Original Paper

Performance of ChatGPT in Ophthalmic Registration and Clinical Diagnosis: Cross-Sectional Study

Shuai Ming[1,2,3*], MD; Xi Yao[1], MD; Xiaohong Guo[1], MD; Qingge Guo[1,2,3], MD; Kunpeng Xie[4], MD; Dandan Chen[1,2,3], MD, PhD; Bo Lei[1,2,3*], MD, PhD

[1]Department of Ophthalmology, Henan Eye Institute, Henan Eye Hospital, Henan Provincial People's Hospital, Zhengzhou, China

[2]Eye Institute, Henan Academy of Innovations in Medical Science, Zhengzhou, China

[3]Henan Clinical Research Center for Ocular Diseases, People's Hospital of Zhengzhou University, Zhengzhou, China

[4]Department of Ophthalmology, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China

[*]these authors contributed equally

Corresponding Author:
Bo Lei, MD, PhD
Department of Ophthalmology
Henan Eye Institute, Henan Eye Hospital
Henan Provincial People's Hospital
No.7 Weiwu Road
Zhengzhou
China
Phone: 86 037167120925
Email: bolei99@126.com

Abstract

Background: Artificial intelligence (AI) chatbots such as ChatGPT are expected to impact vision health care significantly. Their potential to optimize the consultation process and diagnostic capabilities across range of ophthalmic subspecialties have yet to be fully explored.

Objective: This study aims to investigate the performance of AI chatbots in recommending ophthalmic outpatient registration and diagnosing eye diseases within clinical case profiles.

Methods: This cross-sectional study used clinical cases from *Chinese Standardized Resident Training–Ophthalmology (2nd Edition)*. For each case, 2 profiles were created: patient with history (Hx) and patient with history and examination (Hx+Ex). These profiles served as independent queries for GPT-3.5 and GPT-4.0 (accessed from March 5 to 18, 2024). Similarly, 3 ophthalmic residents were posed the same profiles in a questionnaire format. The accuracy of recommending ophthalmic subspecialty registration was primarily evaluated using Hx profiles. The accuracy of the top-ranked diagnosis and the accuracy of the diagnosis within the top 3 suggestions (do-not-miss diagnosis) were assessed using Hx+Ex profiles. The gold standard for judgment was the published, official diagnosis. Characteristics of incorrect diagnoses by ChatGPT were also analyzed.

Results: A total of 208 clinical profiles from 12 ophthalmic subspecialties were analyzed (104 Hx and 104 Hx+Ex profiles). For Hx profiles, GPT-3.5, GPT-4.0, and residents showed comparable accuracy in registration suggestions (66/104, 63.5%; 81/104, 77.9%; and 72/104, 69.2%, respectively; *P*=.07), with *ocular trauma*, *retinal diseases*, and *strabismus and amblyopia* achieving the top 3 accuracies. For Hx+Ex profiles, both GPT-4.0 and residents demonstrated higher diagnostic accuracy than GPT-3.5 (62/104, 59.6% and 63/104, 60.6% vs 41/104, 39.4%; *P*=.003 and *P*=.001, respectively). Accuracy for do-not-miss diagnoses also improved (79/104, 76% and 68/104, 65.4% vs 51/104, 49%; *P*<.001 and *P*=.02, respectively). The highest diagnostic accuracies were observed in *glaucoma*; *lens diseases*; and *eyelid, lacrimal, and orbital diseases*. GPT-4.0 recorded fewer incorrect top-3 diagnoses (25/42, 60% vs 53/63, 84%; *P*=.005) and more partially correct diagnoses (21/42, 50% vs 7/63 11%; *P*<.001) than GPT-3.5, while GPT-3.5 had more completely incorrect (27/63, 43% vs 7/42, 17%; *P*=.005) and less precise diagnoses (22/63, 35% vs 5/42, 12%; *P*=.009).

Conclusions: GPT-3.5 and GPT-4.0 showed intermediate performance in recommending ophthalmic subspecialties for registration. While GPT-3.5 underperformed, GPT-4.0 approached and numerically surpassed residents in differential diagnosis. AI chatbots show promise in facilitating ophthalmic patient registration. However, their integration into diagnostic decision-making requires more validation.

(J Med Internet Res 2024;26:e60226) doi: 10.2196/60226

KEYWORDS

artificial intelligence; chatbot; ChatGPT; ophthalmic registration; clinical diagnosis; AI; cross-sectional study; eye disease; eye disorder; ophthalmology; health care; outpatient registration; clinical; decision-making; generative AI; vision impairment

Introduction

Artificial intelligence (AI) has significantly advanced in health care, particularly in many areas of ophthalmology [1,2]. ChatGPT (OpenAI) [3] is a generative AI featuring a chatbot interface. Benefiting from its expansive knowledge base and complex parameterization, it enables users to input queries and receive responses that showcase advanced, humanlike logic. Since its launch in November 2022, ChatGPT has quickly amassed a substantial user base. It was recognized as having the potential to revolutionize not only ophthalmology but also the entire medical field in diverse aspects [4], including patient care, health care professionals and systems, research, and education and training [5]. Its performance was highlighted in patient triage proficiency [6], scientific writing [7], operative notes writing [8], and passing the ophthalmology specialist licensing examination [9].

Specialized eye hospitals in China, especially tertiary ones, frequently face patient overcapacity. With limited knowledge of eye health, patients could encounter difficulties in choosing the right subspecialty department when registering. User-friendly chatbot such as ChatGPT could provide registration suggestions based on the patients' chief complaints and medical histories and, thus, significantly ease these challenges and reduce health care resource wastage due to unsuitable registrations. However, the role of ChatGPT in classifying diseases into ophthalmic subspecialties and thus facilitating patient registration remains unexplored.

ChatGPT has exhibited encouraging results in diagnosing eye diseases within specific subspecialties, such as corneal and retinal vascular diseases [10,11]. The data for these assessments were sourced from public question banks and case report databases. In diagnosing a diverse range of ophthalmic conditions, ChatGPT failed to match the diagnostic accuracy of ophthalmologists but demonstrated the benefit of a shorter diagnostic time [12]. Further validation studies, particularly in testing ChatGPT's diagnostic effectiveness for a comprehensive range of eye diseases within the context of clinical practice in China, are essential.

Drawing on typical clinical cases from Chinese Standardized Resident Training (SRT) materials, this study aims to evaluate ChatGPT's capacity for classifying ophthalmic subspecialties and its diagnostic potential within the Chinese context. Our research seeks to provide insights into whether ChatGPT can effectively assist patients with appropriate registrations and support ophthalmologists in clinical decision-making.

Methods

Ethical Considerations

The Institutional Review Board of Henan Provincial People's Hospital determined that this in silicon research did not involve direct interaction with real-world human subjects, nor did it require the collection of new human data. Accordingly, an ethics exemption was granted for this study. The case information used was derived from publicly available sources and published materials. At the time of data collection, GPT-4.0 was publicly available by paid subscription through ChatGPT Plus.

Data Source

The clinical cases for our study were sourced from *Chinese Standardized Resident Training–Ophthalmology (2nd Edition)*, which is an official resource conforming to SRT content and standards, as well as the theoretical assessment guidelines of the National Health Commission of China. Unlike traditional undergraduate textbooks, SRT materials are specifically designed to focus on problem-based learning and case-based learning. They feature a variety of typical real-world cases from various ophthalmic subspecialties, making them particularly suitable for interaction with ChatGPT's chatbot interface, which is designed to handle complex, real-life queries.

