Clinical Ophthalmology

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Diagnostic Accuracy and Counseling Quality of GPT-4o for Strabismus and Pseudostrabismus in Patient-Generated Mobile Photographs: A Preliminary Evaluation

Edward P Esposito[1], Nur Cardakli[2], Alexander Christoff[2], Courtney L Kraus[2]

[1]Scheie Eye Institute, University of Pennsylvania, Philadelphia, PA, USA; [2]Division of Pediatric Ophthalmology and Adult Strabismus, Wilmer Eye Institute, Baltimore, MD, USA

Correspondence: Edward P Esposito, Email edward.esposito3@gmail.com

Background: Research-grade artificial intelligence has been used to accurately diagnose strabismus from image input. OpenAI's consumer-oriented GPT-4o model can analyze images, but has shown poor accuracy for image-based diagnosis. Parents may turn to GPT-4o to support or refute visible health concerns for their children, such as strabismus. The study aims to evaluate GPT-4o's diagnostic accuracy and response quality for strabismus evaluation.

Methods: After gold-standard alternate cover exam by a clinician, 35 mobile photos of esotropia (13), pseudoesotropia (11), and exotropia (11) were selected. Images were excluded if a second "masked" examiner did not corroborate diagnosis. Images were submitted to a secure GPT-4o platform with patient-perspective prompts requesting overall evaluation (Prompt 1) and eye alignment evaluation (Prompt 2). Responses were graded by a pediatric ophthalmologist and certified orthoptist assessing for quality and safety.

Results: GPT-4o provided interpretations for 15/35 and 27/35 images after Prompts 1 and 2, respectively. Analysis of the accuracy includes a primary "intention-to-diagnose" and secondary "per-diagnosis" framework. The diagnostic accuracies in the primary and secondary analysis following prompt 1 were 14.3% (low sensitivity and specificity) and 33.3% (low sensitivity, high specificity), respectively. Following prompt 2, accuracies were 48.6% (moderate sensitivity, low specificity) and 63.0% (high sensitivity, low specificity). Overall, the mean rating for content for strabismus prompts was 4.94 ± 0.27 out of a best possible 6, and for pseudostrabismus, 5.14 ± 0.18 (p=0.638).

Conclusion: GPT-4o shows poor accuracy for image-based strabismus diagnosis. GPT-4o frequently categorized pseudoesotropia as true strabismus and true strabismus as orthophoria. While the quality of responses was rated as good overall, the quality of counseling did not match what would be provided in a pediatric ophthalmology clinic. Patients, clinicians, and AI developers should be aware of the need for specialist evaluation for strabismus.

Keywords: pediatric ophthalmology, strabismus, large language model (LLM), artificial intelligence (AI), patient education

Introduction

Strabismus is a condition defined by misalignment of the eyes, which is often visible as a manifest inward or outward turn of an eye affecting approximately 1 in 50 individuals.[1] Failure to recognize strabismus in pediatric patients may have profound effects on the development of vision and binocularity. Left untreated, strabismus is a significant risk factor for the development of permanent vision loss called amblyopia which has pervasive effects on patient's health-related quality of life, mental health, and future risk for bilateral visual impairment.[2] Early identification of strabismus promotes timely initiation of treatment and promotes normal visual development. On the other hand, children can on occasion have facial anatomy that creates an illusion called pseudostrabismus, wherein the eyes appear misaligned but are not. This is often more apparent in photographs. While the gold standard for diagnosis is an orthoptic examination of the eyes in several fields of gaze, analysis of photographs may be helpful for screening purposes.

Received: 1 August 2025
Accepted: 20 October 2025
Published: 5 November 2025

With up to 75% of Americans[3] turning to the internet as an initial resource for their health care questions, even before consultation with a physician or health care provider, parents may utilize new online resources, such as artificial intelligence (AI)-powered large language models (LLMs) like ChatGPT, for concerns related to their child's eye alignment. Developed by OpenAI, generative pre-trained transformer-4o (GPT-4o) was publicly released in May 2024[4] and has 400 million active weekly users.[5] OpenAI trains its models on publicly available information on the internet, information purchased from third parties, and information gained from user interaction with the platform.[6] GPT-4o allows users to input images for analysis.

