

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/382364512>


Development and Evaluation of a Retrieval-Augmented Large Language Model Framework for Ophthalmology

Article in *Jama Ophthalmology* · July 2024
DOI: 10.1001/jamaophthalmol.2024.2513

CITATIONS
4

READS
34


14 authors, including:



Tingxin Cui
Sun Yat-Sen University

8 PUBLICATIONS 80 CITATIONS


SEE PROFILE



Lanqin Zhao
Sun Yat-Sen University

57 PUBLICATIONS 918 CITATIONS


SEE PROFILE



Xiaohang Wu
Sun Yat-Sen University

139 PUBLICATIONS 3,282 CITATIONS

SEE PROFILE



Duoru Lin
Sun Yat-Sen University

108 PUBLICATIONS 1,530 CITATIONS

SEE PROFILE

JAMA Ophthalmology | Original Investigation

Development and Evaluation of a Retrieval-Augmented Large Language Model Framework for Ophthalmology

Ming-Jie Luo, MD; Jianyu Pang, MSc; Shaowei Bi, MD; Yunxi Lai, MD; Jiaman Zhao, BE; Yuanrui Shang, BM; Tingxin Cui, MD; Yahan Yang, PhD, MD; Zhenzhe Lin, MEng; Lanqin Zhao, MSc; Xiaohang Wu, PhD, MD; Duoru Lin, PhD, MD; Jingjing Chen, MD; Haotian Lin, PhD, MD

[+ Invited Commentary](#)

[+ Supplemental content](#)

IMPORTANCE Although augmenting large language models (LLMs) with knowledge bases may improve medical domain-specific performance, practical methods are needed for local implementation of LLMs that address privacy concerns and enhance accessibility for health care professionals.

OBJECTIVE To develop an accurate, cost-effective local implementation of an LLM to mitigate privacy concerns and support their practical deployment in health care settings.

DESIGN, SETTING, AND PARTICIPANTS ChatZOC (Sun Yat-Sen University Zhongshan Ophthalmology Center), a retrieval-augmented LLM framework, was developed by enhancing a baseline LLM with a comprehensive ophthalmic dataset and evaluation framework (CODE), which includes over 30 000 pieces of ophthalmic knowledge. This LLM was benchmarked against 10 representative LLMs, including GPT-4 and GPT-3.5 Turbo (OpenAI), across 300 clinical questions in ophthalmology. The evaluation, involving a panel of medical experts and biomedical researchers, focused on accuracy, utility, and safety. A double-masked approach was used to try to minimize bias assessment across all models. The study used a comprehensive knowledge base derived from ophthalmic clinical practice, without directly involving clinical patients.

EXPOSURES LLM response to clinical questions.

MAIN OUTCOMES AND MEASURES Accuracy, utility, and safety of LLMs in responding to clinical questions.

RESULTS The baseline model achieved a human ranking score of 0.48. The retrieval-augmented LLM had a score of 0.60, a difference of 0.12 (95% CI, 0.02-0.22; $P = .02$) from baseline and not different from GPT-4 with a score of 0.61 (difference = 0.01; 95% CI, -0.11 to 0.13; $P = .89$). For scientific consensus, the retrieval-augmented LLM was 84.0% compared with the baseline model of 46.5% (difference = 37.5%; 95% CI, 29.0%-46.0%; $P < .001$) and not different from GPT-4 with a value of 79.2% (difference = 4.8%; 95% CI, -0.3% to 10.0%; $P = .06$).

CONCLUSIONS AND RELEVANCE Results of this quality improvement study suggest that the integration of high-quality knowledge bases improved the LLM's performance in medical domains. This study highlights the transformative potential of augmented LLMs in clinical practice by providing reliable, safe, and practical clinical information. Further research is needed to explore the broader application of such frameworks in the real world.

JAMA Ophthalmol. doi:10.1001/jamaophthalmol.2024.2513
Published online July 18, 2024.

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Haotian Lin MD, PhD (linht5@mail.sysu.edu.cn), and Jingjing Chen, MD (chenjingjing@gzzoc.com), Zhongshan Ophthalmic Center, Sun Yat-Sen University, #7 Jinsui Rd, Guangzhou 510060, China.

Recently, large language models (LLMs), a branch of deep learning, have garnered global attention. These LLMs are text-in/text-out models capable of following linguistic commands, comprehending contextual information, mimicking existing logical examples, and reasoning step by step to produce accurate conclusions.¹ Their extensive training on diverse data makes them akin to virtual assistants with access to billions of data, offering immense potential in expertise-intensive medical fields. LLM applications range from clinical administrative tasks such as creating informed consent forms, assisting in disease diagnoses, and guiding treatment decisions.² In medical education, LLMs not only excel in passing examinations but also revolutionize interactive teaching methods, such as creating patient case studies or generating examination questions.^{3,4} Additionally, their utility in medical research is of great significance, where they contribute to pharmaceutical discovery by identifying potential drug candidates⁵ and enhancing literature analysis by summarizing research findings or detecting trends in medical studies.^{6,7} These varied applications highlight the extensive impact of LLMs across the medical spectrum.

Despite the potential of LLMs, their integration into clinical practice is impeded by several critical challenges. One key issue is the occurrence of LLM-generated hallucinations, where models generate misleading or inaccurate information.⁸⁻¹² This issue often stems from the lack of medical specialization in current LLMs and the absence of rigorous evaluation standards for LLMs in clinical settings, which underscores the urgency for more reliable assessment frameworks. These challenges are further exacerbated by the quality of the training data,¹³ highlighting the need for medical-specific benchmark datasets to improve LLM accuracy and applicability. Moreover, the majority of LLMs are developed by technology firms, raising privacy and regulatory issues due to the requirement of uploading patient data to external servers.¹⁴ Compounding these concerns, localized LLM deployment in hospitals is further hindered by the expensive computational resources needed. Consequently, there is a growing demand for cost-effective, locally implemented LLMs.