In this study, we gathered 121 cases from 12 ophthalmic subspecialties, creating 2 profiles per case: patient with history (Hx) and patient with history and examination (Hx+Ex) [12]. The "history" portion comprised gender, age, chief complaints, and medical history. When necessary, past medical, familial, and systemic disease details were also added. The "examination" portion covered general ophthalmic assessments such as visual acuity and intraocular pressure, alongside diagnostics such as slit lamp biomicroscopy, a range of ophthalmic imaging (eg, fundus photography, orbital computed tomography [OCT], and fluorescein angiography), and specialized imaging for eye tumors and traumas (OCT and magnetic resonance imaging). The inclusion criteria mandated comprehensive historical and ophthalmology-related chief complaint details, documented examination results, and official case analyses with accurate subspecialty classifications and unique diagnoses.

After excluding 17 cases, a total of 104 cases were retained for further analysis. The reasons for excluding the 17 cases included (1) the lack of medical history or the presence of final diagnostic–like terms in the medical history, which may bias the assessment (6 cases); (2) the lack of textual descriptions of examination results (6 cases); (3) chief complaints not related to ophthalmic symptoms or cases referred from other departments (3 cases); and (4) unclear or non–ophthalmology-related diagnoses (2 cases). The mean age was 36.8 (SD 21.7) years, with male patients comprising 55.8% (58/104) of the cases. Notably, each case had a predominantly unique final diagnosis. The 3 most prevalent diagnoses,

classified into subspecialties, were *eyelid, lacrimal, and orbital diseases*; *retinal diseases*; and *strabismus and amblyopia* (Table 1). Given that the original 104 profiles were ordered based on the ophthalmic subspecialty, a new case numbering system was established by randomly assigning descending numerical values between 0 and 1 and arranging them accordingly.

Table 1. Classification of subspecialties in the 104 clinical cases.

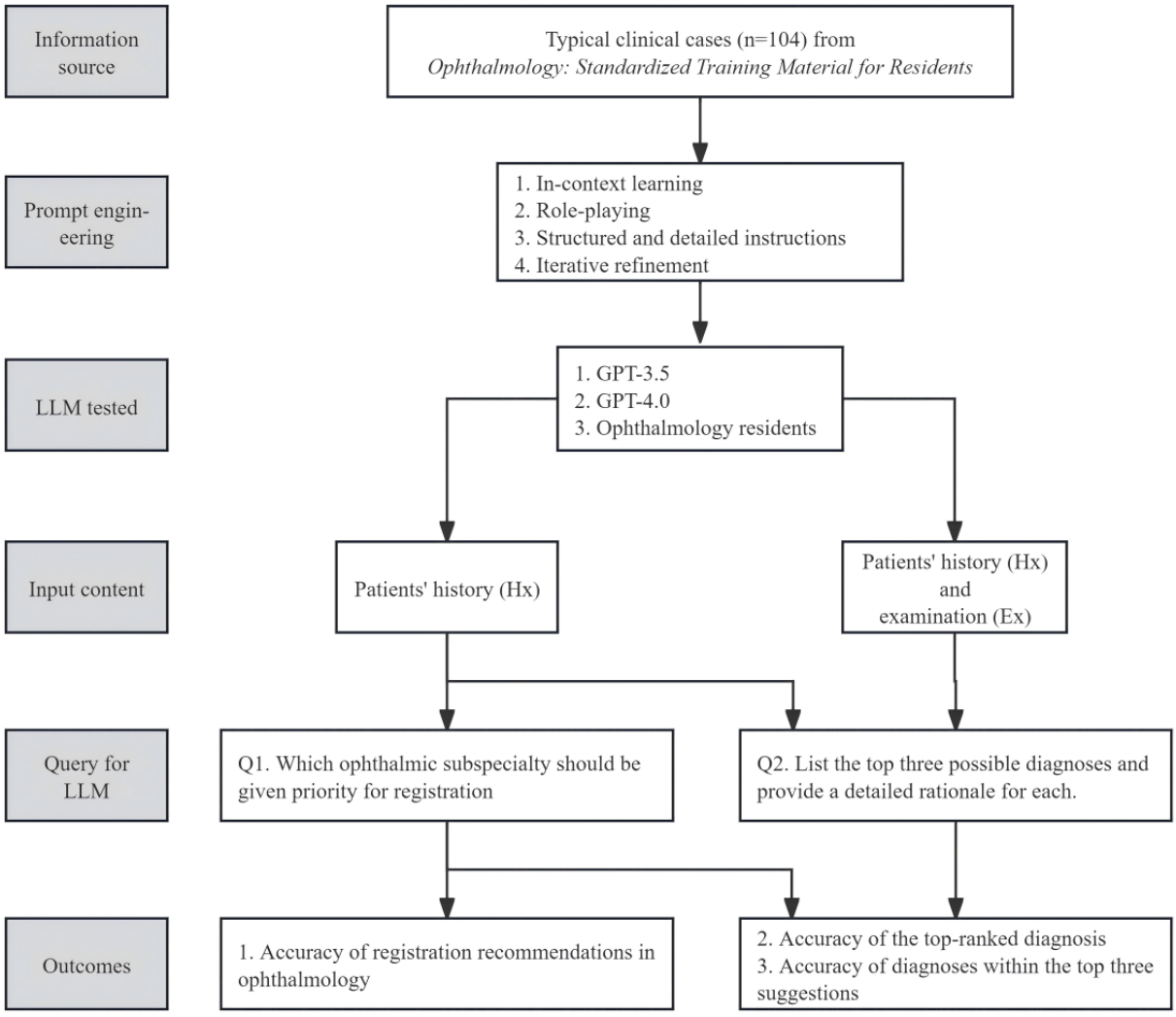
Ophthalmic subspecialty	Clinical cases (n=104), n (%)
Eyelid, lacrimal, and orbital diseases	18 (17)
Retinal diseases	
Nonheritable	15 (13)
Heritable	6 (6)
Strabismus and amblyopia	10 (10)
Corneal and ocular surface diseases	9 (9)
Refractive errors	7 (7)
Scleral and uveal diseases	6 (6)
Ocular trauma	6 (6)
Eye tumors	
Eyelid, lacrimal, and orbital tumors	6 (6)
Scleral and uveal tumors	4 94)
Retinal tumors	1 (1)
Glaucoma	5 (5)
Vitreous diseases	3 (4)
Lens diseases	4 (4)
Neuro-ophthalmology	4 (4)
Total	104 (100)

Testing Process

A total of 208 clinical profiles (104 Hx and 104 Hx+Ex profiles) in Chinese were tested from March 5 to 18, 2024. The tested AI chatbots included ChatGPT versions 3.5 and 4.0 (GPT-3.5 and GPT-4.0, respectively), and 3 ophthalmology residents were also tested. Initially, ChatGPT was assigned a system role to emulate a professional ophthalmologist. Each clinical case scenario was entered in Chinese independently, followed by 2 questions: “Q1. Which ophthalmic subspecialty should be given priority for registration?” and “Q2: List the top three possible diagnoses and provide a detailed rationale for each.” Q1 aimed to elicit the chatbot’s triage recommendations for subspecialty registration, while Q2 focused on extracting the chatbot’s leading and differential diagnosis proposals. For the Hx profiles, both questions were asked, and the AI was informed of the available subspecialties for reference in the prompt. For the Hx+Ex profiles, given that the examination primarily serves for diagnosis, only Q2 was posed. Response history was reset prior to each new query (Figure 1). The detailed prompts and engineering techniques are shown in Multimedia Appendix 1. Based on the structure provided by these prompts, we found that ChatGPT’s responses generally adhered to the fixed template specified in our prompts. Consequently, we opted to input each case continuously in a single chat session. However, for the Hx and Hx+Ex profiles, as well as for tests conducted with both GPT-3.5 and GPT-4.0, each series of tests was initiated in a new chat session.

