• Intelligent Ophthalmology •

Assessing the proficiency of large language models on funduscopic disease knowledge

Jun-Yi Wu[1], Yan-Mei Zeng[2], Xian-Zhe Qian[2], Qi Hong[2], Jin-Yu Hu[2], Hong Wei[2], Jie Zou[2], Cheng Chen[2], Xiao-Yu Wang[2], Xu Chen[3], Yi Shao[4]

[1]Department of Ophthalmology, Wuhan Fourth Hospital, Wuhan 430033, Hubei Province, China
[2]Department of Ophthalmology, the First Affiliated Hospital, Jiangxi Medical College, Nanchang University, Nanchang 330006, Jiangxi Province, China
[3]Ophthalmology Centre of Maastricht University, Maastricht 6200MS, Limburg, Netherlands
[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: Jun-Yi Wu and Yan-Mei Zeng
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
Received: 2024-11-12 Accepted: 2025-03-03

Abstract

● **AIM:** To assess the performance of five distinct large language models (LLMs; ChatGPT-3.5, ChatGPT-4, PaLM2, Claude 2, and SenseNova) in comparison to two human cohorts (a group of funduscopic disease experts and a group of ophthalmologists) on the specialized subject of funduscopic disease.

● **METHODS:** Five distinct LLMs and two distinct human groups independently completed a 100-item funduscopic disease test. The performance of these entities was assessed by comparing their average scores, response stability, and answer confidence, thereby establishing a basis for evaluation.

● **RESULTS:** Among all the LLMs, ChatGPT-4 and PaLM2 exhibited the most substantial average correlation. Additionally, ChatGPT-4 achieved the highest average score and demonstrated the utmost confidence during the exam. In comparison to human cohorts, ChatGPT-4 exhibited comparable performance to ophthalmologists, albeit falling short of the expertise demonstrated by funduscopic disease specialists.

● **CONCLUSION:** The study provides evidence of the exceptional performance of ChatGPT-4 in the domain of funduscopic disease. With continued enhancements, validated LLMs have the potential to yield unforeseen advantages in enhancing healthcare for both patients and physicians.

● **KEYWORDS:** large language models; ChatGPT; funduscopic disease

DOI:10.18240/ijo.2025.07.03

Citation: Wu JY, Zeng YM, Qian XZ, Hong Q, Hu JY, Wei H, Zou J, Chen C, Wang XY, Chen X, Shao Y. Assessing the proficiency of large language models on funduscopic disease knowledge. *Int J Ophthalmol* 2025;18(7):1205-1213

INTRODUCTION

The advent and progression of large language models (LLMs) in recent years have significantly influenced the field of natural language processing (NLP)[1]. NLP enables the extraction, comprehension, and systematic analysis of textual data information in an intelligent and efficient manner[2]. The utilization of NLP in the medical sector has been observed to yield advantageous outcomes[3]. NLP has been employed for the analysis of textual information in studies based on electronic medical records (EMR) and genomics networks[4-5]. NLP methods can be broadly classified into two categories: rule-based approaches and machine learning (ML)-based approaches. Rule-based systems are employed to facilitate decision-making by utilizing pre-established rules. These systems evaluate data according to the predefined rules and execute specific operations based on the corresponding mappings[6]. ML-based systems use algorithms to learn from datas and make predictions or take action without explicit programming[7-8]. In pediatric emergency triage, ML-based systems have a good ability to predict disease outcomes and dispositions, reducing under-triage of critically ill children and over-triage of less ill children[9].

The emergence of powerful language models, such as the LLMs, can be attributed to recent advancements in the field of NLP[10]. Researchers have observed that scaling the models can result in improved performance, in accordance with the scaling law[11]. Consequently, they conducted further investigations

to explore the impact of scaling by increasing the size of the model. The findings revealed that beyond a certain threshold of parameter sizes, the LLM exhibited substantial performance improvements[12]. Additionally, the larger model demonstrated the emergence of novel capabilities, such as context learning, which were absent in the smaller model. These large pre-trained language models (PLMs), commonly known as "LLM" in the research community[13-14], are significantly transforming specific domains of artificial intelligence (AI) research. The research landscape in the field of NLP has been progressively shifting towards the adoption of LLMs, which serve as versatile solutions for various language-related tasks.

An exemplary illustration of the utilization of LLM is observed in ChatGPT, a conversational AI system developed using robust GPT models such as GPT-3.5 and GPT-4. This application showcases an impressive aptitude for engaging in dialogue with humans. The proficiency of ChatGPT in communication is evidenced by its extensive knowledge base, ability to reason through mathematical problems, adeptness in maintaining contextual coherence during multi-round conversations, and alignment with appropriate human values[15]. Consequently, both ChatGPT and GPT-4 represent significant milestones in the advancement of language models, substantially augmenting the capabilities of existing AI systems.