In this quality improvement study, ophthalmology was selected as the primary focus due to the distinctive challenges inherent to the field. Ophthalmology is characterized by a lack of specialized medical centers and a significant deficit in publicly available educational content regarding eye health. This situation results in a constrained public exposure to ophthalmological knowledge, which implies a reduction in the amount of training data available, leading to a diminished capacity of existing LLMs to address issues related to eye care effectively. The practical deployment approach for LLMs in ophthalmology could serve as an example for its application across other specialized areas that are similarly challenged by a scarcity of training resources. To address these challenges, we developed a comprehensive ophthalmic dataset and evaluation framework (CODE) for augmenting and rigorously assessing the LLMs. We then developed a retrieval-augmented LLM (ChatZOC [Sun Yat-Sen University Zhongshan Ophthalmology Center]) to determine its accuracy and cost within a localized LLM framework that may be considered as

Key Points

Question How can the challenges of knowledge inaccuracies and data privacy issues when applying large language models (LLMs) in ophthalmology be overcome?

Findings In this quality improvement study, a retrieval-augmented LLM framework was developed that achieved capabilities for human ranking scores or inappropriate content not different from a commercially available LLM.

Meaning Using knowledge augmentation framework such as that developed here presents a potentially effective approach for deploying LLMs in clinical medicine that may be accurate while preserving privacy when seeking medical consultations or decision-making.

a practical and secure deployment within medical institutions following evaluation by medical experts for medical accuracy, utility, and safety.

Methods

The ethical review of this study was approved by the Zhongshan Ophthalmic Center Ethics Review Committee (2023KYPJ191). From December 2023 to January 2024, we conducted the development and evaluation of our own retrieval-augmented LLM for ophthalmology. This study followed the Standards for Quality Improvement Reporting Excellence (SQUIRE) reporting guidelines.

Comprehensive Ophthalmic Dataset

The comprehensive ophthalmic dataset (COD) comprised the following 4 components (eFigure 5A in [Supplement 1](#)): (1) knowledgeable terminologies and disease descriptions (KTDD) (eFigure 1 in [Supplement 1](#)), (2) single-choice and long-answer case study questions (SLC) (eFigure 2 in [Supplement 1](#)), (3) frequently asked questions on common eye diseases (FAQ) (eFigure 3 in [Supplement 1](#)), and (4) real-world ophthalmology consultations (RWOC) (eFigure 4 in [Supplement 1](#)). KTDD was conducted exclusively in English due to its origin from ophthalmology guidelines and literature, which encompassed a plethora of specialized ophthalmic terminology that currently lacked accurate large-scale translations. On the other hand, both FAQ and RWOC were entirely composed in Chinese, as they derived from authentic dialogues sourced from ophthalmology popular science platforms and online eye hospitals. Given that our study primarily operated in Chinese and aimed to enhance model performance in following Chinese instructions, we did not translate this subset of data into English. Furthermore, the majority of mainstream Chinese LLMs are adapted from English baseline models, enabling them to handle and analyze both English and Chinese data without translation. The datasets were manually reviewed and curated by 30 ophthalmic specialists to ensure content accuracy. The specific details regarding the format and sources of each component are shown in eMethods 1 in [Supplement 1](#). In short, the

KTDD and SLC were primarily derived from ophthalmic guidelines and textbooks, whereas the FAQ originated from educational materials on popular science for ocular diseases. The RWOC comprised colloquial question-answer pairs from online hospital dialogues and hotline dialogues of the Zhongshan Ophthalmic Center (ZOC), and patients provided informed consent for data usage, with subsequent anonymization and cleansing procedures applied.

LLM Development

We initially included 10 LLMs for consideration of the baseline model, namely: GPT-4 (OpenAI), GPT-3.5 Turbo (OpenAI), ChatGLM2 (THUDM), ChatGLM (THUDM), StableVicuna (CarperAI), Chatyuan (ClueAI), Llama2-Chat-70B (Meta), Llama2-Chinese-Chat-13B (FlagAlpha), Baichuan-7B (Baichuan-Inc), and Baichuan-13B (Baichuan-Inc) (eTable 1 and eMethods 2 in [Supplement 1](#)). The majority of these models were reported to provide support for the Chinese language, which was crucial given that the most valuable data in our evaluation dataset was derived from real-world ophthalmic consultations conducted in Chinese. We provided specific version parameters, operating notes, and implementation details of each LLM in eMethods 2 in [Supplement 1](#). Among all of the LLMs, we chose Baichuan-13B as our baseline model because it was comparably small (inference on a server with only 1 Tesla V100 graphics processing unit [GPU]) and showed instruction alignment capabilities in the pilot experiment of this study. The GPT-4 and GPT-3.5 Turbo were not chosen because they were not open-sourced models.

We then used the data generated from COD to fine-tune Baichuan-13B (eMethods 3 in [Supplement 1](#) for details in fine-tuning the LLM) and build a retrieval-augmented generation architecture (eMethods 3 in [Supplement 1](#)). When interacting with our fine-tuned retrieval-augmented LLM, input questions were matched with CODE to identify the top 3 questions with the highest similarity (eFigure 5B in [Supplement 1](#)). The answers to these 3 questions were used as background knowledge in guiding our LLM model to respond in a specific direction and reducing the likelihood of producing hallucinations answers. Because the most important data in the COD came from the ZOC, this retrieval-augmented LLM framework was termed *ChatZOC*.

Rigorous Evaluation Framework

To comprehensively assess the LLM's ability to answer various types of medical questions qualitatively, we conducted automatic evaluations (eFigure 6C in [Supplement 1](#)) and human evaluations (eFigure 6D and E in [Supplement 1](#)) on 300 questions randomly sampled from the COD.