For the residents’ test, 2 sets of questionnaires were created following the Hx + Q1 and Hx+Ex + Q2 format. The residents’ evaluations were independent, which were ensured by implementing a blinded assessment process where the residents did not have information about the performance or responses of the AI systems or each other.

Figure 1. The design and analytical framework of the study. LLM: large language models.



Outcomes and Definition

The study focused on 3 outcomes: accuracy of recommendation for ophthalmic subspecialty registration, accuracy of the top-ranked diagnosis, and accuracy of the diagnosis within the top 3 suggestions (do-not-miss diagnosis). As residents rarely provided 3 possible diagnoses like ChatGPT, they were not evaluated on the accuracy of the do-not-miss diagnosis. The gold standard for judgment was the official diagnosis from *Chinese Standardized Resident Training–Ophthalmology (2nd Edition).*

ChatGPT and residents were evaluated based on the same outcome criteria. For ophthalmic subspecialty registration, overlaps exist in some subspecialties, such as *eye tumors* and *retinal diseases*. Recommendations to either category were considered correct. Precision was crucial for diagnosis suggestions corresponding to Hx+Ex profiles. For instance, if the final diagnosis is sympathetic ophthalmia, responses such as panuveitis or herpetic keratitis were marked as incorrect. Similarly, for acute idiopathic optic neuritis or orbital neurilemmoma, responses of optic neuritis or orbital tumor were also considered incorrect. However, for Hx profiles, an exact diagnosis was not required. All responses in the examples provided were considered correct.

Regarding the residents' performance, a diagnosis was considered correct only if at least 2 out of 3 residents provided the correct diagnosis. This approach emphasizes the importance of consensus in clinical decision-making and reflects the collaborative nature of medical diagnosis in real-world settings.

Statistical Analysis

Data collection and management were performed using Microsoft Excel software. Statistical analyses were mainly conducted in SPSS (version 26.0.0; IBM Corp). To compare the accuracy of triage and diagnosis across different testing strategies, the Pearson chi-square test or Fisher exact test was applied, depending on the expected frequency counts in the contingency tables. Due to limitations in SPSS for conducting the Fisher exact test on the differences in proportions among 3 groups, we switched to R software (version 4.4.1; R Foundation for Statistical Computing) for this analysis. For post hoc analysis, *P* values were adjusted using the Bonferroni method in pairwise comparisons. Unless otherwise specified, differences were considered statistically significant at $P<.05$.

Results

Accuracy of Recommendation for Subspecialty Registration

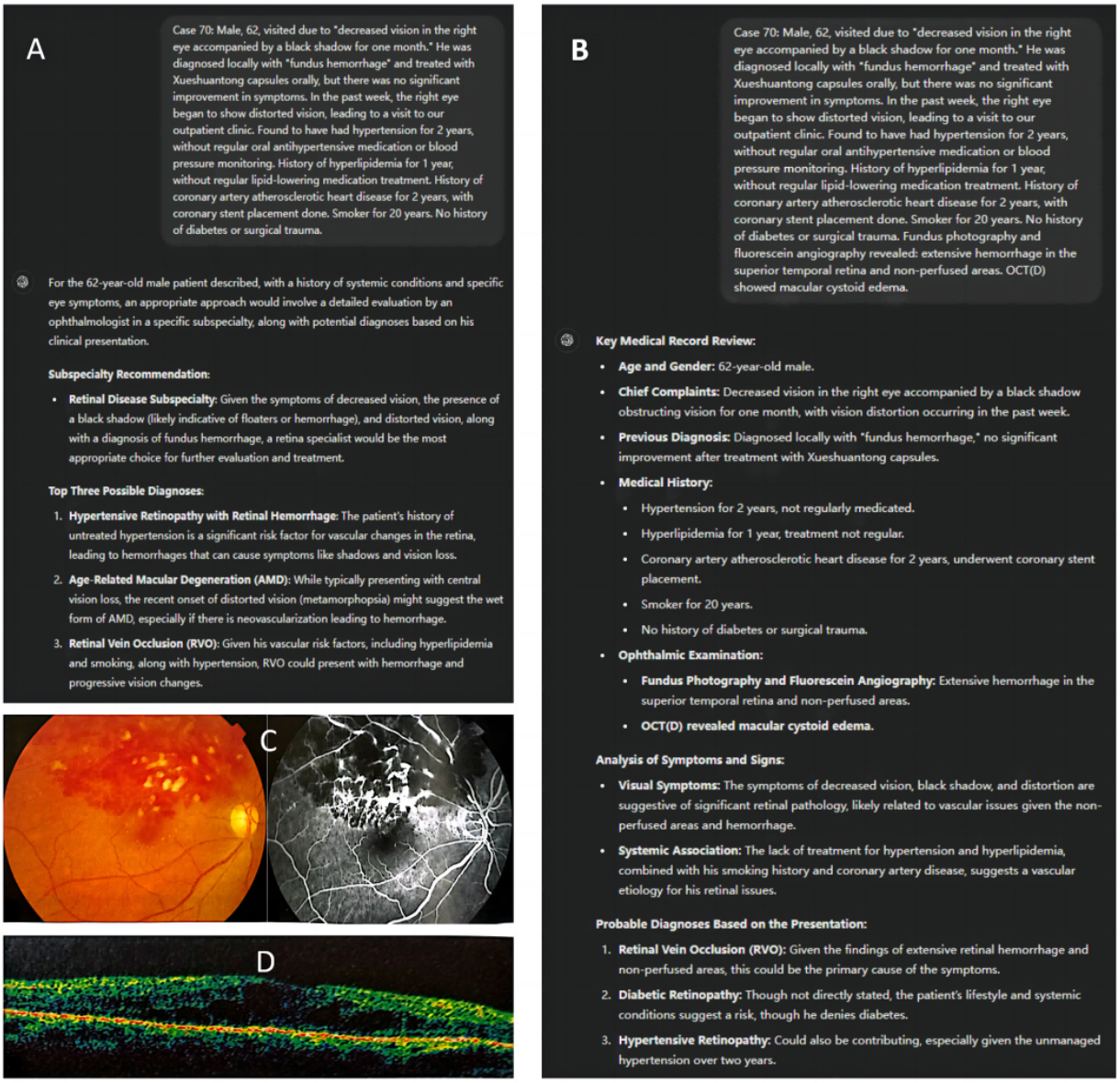
For Hx profiles, GPT-3.5, GPT-4.0, and residents demonstrated moderate accuracy in registering patients to the correct ophthalmic subspecialty, with accuracy of 63.5% (66/104), 77.9% (81/104), and 69.2% (72/104), respectively (*P*=.07). Subgroup analysis revealed that *ocular trauma*, *retinal diseases*, and *strabismus and amblyopia* ranked among the top 3 in overall registration accuracy. In contrast, registration accuracy was low for *glaucoma*, *neuro-ophthalmology*, and *lens diseases* (Table 2). The detailed registration recommendations for GPT-3.5, GPT-4.0, and residents are showed in Multimedia Appendix 2. An example of GPT-4.0's response to subspecialty registration is shown in Figure 2.