In 2020, OpenAI introduced GPT3, a transformer class model with a significant parameter count of 175B. This model has solidified OpenAI's commitment to implementing artificial general intelligence through the LLM approach. GPT3 introduces the concept of contextual learning[16], enabling LLM to comprehend tasks presented in natural language text. Due to its robust capabilities, GPT-3 has served as the foundational model for the development of more advanced LLMs at OpenAI[17]. However, it is important to note that a notable limitation of the original GPT-3 model is its inability to effectively reason about complex tasks. The GPT-3 to ChatGPT iterative pass package encompasses two significant components, namely CodeX[18] and InstructGPT[19]. CodeX, built upon GPT-3, utilizes code data for further training, thereby equipping the model with the capability to comprehend and generate code. On the other hand, InstructGPT introduces two pivotal techniques, namely instruction tuning[20] and reinforcement learning from human feedback (RLHF)[21]. These techniques align the model's output with human preferences, thereby enhancing its performance in practical user-facing situations.

OpenAI refers to GPT4 as a significant advancement in their endeavors to enhance deep learning. GPT4 functions as a substantial multimodal model, capable of processing both image and text inputs and generating text outputs, while demonstrating performance comparable to that of humans across various professional and academic benchmarks. A recent comprehensive investigation into the capabilities of GPT-4, encompassing a diverse array of demanding tasks, unveiled its superior problem-solving proficiency in comparison to earlier iterations of GPT models[22]. Following extensive training and meticulous debugging conducted by OpenAI, GPT-4 has been effectively restricted from engaging in the development of weaponry, problematic mathematical computations, and dispensing harmful advice. Consequently, the security measures surrounding GPT-4 have been significantly enhanced. In comparison to its predecessor, GPT-3, GPT-4 exhibits a notably robust capacity for logical reasoning. Moreover, GPT-4 demonstrates a marked improvement of approximately 80% in the accuracy of answering questions when compared to the previous iteration, ChatGPT. For instance, the previous ChatGPT frequently provided erroneous answers to questions about organizing a meeting based on different people's time windows, but GPT-4's performance in this regard was markedly superior. In the context of the Scholastic Assessment Test (SAT) test competition, it was observed that the GPT-4 achieved a score of 140, surpassing the GPT-3's score of 100. Similarly, in the Uniform Bar Exam, the GPT-3.5 demonstrated a percentile ranks (PR) of approximately 10, whereas the GPT-4 exhibited a significantly higher PR of 90. Comparable outcomes were observed in the Law School Admission Test, where the GPT-3.5 obtained a PR of 40, while the GPT-4 achieved a PR of 88[23].

When evaluating the performance of LLMs in these examinations, the availability and prevalence of the tests, as well as the availability of relevant test preparation resources, serve as significant barriers to achieving a strong performance. Consequently, it is imperative to select more specialized and less widely known subjects when assessing an LLM's performance. In contrast to extensively accessible knowledge repositories, funduscopic disease represents a topic that is relatively unfamiliar to the general public, thereby potentially offering a more equitable means of evaluating LLMs. The exclusion of test questions from the training dataset played a crucial role in evaluating the accuracy of the LLM[22]. To prevent any data contamination, fresh single-choice examinations were developed. Our evaluation will primarily concentrate on ChatGPT (GPT-3.5)[17], ChatGPT (GPT-4)[23], PaLM 2[24], Claude 2[25], and SenseNova[26]. We conducted a comparative analysis of the performance levels of these five models and further examined the stability and confidence of these LLMs in assessing funduscopic disease knowledge. It is expected that our research will contribute to the future incorporation of artificial intelligence within clinical environments and medical training.

MATERIALS 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) and followed the principles of the Declaration of Helsinki (Nanchang, China; No.2021039). All subjects were notified of the objectives and content of the study and latent risks, and then provided written informed consent to participate.

Related Work

LLM In recent times, Google has introduced a transformer model[27] that relies solely on the attention mechanism, leading to significant enhancements in the parallel processing capabilities of sequence models. Various PLMs[28], such as BERT[29], GPT, and BART[30], which are built upon the Transformer architecture, have demonstrated commendable outcomes across numerous NLP tasks. In the initial stage of PLMs, a substantial amount of language knowledge is acquired, thereby enhancing the proficiency in local rewriting and structural transformation during generation of rehearsed text. Following the Scaling Law, researchers persistently explore methods to scale up language models. In recent years, researchers have effectively escalated the parameter scale of PLMs from the billion/billion scale to the billions/billions scale, commonly referred to as LLMs[1]. In contrast to earlier PLMs, LLMs exhibit robust universal language comprehension and generation abilities, enabling them to attain optimal performance across diverse NLP tasks, even in scenarios with limited or no available samples.

The objective of LLMs is to enhance answer precision by encompassing a broader spectrum of knowledge and language contexts. To achieve this, NLP necessitates the utilization of unsupervised training techniques to acquire a comprehensive pre-training model from an extensive unlabeled text corpus. Subsequently, the model is fine-tuned using annotated data from specific downstream tasks, thereby enhancing its performance in those respective tasks[31]. The substantial number of parameters present in LLMs results in significant computational and time expenses for training. Consequently, the conventional approach of pre-training language models, known as "pre-training+fine-tuning", cannot be directly employed for LLMs. To leverage the characteristics of LLMs and circumvent the need to train all model parameters, recent research in the field of NLP has shifted its attention towards "prompt learning"[32]. The utilization of Prompt Learning allows LLMs to attain commendable performance, either by not requiring parameters training or by only necessitating a limited amount of parameters training based on downstream task data[33].

