copilot Pro was statistically significantly superior to the mean score of the 15 LLMs.

Readability data is shown in Table 4 such that higher values are associated with higher reading difficulty. Readability scores were not normally distributed. The median (95% confidence interval) was 14.4 (13.6, 16.0). All but Gemini showed a decrease in readability in the pro version comparative to the free version, although this difference was statistically insignificant across all the AI programs. The differences in readability scores of Meta, Vello, and Vello Pro from the sample median were statistically significant. Meta (10.7) was significantly easier to understand, and both Vello LLMs (17.4, 29.0) were significantly less so.

Discussion

Our focus was to test LLMs as aids to clinicians and to compare them. Ophthalmologists need little help from LLMs in managing simple problems, but assistance would be valuable in addressing complex problems. The clinical challenges published in JAMA Ophthalmology are difficult and would test the skills of clinicians in everyday practice. The format of the challenges reflects real-world situations in the information provided. The answers are detailed and their publication in a peer-reviewed journal suggests trustworthiness. Other studies have used less challenging tests. In a test using American Academy of Ophthalmology Preferred Practice Patterns, the median score for both ChatGPT3.5 and ChatGPT4 was 5 out of a possible 5 (100% correct).[16] In a study of investigator written questions, ChatGPT3.5 answered 83% correctly.[2] In our study, the best score was 35 out of 60 possible points (58%) and the median score of the 15 LLMs was 19 out of 60 points (32%). Others have noted that there is no single answer in many clinical situations.[2] We acknowledge the same. However, in the complicated situations that we posed, there was a best answer as verified by peer review.

Unfortunately, the same LLM can perform differently in different specialties of medicine. ChatGPT3.5 passed the European Exam in Core Cardiology but failed the American Board of Orthopedics Surgery qualifying examination.[17,18] To our knowledge this is the first study comparing multiple LLMs in assisting users facing clinical scenarios in ophthalmology.

Our grading protocol did not assess harm caused by a recommendation, although other studies assigned grade decrements for possible harm.[2] We found grading for possible harm ambiguous because the LLMs were so broad in their responses and were often couched in qualifiers. They never said, “Do this”. Rather they said, “If this is the case, consider doing this”. The user must be aware that the use of the tool is to prevent oversight, not to prescribe actions, and that actions redound to the physician not the LLM, thus the grading for harm seemed beside the point. This would not be the case were lower-level care providers with less knowledge using them and relying on them for unverified accuracy.

LLMs can improve with conversation.[5] So as not to bias our testing, we had no conversation with the LLM. The clinical scenario was presented in a standardized fashion, and the initial response was compared to the others. Because the ability of the LLM to improve in real time is a point of comparison between LLMs, but difficult to assess, it will be a focus of future work. LLMs do not explain how they come up with their answers but will do so if challenged.[4] Useful follow-up questions include challenging the LLM to provide sources and to explain its “thinking”.

Some LLMs are fee-based, and some are free. Three of the fifteen LLMs that we tested were superior to the others. Two of the superior LLMs were fee-based. As one might expect, LLMs with sustaining revenue have greater resources toward continuous improvement. Our results indicate that fee-based LLMs may have a slight advantage in providing a higher standard of assistance to physicians, but well tested free programs may be a more affordable alternative. Not all LLMs get better in all domains as more advanced versions are released. Mihalache et al found that ChatGPT3.5 performed slightly better than ChatGPT4.0 on Preferred Practice Guideline questions involving retinal diseases.[16] We found the same on more difficult questions in a more diverse sample of subspecialty questions.

Previous work has suggested that the readability of LLM responses is too complex for the average person and requires a college or advanced degree for understanding. However, our emphasis is on ophthalmologists using the LLM, which should mitigate poor readability as a criticism. Nevertheless, it is inevitable that the average person will use the

LLMs for his own purposes, and readability will have an impact in this group of users. For this reason, we checked readability and found that the readability of LLM responses is significantly elevated compared to AMA standards, consistent with past studies.[19] This is acceptable considering that the reading level for the input questions was above the recommended level. None of the LLMs that provided answers to the questions yielded reading levels below that of the input questions. The content was easily comprehended by the board-certified grading ophthalmologist. We emphasize caution with respect to patient use of LLMs to investigate their clinical questions in view of hallucinations, lack of contextual understanding, absence of clinical judgment, potential for bias and harm, over-reliance by patients on the material provided with reduced patient-physician interaction, ethical issues regarding responsibility and liability, and lack of regulatory oversight of the responses.

Although one can expect improvement in the performance of LLMs in medical assistance, they currently require assiduous supervision by ophthalmologists. For example, hallucinations, or fabricated answers, are a problem in the use of LLMs by clinicians because of the work involved in fact checking.[20] This effort is a necessity because there are no internal clues in LLM responses regarding truth. LLM responses are regularly confident, which could be erroneously interpreted by a user as a sign of veracity.[3] When challenged and proven wrong, the LLMs do not evince shame, but cheerily respond, "You're correct. Sorry for any confusion", and respond differently, but with equal confidence.

We doubt that ophthalmologist users would use an LLM response to dictate a next clinical action but warn that LLM responses are meant to prevent oversights by informed users, not specify a menu of actions to be executed without reflection and verification. It is incumbent on physician users to check responses. Some LLMs provide citations for fact checking, including Copilot Pro. Others do not. We regard provision of citations as a plus, but the provision alone cannot be taken as equivalent to a fact check. The LLM can misconstrue the source and fabricated citations are part of LLMs' hallucinations. The user must verify the information. Having the citation decreases the time and effort involved in doing so but does not eliminate it.

Responses depend on many variables including ways of asking questions of the LLM.[2] The temperature setting changes responses. Higher temperatures raise the risk of hallucinations.[2] We used default settings of the manufacturers.

The inequality of LLMs is becoming apparent as they are increasingly tested. For example, Patil et al found that ChatGPT3.5 was superior to Bard (now called Gemini) as an aid to providing patients with additional informed consent understanding regarding ophthalmic surgeries.[3] Han et al also found that ChatGPT3.5 performed better than Gemini in answering questions based on clinical vignettes from JAMA and the New England Journal of Medicine.[17]

As LLMs become increasingly sophisticated and resemble humans in cognitive capabilities, one can imagine when grading them against a prespecified scheme that there may be protest that rankings manifest inequity and lack of respect for diversity. To date, LLMs show no signs of an ability to be insulted, but their designers and users have this capacity and therefore these increasingly explored and accommodated aspects of the practice of medicine will have to be considered. LLMs have known biases based on their training data.[21] LMs have also been trained in a set of ethics and have designer guardrails.[5] The guardrails differ across LLMs. We tested one LLM that was programmed not to answer any clinical questions (Meta). As another example, one cannot get ChatGPT3.5 or Gemini to reveal SAT scores broken down by racial categories because of a designer-based rule which seeks to prevent discouraged lines of reasoning. The guardrails can be changed. For instance, Gemini was originally programmed to decline answering clinical inquiries, but this was changed, and it answers them now.[5] There may also be unknown bias in responses because of unrepresentativeness of training data sets.[22]

Multimodal input prompts could risk betraying personal data. For example, a retinal photograph may allow a reader to identify a particular person.[23] Rules have not been designed for the ethics of use of LLMs in actual care.[5]

Huang et al showed that ChatGPT 4 scored better than glaucoma and retina specialists in answering open ended clinical case scenarios.[13] Given this outcome, and others like it, some have worried that LLMs might replace doctors, but we agree with others that this worry is misplaced.[24] The nature of a physician's work changes and we use LLMs to raise the quality of our work, but we still have to do the work. LLM is best viewed as a tool to be used under supervision.[4] Managing patients is a higher order task than answering examination questions. In the diagnosis of common patterns of signs and symptoms, predicting the most prevalent condition is usually correct and LLMs have less to add. They become

more useful when rarer diagnoses are possibilities. In these cases, they suggest to the clinician possibilities that might otherwise be overlooked.

LLMs are by design, dated. Their training data sets are historical.[5] Any question based on newly developing situations is beyond its ken. For example, when brolucizumab was associated with an increased frequency of serious intraocular inflammation, a question regarding the topic to a LLM would have been fruitless. Successive versions of the same LLM have improved.[25,26] Thus, our conclusion that Copilot Pro is the best LLM for an ophthalmologist seeking a clinical assistant applies only for the present and only out of the sample of 15 LLMs that we tested.

There are limitations to this work. We used one grader, rather than several. The reason for this was that the Clinical Challenges had a published correct answer such that interpretational ambiguity was minimized. In other publications, the graders were making the decisions regarding correctness of the LLM responses on a Likert scale. In contrast, we were deciding if the LLM gave the published correct response, a more straightforward task. Nevertheless, having multiple masked graders would have increased the strength of the study.

It is possible that our results reflect sampling bias. Had we chosen a different set of 20 Clinical Challenges, the results could have been different. ChatGPT3.5 has shown differential deficits across subspecialties with the greatest deficit in the subspecialty area of retina in which it answered 0% correctly compared to general medicine for which questions it answered 79% correctly.[24] LLM performance generally depends on the bank of questions used. Using the BCBS sample questions Lin et al found that ChatGPT4 answered 86.7% of the retina and vitreous questions correctly.[27]

It is possible that some of the Clinical Challenges were used in training the LLMs, although we tried to minimize this risk by choosing ones published in 2023–2024. We did not test all LLMs available. We tested text only LLMs but increasingly LLMs that can incorporate imaging are becoming available. ChatGPT4Vision and Gemini Pro allow incorporation of visual images in prompts. We did not test these as they are not available to physicians. When images can be input, the usefulness of LLMs for ophthalmologists will increase.

It is possible that our results were biased by internet speed, online traffic, and delays in response time, which we did not assess, but which have been suggested as sources of variability.[28] None of the LLMs that we tested allowed the user to input images although this is a feature under development.

In conclusion, we found that 3 LLMs - ChatGPT 3.5, Claude Pro, and Copilot Pro – were more accurate than the others, and that Copilot Pro was the most accurate against the complex clinical cases were tested. We do not recommend LLMs to diagnose and treat patients in ophthalmology. The problems of accuracy and readability require further work. Still, the three top rated LLMs cover gaps in the knowledge base of user ophthalmologists and could be valuable assistants in clinical practice. All suggestions from LLMs require fact checking.

Funding

There is no funding to report.

Disclosure

David Browning has an equity interest in Zeiss-Meditec. He receives royalties from Springer Inc. Sujeeth declares no conflicts of interest.

References

1. Wu T, He S, Liu J, et al. A brief overview of Chatgpt: the history, status quo and potential future development. *IEEE/CAA J Automatica Sinica*. 2023;10(5):1122–1136. doi:10.1109/JAS.2023.123618
2. Maywood MJ, Parikh R, Deobhakta A, Begaj T. Performance assessment of an artificial intelligence chatbot in clinical vitreoretinal scenarios. *Retina*. 2024;44(6):954–964. doi:10.1097/IAE.0000000000004053
3. Patil NS, Huang R, Mihalache A, et al. The ability of artificial intelligence chatbots Chatgpt and google bard to accurately convey preoperative information for patients undergoing ophthalmic surgeries. *Retina*. 2024;44(6):950–953. doi:10.1097/IAE.0000000000004044
4. Lin JC, Younessi DN, Kurapati SS, Tang OY, Scott IU. Comparison of GPT-3.5, GPT-4, and human user performance on a practice ophthalmology written examination. *Eye*. 2023;37(17):3694–3695. doi:10.1038/s41433-023-02564-2
5. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930–1940. doi:10.1038/s41591-023-02448-8

6. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198
7. Biswas S, Logan NS, Davies LN, Sheppard AL, Wolffsohn JS. Assessing the utility of ChatGPT as an artificial intelligence-based large language model for information to answer questions on myopia. *Ophthalmic Physiol Opt*. 2023;43:1562–1570. doi:10.1111/opo.13207
8. Biswas S, Davies LN, Sheppard AL, Logan NS, Wolffsohn JS. Utility of artificial intelligence-based large language models in ophthalmic care. *Ophthalmic Physiol Opt*. 2024;44(3):641–671. doi:10.1111/opo.13284
9. Liu T, Hetherington TC, Stephens C, et al. AI-powered clinical documentation and clinicians' electronic health record experience: a nonrandomized clinical trial. *JAMA Network Open*. 2024;7(9):e2432460. doi:10.1001/jamanetworkopen.2024.32460
10. Johnson D, Goodman R, Patrinely J, et al. Assessing the accuracy and reliability of ai-generated medical responses: an evaluation of the Chat-GPT model. *Res Sq*. 2023. doi:10.21203/rs.3.rs-2566942/v1
11. Goodman RS, Patrinely JR, Stone CA, et al. Accuracy and reliability of chatbot responses to physician questions. *JAMA Network Open*. 2023;6(10):e2336483. doi:10.1001/jamanetworkopen.2023.36483
12. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology. *Ophthalmol Sci*. 2023;3(4):100324. doi:10.1016/j.xops.2023.100324
13. Huang AS, Hirabayashi K, Barna L, Parikh D, Pasquale LR. Assessment of a large language model's responses to questions and cases about glaucoma and retina management. *JAMA Ophthalmol*. 2024;142(4):371. doi:10.1001/jamaophthalmol.2023.6917
14. Han T, Adams LC, Bressem KK, Busch F, Nebelung S, Truhn D. Comparative analysis of multimodal large language model performance on clinical vignette questions. *JAMA*. 2024;331(15):1320–1321. doi:10.1001/jama.2023.27861
15. Solnyshkina M, Zamaletdinov R, Gorodetskaya L, Gabitov A. Evaluating text complexity and flesch-kincaid grade level. *J Soc Studies Educ Res*. 2017;8(3):238–248.
16. Mihalache A, Huang RS, Patil NS, et al. Chatbot and academy preferred practice pattern guidelines on retinal diseases. *Ophthalmol Retina*. 2024;8(7):723–725. doi:10.1016/j.oret.2024.03.013
17. Skalidis I, Cagnina A, Luangphiphat W, et al. ChatGPT takes on the European exam in core cardiology: an artificial intelligence success story? *Eur Heart J Digit Health*. 2023;4(3):279–281. doi:10.1093/ehjdh/ztad029
18. Lum ZC. Can artificial intelligence pass the American board of orthopaedic surgery examination? Orthopaedic residents versus ChatGPT. *Clin Orthop Relat Res*. 2023;481(8):1623–1630. doi:10.1097/CORR.0000000000002704
19. Kianian R, Sun D, Crowell EL, Tsui E. The use of large language models to generate education materials about uveitis. *Ophthalmol Retina*. 2024;8(2):195–201. doi:10.1016/j.oret.2023.09.008
20. Chen JS, Reddy AJ, Al-Sharif E, et al. Analysis of ChatGPT responses to ophthalmic cases: can ChatGPT think like an ophthalmologist? *Ophthalmol Sci*. 2024;5(1):100600. doi:10.1016/j.xops.2024.100600
21. Feng S, Park CY, Liu Y, Tsvetkov Y (2023). From pretraining data to language models to downstream tasks: tracking the trails of political biases leading to unfair NLP models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11737–11762.
22. Vaughan IP, Ormerod SJ. Improving the quality of distribution models for conservation by addressing shortcomings in the field collection of training data. *Conserv Biol*. 2003;17(6):1601–1611. doi:10.1111/j.1523-1739.2003.00359.x
23. Hill R. Retina Identification. In: Jain AK, Bolle R, Pankanti S, editors. *Biometrics*. Boston, MA: Springer; 1996:123–141. doi:10.1007/0-306-47044-6_6.
24. Jha S, Topol EJ. Adapting to artificial intelligence: radiologists and pathologists as information specialists. *JAMA*. 2016;316(22):2353–2354. doi:10.1001/jama.2016.17438
25. Moshirfar M, Altaf AW, Stoakes IM, Tuttle JJ, Hoopes PC. Artificial intelligence in ophthalmology: a comparative analysis of GPT-3.5, GPT-4, and human expertise in answering statpearls questions. *Cureus*. 2023;15(6):e40822. doi:10.7759/cureus.40822
26. Mihalache A, Huang RS, Popovic MM, Muni RH. ChatGPT-4: an assessment of an upgraded artificial intelligence chatbot in the United States medical licensing examination. *Med Teach*. 2024;46(3):366–372. doi:10.1080/0142159X.2023.2249588
27. Lin JC, Kurapati SS, Scott IU. Advances in artificial intelligence chatbot technology in ophthalmology. *JAMA Ophthalmol*. 2023;141(11):1088. doi:10.1001/jamaophthalmol.2023.4619
28. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. *JAMA Ophthalmol*. 2023;141(6):589–597. doi:10.1001/jamaophthalmol.2023.1144

Clinical Ophthalmology Dovepress

Publish your work in this journal

Clinical Ophthalmology is an international, peer-reviewed journal covering all subspecialties within ophthalmology. Key topics include: Optometry; Visual science; Pharmacology and drug therapy in eye diseases; Basic Sciences; Primary and Secondary eye care; Patient Safety and Quality of Care Improvements. This journal is indexed on PubMed Central and CAS, and is the official journal of The Society of Clinical Ophthalmology (SCO). The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use. Visit http://www.dovepress.com/testimonials.php to read real quotes from published authors.

Submit your manuscript here: https://www.dovepress.com/clinical-ophthalmology-journal