Table 2. Correct triage recommendations for subspecialty registration for patient history (Hx) profiles.

Ophthalmic subspecialty	GPT-3.5, n (%)	GPT-4.0, n (%)	Residents, n (%)	*P* value
Total (n=104)	66 (63.5)	81 (77.9)	72 (69.2)	.07
Retinal diseases (n=21)	18 (86)	20 (95)	19 (91)	.86
Eyelid, lacrimal, and orbital diseases (n=18)	14 (78)	14 (78)	11 (61)	.44
Eye tumors (n=11)	6 (55)	8 (73)	9 (82)	.52
Strabismus and amblyopia (n=10)	8 (80)	9 (90)	7 (70)	.85
Corneal and ocular surface diseases (n=9)	4 (44)	7 (78)	8 (89)	.10
Refractive errors (n=7)	6 (86)	5 (71)	4 (57)	.83
Ocular trauma (n=6)	6 (100)	6 (100)	6 (100)	—[a]
Scleral and uveal diseases (n=6)	1 (17)	4 (67)	3 (50)	.36
Glaucoma (n=5)	1 (20)	3 (60)	1 (20)	.50
Lens diseases (n=4)	1 (25)	2 (50)	1 (25)	.69
Neuro-ophthalmology (n=4)	1 (25)	1 (25)	1 (25)	>.99
Vitreous diseases (n=3)	0 (0)	2 (67)	2 (67)	.36

[a]Not applicable.

Figure 2. Example profile 70: interaction with and responses of the GPT-4.0 chatbot. (A) When provided with Hx information, GPT-4.0 correctly recommended the ophthalmic subspecialty of “retinal diseases.” (B) When provided with Hx+Ex information, GPT-4.0 gave the top 3 diagnostic suggestions and accurately identified “retinal vein occlusion” as the top-ranked diagnosis. (C) Fundus photography and fluorescein angiography. (D) OCT imaging. (C) and (D) were presented to ChatGPT as textual descriptions of the examination results. Hx: patient with history; Hx+Ex: patient with history and examination; OCT: orbital computed tomography.



Accuracy of Diagnosis

For Hx+Ex profiles, both GPT-4.0 and residents demonstrated higher diagnostic accuracy compared to GPT-3.5 (62/104, 59.6% vs 41/104, 39.4%; *P*=.003; and 63/104, 60.6% vs 41/104, 39.4%; *P*=.001, respectively). Similarly, the accuracy of diagnoses within the top 3 suggestions were also higher (79/104, 76% vs 51/104, 49%; *P*<.001; and 68/104, 65.4% vs 51/104, 49%; *P*=.02, respectively). However, there was no statistically significant difference in diagnostic accuracy between GPT-4.0 and residents (79/104, 76% vs 68/104, 65.4%; *P*=.09). Compared to Hx profiles, GPT-4.0 showed improved diagnostic accuracy for Hx+Ex profiles and diagnoses within the top 3 suggestions (62/104, 59.6% vs 42/104, 40.4%; *P*=.007; and 79/104, 76% vs 63/104, 60.6%; *P*=.02, respectively; Table 3). The detailed top-3 predicted diagnoses by GPT-3.5 and GPT-4.0, alongside the composite diagnosis by the 3 residents, are showed in Multimedia Appendix 3. An example of GPT-4.0’s response to diagnosis is shown in Figure 2.

In the subgroup analysis, both GPT-3.5 and GPT-4.0 exhibited generally lower diagnostic accuracy for Hx profiles. However, for Hx+Ex profiles, there was an overall improvement in diagnosis, particularly for glaucoma. The top 3 subspecialties in overall accuracy were *glaucoma*; *lens diseases*; and *eyelid, lacrimal, and orbital diseases*. In the subspecialties of *eye tumors* and *scleral and uveal diseases*, significant differences were observed in the top-ranked diagnosis accuracy for Hx+Ex profiles among GPT-3.5 , GPT-4.0, and residents (4/11, 36% vs 5/11, 46% vs 10/11, 91%; *P*=.03; and 1/6, 17% vs 5/6, 83% vs 5/6, 83%; *P*=.02, respectively). Notably, for *scleral and uveal diseases*, GPT-3.5 demonstrated a lower accuracy of 17% (1/6)

both in the top-ranked diagnosis (*P*=.02) and within the top 3 diagnoses (*P*=.02) compared to GPT-4.0 (5/6, 83%) and the residents (5/6, 83%; Figure 3).

Table 3. Secondary outcomes for patient with history (Hx) and patient with history and examination (Hx+Ex) profiles (n=104).

Diagnosis accuracy	GPT-3.5, n (%)	GPT-4.0, n (%)	Residents, n (%)	*P* value
Accuracy A: the top-ranked diagnosis is correct for Hx profiles	37 (35.6)	42 (40.4)[a,b]	—[c]	.48
Accuracy B: the diagnosis is within the top 3 suggestions for Hx profiles	50 (48.1)	63 (60.6)[a]	—	.07
Accuracy C: the top-ranked diagnosis is correct for Hx+Ex profiles	41 (39.4)[d,e]	62 (59.6)[b]	63 (60.6)	.002
Accuracy D: the diagnosis is within the top 3 suggestions for Hx+Ex profiles	51 (49)[d,e]	79 (76)	68 (65.4)	<.001

[a]Statistically significant differences for GPT-4.0 when comparing accuracy A vs accuracy C (*P*=.007) and accuracy B vs accuracy D (*P*=.02).

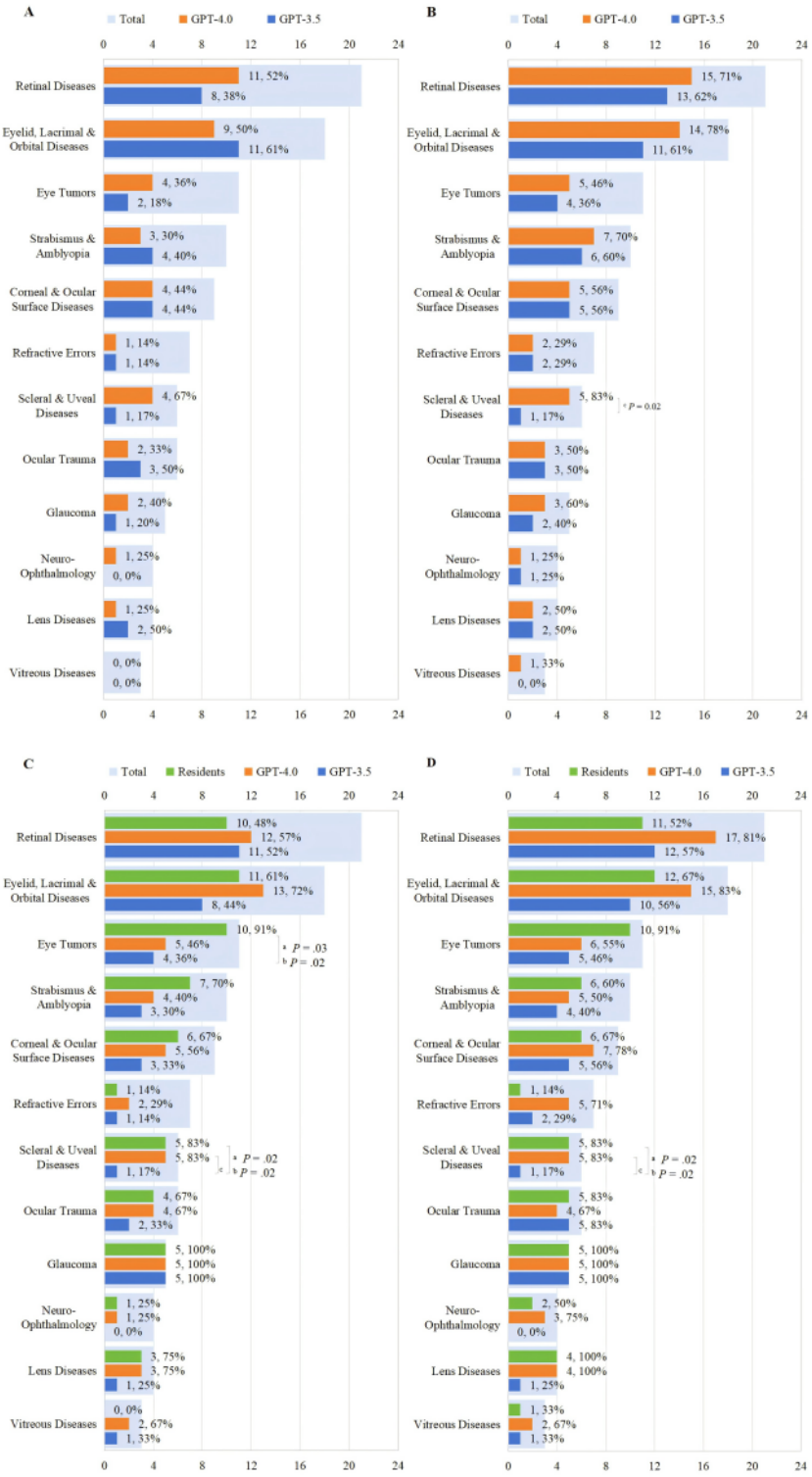
[b]Statistically significant differences for GPT-4.0 when comparing accuracy A vs accuracy B (*P*=.004) and accuracy C vs accuracy D (*P*=.01).