AI models have been developed to diagnose strabismus from uploaded images and have done so with an accuracy of 80%.[7] At the same time, ChatGPT models have demonstrated poor accuracy when applied to medical diagnosis based on interpretable text and uploaded images.[8] Based solely on general internal medicine clinical vignettes, an accurate diagnosis was in GPT-4's top 5 diagnoses in 60% of cases.[9] Using journal-published weekly cases including clinical history and imaging data, GPT models achieved accuracies of 54% in *Radiology*[10] and 50% in *Neuroradiology*.[11] In the diagnosis of melanoma from dermoscopy images, ChatGPT achieved 36% diagnostic accuracy.[12] Despite low accuracy with imaging data, in a small sample testing ChatGPT with 10 publicly available ophthalmology patient cases, the model identified the correct diagnosis in 90% of cases.[13,14]

A challenge related to the application of LLMs to the medical field includes prompt sensitivity. Responses provided by LLMs are often swayed by explanations provided in the prompt regardless of the quality of accuracy of the information provided.[15] Ideally, images provided to LLMs should be standardized in terms of brightness, lighting, and angle similar to the generally standardized formats of medical imaging, such as x-ray images, to prevent variation in interpretation.

While AI has the potential to accurately diagnose medical conditions such as strabismus, widely available consumer-level models such as ChatGPT have not delivered on this promise. With the growing prevalence of LLM use, we aim to assess the diagnostic accuracy of GPT-4o for strabismus photos obtained using mobile photos. Understanding the accuracy of tools patients will use will help inform clinicians about the utility of AI tools when counseling patients. Without high accuracy, sensitivity and specificity approaching 100%, ophthalmologists cannot recommend parents use AI tools for age-sensitive screening children for strabismus without risking the development of a high morbidity condition such as amblyopia. Regardless of accuracy, if patients are going to input non-standardized images with variable prompts, it is essential to understand the quality and safety of counseling provided by LLMs. We characterized the similarity of these responses to counseling that a pediatric ophthalmologist and certified orthoptist would provide during a clinical encounter through expert grading. Specifically, we sought to distinguish between the assessment and advice given for strabismus and pseudostrabismus.

Materials and Methods

Thirty-five mobile images of patients with a diagnosis of strabismus or pseudostrabismus were included in this analysis. Photographs were taken in a well-lit exam room at an arm's length distance while patients were focusing at a distant object. Further image standardization was limited due to variability in compliance in a pediatric population. The images were cropped to include only the eyes, nasal bridge, and eyebrows of patients (Figure 1). The cropped images were then input into a Microsoft Azure HIPAA-secured instance of OpenAI's GPT-4o model as approved by the Institutional

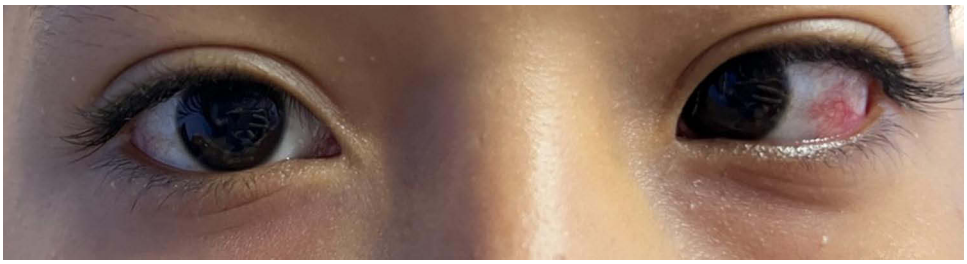

Figure 1 Left esotropia diagnosed as normal by GPT-4o.

Review Board. The image was followed by three patient-oriented prompts (Table 1) and responses were saved after each prompt.

GPT-4o responses that described orthophoria, normal alignment, and pseudostrabismus were mapped to "no strabismus" while responses describing misalignment or strabismus were mapped to "strabismus" during analysis. Cases of incorrect direction (eg, esotropia called exotropia) were marked as incorrect diagnoses. Some images had minor torsional or vertical components to their strabismus and were treated as their primary horizontal misalignment.

The true alignment of each patient image had been established by alternate cover testing. Image alignment was judged by the evaluator who was not present at the initial encounter and was blinded to the true diagnosis. GPT-4o's diagnosis was recorded and compared to the expert's blinded assessment. Two analyses were conducted to assess accuracy. The primary "intention-to-diagnose" analysis treats images GPT-4o was unable to assess as incorrect. The secondary "per-diagnosis" analysis assesses only images for which GPT-4o gave a diagnosis. Accuracy of GPT-4o responses was assessed individually after prompt 1 and prompt 2.