Despite the impressive natural language generation abilities exhibited by GPT-3 and similar LLMs, their generated text often falls short of meeting human expectations. Consequently, OpenAI has put forth a methodology to align model output with human preferences, encompassing two pivotal techniques: "Instruction Tuning"[20] and "RLHF"[21]. RLHF, a reinforcement learning algorithm, employs human feedback as a reward signal to guide the behavior of these LLMs. The utilization of RLHF has been extensively employed in ChatGPT to enhance the efficacy of conversational generation models, thereby facilitating improved comprehension and generation of natural language. Through the conversion of user feedback into reward or penalty values, RLHF algorithms enable the continual optimization of dialogue strategies within the dialogue generation model, resulting in enhanced ability to cater to user requirements[34].

Language models and examination LLMs employ deep learning models that have been trained on extensive text data to generate natural language text or comprehend the semantic meaning of textual content. The proliferation of training data and advancements in computing power have contributed to the rapid growth of LLMs. Notably, GPT-4, a prominent LLM, has demonstrated impressive performance on various standardized assessments. For instance, GPT-4 has achieved scores within the top 10 percent on the Uniform Bar Examination (UBE), the top 7 percent on the SAT reading test, and the top 11 percent on the SAT math test[23]. This study serves as an introductory evaluation of LLMs in the domain of Funduscopic Disease Knowledge, with the aim of inspiring further research in the assessment of LLMs within highly specialized areas of medicine.

Methods A 100-question single-choice examination on funduscopic disease was developed by a group of seasoned ophthalmology professors for the purpose of assessing the performance of both LLMs and human participants in answering ophthalmology specialty exam questions. The study included the evaluation of five LLMs: ChatGPT (GPT-3.5), ChatGPT (GPT-4), PaLM2, Claude 2, and SenseNova. The exam encompasses questions on the following topics: vitreous diseases (20 questions), optic nerve diseases (20 questions), retinal vascular diseases (20 questions), retinal detachment diseases (20 questions), and macular diseases (20 questions).

The study involved conducting five distinct trials (designated as Trial 1 through Trial 5), wherein each LLM was tested with a set of 100 single-choice questions pertaining to funduscopic disease. In each test trial, the global prompt and instructions prompt were phrased differently in order to account for response-noise due to prompt-noise. At the commencement of each trial, an initialization prompt was provided to the LLM, regardless of whether it began on a new thread or following a reset. The LLM received instructions and questions until the completion of the test. The LLM participants were instructed

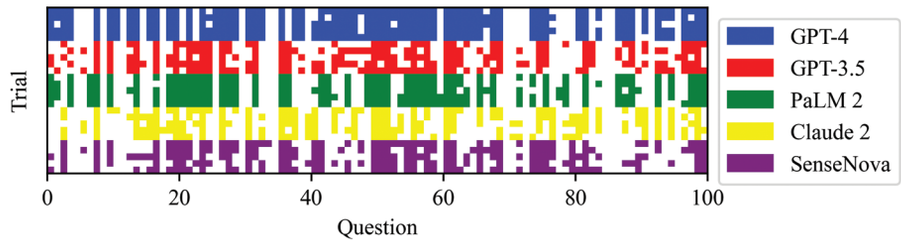

Figure 1 Raw average scores for every LLM test Test questions were in the columns, and separate LLMs were in the rows with different colors. The correct answers are indicated with dark squares. LLM: Large language model.

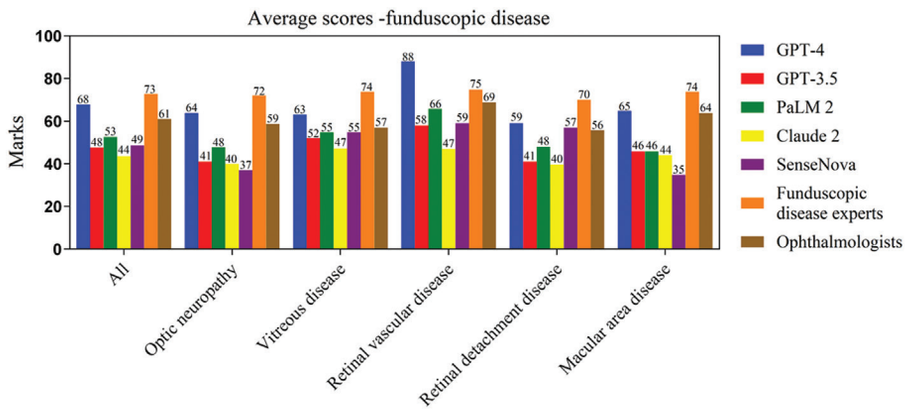

Figure 2 Average test scores for each large language models and humans.

to solely provide the accurate response without offering any accompanying explanations. Each question underwent five trials, resulting in each LLM electing the answer to the same question five times.