[c]Not applicable.

[d]Statistically significant differences for accuracy C and accuracy D when comparing GPT-3.5 with GPT-4.0 (*P*=.004 and *P*<.001, respectively).

[e]Statistically significant differences for accuracy C and accuracy D when comparing GPT-3.5 with residents (*P*=.002 and *P*=.02, respectively).

Figure 3. The diagnostic accuracy of GPT-3.5, GPT-4.0, and residents across various ophthalmic subspecialties: (A) accuracy of the top-ranked diagnosis for Hx profiles; (B) accuracy of the diagnosis within the top 3 suggestions for Hx profiles; (C) accuracy of the correct top-ranked diagnosis for Hx+Ex profiles; and (D) accuracy of the diagnosis within the top 3 suggestions for Hx+Ex profiles. “a” indicates significant statistical differences across all 3 groups (GPT-3.5, GPT-4.0, and residents), “b” denotes a significant difference between residents and GPT-3.5, and “c” represents a significant difference between GPT-4.0 and GPT-3.5. Hx: patient with history; Hx+Ex: patient with history and examination.



Case Description and Accuracy

For Hx+Ex profiles where the medical history provided past diagnoses related to the final diagnosis, a higher top-ranked diagnosis accuracy was observed in the GPT-4.0 model (7/7, 100% vs 55/97, 57%; *P*=.04). However, case descriptions including past diagnoses unrelated to the final diagnosis and cases requiring ophthalmic examination for a definitive diagnosis did not significantly affect the top-ranked diagnosis accuracy for GPT-3.5, GPT-4.0, and the residents (all *P*>.21; Table 4).

Table 4. Case characteristics and their association with the top-ranked diagnosis accuracy for patient with history and examination (Hx+Ex) profiles.

Cases characteristics	GPT-3.5	*P* value	GPT-4.0	*P* value	Residents	*P* value
Prescence of medical history A[a], n (%)		.43		.04		.70
No (n=97)	36 (37)		55 (57)		58 (60)	
Yes (n=7)	4 (57)		7 (100)		5 (71)	
Prescence of medical history B[b], n (%)		.73		.74		>.99
No (n=95)	36 (38)		56 (59)		57 (60)	
Yes (n=9)	4 (44)		6 (67)		6 (67)	
Prescence of characteristics C[c], n (%)		.21		.29		>.99
No (n=51)	24 (47.1)		31 (60.8)		31 (61)	
Partly (n=20)	6 (30)		9 (45)		12 (60)	
Yes (n=33)	10 (30.3)		22 (66.7)		20 (61)	

[a]Medical history A: descriptions of past diagnoses related to the final diagnosis.

[b]Medical history B: description of past diagnoses unrelated to the final diagnosis.

[c]Characteristic C: official diagnosis states that the diagnosis must be made in conjunction with an ophthalmic examination.

Characteristic of Incorrect Diagnoses

In the analysis of incorrect top-ranked diagnoses from Hx+Ex profiles, GPT-4.0 exhibited fewer incorrect top-3 diagnoses than GPT-3.5 (25/42, 60% vs 53/63, 84%; *P*=.005), making partially correct diagnoses with incorrect lesion nature (21/42, 50% vs 7/63, 11%; *P*<.001). In contrast, GPT-3.5 often made completely incorrect diagnoses about lesion nature (27/63, 43% vs 7/42, 17%; *P*=.005) and exhibited less precision with no further diagnosis more frequently than GPT-4.0 (22/63, 35% vs 5/42, 12%; *P*=.009; Table 5).

Table 5. Analysis of ChatGPT's incorrect top-ranked diagnoses in patient with history and examination (Hx+Ex) profiles.

	GPT-3.5 (n=63)	GPT-4.0 (n=42)	*P* value
Proportion of incorrect top-3 diagnosis, n (%)	53 (84)	25 (60)	.005
Reasons for incorrect top-ranked diagnosis, n (%)			<.001
Incorrect lesion nature	27 (43)	7 (17)	.005[a]
Incorrect etiology	1 (2)	2 (5)	.57
Lacks precision, no further diagnosis	22 (35)	5 (12)	.009[a]
Partially correct with incorrect lesion location	6 (10)	7 (17)	.28
Partially correct with incorrect lesion nature	7 (11)	21 (50)	<.001[a]

[a]Statistically significant difference between GPT-4.0 and GPT-3.5 at an adjusted *P* value of .01.

Discussion

Principal Finding

AI has demonstrated its potential in facilitating accurate patient registration and health care services [13,14]. With iterations, ChatGPT—a chatbot powered by AI—has shown promise for triage in ophthalmic emergencies and in achieving diagnostic accuracy in simulated vignettes [6,15]. Our study focused on investigating the triage and diagnostic value of ChatGPT. To our knowledge, this was the first study to explore the role of ChatGPT in ophthalmic registration. Additionally, we designed 2 types of information inputs, Hx and Hx+Ex profiles, and simultaneously tested the accuracy of leading diagnoses and do-not-miss diagnoses. This approach helped provide a comprehensive understanding of ChatGPT's performance. While existing studies mainly focus on the application within a single ophthalmic subspecialty [10,11,16,17], another strength of our study was the use of a diverse set of diagnostic cases, covering 12 ophthalmic subspecialties and 104 distinct cases. This breadth enhances the AI system's evaluation across varied clinical scenarios.