Automatic Evaluations

In general, our data were presented in 2 distinct forms. Data obtained from SLC included single- and multiple-choice questions with their reference answers and explanations. For data from SLC, the accuracy of the answer was automatically assessed first by comparing the options of multiple-choice questions in the model answer with the standard answer, and the explanations were then evaluated by human experts, to de-

Table 1. Three Axes for Rigorous Large Language Model Evaluation Framework in Ophthalmic Context

Evaluation aspect	Specific question	Evaluation option (description)
Scientific consensus	How is the answer correlated with the scientific and clinical consensus? (No consensus involved means the question is not relevant to scientific knowledge)	<ul style="list-style-type: none"> Aligned with consensus Opposed to consensus No consensus involved
Accuracy		
Missing content	Does the answer omit any content that should not be omitted?	<ul style="list-style-type: none"> No Yes, little clinical significance Yes, great clinical significance
Possible bias	Does the answer contain any discriminatory or prejudiced elements?	<ul style="list-style-type: none"> No Yes
Correct understanding	Does the answer contain evidence of correct reading comprehension? (indicating the question has been understood)	<ul style="list-style-type: none"> Yes No
Utility		
Correct retrieval	Does the answer contain evidence of correct knowledge recall? (involving relevant facts and/or correct information to the question)	<ul style="list-style-type: none"> Yes No
Correct reasoning	Does the answer contain evidence of correct reasoning? (right justification for the answer)	<ul style="list-style-type: none"> Yes No
Inappropriate/wrong content	Does the answer contain inappropriate or incorrect content?	<ul style="list-style-type: none"> No Yes, little clinical significance Yes, great clinical significance
Safety		
Possible hazard	How likely is the potential harm? (low: there is minimal risk of subsequent eye disease; medium: there is a certain probability of subsequent eye disease, influenced by individual factors; high: subsequent eye disease is almost inevitable.)	<ul style="list-style-type: none"> Low Medium High
Hazard potential	To what extent is the potential harm? (moderate harm: slight eye discomfort or irritation, minor allergic reactions, etc, and do not result in any long-term effects or impairments to the patient's vision; severe harm: continued impairment of vision or acute, severe vision impact requires further diagnosis and treatment.)	<ul style="list-style-type: none"> No harm Moderate harm Severe harm

termine whether the reasoning and content in the answer were correct. Specifically, a reasonable explanation with an incorrect option for the single- or multiple-choice questions did not count. On the other hand, data from KTDD, FAQ, and RWOC consist of questions and answers from various sources, which were simultaneously evaluated by both human and automatic scores. We used the BLEU, ROUGE and Sentence-BERT (SBERT) embeddings (distiluse-base-multilingual-cased-v1 [Sentence Transformers]),¹⁵ to automatically score the similarity between the ground truths and answers of the models.

Human Evaluations

The 300 questions were sampled from 4 data sources (FLC, KTDD, FAQ, and RWOC), each with 75 questions, to avoid se-

Table 2. Model Performance Comparison of Our Ophthalmic Large Language Model (LLM) With Other LLMs

Model name ^a	Human evaluation score (95% CI)	Bleu score (95% CI)	Rouge score (95% CI)	Embedding score (95% CI)
GPT-4	0.61 (0.53-0.68)	0.32 (0.28-0.36)	0.51 (0.49-0.53)	0.59 (0.57-0.60)
GPT-3.5 Turbo	0.56 (0.44-0.67)	0.40 (0.36-0.46)	0.49 (0.48-0.51)	0.60 (0.59-0.62)
ChatGLM2-6B	0.51 (0.39-0.62)	0.27 (0.24-0.31)	0.42 (0.40-0.44)	0.54 (0.53-0.56)
ChatGLM-6B	0.53 (0.43-0.63)	0.33 (0.29-0.38)	0.44 (0.42-0.46)	0.55 (0.53-0.57)
StableVicuna-13B	0.48 (0.39-0.59)	0.39 (0.34-0.44)	0.41 (0.39-0.44)	0.49 (0.46-0.51)
Chatyuan	0.53 (0.44-0.62)	0.26 (0.22-0.30)	0.44 (0.41-0.46)	0.49 (0.47-0.51)
Llama2-Chat-70B	0.35 (0.26-0.45)	0.54 (0.49-0.60)	0.41 (0.39-0.43)	0.49 (0.46-0.51)
Llama2-Chinese-Chat-13B	0.48 (0.41-0.56)	0.23 (0.19-0.28)	0.49 (0.47-0.51)	0.56 (0.54-0.58)
Baichuan-7B	0.37 (0.28-0.48)	0.26 (0.23-0.30)	0.42 (0.40-0.44)	0.51 (0.49-0.53)
Baichuan-13B	0.48 (0.36-0.59)	0.12 (0.10-0.14)	0.42 (0.40-0.45)	0.45 (0.43-0.47)
Baichuan-13B+COD ^b	0.60 (0.54-0.67)	0.94 (0.84-1.04)	0.86 (0.81-0.91)	0.64 (0.62-0.65)

^a The LLMs featured in this table are as follows: Baichuan-7B (Baichuan-Inc), Baichuan-13B (Baichuan-Inc), Baichuan-13B+COD (ZOC), ChatGLM-6B (THUDM), ChatGLM2-6B (THUDM), Chatyuan (ClueAI), GPT-4 (OpenAI), GPT-3.5 Turbo (OpenAI), Llama2-Chat-70B (Meta), Llama2-Chinese-Chat-13B (FlagAlpha), StableVicuna-13B (CarperAI).

^b The authors' LLM.

lection bias and ensure a comprehensive assessment. The 11 models were anonymized, and the order of the answers were randomized during evaluation. Our review panel consisted of 3 teams each consisting of 2 members, who were ophthalmologists and/or biomedical researchers, for a total of 6 individuals. In case of disagreements, the judgment team, which consisted of 2 senior ophthalmologists who were independent from the review panel, would reassess the case until reaching a consensus. The selected questions were not previously exposed to the LLMs, preventing prior test data leakage.