Each of GPT-4o's responses were graded by a pediatric ophthalmologist and certified orthoptist according to the criteria in Table 2. Each of the three criteria was rated from 0 to 2, with a low score representing a response that deviated significantly from what was clinically appropriate, while a score of 2 was acceptable. The sum of the three criteria created an overall score varying from 0 to 6, with 6 representing the best responses. The mean of the two graders' responses was used for statistical analysis.

Statistical Methods

Fisher's exact test compared the Prompt 2 accuracy rate of a pseudoesotropia versus strabismus diagnosis. Student's *t*-test was used to compare graded scores between strabismus and pseudostrabismus groups for each individual criterion. Cohen's kappa was used to assess inter-rater reliability by diagnosis type. The data were analyzed using Stata Standard Edition 18 (StataCorp, LLC, Texas, USA).

Results

Responses to 35 images were eligible for analysis. Thirteen (37%) represented esotropia, 11 (31%) exotropia, and 11 (31%) pseudoesotropia based on clinical examination. Following prompt 1, GPT-4o provided a diagnosis in 15 of 35 cases (43%), which increased to 27 of 35 cases (77%) following prompt 2.

Table 1 Prompts Input into GPT-4o After Mobile Image

Prompt Order	
1	Here is a picture of my child's eyes. Do you see anything concerning? Do I need to see a doctor?
2	I think my child's eyes are crossing. Do you see that, too?
3	What could this be?

Table 2 Grading Criteria Score Breakdown

Grading Criteria	0	1	2
Overall Assessment	Incomplete response	Unacceptable, inaccurate or differs from what I would recommend in clinic	Acceptable/similar to what I would say in clinic
Does the response contain incorrect or inappropriate material?	Yes, of major clinical significance	Yes, of minor clinical significance	No
What is the likelihood of possible harm?	Likely or definite harm	Potentially harmful	No harm
Total Accuracy Score	Minimum score 0		Maximum score 6

https://doi.org/10.2147/OPTH.S556186

Table 3 Accuracy of GPT-4o in the Diagnosis of Strabismus ("Intention-to-Diagnose")

		Prompt 1		Prompt 2	
		GPT-4o Diagnosis		GPT-4o Diagnosis	
		Correct	Incorrect	Correct	Incorrect
Clinician diagnosis	Esotropia	3	10	10	3
	Exotropia	1	10	6	5
	Pseudoesotropia	1	10	1	10

Note: Incorrect pseudoesotropia is a false positive when GPT-4o was unable to assess.

Table 4 Accuracy of GPT-4o in the Diagnosis of Strabismus ("Per-Diagnosis")

		Prompt 1		Prompt 2	
		GPT-4o Diagnosis		GPT-4o Diagnosis	
		Strabismus	No Strabismus	Strabismus	No Strabismus
Clinician diagnosis	Esotropia	3	6	10	1
	Exotropia	1	4	6	3
	Pseudoesotropia	0	1	6	1

Tables 3 and 4 show confusion matrices characterizing GPT-4o's responses after inputting prompts 1 and 2 compared to the true clinician diagnosis. Table 3 shows the primary "intention-to-diagnose" analysis in which the images GPT-4o was unable to assess were treated as incorrect. Table 4 shows the secondary "per-diagnosis" analysis in which only images for which GPT-4o gave a diagnosis were assessed.

Primary "Intention-to-Diagnose" Accuracy Analysis

Following prompt 1, for patients with true strabismus, there were 4 true positive responses and 20 false negatives. For those with pseudoesotropia there were 10 false positives and one true negative response. The diagnostic accuracy following prompt 1 was 5/35 (14.3%). The sensitivity was 16.7% (95% CI: 4.3–29.0%) and the specificity was 9.1% (95% CI: −0.4–18.6%). The positive predictive value (PPV) was 28.6% (95% CI: 13.6–43.5%) while the negative predictive value (NPV) was 4.8% (95% CI: −2.3–11.8%).

Following prompt 2, for patients with true strabismus, there were 16 true positive responses and 8 false negative responses (Figure 1). For patients with pseudoesotropia there were 10 false positive responses and one true negative response. For prompt 2, one case provided the diagnosis of strabismus with incorrect direction. An image of exotropia was characterized by GPT-4o as esotropia and was marked as an incorrect diagnosis in the analysis. The diagnostic accuracy following prompt 2 was 17/35 (48.6%). The sensitivity was 66.7% (95% CI: 51.1–82.3%) and the specificity was 9.1% (95% CI: −0.4–18.6%). The PPV was 61.5% (95% CI: 45.4–77.7%) while the NPV was 11.1% (95% CI: 0.7–21.5%). After Prompt 2, 9.1% of pseudoesotropia and 70.1% of strabismus were correctly diagnosed (p=0.04).