From the patients' perspective, AI chatbots facilitate ophthalmic triage and appointment [18]. Patients know their own complaints and medical histories well. This knowledge enables them to directly interact with AI chatbots, seeking advice on appropriate ophthalmic subspecialties for registration [4]. Our study used Hx profiles to simulate this self-service interaction, highlighting the practical utility of chatbots in patient-initiated health care navigation. The findings demonstrated that GPT-4.0 directed patients to the correct registration with 78% accuracy, which numerically surpassed the 69% accuracy achieved by medically trained residents. In China, where major tertiary hospitals have

implemented web-based registration systems, such as through a WeChat-based medical platform [19], the integration of a user-friendly and accessible AI chatbot significantly streamlines the consultation process for patients, particularly those unsure of which ophthalmic subspecialty to choose. This study provided the first empirical evidence of how AI chatbots can facilitate more accurate and efficient patient registration in clinical settings.

For ophthalmic diagnoses, studies had shown that GPT-4.0 exhibited lower than 50% accuracy in deriving leading diagnoses from complaint records alone [6], while the free version of GPT-3.5 underperforms compared to medical residents [12]. These findings were consistent with our study. In a pilot study by Hu et al [20], GPT-4.0 was tested on its capability to diagnose rare eye diseases, revealing that more comprehensive information provided to the model resulted in considerably more "right" diagnoses. The reduced accuracy when used with limited patient information (such as Hx only) indicated that GPT-4.0 was not yet suitable as a stand-alone diagnostic tool. However, adding details from patient examinations significantly enhanced its performance, making it comparable to that of residents. In real-world clinical settings, where physicians have access to comprehensive case information, GPT-4.0 achieved diagnostic accuracies between 60% to 76%, demonstrating its potential to support clinical decision-making. Similar performance had been observed in uveitis studies, with accuracy rates ranging from 60% to 66% [21,22]. Unlike GPT-3.5, which provided imprecise diagnoses, GPT-4.0, even when erring in its initial diagnosis, tended to include the correct diagnosis within the top 3 suggestions. This illustrated GPT-4.0's superior information retrieval capabilities and its role as a diagnostic aid [23]. It is expected that with continuous improvements, AI chatbots will play an increasingly vital role in enhancing health care efficiency [24,25].

Our findings revealed that the accuracy of recommendations for ophthalmic subspecialty registration did not consistently correlate with diagnostic accuracy. For example, while the diagnostic accuracy for glaucoma reached 100%, the accuracy of leading patients to register in the glaucoma department was notably low. This discrepancy was likely attributed to the frequently nonspecific initial complaints associated with glaucoma, which generally require additional clinical examinations to establish a definitive diagnosis, such as intraocular pressure, OCT, and visual field tests. Therefore, relying solely on patient complaints and medical histories proved insufficient for accurately guiding patients to the appropriate glaucoma specialty for initial registration. Conversely, when patient histories were supplemented with specific ophthalmic examinations, the accuracy of differential diagnosis improved, thereby enhancing overall diagnostic performance.

Compared to the accuracy of GPT-4.0 in answering ophthalmic questions [26-28] and suggesting surgical plan [29], which can exceed 80%, its performance in registration recommendations and clinical diagnosis was intermediate or inferior. This discrepancy may be attributed to the conservative diagnostic judgment standards and the inclusion of less common diseases. Although the evaluated profiles were derived from real-world clinical reports in the published Chinese SRT database, the inclusion of uncommon and atypical cases was inevitable. Such heightened complexity poses significant challenges not only for AI chatbots but also for residents in the early stages of their medical careers [30]. In real-world clinical settings, the uneven distribution of diseases across various ophthalmic subspecialties typically leads to variability in diagnostic outcomes.

Given that the performance of AI chatbots is contingent upon the volume of information provided, future AI chatbots may still require upgrades and iterations to mitigate information asymmetry between patients and health care professionals, thereby enhancing the delivery of more effective and professional ophthalmic care. For example, AI chatbots specifically designed and trained for ophthalmic care [31,32]; chatbots that can proactively solicit information not provided by end users, similar to the process used by ophthalmologists; and those capable of directly accessing and interpreting imaging data like a multimodal AI chatbot, are essential. Unlike general-purpose AI chatbots such as ChatGPT, Zheng et al [33] have developed a Chinese large language model for ophthalmology using a corpus with extensive clinical vignettes (Hx+Ex). This model demonstrated a diagnostic accuracy of 81.1% across 6 common ophthalmic subspecialties, surpassing the performance of GPT-4.0 (59.6%) in our study. However, the current capability of multimodal GPT-4.0 to diagnose vitreoretinal diseases through the analysis of retinal images remains less than optimal [34].

ChatGPT allows users to customize responses based on personalized prompts, as illustrated in this study by providing 3 differential diagnostic recommendations and the corresponding rationale [35,36]. Research showed that ChatGPT consistently offers a broader range of differential diagnoses than ophthalmology residents [30]; this tendency was also observed in our data collection. The ability to organize key case information through the explanation of diagnostic reasoning not only enhances the knowledge structure of physicians but also underscores the significant educational value of AI chatbots in medical training [37-39]. The proficiency of AI chatbots in responding to ophthalmic examination questions and addressing eye disease queries have been confirmed by recent studies [40-42]. However, it is crucial to note that current AI chatbots do not necessarily replace the clinical judgment of professional ophthalmologists. Their application is still subject to ethical considerations [43,44] and concerns about hallucinations [45]. Physicians responsible for diagnosis need to remain cautious when considering information provided by ChatGPT. The extent to which AI can serve as an adjunct tool in health care still requires further real-world testing.

Limitations

In clinical settings, ophthalmologists typically rely on direct observation of patients' symptoms and examinations for intuitive face-to-face diagnoses. However, the text-based format of medical records may not fully reflect ophthalmic residents' capabilities and might even underestimate them. Additionally, while images are crucial for ophthalmic diagnosis, GPT-4.0's support for multimodal data still shows suboptimal performance in image-based cases [46]. This explains our decision to exclude images from our evaluation of ChatGPT. However, textual

descriptions of these images could have impacted the outcomes. Moreover, although ChatGPT supports multiple languages, differences in language use in diagnosing eye diseases have been observed [11]. Our study, conducted in Mandarin Chinese, may affect the generalization of the results. Despite these limitations, our study contributes to understanding AI's role as a tool in ophthalmic health care.

Conclusion

Our study showed that GPT-3.5 and GPT-4.0 demonstrated moderate performance in directing patients to appropriate ophthalmic subspecialties for registration. While GPT-3.5 was less effective, GPT-4.0 approached and even numerically surpassed residents in differential diagnosis when presented with patient histories and examination results. AI chatbots merit emphasis for their potential to facilitate patient registration and optimize consultations in ophthalmology. While their diagnostic capabilities could benefit ophthalmologists, integrating them into diagnostic decision-making still requires further validation.

Acknowledgments

This research was supported by the Medical Science and Technology Tackling Plan of Henan Province (LHGJ20210078).

Data Availability

The raw data supporting the conclusions of this paper will be made available by the authors, without undue reservation.