In this study, we conducted a comparative analysis of LLM test scores between each other as well as with scores from two human groups (funduscopic disease experts group and ophthalmologists group) and evaluated the average scores, score consistency, and confidence in correct answers. The funduscopic disease experts group consisted of three experienced directors of funduscopic and two attending doctors of the funduscopic. The ophthalmologists group comprised five ophthalmic residents who were not specialized in funduscopic. In order to assess the overall consistency of the scoring, we calculated for average correlations and standard deviations. The average correlation value indicated the level of consistency in the accurate scores obtained from the experiment: a value of 1 indicated identical distributions, 0 indicated completely random distributions, and -1 indicated distributions with a complete inverse correlation. The total number of correct answers for each question was tallied across all trials to ascertain the level of confidence in the answers provided by the LLMs. For example, in the case of a test consisting of 100 questions, the proportion of questions in which all five answers were correctly answered increased by 1% when each LLM responded accurately to the same question five times. Additionally, the test results were compared to the expected distribution when candidates made random guesses. When guessing randomly, the predicted number of correct answers in five trials is roughly 0.2×5=1.0 on average (100 questions contain four alternatives). The aforementioned value can be employed to estimate the frequency of correct responses for each question through the utilization of the Poisson distribution.

Ultimately, the scores obtained from the cumulative calculations of five LLMs and two human groups were subjected to comparison.

RESULTS

Comparison Between LLMs Scores Figures 1 and 2 depict the raw marks and mean test scores, respectively. Upon examining the raw marks in Figure 1, it becomes evident that each LLM exhibited variability across trials, not only in terms of the uncertainty surrounding the total score, but also in relation to the frequency of correctly answered questions. Notably, the ChatGPT-4 displayed the highest number of dark squares, indicating correct answers. In Figure 2, the average score is presented, wherein the LLM mean test score represents the average of five distinct trials, while the average scores for human groups represent the mean scores of each individual within their respective groups. The average scores of ChatGPT-4, PaLM2, SenseNova, ChatGPT-3.5, and Claude 2 were 68, 53, 49, 48, and 44, respectively, arranged in descending order. In contrast, ChatGPT-4 exhibited the highest performance in terms of LLM. The average scores of the group consisting of funduscopic disease experts and ophthalmologists were 73 and 61, respectively. Overall,

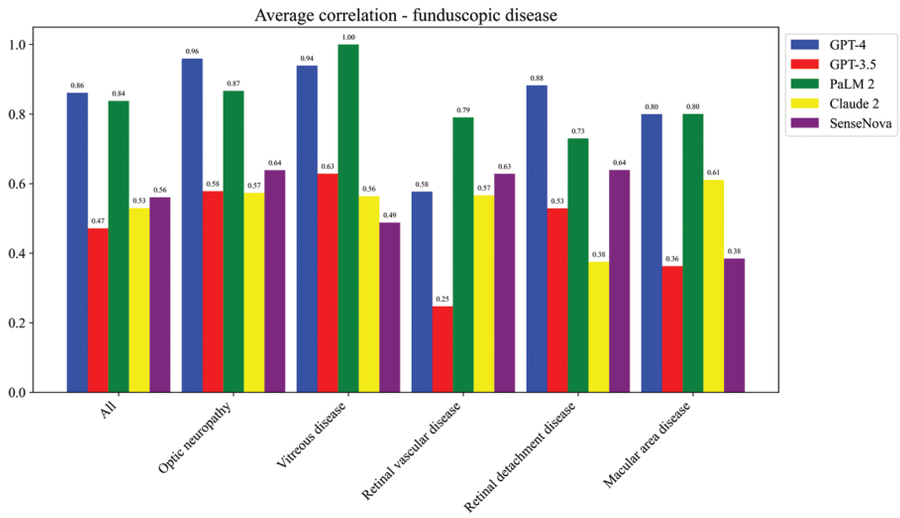

Figure 3 Correlation in scoring for large language models by category.

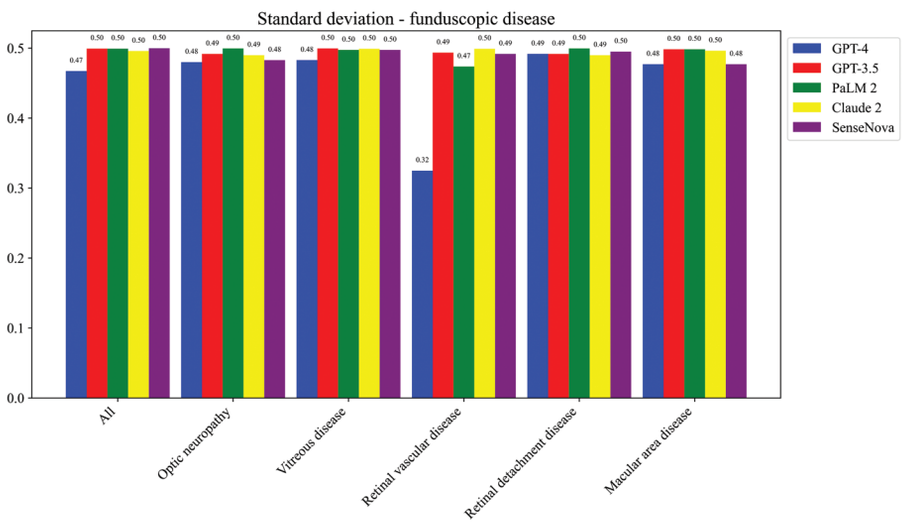

Figure 4 Standard deviation in scoring for large language models by category.