The human evaluations involved ranking the models based on their responses and detailed analysis across 3 axes that contained 9 aspects (Table 1), adapted from previous works¹⁶⁻²⁰ and tailored for ophthalmology. Special emphasis was placed on criteria critical to disease diagnosis and treatment recommendations, including hazard potential and the accuracy of content, to minimize risk in medical settings. For the relatively vague judgments in the standard, such as possible hazard and hazard potential, we held a meeting with the review panel in advance to unify the evaluation criteria and listed them in detail in Table 1.

Statistical Analysis

To calculate the 95% CI of the scores of automatic evaluation, human ranking, and human detail evaluation, we applied the nonparametric bootstrap procedure for the 300 samples extracted from the COD. To evaluate the performance of the models, we conducted McNemar tests on the human evaluation results to evaluate variance. All *P* values were 2-sided and not adjusted for multiple analyses. A *P* value <.05 was considered significant. All statistical analyses were conducted using R, version 4.3.1 (R Foundation for Statistical Computing).

Results

We first introduced the COD, which was specially designed, to evaluate and enhance LLMs in ophthalmology. The COD contained 4 categories: a theoretical ophthalmic knowledge base with 17 944 entries, 793 educational textbook entries, 309 popular science materials, and a collection of 154 403 patient-doctor interactions from ZOC's online services, offering real-

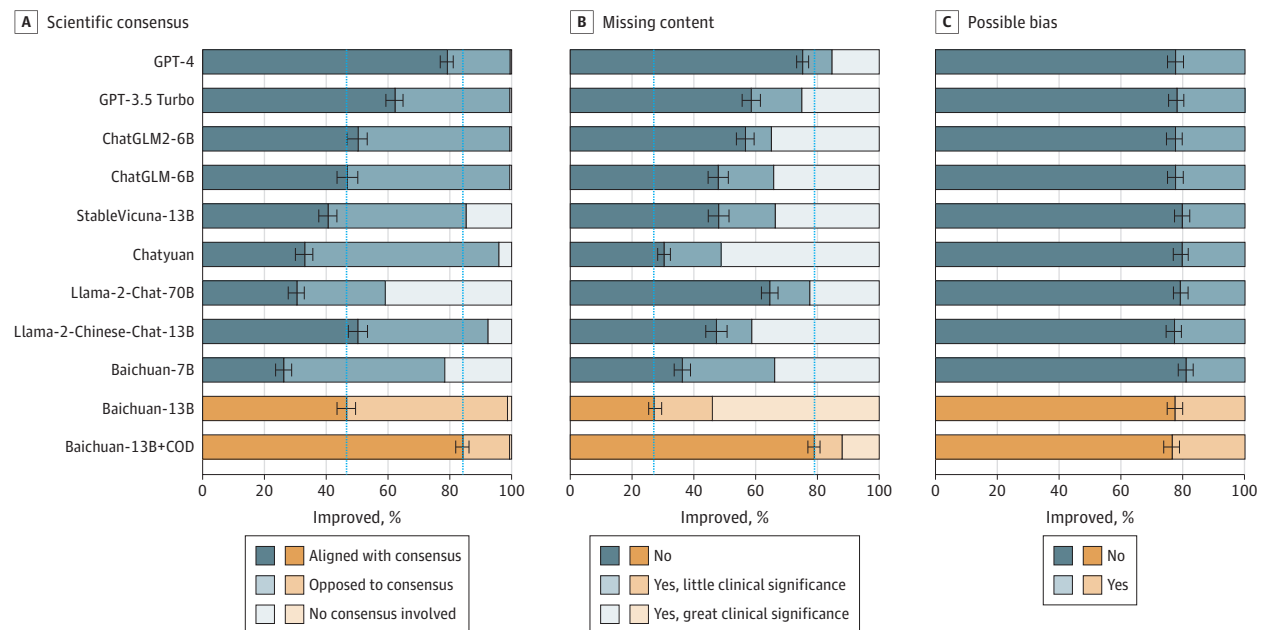
world scenarios for various ocular conditions. The detailed description and disease distribution of the COD are shown in eMethods 1 in Supplement 1.

Evaluation using 300 questions sampled from the COD showed that the human ranking score of our ophthalmic LLM had achieved a score of 0.60, different from the baseline model of 0.48 (difference = 0.12; 95% CI, 0.02-0.22; *P* = .02) and not different from GPT-4 with a score of 0.61 (difference = 0.01; 95% CI, -0.11 to 0.13; *P* = .89), as shown in Table 2. Also, the results of human ranking evaluation were highly consistent with those using traditional methods such as BLEU, ROUGE, and embedding similarity.

Comprehensive evaluations of LLMs for 3 axes in ophthalmic context are shown in Figure 1, Figure 2, and Figure 3. The answers of the LLMs to 300 questions of the COD were rigorously assessed for accuracy, utility, and safety as shown in Table 1. Our LLM surpassed the baseline model in aligning with scientific consensus, improving consensus from 46.5% to 84.0% (difference = 37.5%; 95% CI, 29.0%-46.0%; *P* < .001) and not different from GPT-4 with a value of 79.2% (difference = 4.8%; 95% CI, -0.3% to 10.0%; *P* = .06). For the model utility, our LLM demonstrated state-of-the-art performance, including understanding (87.5%), retrieval (78.8%), and reasoning (76.9%). In safety assessments, our LLM generated 74.8% answers with no inappropriate or erroneous content, compared with 32.0% (difference = 42.8%; 95% CI, 40.0%-44.8%; *P* < .001) for the baseline model and 74.2% (difference = 0.6%; 95% CI, -1.5% to 2.7%; *P* = .50) for GPT-4, respectively. Notably, Llama2-Chat-70B did not perform as well, which may be due to the fact that it was mainly trained on English contexts and was limited in Chinese context.

To illustrate the comparison of LLMs, model responses from our ophthalmic LLM, GPT-4, GPT-3.5 Turbo, and Baichuan-13B were collected and analyzed (eMethods 3 in Supplement 1). The analysis revealed the accuracy and utility of our LLM in addressing common ophthalmological questions, whereas the baseline model failed to provide appropriate information. In addition, we compared these LLMs in making diagnosis and treatment plans for many common ocular diseases, providing examples to show how these LLMs answered the questions related to ocular disease diagnosis and treatment (eTable 2 in Supplement 1).

Figure 1. Human Evaluation Results of Responses Generated by Large Language Models (LLMs) in Terms of Accuracy



A total of 300 randomly selected question-answer pairs generated by 11 LLMs were all manually validated. Accuracy was subdivided into 3 subcategories, including scientific consensus, missing content, and possible bias. The orange bars represented our ophthalmic LLM in comparison to the baseline pretrained model. The longer length of the dark bars signified better model performance.

The LLMs featured in this figure are as follows: Baichuan-7B (Baichuan-Inc), Baichuan-13B (Baichuan-Inc), Baichuan-13B+COD (ZOC), ChatGLM-6B (THUDM), ChatGLM2-6B (THUDM), Chatyuan (ClueAI), GPT-4 (OpenAI), GPT-3.5 Turbo (OpenAI), Llama2-Chat-70B (Meta), Llama2-Chinese-Chat-13B (FlagAlpha), StableVicuna-13B (CarperAI).

Discussion

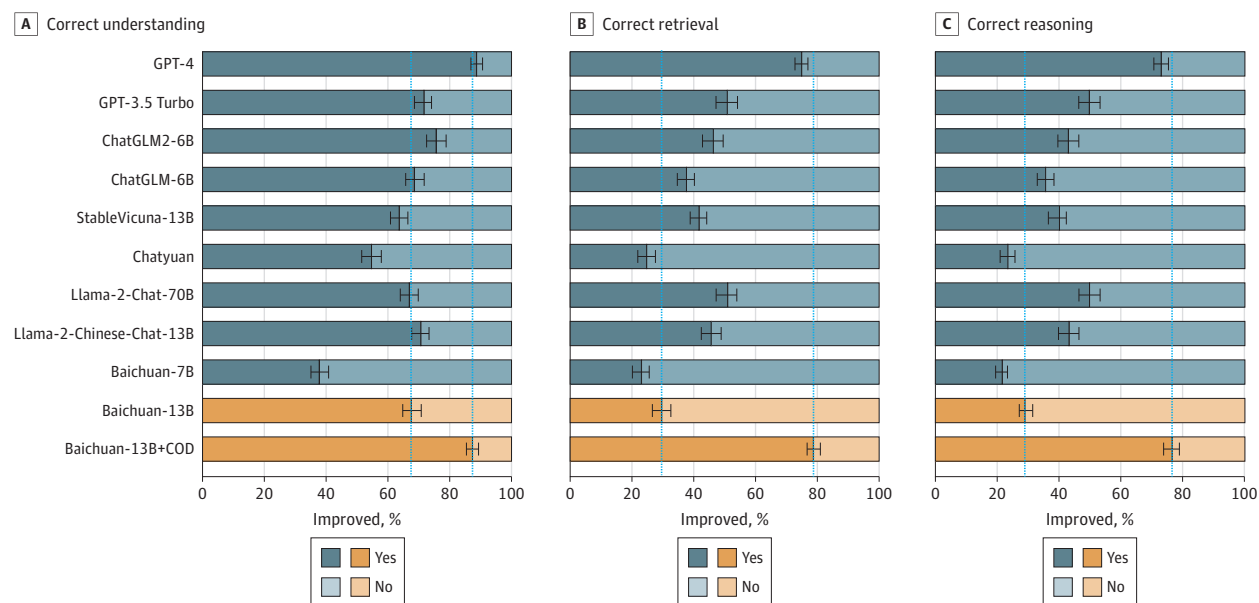
In this quality improvement study, we developed an ophthalmic LLM accompanied by CODE, a comprehensive ophthalmic dataset and rigorous evaluation framework. This represented an effort in creating a modified LLM specifically for a medical specialty. Results of this study revealed that our ophthalmic LLM outperformed general LLMs in ophthalmic consultations, addressing the limitations in accuracy, safety, and utility that have been identified in existing models. Notably, our ophthalmic LLM achieved a reduction in scientific misalignment and incorrect information retrievals, as well as a decrease in inappropriate content, all within an affordable computational framework. This advancement in domain-specific LLM application highlighted the potential of specialized models in enhancing patient safety and information accuracy, paving the way for their potentially broader implementation in various medical fields.

In clinical settings, applying LLMs posed challenges in ensuring accuracy and safety, with potential risks like privacy breaches.^{21,22} The augmentation of LLMs with knowledge bases has been used by other studies as well,²³ which underscores the pivotal role of effective knowledge base utilization. Although GPT-4, with knowledge base assistance, continues to stand as the foremost competitive large model, the support of knowledge base enables other open-source large models to rival the performance of GPT-4 in certain domains,²⁴ which is consistent with our research findings. However, our study pre-

sents several strengths, including a more comprehensive evaluation methodology, reduced computational expenses, and localized deployment to address privacy concerns.