Authors' Contributions

SM conceived and designed the study, carried out data analysis, drafted the manuscript, and responded to review comments, therefore is the first corresponding author. SM, QG, and XG evaluated ChatGPT's responses and assisted in interpreting the results. KX, DC, and XY were responsible for quality control and assisted in data cleaning. BL participated in the design, reviewed the manuscript, and provided reimbursement for the publication fee, therefore shares the role of corresponding author with SM. All authors contributed to the paper and approved the submitted version.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Prompts for the Hx and Hx+Ex profiles before testing ChatGPT's registration and diagnosis responses. Hx+Ex: patient with history and examination; Hx: patient with history.
[PDF File (Adobe PDF File), 171 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Detailed official subspecialty classification for each clinical profile (Hx), alongside registration recommendations for GPT-3.5, GPT-4.0, and residents. Hx: patient with history.
[DOCX File , 24 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Detailed official diagnosis and top 3 predicted diagnoses by GPT-3.5 and GPT-4.0, alongside the composite diagnosis by 3 residents (1=correct and 0=incorrect) for each clinical profile (Hx+Ex). Hx+Ex: patient with history and examination.
[DOCX File , 41 KB-Multimedia Appendix 3]

References

1. Ting DSW, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol. Feb 25, 2019;103(2):167-175. [FREE Full text] [doi: 10.1136/bjophthalmol-2018-313173] [Medline: 30361278]
2. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng. Oct 10, 2018;2(10):719-731. [doi: 10.1038/s41551-018-0305-z] [Medline: 31015651]
3. ChatGPT. OpenAI. URL: https://chatgpt.com/ [accessed 2024-10-24]
4. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. Aug 17, 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
5. Ting DSJ, Tan TF, Ting DSW. ChatGPT in ophthalmology: the dawn of a new era? Eye (Lond). Jan 27, 2024;38(1):4-7. [doi: 10.1038/s41433-023-02619-4] [Medline: 37369764]

6. Zandi R, Fahey JD, Drakopoulos M, Bryan JM, Dong S, Bryar PJ, et al. Exploring diagnostic precision and triage proficiency: a comparative study of GPT-4 and Bard in addressing common ophthalmic complaints. Bioengineering (Basel). Jan 26, 2024;11(2):120. [FREE Full text] [doi: 10.3390/bioengineering11020120] [Medline: 38391606]
7. Salimi A, Saheb H. Large language models in ophthalmology scientific writing: ethical considerations blurred lines or not at all? Am J Ophthalmol. Oct 2023;254:177-181. [doi: 10.1016/j.ajo.2023.06.004] [Medline: 37348667]
8. Waisberg E, Ong J, Masalkhi M, Kamran SA, Zaman N, Sarker P, et al. GPT-4 and ophthalmology operative notes. Ann Biomed Eng. Nov 02, 2023;51(11):2353-2355. [doi: 10.1007/s10439-023-03263-5] [Medline: 37266720]
9. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. JAMA Ophthalmol. Jun 01, 2023;141(6):589-597. [FREE Full text] [doi: 10.1001/jamaophthalmol.2023.1144] [Medline: 37103928]
10. Delsoz M, Madadi Y, Raja H, Munir WM, Tamm B, Mehravaran S, et al. Performance of ChatGPT in diagnosis of corneal eye diseases. Cornea. May 01, 2024;43(5):664-670. [doi: 10.1097/ICO.0000000000003492] [Medline: 38391243]
11. Liu X, Wu J, Shao A, Shen W, Ye P, Wang Y, et al. Uncovering language disparity of ChatGPT on retinal vascular disease classification: cross-sectional study. J Med Internet Res. Jan 22, 2024;26:e51926. [FREE Full text] [doi: 10.2196/51926] [Medline: 38252483]
12. Shemer A, Cohen M, Altarescu A, Atar-Vardi M, Hecht I, Dubinsky-Pertzov B, et al. Diagnostic capabilities of ChatGPT in ophthalmology. Graefes Arch Clin Exp Ophthalmol. Jul 06, 2024;262(7):2345-2352. [doi: 10.1007/s00417-023-06363-z] [Medline: 38183467]
13. Wang J, Zhang G, Wang W, Zhang K, Sheng Y. Cloud-based intelligent self-diagnosis and department recommendation service using Chinese medical BERT. J Cloud Comp. Jan 15, 2021;10(1):4. [doi: 10.1186/s13677-020-00218-2]
14. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: development, applications, and challenges. Health Care Sci. Aug 24, 2023;2(4):255-263. [FREE Full text] [doi: 10.1002/hcs2.61] [Medline: 38939520]
15. Lyons RJ, Arepalli SR, Fromal O, Choi JD, Jain N. Artificial intelligence chatbot performance in triage of ophthalmic conditions. Can J Ophthalmol. Aug 2024;59(4):e301-e308. [doi: 10.1016/j.jcjo.2023.07.016] [Medline: 37572695]
16. Carlà MM, Gambini G, Baldascino A, Boselli F, Giannuzzi F, Margollicci F, et al. Large language models as assistance for glaucoma surgical cases: a ChatGPT vs. Google Gemini comparison. Graefes Arch Clin Exp Ophthalmol. Sep 2024;262(9):2945-2959. [doi: 10.1007/s00417-024-06470-5] [Medline: 38573349]
17. Maywood M, Parikh R, Deobhakta A, Begaj T. Performance assessment of an artificial intelligence chatbot in clinical vitreoretinal scenarios. Retina. Jun 01, 2024;44(6):954-964. [doi: 10.1097/IAE.0000000000004053] [Medline: 38271674]
18. Betzler BK, Chen H, Cheng C, Lee CS, Ning G, Song SJ, et al. Large language models and their impact in ophthalmology. Lancet Digit Health. Dec 2023;5(12):e917-e924. [FREE Full text] [doi: 10.1016/S2589-7500(23)00201-7] [Medline: 38000875]
19. Yang X, Kovarik CL. A systematic review of mobile health interventions in China: identifying gaps in care. J Telemed Telecare. Jan 2021;27(1):3-22. [doi: 10.1177/1357633X19856746] [Medline: 31319759]
20. Hu X, Ran AR, Nguyen TX, Szeto S, Yam JC, Chan CKM, et al. What can GPT-4 do for diagnosing rare eye diseases? a pilot study. Ophthalmol Ther. Dec 01, 2023;12(6):3395-3402. [FREE Full text] [doi: 10.1007/s40123-023-00789-8] [Medline: 37656399]
21. Rojas-Carabali W, Cifuentes-González C, Wei X, Putera I, Sen A, Thng ZX, et al. Evaluating the diagnostic accuracy and management recommendations of ChatGPT in uveitis. Ocul Immunol Inflamm. Oct 18, 2024;32(8):1526-1531. [doi: 10.1080/09273948.2023.2253471] [Medline: 37722842]
22. Rojas-Carabali W, Sen A, Agarwal A, Tan G, Cheung CY, Rousselot A, et al. Chatbots vs. human experts: evaluating diagnostic performance of chatbots in uveitis and the perspectives on AI adoption in ophthalmology. Ocul Immunol Inflamm. Oct 13, 2024;32(8):1591-1598. [doi: 10.1080/09273948.2023.2266730] [Medline: 37831553]
23. Lim ZW, Pushpanathan K, Yew SME, Lai Y, Sun C, Lam JSH, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine. Sep 2023;95:104770. [FREE Full text] [doi: 10.1016/j.ebiom.2023.104770] [Medline: 37625267]
24. Biswas S, Davies LN, Sheppard AL, Logan NS, Wolffsohn JS. Utility of artificial intelligence-based large language models in ophthalmic care. Ophthalmic Physiol Opt. May 25, 2024;44(3):641-671. [doi: 10.1111/opo.13284] [Medline: 38404172]
25. Madadi Y, Delsoz M, Khouri A, Boland M, Grzybowski A, Yousefi S. Applications of artificial intelligence-enabled robots and chatbots in ophthalmology: recent advances and future trends. Curr Opin Ophthalmol. May 01, 2024;35(3):238-243. [doi: 10.1097/ICU.0000000000001035] [Medline: 38277274]
26. Pushpanathan K, Lim ZW, Er Yew SM, Chen DZ, Hui'En Lin HA, Lin Goh JH, et al. Popular large language model chatbots' accuracy, comprehensiveness, and self-awareness in answering ocular symptom queries. iScience. Nov 17, 2023;26(11):108163. [FREE Full text] [doi: 10.1016/j.isci.2023.108163] [Medline: 37915603]
27. Taloni A, Borselli M, Scarsi V, Rossi C, Coco G, Scorcia V, et al. Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology. Sci Rep. Oct 29, 2023;13(1):18562. [FREE Full text] [doi: 10.1038/s41598-023-45837-2] [Medline: 37899405]