ChatGPT-4 outperformed the other LLMs and demonstrated comparable performance to the ophthalmologists' group, albeit falling short of the expertise exhibited by the funduscopic disease experts' group.

Comparison of LLMs Answer Stability Figures 3 and 4 depict the correlation and standard deviation across 5 trials. The LLMs exhibited a significantly higher degree of consistency in their answers and scores, as evidenced by their low standard deviation in scoring and high average correlation between trials. Notably, both ChatGPT-4 and PaLM2 consistently demonstrated a high average correlation in the tests, surpassing 0.8. ChatGPT-4 exhibited a significantly lower standard deviation of 0.47, in contrast to the remaining four LLMs which displayed a standard deviation of 0.50. Overall, ChatGPT-4 demonstrated slightly superior stability when addressing professional single-choice questions.

Comparison of LLM Answer Confidence Based on the data presented in Figure 5, it is apparent that both ChatGPT-4 and PaLM2 exhibited a very little chance of guessing answers. ChatGPT-4 demonstrated a higher level of confidence with a 59% success rate in answering questions, yet it also displayed a tendency towards perplexity by providing incorrect responses in 28% of questions (Figure 5A). On the other hand, PaLM2 either had confidence, correctly answering 44% of each trial, or confusion, answering incorrectly 38% each trial (Figure 5C). The SenseNova model demonstrated a moderate level of confidence, achieving a 29% accuracy rate for correctly answered questions and a 26% error rate (Figure 5E). In contrast, ChatGPT-3.5 exhibited a lower level of confidence and displayed a higher inclination towards confusion, with a 23% accuracy rate for correct answers and a 26% error rate (Figure 5B). Claude 2 exhibited the lowest level of confidence, with only 21% of answers being correct and 32% being incorrect (Figure 5D).

DISCUSSION

This study designed a 100-question single-choice exam centered on funduscopic disease, with the purpose of assessing the proficiency of LLMs in a profoundly specialized

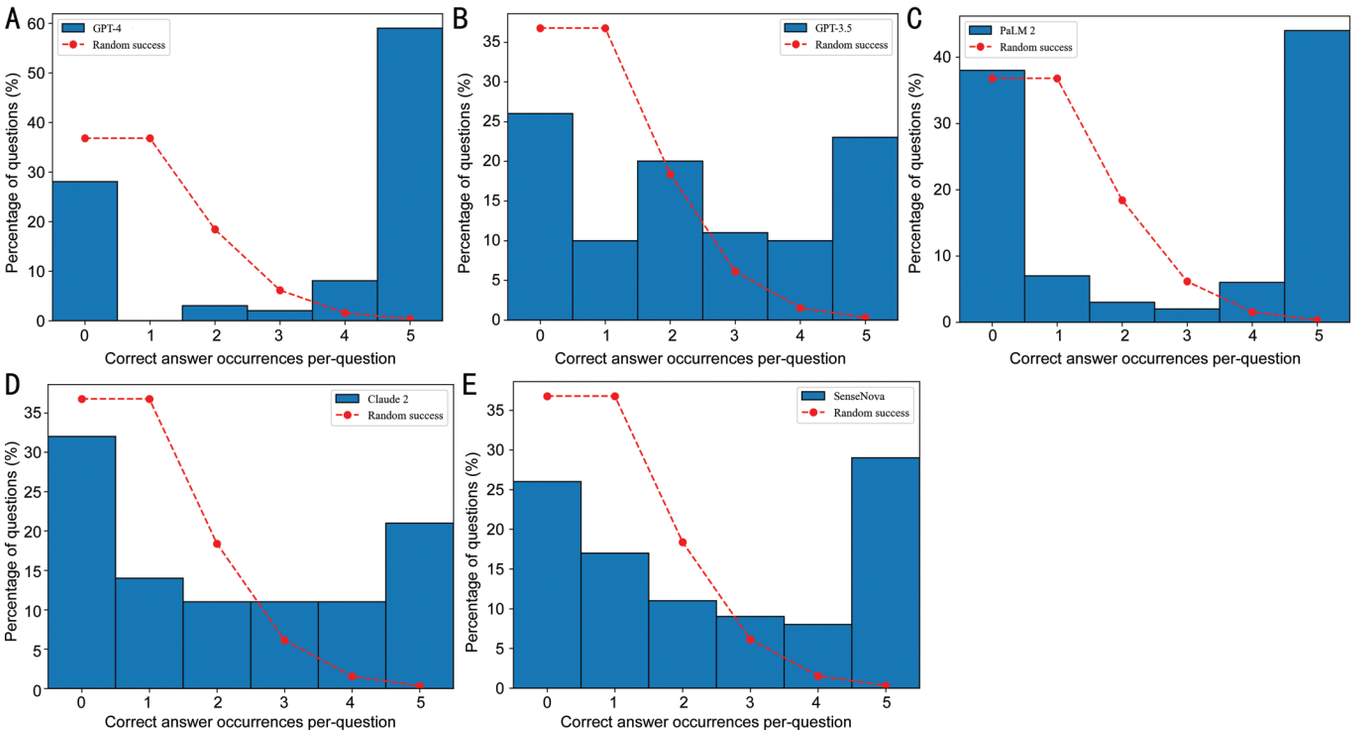

Figure 5 Confidence in answers The number of correct answers occurrences per-question for each LLMs. The dashed red curve indicates the expected distribution if the answers were randomly selected based on the Poisson distribution. LLM: Large language models.

subject matter. Additionally, the study aimed to compare the performance of five distinct LLMs against one another, as well as two human cohorts. Remarkably, the ChatGPT-4 model exhibited exceptional performance and consistency in terms of answer correlations and answer confidence within these highly specialized tests. On the other hand, PaLM2 exhibited a substantial positive correlation (exceeding 0.8) in terms of answer accuracy during the exam. PaLM2 ranks second only to ChatGPT-4 in terms of confidence in answer, surpassing Claude 2, SenseNova, and ChatGPT-3.5. Although funduscopic disease is a highly specialized subject, ChatGPT-4 demonstrates exceptional performance and is anticipated to undergo further enhancements, suggesting its potential as a valuable resource for medical students in the foreseeable future. Within the realm of highly specialized medical domains, ChatGPT possesses the capacity to swiftly gather and address diverse medical information, rendering it a valuable pedagogical tool for students[35]. Chat-GPT possesses the capability to engage in interactive question and answer sessions, promptly offering feedback on medical inquiries. Additionally, it has the ability to simulate interactive scenarios, fostering increased student engagement within educational settings and enhancing their aptitude for self-directed learning. Furthermore, ChatGPT can recommend supplementary educational resources, thereby enabling medical students to allocate less time to traditional classroom instruction and devote more time to refining their practical skills. ChatGPT-4 demonstrated exceptional performance overall, its response to questions was found to be comparable to that of a group of ophthalmologists, but not as proficient as a group of funduscopic disease experts (Figure 2). Moreover, even among professionals with similar backgrounds, ophthalmologists possess greater flexibility in selecting diverse diagnostic and treatment approaches based on individual patient circumstances. Consequently, it is evident that GPT-4 cannot entirely substitute for the expertise and decision-making capabilities of ophthalmologists.

Application of LLMs in Ophthalmology The utilization of LLMs in the field of ophthalmology holds potential benefits, although the existing body of published research on this subject appears to be relatively limited in comparison to other medical specialties. This study serves as an extension of previous investigations that have employed cutting-edge LLM technology within the realm of ophthalmology. Antaki *et al*[36] conducted an assessment to evaluate the precision of ChatGPT in the Ophthalmic Knowledge Assessment Program (OKAP) exam. The performance of the LLMs in different subspecialties of ophthalmology varied, with the best outcomes observed in general medicine and the worst in neuroophthalmology, ophthalmic pathology, and intraocular tumors. This suggests that specialized LLMs with prior training in specific ophthalmic fields may be necessary to enhance their performance in the respective subspecialties. A significant development in this regard was witnessed at the Vision China 2023 Conference in May 2023, where the Optometry Hospital affiliated to Wenzhou

Medical University and China Eye Valley pioneered the launch of Neuro-OphGPT. Neuroophthalmology, as a highly challenging field within ophthalmology, encompasses the intricate task of visually perceiving and interpreting observed situations or objects, while also finishing intricate interplay between the eye and the brain. The implementation of Neuro-OphGPT can assist the diagnosis of neuroophthalmological disorders, enabling physicians to progressively enhance their comprehension of disease diagnosis and treatment within the realm of neuroophthalmology. Furthermore, a study revealed that ChatGPT exhibited proficiency in analyzing eye care inquiries written by patients and subsequently generating suitable responses that were comparable to written responses from doctors in terms of accuracy of information, perceived consensus within the medical community, and the potential likelihood and severity of adverse outcomes[37].

Application of LLMs in Medical Our findings align with previous research indicating that LLMs have the capability to effectively handle various medical tasks. ChatGPT challenged a tough USMLE, and as a result, ChatGPT scores passed or came close to passing in all three sections of the exam. Furthermore, GPT-4 exhibits significantly improved performance compared to its predecessor, GPT-3.5, in the USMLE[38]. LLMs have demonstrated their ability to successfully pass examinations in specialized fields such as pathology[39], radiation oncology physics[40], and ophthalmology[41]. Additional applications of LLMs encompass the collection of clinical histories and medical records. ChatGPT can ask patients questions about symptoms, and gather, sort, and integrate clinical information from multiple sources, surpassing human capabilities in terms of speed. When combined with speech recognition technology, ChatGPT holds promise for automating the compilation of medical histories[42]. Furthermore, ChatGPT can contribute to clinical decision making by facilitating evidence-based assessments of participants/patients, intervention, control/comparison, and outcomes (PICO)[43]. The utilization of ChatGPT facilitates the examination of linguistic patterns in both spoken and written communication that undergo alterations during the initial phases of Alzheimer's disease, thereby offering the possibility of early diagnosis of dementia[44]. A recent advancement is Med-PaLM 2, which has demonstrated remarkable performance, nearing the proficiency level of human experts[45].

Challenge This have sparked speculations regarding the potential substitution of doctors by AI, although the truth is not nearly that dramatic[46]. It has been demonstrated that ChatGPT gives false information in response to reasonable patient inquiries about the prevention of cardiovascular disease[47]. According to Howard's research, ChatGPT was capable of proposing suitable dosages; nevertheless, it exhibited inconsistent drug recommendations and failed to explicitly mention any contraindications. ChatGPT can be utilized by physicians to obtain preliminary diagnostic and treatment suggestions; however, it lacks the capability to analyze specific cases individually[48]. Furthermore, the medical profession necessitates not only theoretical medical expertise, but also practical clinical experience, as each patient's physiological state differs and diverse treatment approaches are warranted for the same ailment. Additionally, medicine encompasses aspects of humanities, whereby the efficacy of treatment relies not solely on technological advancements, but also on compassionate care. In the process of disease treatment, humanistic care plays a very important role. However, due to the imperfect performance of LLMs in their specialized domain, concerns regarding uncertainty and inaccuracy arise, thereby rendering them inadequate substitutes for doctors.

The primary objective of medical LLMs should be to provide support to physicians rather than to replace them. Through model training, LLMs continuously acquire new knowledge and thinking approaches in the medical domain, relying on extensive computing capabilities for integration. AI has the potential to aid physicians in swiftly executing straightforward and repetitive tasks, thereby enhancing efficiency, elevating the quality of their work, advancing the level of treatment, and alleviating the burden on medical personnel. Nevertheless, ultimate judgment and decision must still be conducted under the supervision of a doctor. Especially in medical scenarios, the utilization of LLMs encompasses various considerations, including ethical implications, policy considerations, and the need for demonstration.

In addition, poor model accuracy for users not represented in the training data, model openness and construction, model output accountability, potential model bias, and the possibility of privacy and confidentiality violations are the main ethical concerns of LLMs in medicine. Effectively regulating LLM is essential to master the fundamental ethical issues inherent in its design and use. A comprehensive framework and mitigating strategies will be imperative for the responsible integration of LLMs into medical practice, ensuring alignment with ethical principles and safeguarding against potential societal risks.

In conclusion, this study presents the pioneering evidence that ChatGPT-4 is excellent at answering the highly specialized question of funduscopic disease. The significance and potential of this finding within the domain of ophthalmology are noteworthy. However, despite the encouraging performance of ChatGPT, its practical application in funduscopic disease might be limited due to its inability to process images. The diagnosis and treatment of funduscopic disease heavily depend on examination and imaging techniques. In the future, it may be necessary for LLMs such as ChatGPT to integrate

additional transformers capable of handling diverse data types in order to facilitate image processing. As long as ethical and technical considerations are adequately addressed, we are of the opinion that validated LLMs can play a crucial role in enhancing healthcare for both patients and physicians.

ACKNOWLEDGEMENTS

Foundations: Supported by National Natural Science Foundation of China (No.82160195); Science and Technology Project of Jiangxi Provincial Department of Education (No. GJJ200169); Science and Technology Project of Jiangxi Province Health Commission of Traditional Chinese Medicine (No.2020A0087); Science and Technology Project of Jiangxi Health Commission (No.202130210).

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

REFERENCES

1 Zhao WX, Zhou K, Li JY, *et al*. A survey of large language models. arXiv:2303.18223.

2 Hirschberg J, Manning CD. Advances in natural language processing. *Science* 2015;349(6245):261-266.

3 Kreimeyer K, Foster M, Pandey A, *et al*. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *J Biomed Inform* 2017;73:14-29.

4 Thomas AA, Zheng CY, Jung H, *et al*. Extracting data from electronic medical records: validation of a natural language processing program to assess prostate biopsy results. *World J Urol* 2014;32(1):99-103.

5 Kho AN, Pacheco JA, Peissig PL, *et al*. Electronic medical records for genetic research: results of the eMERGE consortium. *Sci Transl Med* 2011;3(79):79re1.

6 Waites W, Cavaliere M, Manheim D, *et al*. Rule-based epidemic models. *J Theor Biol* 2021;530:110851.

7 Shah P, Kendall F, Khozin S, *et al*. Artificial intelligence and machine learning in clinical development: a translational perspective. *NPJ Digit Med* 2019;2:69.

8 Helm JM, Swiergosz AM, Haeberle HS, *et al*. Machine learning and artificial intelligence: definitions, applications, and future directions. *Curr Rev Musculoskelet Med* 2020;13(1):69-76.

9 Goto T, Camargo CA Jr, Faridi MK, *et al*. Machine learning-based prediction of clinical outcomes for children during emergency department triage. *JAMA Netw Open* 2019;2(1):e186937.

10 Zhao L, Zhang L, Wu ZH, *et al*. When brain-inspired AI meets AGI. *Meta-Radiology* 2023;1(1):100005.

11 Kaplan J, McCandlish S, Henighan T, *et al*. Scaling laws for neural language models. arXiv:2001.08361.

12 Wei J, Tay Y, Bommasani R, *et al*. Emergent abilities of large language models. arXiv:2206.07682.

13 Hoffmann H, Borgeaud S, Mensch A, *et al*. Training compute-optimal large language models. arXiv:2203.15556.

14 Taylor R, Kardas M, Cucurull G, *et al*. Galactica: a large language model for science. arXiv:2211.09085.

15 Qin CW, Zhang A, Zhang ZS, *et al*. Is ChatGPT a general-purpose natural language processing task solver. arXiv:2302.06476.

16 Dong QX, Li L, Dai DM, *et al*. A survey on in-context learning. arXiv:2301.00234.

17 Brown TB, Mann B, Ryder N, *et al*. Language models are few-shot learners. arXiv:2005.14165.

18 Chen M, Tworek J, Jun H, *et al*. Evaluating large language models trained on code. arXiv:2107.03374.

19 Ouyang L, Wu J, Xu J, *et al*. Training language models to follow instructions with human feedback. arXiv:2203.02155.

20 Wei J, Bosma M, Zhao VY, *et al*. Finetuned language models are zero-shot learners. arXiv:2109.01652.

21 Stiennon N, Ouyang L, Wu J, *et al*. Learning to summarize from human feedback. arXiv:2009.01325.

22 Bubeck S, Chandrasekaran V, Eldan R, *et al*. Sparks of artificial general intelligence: early experiments with GPT-4. arXiv:2303.12712.

23 OpenAI, Achiam J, Adler S, *et al*. GPT-4 technical report. arXiv:2303.08774.

24 Anil R, Dai AM, Firat O, *et al*. Palm 2 technical report. arXiv:2305.10403.

25 Bai YT, Kadavath S, Kundu S, *et al*. Constitutional AI: harmlessness from AI feedback. arXiv:2212.08073.

26 Xue ZY, Song GL, Guo QS, *et al*. RAPHAEL: text-to-image generation *via* large mixture of diffusion paths. arXiv:2305.18295.

27 Vaswani A, Noam Shazeer N, ParmarN, *et al*. Attention is all you need. arXiv:1706.03762.

28 Wu WQ, Jiang CY, Jiang Y, *et al*. Do PLMs know and understand ontological knowledge *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Toronto, Canada. Stroudsburg, PA, USA: ACL, 2023:3080-3101.

29 Devlin J, Chang MW, Lee K, *et al*. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805.

30 Lewis M, Liu YH, Goyal N, *et al*. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Online. Stroudsburg, PA, USA: ACL, 2020:7871-7880.

31 Kalyan KS, Rajasekharan A, Sangeetha S. AMMUS: a survey of transformer-based pretrained models in natural language processing. arXiv:2108.05542.

32 Liu PF, Yuan WZ, Fu JL, *et al*. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput Surv* 2023;55(9):1-35.

33 Wei J, Wang XZ, Schuurmans D, *et al*. Chain-of-thought prompting elicits reasoning in large language models. arXiv:2201.11903.

34 Bai YT, Jones A, Ndousse K, *et al*. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv:2204.05862.

35 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.

36 Antaki F, Touma S, Milad D, *et al*. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. *Ophthalmol Sci* 2023;3(4):100324.

37 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.

38 Nori H, King N, McKinney SM, *et al*. Capabilities of GPT-4 on medical challenge problems. arXiv:2303.13375.

39 Sinha RK, Deb Roy A, Kumar N, *et al*. Applicability of ChatGPT in assisting to solve higher order problems in pathology. *Cureus* 2023;15(2):e35237.

40 Holmes J, Liu ZL, Zhang L, *et al*. Evaluating large language models on a highly-specialized topic, radiation oncology physics. *Front Oncol* 2023;13:1219326.

41 Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. *JAMA Ophthalmol* 2023;141(6):589-597.

42 DiGiorgio AM, Ehrenfeld JM. Artificial intelligence in medicine & ChatGPT: de-tether the physician. *J Med Syst* 2023;47(1):32.

43 Fatani B. ChatGPT for future medical and dental research. *Cureus* 2023;15(4):e37285.

44 Agbavor F, Liang HL. Predicting dementia from spontaneous speech using large language models. *PLoS Digit Health* 2022;1(12):e0000168.

45 Singhal K, Tu T, Gottweis J, *et al*. Toward expert-level medical question answering with large language models. *Nat Med* 2025;31(3):943-950.

46 Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med* 2023;388(13):1233-1239.

47 Sarraju A, Bruemmer D, van Iterson E, *et al*. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. *JAMA* 2023;329(10):842-844.

48 Howard A, Hope W, Gerada A. ChatGPT and antimicrobial advice: the end of the consulting infection doctor *Lancet Infect Dis* 2023;23(4):405-406.