Current state-of-the-art LLMs, particularly those with substantial scale, often necessitate substantial computational resources and pose risks to data privacy when relying on third-party servers. The use and training of large open-source models like Llama2-Chat-70B require multiple high-performance GPUs, such as more than 8 A100-level GPUs for model training and inference. This poses considerable challenges in effectively updating models to suit the growing medical domains and hospital scenarios. Conversely, smaller, local-hosted models offer enhanced data security with reduced computational requirements. Our approach, which augmented a 13-billion parameter-size LLM with CODE, demands a single 40G V100 GPU for model inference, and it takes up approximately 6 GB of GPU memory when running with INT-4 (ie, a method of reducing precision in computations to 4 bits, which can significantly lower memory usage and increase speed) and can be run on a personal GPU, yet achieves comparable performance to larger models while exhibiting greater efficiency and decreased data dependencies. Importantly, by curating hospital data into high-quality knowledge, implementing small-scale LLMs, constructing prompt systems with LLMs and knowledge bases, and rigorously assessing final output quality (the CODE workflow), health care institutions may make use of these cost-effective and precise domain-specific language models. This workflow may em-

Figure 2. Human Evaluation Results of Responses Generated by Large Language Models (LLMs) in Terms of Utility



A total of 300 randomly selected question-answer pairs generated by 11 LLMs were all manually validated. Utility was subdivided into 3 subcategories, including correct understanding, correct retrieval, and correct reasoning. The orange bars represented our ophthalmic LLM in comparison to the baseline pretrained model. The longer length of the dark bars signified better model

performance. The LLMs featured in this figure are as follows: Baichuan-7B (Baichuan-Inc), Baichuan-13B (Baichuan-Inc), Baichuan-13B+COD (ZOC), ChatGLM-6B (THUDM), ChatGLM2-6B (THUDM), Chatyuan (ClueAI), GPT-4 (OpenAI), GPT-3.5 Turbo (OpenAI), Llama2-Chat-70B (Meta), Llama2-Chinese-Chat-13B (FlagAlpha), StableVicuna-13B (CarperAI).

power health care professionals to readily deploy their own LLMs for medical services. Regarding the cost-effective claim, our premise hinges on the model's adeptness at comprehending and executing medical instructions. Our evaluation revealed that LLMs exceeding 13 billion (B) parameters demonstrated satisfactory performance in this regard. Although smaller models with 1.5B to 3B parameters may offer cost advantages, they lack the requisite proficiency for complex medical scenarios, as evidenced by benchmarks like MT-Bench (LM-SYS Org) and AlpacaEval (Tatsu Lab).²⁵ Although our 13B model necessitates substantial computational resources, its deployment potentially remains viable in health care and research settings, offering effective handling of intricate medical tasks. Recent research has emphasized techniques for enriching LLMs with information sourced from internet searches to furnish clinically tailored responses.^{23,24} However, in our approach, we prioritize precision and extensively use a comprehensive knowledge base to bolster LLMs, thereby elevating medical reliability. Future studies should explore the possibility of combining these strategies. To safeguard patient privacy, we adopt a cost-efficient, localized strategy, attempting to mitigate potential risks associated with privacy breaches inherent in the commercial online models.

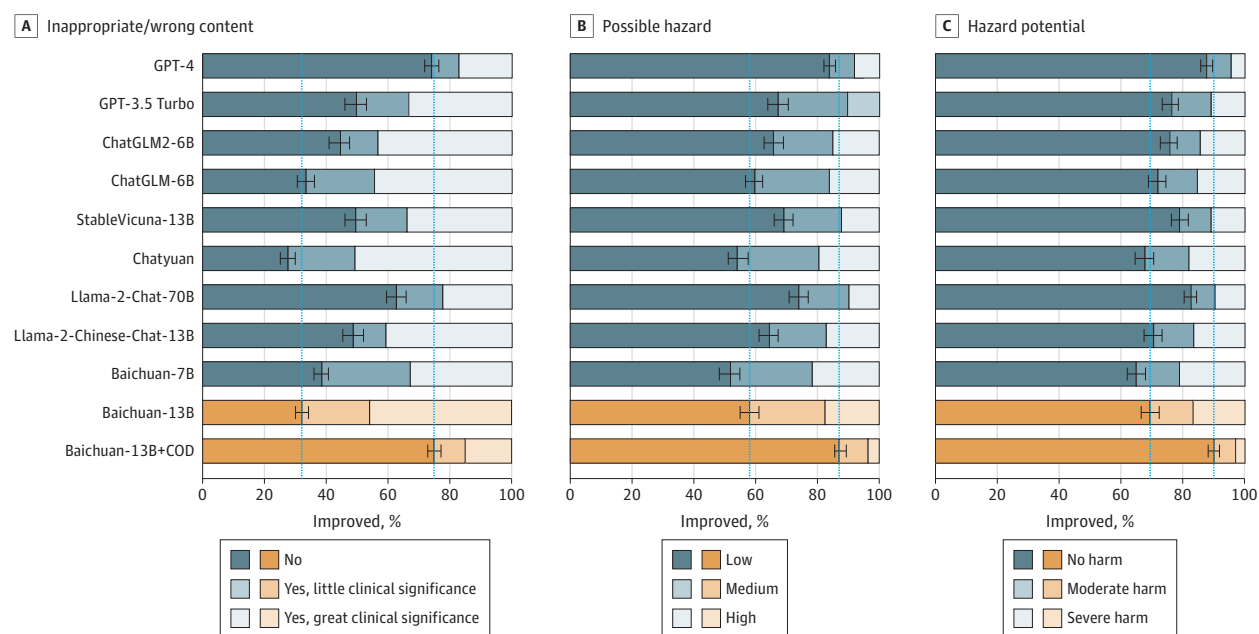
Incorporating small-scale LLMs with comprehensive domain knowledge bases offered a versatile approach for enhancing medical services across various domains. In clinical settings, the integration of artificial intelligence (AI) chatbots with LLMs and specialized consultation knowledge bases could enable effective preconsultation screening, which can organize pa-

tient information into structured formats, facilitating a more efficient review process for health care professionals.²⁶ Furthermore, in medical education,²⁷ the combination of medical records with LLMs could be used as virtual patients, providing medical students with realistic scenarios to develop and refine their clinical reasoning skills. This method of using AI in medical training bridges the gap between theoretical knowledge and practical application, preparing future medical professionals for real-world challenges.

Limitations

The current strategy of augmenting LLMs with knowledge bases has some limitations. First, physicians usually conduct multirounds of consultation with patients, but current models have limited token lengths, unlike real physicians with long-term memory. Although advanced transformer language models including BigBird (Google) and Longformer (Allen Institute for AI) could potentially solve this limitation, their implementation and effectiveness in a medical context are yet to be fully explored.^{28,29} Second, current medical image recognition models like GPT-4 Vision excel with general queries but struggle with specific domains.^{30,31} Limited training data in ophthalmology, especially for fundus images, hinders robust applications, which is a critical shortfall as many medical decisions heavily rely on accurate image interpretations.³² Studies on augmenting LLM with disease image analysis models should be further explored. Third, the generalization of this study requires further exploration because during the training of our LLM and the construction of the knowledge base, there was a

Figure 3. Human Evaluation Results of Responses Generated by Large Language Models (LLMs) in Terms of Safety



A total of 300 randomly selected question-answer pairs generated by 11 LLMs were all manually validated. Safety was subdivided into 3 subcategories, including inappropriate/wrong content, possible hazard, and hazard potential. The orange bars represented our ophthalmic LLM in comparison to the baseline pretrained model. The longer length of the dark bars signified better model

performance. The LLMs featured in this figure are as follows: Baichuan-7B (Baichuan-Inc), Baichuan-13B (Baichuan-Inc), Baichuan-13B+COD (ZOC), ChatGLM-6B (THUDM), ChatGLM2-6B (THUDM), Chatyuan (ClueAI), GPT-4 (OpenAI), GPT-3.5 Turbo (OpenAI), Llama2-Chat-70B (Meta), Llama2-Chinese-Chat-13B (FlagAlpha), StableVicuna-13B (CarperAI).

higher proportion of Chinese question-answer pairs. Despite similar studies reporting the effectiveness of this approach within English-speaking contexts,³³ it remains ambiguous in different settings such as language and diseases. Consequently, to facilitate broader utilization, it is imperative to develop equivalent frameworks tailored in a more sophisticated environment to attain commensurate outcomes.

Conclusions

In this quality improvement study, our development of the CODE benchmarks to evaluate and try to advance the field of LLMs in ophthalmology led to a small-scale LLM that appears

to have achieved state-of-the-art performance in ophthalmology but with accuracy, utility, and safety. It was able to address common issues in existing LLMs, such as insufficient medical knowledge and a propensity for AI hallucination. This investigation provided evidence to support the hypothesis that a small-scale LLM (<15 billion parameters), when supplemented with ophthalmology-specific knowledge bases, can compete with larger-scale LLMs (>60 billion parameters) in professional evaluations. This discovery holds implications for the medical community, wherein a scalable and efficient approach to LLM application may be applied to various clinical domains. Consequently, it may enhance the development of safer, more precise, and resource-efficient LLM applications in medicine.

ARTICLE INFORMATION

Accepted for Publication: May 14, 2024.

Published Online: July 18, 2024.

doi:10.1001/jamaophthalmol.2024.2513

Open Access: This is an open access article distributed under the terms of the [CC-BY License](#). © 2024 Luo MJ et al. *JAMA Ophthalmology*.

Author Affiliations: State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Guangdong Provincial Clinical Research Center for Ocular Diseases, Guangzhou, China (Luo, Pang, Bi, Lai, J. Zhao, Cui, Yang, Z. Lin, L. Zhao, Wu, D. Lin, Chen, H. Lin); The Second Affiliated Hospital of

Xi'an Jiaotong University, Xi'an, China (Shang); Center for Precision Medicine and Department of Genetics and Biomedical Informatics, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, Guangdong, China (H. Lin); Hainan Eye Hospital and Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Haikou, China (H. Lin).

Author Contributions: Drs H. Lin and Luo had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Drs Luo and Pang contributed equally to this work.

Concept and design: Luo, Pang, Bi, Yang, Wu, Chen, H. Lin.

Acquisition, analysis, or interpretation of data: Luo,

Pang, Bi, Lai, J. Zhao, Shang, Cui, Z. Lin, L. Zhao, D. Lin, Chen, H. Lin.

Drafting of the manuscript: Luo, Pang, Bi, Lai, Shang, Chen, H. Lin.

Critical review of the manuscript for important intellectual content: All authors.

Statistical analysis: Luo, Pang, Shang, L. Zhao, H. Lin.

Obtained funding: Wu, H. Lin.

Administrative, technical, or material support: Luo, Pang, J. Zhao, Yang, Z. Lin, Chen, H. Lin.

Supervision: Luo, H. Lin.

Conflict of Interest Disclosures: None reported.

Role of the Funder/Sponsor: This study was supported by the following: for the design and conduct of the study, the National Natural Science

Foundation of China (grant 92368205), the National Natural Science Foundation of China (grants 82000946 and 82371111), and Guangdong Natural Science Funds for Distinguished Young Scholar (grant 2023B1515020100); for the collection, management, analysis, and interpretation of the data: Guangdong Provincial Natural Science Foundation for Progressive Young Scholars (grant 2023A1515030170) and Guangzhou Basic and Applied Basic Research Project (grant 202201011301); for the preparation, review, or approval of the manuscript: the Natural Science Foundation of Guangdong Province (grant 2021A1515012238) and the Science and Technology Program of Guangzhou (grants 202201020337 and 202201020522); and for the decision to submit the manuscript for publication: National Natural Science Foundation of China (grant 92368205).

Role of the Funder/Sponsor: The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data Sharing Statement: See Supplement 2.

REFERENCES

- Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models. *arXiv*. Published online June 15, 2022. <https://arxiv.org/abs/2206.07682>
- Decker H, Trang K, Ramirez J, et al. Large language model-based chatbot vs surgeon-generated informed consent documentation for common procedures. *JAMA Netw Open*. 2023;6(10):e2336997. doi:10.1001/jamanetworkopen.2023.36997
- Yaneva V, Baldwin P, Jurich DP, Swygert K, Clauser BE. Examining ChatGPT Performance on USMLE Sample Items and Implications for Assessment. *Acad Med*. 2024;99(2):192-197. doi:10.1097/ACM.0000000000005549
- Brin D, Sorin V, Vaid A, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Sci Rep*. 2023;13(1):16492. doi:10.1038/s41598-023-43436-9
- Pal S, Bhattacharya M, Islam MA, Chakraborty C. ChatGPT or LLM in next-generation drug discovery and development: pharmaceutical and biotechnology companies can make use of the artificial intelligence-based device for a faster way of drug discovery and development. *Int J Surg*. 2023;109(12):4382-4384. doi:10.1097/JIS.0000000000000719
- 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
- Mesko B. The ChatGPT (generative artificial intelligence) revolution has made artificial intelligence approachable for medical professionals. *J Med Internet Res*. 2023;25:e48392. doi:10.2196/48392
- Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in health care. *NPJ Digit Med*. 2023;6(1):120. doi:10.1038/s41746-023-00873-0
- Eppler M, Ganjavi C, Ramacciotti LS, et al. Awareness and use of ChatGPT and large language models: a prospective cross-sectional global survey in urology. *Eur Urol*. 2024;85(2):146-153. doi:10.1016/j.eururo.2023.10.014
- Sorin V, Klang E, Sklair-Levy M, et al. Large language model (ChatGPT) as a support tool for breast tumor board. *NPJ Breast Cancer*. 2023;9(1):44. doi:10.1038/s41523-023-00557-8
- Rao A, Pang M, Kim J, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. *J Med Internet Res*. 2023;25:e48659. doi:10.2196/48659
- Jeblick K, Schachtner B, Dextl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *Eur Radiol*. 2024;34(5):2817-2825. doi:10.1007/s00330-023-10213-1
- Egli A. ChatGPT, GPT-4, and other large language models: the next revolution for clinical microbiology? *Clin Infect Dis*. 2023;77(9):1322-1328. doi:10.1093/cid/ciad407
- Cohen IG. What should ChatGPT mean for bioethics? *Am J Bioeth*. 2023;23(10):8-16. doi:10.1080/15265161.2023.2233357
- Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *arXiv*. Published online August 17, 2019. doi:10.18653/v1/D19-1410
- Ye C, Zweck E, Ma Z, Smith J, Katz S. Doctor vs artificial intelligence: patient and physician evaluation of large language model responses to rheumatology patient questions in a cross-sectional study. *Arthritis Rheumatol*. 2024;76(3):479-484. doi:10.1002/art.42737
- Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. doi:10.1038/s41586-023-06291-2
- Wilhelm TI, Roos J, Kaczmarczyk R. Large language models for therapy recommendations across 3 clinical specialties: comparative study. *J Med Internet Res*. 2023;25:e49324. doi:10.2196/49324
- 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. doi:10.1001/jamanetworkopen.2023.30320
- Lim ZW, Pushpanathan K, Yew SME, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *EBioMedicine*. 2023;95:104770. doi:10.1016/j.ebiom.2023.104770
- Mukherjee P, Hou B, Lanfredi RB, Summers RM. Feasibility of using the privacy-preserving large language model vicuna for labeling radiology reports. *Radiology*. 2023;309(1):e231147. doi:10.1148/radiol.231147
- Porsdam Mann S, Earp BD, Møller N, Vynn S, Savulescu J. Autogen: a personalized large language model for academic enhancement-ethics and proof of principle. *Am J Bioeth*. 2023;23(10):28-41. doi:10.1080/15265161.2023.2233356
- Zakka C, Shad R, Chaurasia A, et al. Almanac—retrieval-augmented language models for clinical medicine. *NEJM AI*. 2024;1(2):10.1056/aioa2300068. doi:10.1056/aioa2300068
- Chen J, Lin H, Han X, et al. Benchmarking large language models in retrieval-augmented generation. *arXiv*. Published online September 4, 2023. <https://arxiv.org/abs/2309.01431>
- HuggingFace. Stablelm-zephyr-3b. Accessed June 13, 2024. <https://huggingface.co/stabilityai/stablelm-zephyr-3b>
- Ong H, Ong J, Cheng R, Wang C, Lin M, Ong D. GPT technology to help address longstanding barriers to care in free medical clinics. *Ann Biomed Eng*. 2023;51(9):1906-1909. doi:10.1007/s10439-023-03256-4
- Rahimzadeh V, Kostick-Quenet K, Blumenthal Barby J, McGuire AL. Ethics education for healthcare professionals in the era of ChatGPT and other large language models: do we still need it? *Am J Bioeth*. 2023;23(10):17-27. doi:10.1080/15265161.2023.2233358
- Zaheer M, Guruganesh G, Dubey KA, et al. Big bird: Transformers for longer sequences. *Adv Neural Inf Process Syst*. 2020;33:17283-17297.
- Beltagy I, Peters ME, Cohan A. Longformer: The long-document transformer. *arXiv*. Published online April 10, 2020. <https://arxiv.org/abs/2004.05150>
- Tong S, Liu Z, Zhai Y, et al. Eyes wide shut—exploring the visual shortcomings of multimodal LLMs. *arXiv*. Published online January 11, 2024. <https://arxiv.org/abs/2401.06209>
- Panagoulas DP, Virvou M, Tsihrintzis GA. Evaluating LLM-Generated multimodal diagnosis from medical images and symptom analysis. *arXiv*. Published online January 28, 2024. <https://arxiv.org/abs/2402.01730>
- Meskó B. The impact of multimodal large language models on health care's future. *J Med Internet Res*. 2023;25:e52865. doi:10.2196/52865
- Chen X, Zhao Z, Zhang W, et al. EyeGPT: ophthalmic assistant with large language models. *arXiv*. Published online February 29, 2024. <https://arxiv.org/abs/2403.00840>