28. Momenaei B, Wakabayashi T, Shahlaee A, Durrani AF, Pandit SA, Wang K, et al. Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases. Ophthalmol Retina. Oct 2023;7(10):862-868. [doi: 10.1016/j.oret.2023.05.022] [Medline: 37277096]
29. Carlà MM, Gambini G, Baldascino A, Giannuzzi F, Boselli F, Crincoli E, et al. Exploring AI-chatbots' capability to suggest surgical planning in ophthalmology: ChatGPT versus Google Gemini analysis of retinal detachment cases. Br J Ophthalmol. Sep 20, 2024;108(10):1457-1469. [doi: 10.1136/bjo-2023-325143] [Medline: 38448201]
30. Delsoz M, Raja H, Madadi Y, Tang AA, Wirostko BM, Kahook MY, et al. The use of ChatGPT to assist in diagnosing glaucoma based on clinical case reports. Ophthalmol Ther. Dec 14, 2023;12(6):3121-3132. [FREE Full text] [doi: 10.1007/s40123-023-00805-x] [Medline: 37707707]
31. Li F, Wang D, Yang Z, Zhang Y, Jiang J, Liu X, et al. The AI revolution in glaucoma: bridging challenges with opportunities. Prog Retin Eye Res. Aug 24, 2024;103:101291. [doi: 10.1016/j.preteyeres.2024.101291] [Medline: 39186968]
32. Chen X, Zhang W, Xu P, Zhao Z, Zheng Y, Shi D, et al. FFA-GPT: an automated pipeline for fundus fluorescein angiography interpretation and question-answer. NPJ Digit Med. May 03, 2024;7(1):111. [FREE Full text] [doi: 10.1038/s41746-024-01101-z] [Medline: 38702471]
33. Zheng C, Ye H, Guo J, Yang J, Fei P, Yuan Y, et al. Development and evaluation of a large language model of ophthalmology in Chinese. Br J Ophthalmol. Sep 20, 2024;108(10):1390-1397. [FREE Full text] [doi: 10.1136/bjo-2023-324526] [Medline: 39019566]
34. Ghalibafan S, Taylor Gonzalez DJ, Cai L, Graham Chou B, Panneerselvam S, Conrad Barrett S, et al. Applications of multimodal generative artificial intelligence in a real-world retina clinic setting. Retina. Oct 01, 2024;44(10):1732-1740. [doi: 10.1097/IAE.0000000000004204] [Medline: 39287535]
35. Kleinig O, Gao C, Kovoor JG, Gupta AK, Bacchi S, Chan WO. How to use large language models in ophthalmology: from prompt engineering to protecting confidentiality. Eye (Lond). Mar 05, 2024;38(4):649-653. [FREE Full text] [doi: 10.1038/s41433-023-02772-w] [Medline: 37798360]
36. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res. Oct 04, 2023;25:e50638. [FREE Full text] [doi: 10.2196/50638] [Medline: 37792434]
37. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ. Nov 19, 2024;58(11):1276-1285. [doi: 10.1111/medu.15402] [Medline: 38639098]
38. Gurnani B, Kaur K. Leveraging ChatGPT for ophthalmic education: a critical appraisal. Eur J Ophthalmol. Mar 16, 2024;34(2):323-327. [doi: 10.1177/11206721231215862] [Medline: 37974429]
39. Sevgi M, Antaki F, Keane PA. Medical education with large language models in ophthalmology: custom instructions and enhanced retrieval capabilities. Br J Ophthalmol. Sep 20, 2024;108(10):1354-1361. [FREE Full text] [doi: 10.1136/bjo-2023-325046] [Medline: 38719344]
40. Mihalache A, Huang RS, Popovic MM, Muni RH. Performance of an upgraded artificial intelligence chatbot for ophthalmic knowledge assessment. JAMA Ophthalmol. Aug 01, 2023;141(8):798-800. [FREE Full text] [doi: 10.1001/jamaophthalmol.2023.2754] [Medline: 37440220]
41. Antaki F, Milad D, Chia MA, Giguère C-É, Touma S, El-Khoury J, et al. Capabilities of GPT-4 in ophthalmology: an analysis of model entropy and progress towards human-level medical question answering. Br J Ophthalmol. Sep 20, 2024;108(10):1371-1378. [doi: 10.1136/bjo-2023-324438] [Medline: 37923374]
42. Huang AS, Hirabayashi K, Barna L, Parikh D, Pasquale LR. Assessment of a large language model's responses to questions and cases about glaucoma and retina management. JAMA Ophthalmol. Apr 01, 2024;142(4):371-375. [FREE Full text] [doi: 10.1001/jamaophthalmol.2023.6917] [Medline: 38386351]
43. Ning Y, Teixayavong S, Shang Y, Savulescu J, Nagaraj V, Miao D, et al. Generative artificial intelligence and ethical considerations in health care: a scoping review and ethics checklist. Lancet Digit Health. Nov 2024;6(11):e848-e856. [FREE Full text] [doi: 10.1016/S2589-7500(24)00143-2] [Medline: 39294061]
44. Kalaw FGP, Baxter SL. Ethical considerations for large language models in ophthalmology. Curr Opin Ophthalmol. Nov 01, 2024;35(6):438-446. [doi: 10.1097/ICU.0000000000001083] [Medline: 39259616]
45. Cai LZ, Shaheen A, Jin A, Fukui R, Yi JS, Yannuzzi N, et al. Performance of generative large language models on ophthalmology board-style questions. Am J Ophthalmol. Oct 2023;254:141-149. [doi: 10.1016/j.ajo.2023.05.024] [Medline: 37339728]
46. Mihalache A, Huang RS, Popovic MM, Patil NS, Pandya BU, Shor R, et al. Accuracy of an artificial intelligence chatbot's interpretation of clinical ophthalmic images. JAMA Ophthalmol. Apr 01, 2024;142(4):321-326. [doi: 10.1001/jamaophthalmol.2024.0017] [Medline: 38421670]

Abbreviations

AI: artificial intelligence
Hx: patient with history
Hx+Ex: patient with history and examination
OCT: orbital computed tomography

SRT: Standardized Resident Training

Edited by A Schwartz; submitted 05.05.24; peer-reviewed by MO Khursheed, C Cifuentes-Gonzalez, W Rojas-Carabali; comments to author 20.06.24; revised version received 05.08.24; accepted 15.10.24; published 14.11.24

Please cite as:
Ming S, Yao X, Guo X, Guo Q, Xie K, Chen D, Lei B
Performance of ChatGPT in Ophthalmic Registration and Clinical Diagnosis: Cross-Sectional Study
J Med Internet Res 2024;26:e60226
URL: https://www.jmir.org/2024/1/e60226
doi: 10.2196/60226
PMID: