• Intelligent Ophthalmology •

Assessing the possibility of using large language models in ocular surface diseases

Qian Ling[1], Zi-Song Xu[1], Yan-Mei Zeng[1], Qi Hong[1], Xian-Zhe Qian[1], Jin-Yu Hu[1], Chong-Gang Pei[1], Hong Wei[1], Jie Zou[1], Cheng Chen[1], Xiao-Yu Wang[1], Xu Chen[2], Zhen-Kai Wu[3], Yi Shao[4]

[1]Department of Ophthalmology, the First Affiliated Hospital, Jiangxi Medical College, Nanchang University, Nanchang 330006, Jiangxi Province, China
[2]Ophthalmology Centre of Maastricht University, Maastricht 6200MS, Limburg, Netherlands
[3]Changde Hospital, Xiangya School of Medicine, Central South University (the First People's Hospital of Changde City), Changde 415000, Hunan Province, China
[4]Department of Ophthalmology, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, National Clinical Research Center for Eye Diseases, Shanghai 200080, China

Co-first authors: Qian Ling and Zi-Song Xu

Correspondence to: Yi Shao. Department of Ophthalmology, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, National Clinical Research Center for Eye Diseases, Shanghai 200080, China. freebee99@163.com; Zhen-Kai Wu. Changde Hospital, Xiangya School of Medicine, Central South University (the First People's Hospital of Changde City), Changde 415000, Hunan Province, China. 18907421671@163.com

Received: 2024-04-20 Accepted: 2024-09-05

Abstract

● **AIM:** To assess the possibility of using different large language models (LLMs) in ocular surface diseases by selecting five different LLMS to test their accuracy in answering specialized questions related to ocular surface diseases: ChatGPT-4, ChatGPT-3.5, Claude 2, PaLM2, and SenseNova.

● **METHODS:** A group of experienced ophthalmology professors were asked to develop a 100-question single-choice question on ocular surface diseases designed to assess the performance of LLMs and human participants in answering ophthalmology specialty exam questions. The exam includes questions on the following topics: keratitis disease (20 questions), keratoconus, keratomalaciac, corneal dystrophy, corneal degeneration, erosive corneal ulcers, and corneal lesions associated with systemic diseases (20 questions), conjunctivitis disease (20 questions), trachoma, pterygoid and conjunctival tumor diseases (20 questions), and dry eye disease (20 questions). Then the total score of each LLMs and compared their mean score, mean correlation, variance, and confidence were calculated.

● **RESULTS:** GPT-4 exhibited the highest performance in terms of LLMs. Comparing the average scores of the LLMs group with the four human groups, chief physician, attending physician, regular trainee, and graduate student, it was found that except for ChatGPT-4, the total score of the rest of the LLMs is lower than that of the graduate student group, which had the lowest score in the human group. Both ChatGPT-4 and PaLM2 were more likely to give exact and correct answers, giving very little chance of an incorrect answer. ChatGPT-4 showed higher credibility when answering questions, with a success rate of 59%, but gave the wrong answer to the question 28% of the time.

● **CONCLUSION:** GPT-4 model exhibits excellent performance in both answer relevance and confidence. PaLM2 shows a positive correlation (up to 0.8) in terms of answer accuracy during the exam. In terms of answer confidence, PaLM2 is second only to GPT4 and surpasses Claude 2, SenseNova, and GPT-3.5. Despite the fact that ocular surface disease is a highly specialized discipline, GPT-4 still exhibits superior performance, suggesting that its potential and ability to be applied in this field is enormous, perhaps with the potential to be a valuable resource for medical students and clinicians in the future.

● **KEYWORDS:** ChatGPT-4.0; ChatGPT-3.5; large language models; ocular surface diseases

DOI:10.18240/ijo.2025.01.01

Citation: Ling Q, Xu ZS, Zeng YM, Hong Q, Qian XZ, Hu JY, Pei CG, Wei H, Zou J, Chen C, Wang XY, Chen X, Wu ZK, Shao Y. Assessing the possibility of using large language models in ocular surface diseases. *Int J Ophthalmol* 2025;18(1):1-8

INTRODUCTION

In recent years, rapid advances in artificial intelligence (AI) have led to the development of complex large language models (LLMs), such as OpenAI's GPT-4 and Google's

Bard[1]. AI work has been dominated by the success of neural networks, most notably LLMs[2]. LLMs are built on top of enlarged pre-trained language models. The researchers found that when the language model was scaled up to a certain extent, it exhibited different behaviors and showed different behaviors when solving complex tasks[3]. LLMs have the ability to perform various natural language processing (NLP) tasks with excellent performance[4]. These models, such as GPT3 and GPT-4, are trained on large amounts of text data and can be fine-tuned for specific downstream tasks, such as language translation or question answering[5-6]. LLMs can respond to free-text queries without the need for specialized training in related tasks, which makes people excited about the prospect of their use in healthcare settings[7]. ChatGPT-4 scored GPT-4 in the top 10% of examinees on the mock bar exam. This is in stark contrast to GPT3.5, which scored in the bottom 10%. With 57 subject multiple choice questions in English, GPT-4 not only performs far better than existing models in English, but also shows strong performance in other languages. On the translation variant of Massive Multitask Language Understanding (MMLU), GPT-4 outpaces English as the most advanced language in 24 of the 26 languages considered[3]. In fact, LLMs have been gradually used in all parts of life. GPT-4 is currently considered to be one of the most powerful LLMs, and it performs well in several aspects. LLMs have the potential to help in various areas of medicine as they are able to deal with complex concepts and respond to different requests and questions[8-9]. Currently, its applications extend to the medical field, such as BioBERT, a specialized NLP model pre-trained on an extensive biomedical corpus. Its capabilities in biomedical text mining tasks are truly outstanding[10]. ChatGPT even met the criteria for passing the US Medical Licensing Examination (USMLE)[11].

In the field of ophthalmology, people are also gradually beginning to realize the development trend of combining AI and ophthalmology. In the field of ophthalmology, AI systems have demonstrated comparable or even better performance than experienced ophthalmologists in tasks such as diabetic retinopathy detection and grading[12]. LLMs are also widely used to improve scientific writing, increase research equity, streamline healthcare workflows, save costs, and improve personalized learning in medical education[13-14]. However, the application of LLMs also raises questions about misinformation, privacy, bias in training data, and the potential for abuse[15]. This is also something we need to pay attention to. To date, the use of LLM in medicine still has potential risks, including differences in answers to medical questions due to failure to learn the latest medical data in a timely manner[16]. A study by Potapenko *et al*[17] involving the training of LLMs in retina-related diseases showed that the recognition accuracy was only 45% when it came to information sources from patients with retinal diseases. This suggests that there is a significant gap in the application of AI in the clinical setting of ophthalmology. Model applications are particularly valuable in the medical and scientific fields, where limited data is often a challenge[18-20].

In order to cope with the above situation, to explore and analyze the performance of the model more scientifically and systematically, we need to select more detailed and specialized fields, and at the same time ensure that the test content is not included in the training data[21]. We therefore created 100 multiple-choice questions for ocular surface diseases to evaluate the performance of 5 different LLMs (GPT-3.5, GPT-4, PaLM2, Claude2, SenseNova) in answering questions for ocular surface diseases. We further examine the stability and confidence of these LLMS in assessing fundus disease knowledge, explore the stability of different models in tests, and will continue to explore the reliability of ChatGPT-4 for medical education and clinical decision making.

PARTICIPANTS AND METHODS

Ethical Approval The study methods and protocols were approved by the Medical Ethics Committee of the First Affiliated Hospital of Nanchang University (Nanchang, China. No.2021039) and followed the principles of the Declaration of Helsinki. All subjects were notified of the objectives and content of the study and latent risks, and then provided written informed consent to participate.

We asked a group of experienced ophthalmology professors to develop a 100-question single-choice question on ocular surface diseases designed to assess the performance of LLM and human participants in answering ophthalmology specialty exam questions. The study included an evaluation of five LLMs: ChatGPT (GPT-3.5), ChatGPT (GPT-4), PaLM2, Claude 2, and SenseNova. The exam includes questions on the following topics: keratitis disease (20 questions), keratoconus, keratomalaciac, corneal dystrophy, corneal degeneration, erosive corneal ulcers, and corneal lesions associated with systemic diseases (20 questions), conjunctivitis disease (20 questions), trachoma, pterygoid and conjunctival tumor diseases (20 questions), and dry eye disease (20 questions). The questions for the exam can be found in the Appendix section. Each of our experiments is repeated five times, with each LLM tested with a set of 100 multiple-choice questions related to ocular surfayce disease. At the beginning of each trial, we ask for LLM initialization. The LLM receives instructions and questions prior to the completion of the test, and the LLM participant is instructed to provide only accurate answers without any accompanying explanations. Each question was experimented with five times, and when each LLM answered the same question exactly 5 times, the percentage of questions

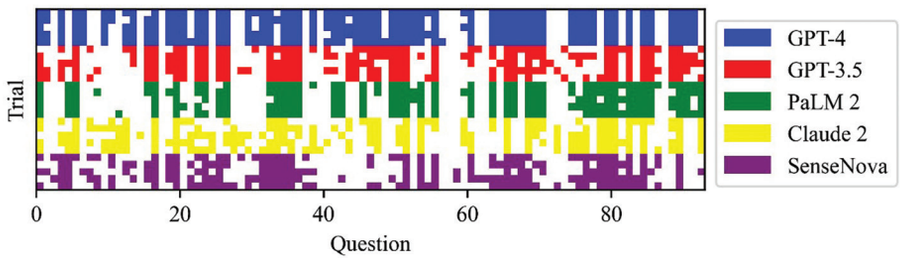

Figure 1 Raw average scores for 5 LLM tests: different colors represent different LLMs, and correct answers are represented by color squares
LLM: Large language model.

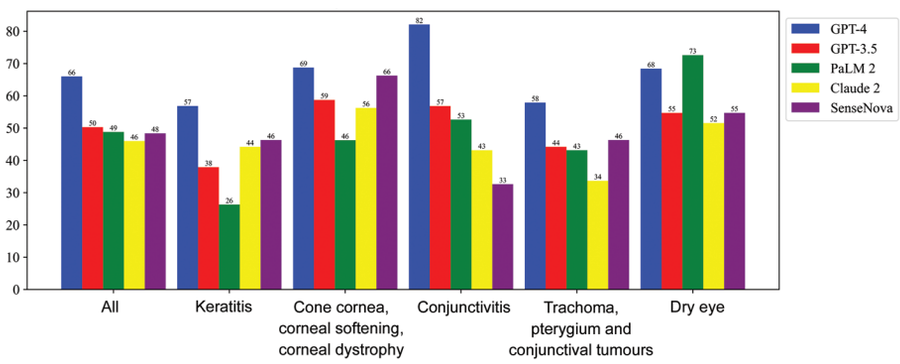

Figure 2 Scores of five large language models on tests in five ocular surface disease domains.

that answered all 5 answers correctly increased by 1%. Also, the test results are compared to the expected distribution that might occur if the candidate guessed randomly. When guessed randomly, the average number of expected correct answers in the 5 trials was about 0.25×5=1.25, and the multiple-choice questions were all 4 choices. Using this value, the number of occurrences of the correct answer to each question can be estimated based on the resulting poison distribution, and the score can be calculated based on the correct response rate of multiple trials of the LLM and each seniority test question process.

At the same time, in order to evaluate the level of clinical knowledge reserve of LLMs, we also selected 5 chief physicians, attending physicians, trainees and graduate students who are all ophthalmology specialties, and examined them in the form of questionnaires on the same topics. We collected the selected person's 100-question test total score and calculated the average of the total scores for each level, comparing their total average scores to the average of the total scores of the five LLMs.

RESULTS

Figures 1 and 2 depict raw and average test scores, respectively. The five colors (Figure 1) represent five different LLMS, which are composed of 100 points from left to right, corresponding to 100 multiple choice questions, and each correct answer of the LLM is marked with a color dot. Five lines of each color from top to bottom represent five repetitions. After examining the raw scores in Figure 1, it is clear that each LLM exhibits variability in the experiment,

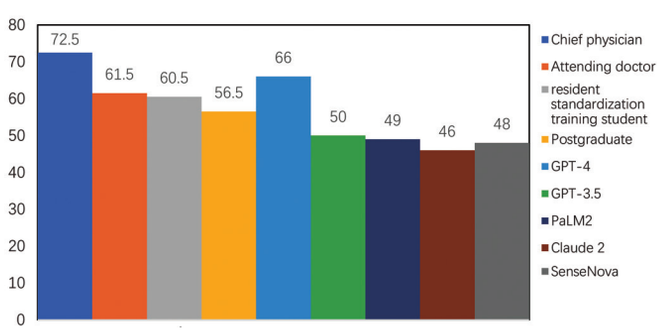

Figure 3 Comparison of the average total scores of the 4 human and 5 large language model groups.

both in terms of uncertainty in the total score and in terms of how often the questions are answered correctly. It is worth noting that GPT-4 displays the largest number of color squares, indicating that GPT-4 has the highest probability of correct answers, and is also the LLM that can consistently produce correct answers. In Figure 2, average scores are given, where the average LLM test score represents the average of five different trials. GPT-4, GPT-3.5, PaLM2, SenseNova, and Claude 2 have average scores of 66, 50, 49, 48, and 46, in descending order, respectively. In contrast, GPT-4 exhibits the highest performance in terms of LLMs. Comparing the average scores of the LLM group with the four human groups, chief physician, attending physician, regular trainee, and graduate student, it is found that except for ChatGPT-4, the total score of the rest of the LLM is lower than that of the graduate student group, which has the lowest score in the human group (56.5) (Figure 3). But what is surprising is that the average score of ChatGPT-4 is as high as 66, which even exceeds the average

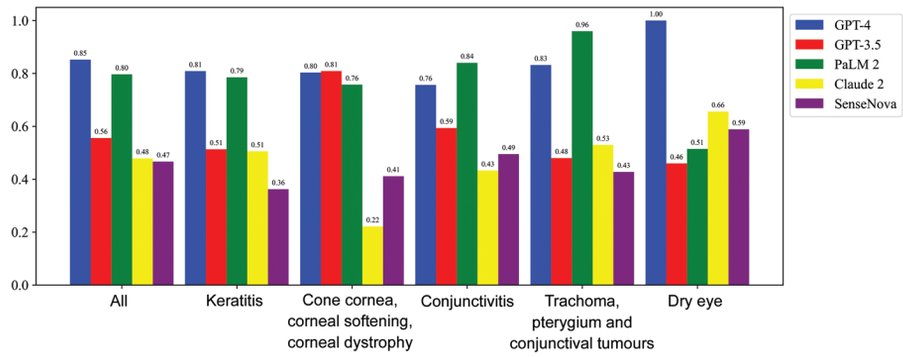

Figure 4 Average correlation of large language model scores by category.

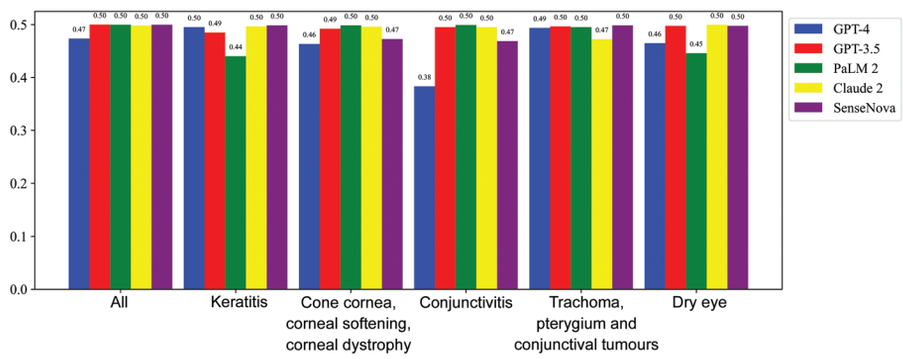

Figure 5 Standard deviation of large language model scores by category.

score of 61.5 of the attending physician and is second only to the average score of the chief physician of 72.5. This indicates that although the professional level of other LLMs needs to be improved, the professional level of ChatGPT-4 is even higher than that of the attending physician in terms of ocular surface diseases. The results indicate that GPT-4 has great potential for application in the field of ocular surface diseases, and it has a certain professional quality.

We described the mean correlation and standard deviation of five LLMs in five trials in Figures 4 and 5, and found that their score standard deviation was low and the mean correlation between trials was high, indicating a high consistency between their answers and scores. It can be found that ChatGPT-4 and PaLM have a higher overall mean correlation than the other three LLMS, all above 0.80. In addition, the standard deviation of ChatGPT-4 is lower than that of other LLMs most of the time, which is 0.47, and the other LLMs are all 0.50, indicating that the probability of getting the correct answer is more stable, indicating that ChatGPT-4 has better stability than other LLMs in solving ocular surface diseases.

Comparison of LLM Answer Confidence According to the data presented in Figure 6, both ChatGPT-4 and PaLM2 are more likely to give exact and correct answers, giving very little chance of an incorrect answer. ChatGPT-4 showed higher credibility when answering questions, with a success rate of 59%, but gave the wrong answer to the question 28% of the time (Figure 6A). In contrast, the results of PaLM2 were more polarized, either answering all correct answers, with 40% correct answers per trial, or wrong answers, with 39% correct answers per trial (Figure 6C). The ChatGPT-3.5 model showed a moderate level of performance, answering questions correctly with 32% accuracy and 23% error (Figure 6B). Claude2 and SenseNova showed a lower level of performance and a higher tendency to be confused, with only 18% and 25% accuracy respectively, and 27% and 24% probability of the answer being wrong (Figure 6B, 6D).

DISCUSSION

Our study designed a 100-question multiple-choice exam centered on ocular surface diseases to assess proficiency in a highly specialized topic. The study aimed to compare the performance of five different LLMs. The test results show that the GPT-4 model exhibits excellent performance in both answer relevance and confidence. On the other hand, PaLM2 showed a positive correlation (up to 0.8) in terms of answer accuracy during the exam. In terms of answer confidence, PaLM2 is second only to GPT4 and surpasses Claude 2, SenseNova, and GPT-3.5. Despite the fact that ocular surface disease is a highly specialized discipline, GPT-4 still exhibits superior performance, suggesting that its potential and ability to be applied in this field is enormous, perhaps with the potential to be a valuable resource for medical students and clinicians in the future.

Actually, digital technology is revolutionizing healthcare. This is manifested in several ways, notably the increase in telemedicine, medical Internet of Things, and AI health diagnostics[22]. Ophthalmology is a leading medical specialty

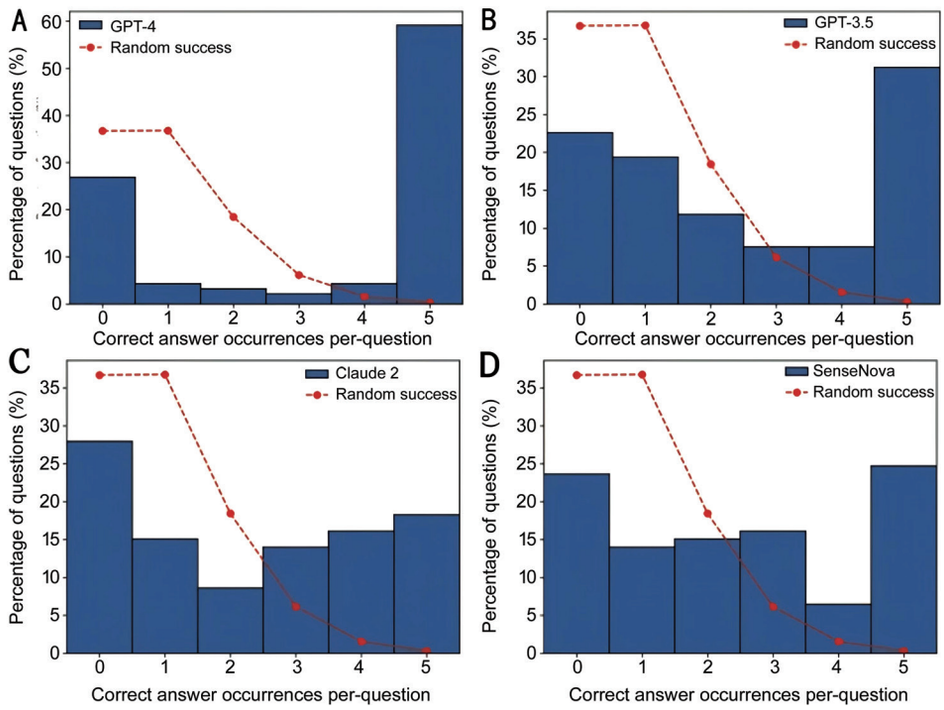

Figure 6 Answer confidence, the number of correct answers that appear for each question per LLM The red dotted line represents the expected distribution when the answer is randomly selected based on the Poisson distribution. A: GPT-4 random success rate; B: GPT-3.5 random success rate; C: Claude 2 random success rate of answers; D: PaLM2 and SenseNava random success rate of answers. LLM: Large language model.

applying AI to screening, diagnosis, and treatment[23]. AI is now being used to treat keratoconus, infectious keratitis, refractive surgery, corneal transplantation, cataracts in adults and children, angle-closure glaucoma, and iris tumors, among others[24]. As far as I'm concerned, in the highly specialized field of medicine, ChatGPT has the ability to quickly collect and process a variety of medical information, making it a valuable teaching tool for students[25]. ChatGPT Rapid retrieval and generalization capabilities can be used as a supplement to educational resources to reduce the time that medical students spend receiving knowledge in traditional classrooms, improve the efficiency of knowledge acquisition, and allocate more time to clinical learning and skills training. LLM can generate teaching content based on medical education needs, such as classroom presentations, textbooks, case studies, etc. It provides personalized learning resources based on the needs and knowledge level of students, helping them better understand and master medical knowledge; Automatically generate medical exam questions based on the set knowledge points and exam requirements, reducing the workload of teachers and ensuring the quality and accuracy of exam questions. In addition, real-time feedback and evaluation can be provided based on student responses, helping students identify learning progress and weak areas. Interact with students and provide personalized educational guidance based on their questions and needs[26]. It can also answer students' questions, provide supplementary explanations and examples, and help students gain a deeper understanding of medical concepts and principles. At the same time, relevant learning resources can be recommended based on students' learning situations and preferences, achieving personalized learning path recommendations.

In the field of ophthalmology, deep learning models have shown excellent diagnostic performance in eye disease screening[27]. Due to the popularization of video display terminal devices such as mobile phones and computers, the incidence of ophthalmic diseases has shown a very obvious upward trend and younger trend compared with before, which makes the clinical work of ophthalmology gradually bear huge pressure and burden. However, the uneven distribution of medical resources makes it more difficult for rural areas to obtain high-quality ophthalmic medical resources. If the diagnosis and treatment ability of primary ophthalmologists can be effectively improved, the prognosis of patients with eye diseases can be greatly improved. In fact, a considerable number of AI-assisted diagnostic medical device products have been launched into the ophthalmic market in terms of intelligent assisted diagnosis, intelligent assisted treatment, intelligent monitoring and life support, intelligent rehabilitation physiotherapy and intelligent traditional Chinese medicine diagnosis and treatment[28]. Thanks to the easy training and refinement of LLM, untrained patients may use LLM (such as

ChatGPT) as a virtual assistant to classify and self diagnose ophthalmic diseases ranging from harmless to potential visual threats. In addition, LLM can effectively produce patient education materials, translate a large number of professional terms into simplified and empathetic language suitable for outsiders, or act as a "therapist" to provide consultation for patients with mental health disorders. Patients with impaired vision are more likely to suffer from psychological distress, which would be very valuable. The potential application of LLMs in healthcare education may be promising[29]. Currently, ChatGPT has been tested in the field of ophthalmology[30]. Gao *et al*[31]. used retinal images as a starting point for disease assessment and diagnosis, and realized the diagnosis and segmentation of common ophthalmic diseases. Subsequently, they established a new multimodal teaching approach in ophthalmology and interacted with disease-related knowledge data to collect available real-world medical plans for learning, leading to the development of an LLM specifically for ophthalmology, the OphGLM, a large language model that showed robust functionality in subsequent experiments. Although there appears to be relatively limited published research on the topic compared to other medical specialties.

In addition to the extremely specialized areas of knowledge in diagnosis and treatment, the potential for clinical application of LLMs remains not low. Compared to clinicians who are busy and have a valuable amount of time, LLMs can provide more inexpensive and patient consultations. LLMs can be used as a platform to provide useful insights across language barriers when patients have eye problems and to meet their consultation needs[32]. Bernstein *et al*[33] conducted a study evaluating the quality of ophthalmic advice generated by LLM chatbots compared to those written by ophthalmologists and found that the answers generated by the chatbot did not differ from human answers in terms of guidance, safety, and reliability. A study by Tan Yip Ming *et al*[34] suggest the potential application of LLMs in uveitis patients, indicating their potential to assist in uveitis consultation and management, as a diagnostic support tool for uveitis, and in uveitis research. Due to the fact that ChatGPT also has multilingual translation capabilities, it can even meet the needs of cross ethnic and cultural patient groups[34]. It can also achieve visual and auditory text to meet the needs of hearing-impaired patients; It can be integrated through text to image or video generation platforms to enhance the patient experience. Kianian *et al*'s[35] research suggest that ChatGPT can respond to less complex words when reading for uveitis patients, helping them better understand. At the same time, LLM can also improve the efficiency of doctors by helping doctors quickly compile heavy paperwork such as surgical records and discharge summaries[36-38].

LLM require a large amount of sample training to improve model performance. A successful LLM often requires a large dataset and more complex and accurate algorithms that can be applied. However, the clinical data collection process involves multiple participants, and differences in data quality are inevitable, which may further lead to poor performance of the model on specific data or inability to accurately predict certain outcomes. In the field of healthcare, the interpretability of models is crucial. Doctors and patients need to understand the reasoning process of the model and the basis for producing results. LLM is considered a black box model, and the logic and foundation of its generated results are difficult to explain, which greatly reduces trust in the model's generated results and affects its reliability and acceptability in practical applications. Also, ChatGPT-4 may be inferior than experienced ophthalmologists in choosing different diagnosis and treatment methods according to individual patients, and it is difficult for GPT-4 to completely replace the professional knowledge and decision-making ability of ophthalmologists at present. GPT-4 has a high degree of confidence in its answers, regardless of whether the answer it ends up with is correct or wrong, but if a human doctor is confronted with a question that he does not know, he may be more deliberate and avoid the most wrong answer.

The bias of AI in answering questions is part of a broader ethical issue that has many consequences[23]. The existence of such biases is based on a number of factors, such as race, genetics, region, or gender. Here are some ways to address bias: 1) establish a gold standard in all areas of medicine[39]; 2) thoroughly check the training data for bias[40]; 3) collaboration between physicians and AI to eliminate sources of bias[41]; 4) the use of machine learning as a tool to assess and classify the risk of bias in randomized trials[42-43].

At present, there are still no perfect laws related to AI intervention in medical treatment in the world, and how to judge its ethics, correctness and responsibility attribution is still an urgent problem to be solved. Health care related data resources need to be formatted and integrated in order to play an optimal role in advancing the application of AI in the health field. Therefore, how to protect patients' privacy information is a major problem on the road of LLM applications. The patient's private information may be exposed twice when entering the electronic medical record and the electronic medical record are connected to the AI system[44]. Data privacy violations can lead to many issues, including but not limited to discrimination or denial of insurance or employment, emotional stress from exposure of sensitive health data, mental health consequences such as embarrassment, paranoia and mental distress, deontological concerns about the vulnerability of personal data, erosion of trust, failure to seek health care services or withhold information to protect privacy, and group-

based harm[45]. As a black box tool, the opaque, complex and lack of transparency of AI intervention in medical output is a major factor in their fear and suspicion[46]. The existence of this phenomenon makes the certification process difficult, difficult to govern, and difficult to study and evaluate its safety and effectiveness[47]. Hacker *et al*[48] argue that regulation should focus on specific high-risk applications, rather than on the pre-trained model itself, and should include transparency obligations, risk management, non-discrimination provisions, and content moderation rules. Mökander *et al*[49] noted that existing audit procedures do not address the governance challenges posed by LLMs, but can be further improved in three ways: 1) identifying the need to develop new audit procedures to capture the risks posed by LLMs; 2) draw on best practices in IT governance and systems engineering to outline a blueprint for auditing LLMs in a feasible and effective manner; 3) discuss the limitations of the prospects for an LLM in auditing. The LLM is still evolving rapidly, so regulators and lawmakers need to act quickly to improve the relevant legal provisions before the LLM is formally applied to the industry to determine how to use the LLM as a human helper.

ACKNOWLEDGEMENTS

Foundations: Supported by National Natural Science Foundation of China (No.82160195; No.82460203); Degree and Postgraduate Education Teaching Reform Project of Jiangxi Province (No.JXYJG-2020-026).

Conflicts of Interest: Ling Q, None; **Xu ZS,** None; **Zeng YM,** None; **Hong Q,** None; **Qian XZ,** None; **Hu JY,** None; **Pei CG,** None; **Wei H,** None; **Zou J,** None; **Chen C,** None; **Wang XY,** None; **Chen X,** None; **Wu ZK,** None; **Shao Y,** None.

REFERENCES

1 Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med* 2023;6(1):120.

2 Pavlick E. Symbols and grounding in large language models. *Philos Trans A Math Phys Eng Sci* 2023;381(2251):20220041.

3 OpenAI, Josh A, Steven A, *et al*. GPT-4 technical report. arXiv:2303.08774. https://arxiv.org/abs/2303.08774

4 Wei J, Wang XZ, Schuurmans D, Bosma M, Ichter B, Xia F, Chi E, Le Q, Zhou D. Chain-of-thought prompting elicits reasoning in large language models. arXiv:2201.11903. https://arxiv.org/abs/2201.11903v6

5 Yang JF, Jin HY, Tang RX, Han XT, Feng QZ, Jiang HM, Zhong SC, Yin B, Hu X. Harnessing the power of LLMs in practice: a survey on ChatGPT and beyond. *ACM Trans Knowl Discov Data* 2024;18(6):1-32.

6 Hadi MU, Tashi A, Qureshi R, *et al*. A survey on large language models: applications, challenges, limitations, and practical usage. *TechRxiv* 2023.

7 Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med* 2023;29(8):1930-1940.

8 Singhal K, Azizi S, Tu T, *et al*. Large language models encode clinical knowledge. *Nature* 2023;620(7972):172-180.

9 Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv:2303.13375.https://arxiv.org/abs/2303.13375v2

10 Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;36(4):1234-1240.

11 Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ* 2023;9:e45312.

12 Li ZW, Wang L, Wu XF, Jiang JW, Qiang W, Xie H, Zhou HJ, Wu SJ, Shao Y, Chen W. Artificial intelligence in ophthalmology: the path to the real-world clinic. *Cell Rep Med* 2023;4(7):101095.

13 Sallam M. The utility of ChatGPT as an example of large language models in healthcare education, research and practice: systematic review on the future perspectives and potential limitations. *MedRxi* 2023.02.19.23286155.

14 Li JN, Dada AM, Puladi B, Kleesiek J, Egger J. ChatGPT in healthcare: a taxonomy and systematic review. *Comput Methods Programs Biomed* 2024;245:108013.

15 Clusmann J, Kolbinger FR, Muti HS, *et al*. The future landscape of large language models in medicine. *Commun Med* 2023;3(1):141.

16 Ayers JW, Poliak A, Dredze M, *et al*. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med* 2023;183(6):589-596.

17 Potapenko I, Boberg-Ans LC, Stormly Hansen M, Klefter ON, van Dijk EHC, Subhi Y. Artificial intelligence-based chatbot patient information on common retinal diseases using ChatGPT. *Acta Ophthalmol* 2023;101(7):829-831.

18 Kagawa R, Shirasuna M, Ikeda A, *et al*. One-second boosting: a simple and cost-effective intervention for data annotation in machine learning. *Proceedings of the 44th Annual Conference of the Cognitive Science Society* 2022.

19 Rezayi S, Dai HX, Liu ZL, *et al*. ClinicalRadioBERT: knowledge-infused few shot learning for clinical notes named entity recognition. In: Lian C, Cao X, Rekik I, Xu X, Cui Z. (eds) *Machine Learning in Medical Imaging. MLMI 2022*. Lecture Notes in Computer Science, vol 13583. Springer, Cham.

20 Liu ZL, Huang Y, Yu XW, *et al*. Deid-GPT: zero-shot medical text de-identification by GPT-4. arXiv:2303.11032. https://arxiv.org/pdf/2303.11032

21 Kalyan KS, Rajasekharan A, Sangeetha S. AMMUS: a survey of transformer-based pretrained models in natural language processing. arXiv:2108.05542. https://arxiv.org/abs/2108.05542v2

22 Ting DSW, Carin L, Dzau V, Wong TY. Digital technology and COVID-19. *Nat Med* 2020;26(4):459-461.

23 Abdullah YI, Schuman JS, Shabsigh R, Caplan A, Al-Aswad LA. Ethics of artificial intelligence in medicine and ophthalmology. *Asia*

Pac J Ophthalmol (Phila) 2021;10(3):289-298.

24 Ting DSJ, Foo VH, Yang LWY, Sia JT, Ang M, Lin HT, Chodosh J, Mehta JS, Ting DSW. Artificial intelligence for anterior segment diseases: emerging applications in ophthalmology. *Br J Ophthalmol* 2021;105(2):158-168.

25 Kung TH, Cheatham M, Medenilla A, Sillos C, de Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit Health* 2023;2(2):e0000198.

26 Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. *JMIR Med Educ* 2019;5(1):e13930.

27 Teo ZL, Ting DSW. AI telemedicine screening in ophthalmology: health economic considerations. *Lancet Glob Health* 2023;11(3): e318-e320.

28 Qiao Z, Sirui H, Nan Z, *et al.* Analysis of the development status of domestic artificial intelligence medical devices. *Medical and Health Equipment* 2023,44(05):64-68.

29 Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare* 2023;11(6):887.

30 Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. *Ophthalmol Sci* 2023;3(4):100324.

31 Gao WH, Deng Z, Niu ZY, *et al*. OphGLM: training an ophthalmology large language-and-vision assistant based on instructions and dialogue. arXiv:2306.12174. https://arxiv.org/pdf/2306.12174

32 Will ChatGPT transform healthcare? *Nat Med* 2023;29(3):505-506.

33 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(8):e2330320.

34 Tan Yip Ming C, Rojas-Carabali W, Cifuentes-González C, *et al*. The potential role of large language models in uveitis care: perspectives after ChatGPT and bard launch. *Ocul Immunol Inflamm* 2024;32(7):1435-1439.

35 Kianian R, Sun DY, Crowell EL, Tsui E. The use of large language models to generate education materials about uveitis. *Ophthalmol Retina* 2024;8(2):195-201.

36 Singh S, Djalilian A, Ali MJ. ChatGPT and ophthalmology: exploring its potential with discharge summaries and operative notes. *Semin Ophthalmol* 2023;38(5):503-507.

37 Bhattacharya K, Bhattacharya AS, Bhattacharya N, *et al*. ChatGPT in surgical practice—a new kid on the block. *Indian Journal of Surgery* 2023;85(6):1346-1349.

38 Waisberg E, Ong J, Masalkhi M, Kamran SA, Zaman N, Sarker P, Lee AG, Tavakkoli A. GPT-4: a new era of artificial intelligence in medicine. *Ir J Med Sci* 2023;192(6):3197-3200.

39 Ting DSW, Peng L, Varadarajan AV, Keane PA, Burlina PM, Chiang MF, Schmetterer L, Pasquale LR, Bressler NM, Webster DR, Abramoff M, Wong TY. Deep learning in ophthalmology: the technical and clinical considerations. *Prog Retin Eye Res* 2019;72:100759.

40 He JX, Baxter SL, Xu J, Xu JM, Zhou XT, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med* 2019;25(1):30-36.

41 Yarnell CJ, Fu LD, Manuel D, Tanuseputro P, Stukel T, Pinto R, Scales DC, Laupacis A, Fowler RA. Association between immigrant status and end-of-life care in Ontario, Canada. *JAMA* 2017;318(15):1479-1488.

42 Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. *N Engl J Med* 2018;378(11):981-983.

43 Higgins JPT, Green S. Cochrane handbook for systematic reviews of interventions. *Journal of MultiDisciplinary Evaluation* 2010;6(14): 142-148.

44 Pesapane F, Volonté C, Codari M, Sardanelli F. Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. *Insights Imaging* 2018;9(5):745-753.

45 Price WN 2nd, Cohen IG. Privacy in the age of medical big data. *Nat Med* 2019;25(1):37-43.

46 Sullivan HR, Schweikart SJ. Are current tort liability doctrines adequate for addressing injury caused by AI? *AMA J Ethics* 2019;21(2):E160-E166.

47 Ting DSW, Pasquale LR, Peng L, *et al*. Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol* 2019;103(2):167-175.

48 Hacker P, Engel A, Mauer M. Regulating ChatGPT and other large generative AI models. arXiv:2302.02337. https://arxiv.org/abs/2302.02337

49 Mökander J, Schuett J, Kirk HR, Floridi L. Auditing large language models: a three-layered approach. *AI Ethics* 2024;4(4):1085-1115.