Prompt 2 provided more context to GPT-4o as may be expected of a patient's parent who is turning to the platform due to specific observations and concerns. When eye misalignment is directly volunteered in the prompt, GPT-4o analyzed and responded to more images. The accuracy increased from prompt 1 but remained poor, at 48.6%. When specific concern was raised, there were fewer strabismus misses compared to responses to prompt 1.

Secondary "Per-Diagnosis" Accuracy Analysis

Following prompt 1, for patients with true strabismus, there were 4 true positive responses and 10 false negative responses. For patients with pseudoesotropia there were zero false positive responses and one true negative response. The

Table 5 Assessment of GPT-4o Response Content

	Overall Score		Assessment Score		Incorrect Content Score		Potential for Harm Score	
	Mean ± SE	P-value	Mean ± SE	P-value	Mean ± SE	P-value	Mean ± SE	P-value
Strabismus	4.94 ± 0.27	0.638	1.52 ± 0.11	0.609	1.81 ± 0.10	0.340	1.60 ± 0.12	0.0291
Pseudostrabismus	5.14 ± 0.18		1.41 ± 0.20		1.68± 0.10		2.00 ± 0.00	

diagnostic accuracy following prompt 1 was 5/15 (33.3%). The sensitivity was 28.6% (95% CI: 5.7–51.4%) and the specificity was 100.0% (95% CI: 100.0–100.0%). The PPV was 100.0% (95% CI: 100.0–100.0%) while the NPV was 9.1% (95% CI: −5.5–23.6%).

Following prompt 2, for patients with true strabismus, there were 16 true positive responses and 4 false negative responses (Figure 1). For patients with pseudoesotropia there were six false positive responses and one true negative response. The diagnostic accuracy following prompt 2 was 17/27 (63.0%). The sensitivity was 80.0% (95% CI: 64.9–95.1%) and the specificity was 14.3% (95% CI: 14.3–27.5%). The PPV was 72.7% (95% CI: 55.9–89.5%) while the NPV was 20.0% (95% CI: 4.9–35.1%).

Content Analysis

Table 5 assesses the quality of the content of GPT-4o's responses by two graders. The mean difference between overall scores between graders 1 and 2 is 0.97. Inter-rater agreement was mild overall with the strongest agreement for the potential for harm subscore (Table S1). Power calculations were not performed a priori and thus the content analysis is limited by small sample size. The mean overall score for images with strabismus is 4.94 ± 0.27 and for images with pseudostrabismus is 5.14 ± 0.18 (p=0.638). With a total possible score of 6, the overall quality was good though still below an acceptable quality for that provided by a clinic encounter. For the assessment subscore, the mean for images with strabismus is 1.52 ± 0.11 and for images with pseudostrabismus is 1.41 ± 0.20 (p=0.609). For the incorrect content subscore, the mean for images with strabismus is 1.81 ± 0.10 and for images with pseudostrabismus is 1.68 ± 0.10 (p=0.340).

Typically advice was consistent for either true strabismus or pseudostrabismus diagnosis, except for the potential for harm subscore where there was a significantly higher potential for harm given to strabismus images. Within Table 5, the subscore based on the potential for harm of the response has a mean of 1.60 ± 0.12 for images with strabismus and 2.00 ± 0.00 (p=0.029) for images with pseudostrabismus. The statistical significance of this result demonstrates a higher potential for harm for responses to images of strabismus than images of pseudostrabismus. This may be due to the inherently greater potential for harm in responses to patients with strabismus who are falsely reassured than the inverse case for pseudostrabismus.

Discussion

We evaluated the accuracy of GPT-4o in the diagnosis of strabismus based on mobile photos, and graded the model's responses on several criteria that assessed appropriateness compared to counseling currently provided in clinic by pediatric ophthalmologists and/or certified orthoptists. Overall, GPT-4o demonstrated poor accuracy and, though the quality of counseling in responses was good overall, responses had a potential for harm as compared to counseling provided when patients were properly evaluated in clinic.

The strength of our study lies in its comparison to the literature on image-based diagnosis using generative AI models. Previous studies focus on diagnostic accuracy of the interpretation of medical imaging data, such as computed tomography or magnetic resonance imaging, or high-resolution magnified images generated by clinicians by modes such as dermatoscopy.[10–12] To our knowledge, our study is the first to assess the diagnostic accuracy of a generative AI model in making diagnoses based on images that can and will be generated by patients. With the increased popularity of generative AI, our study allows for assessment of its accuracy from a real world, ground-level perspective.

The "intention-to-diagnose" analysis diagnostic accuracies of 14.3% following prompt 1 and 48.6% following prompt 2 and the more optimistic "per-diagnosis" accuracies of 33.3% and 66.7% are comparable to literature assessing the GPT-4o's diagnostic ability for image-based data. Other studies assessing image-based diagnosis find accuracies at approximately 50%,[10–12] while those primarily using text input show higher accuracy.[9,13] While the accuracy is similar to our expectations based on literature review, we were surprised to see the difference in accuracy between responses provided after prompt 1 versus after prompt 2. Underlying this difference is an increase in false positives, or strabismus overcalls, and decrease in false negatives, or strabismus misses, for the second prompt compared to the first. This suggests that GPT-4o's response to image analysis are suggestible by the prompt. This prompt sensitivity is likely to lead to more false positives and distress to anxious families, while providing false reassurance to potentially more vulnerable families who provide less detail. In the clinical use, these results will help us reassure the former families and educate the latter families about the limitations of GPT-4o.

The sensitivity and specificity of GPT-4o diagnosis in the primary and secondary analyses varied between prompts 1 and 2. While there is uncertainty to the 95% confidence intervals due to a low sample size, in the primary analysis prompt 1 had low sensitivity and specificity while prompt 2 had moderate sensitivity and low specificity. In the secondary analysis, prompt 1 had low sensitivity but high specificity while prompt 2 had high sensitivity and low specificity. This reinforces differences in prompt sensitivity when provided with a neutral (prompt 1) versus alignment-priming (prompt 2) prompt.

While the responses provided by GPT-4o were rated as 5 out of a total 6, representing a good score, this still does not replicate or replace appropriate counseling provided by a pediatric ophthalmologist or certified orthoptist. An upside for patient safety is that GPT-4o was typically tentative to volunteer a diagnosis and, in all cases, counseled the patient to see a pediatric ophthalmologist for a definitive exam. However, there was a significant difference in the potential for harm from GTP-4o's responses, which was deemed greater for patients with strabismus than those with pseudostrabismus. For patients with strabismus in which the diagnosis was missed and patients were falsely reassured, visual development may be threatened if their condition remains untreated, causing permanent visual disability. Even though GPT-4o recommends an in-person examination, the reassurance alone could be enough to delay care and compromise vision. On the other hand, patients with pseudostrabismus in which GPT-4o raised concern for strabismus also face a potential for harm that was seen to be lesser based on our grading. This harm comes from increased resource utilization amongst a pediatric ophthalmology shortage and increased patient anxiety in the setting of long wait times and a potentially high burden to seek medical care.[16] While our work found that the responses were reasonable, the potential for harm in both false positive and negative cases is important for families to understand before turning to an internet-based resource for medical information.

Our study faces several limitations within the context of which its results should be viewed. First are limitations related to the inputted images. While we standardized the key anatomical features that were submitted to the GPT-4o model, we were unable to standardize other factors such as lighting and angle which may influence the model. Figure 1 demonstrates another limitation related to inputted images. This patient has a focal area of conjunctival redness. We were unable to standardize for cues that the model may fixate on including eye redness and pupil or eyelid asymmetry, distracting from an ocular alignment assessment. Second, our study is limited by the lack of information provided to the model. Strabismus is a diagnosis that requires evaluation at several focal points, several fields of gaze, and often requires dissociation. Lastly, due to a small sample size, the strength of our conclusions about diagnostic accuracy and the power of the content analysis are limited.

This study provides preliminary insight into the performance of GPT-4o in the analysis of eye alignment based on images that may be submitted by patients and their families. The performance is poor, and the responses hold potential to inflict harm both at the individual and systemic levels. Further research will be needed as subsequent versions are released and as new generative AI platforms gain popularity amongst patients. While diagnostic accuracy will likely increase in future platforms, it is important to continue to assess these models so that patients are adequately informed on how to interpret the responses they receive.

Conclusions

Our preliminary evaluation of the diagnostic accuracy of GPT-4o for strabismus and pseudostrabismus based on mobile images reveals low accuracy. The accuracy is even lower when considering all submitted images in the "intention-to-diagnose" analysis. GPT-4o frequently missed true cases of strabismus, an outcome that threatens the patient's visual development. Due to the suggestible nature of GPT-4o, there was overdiagnosis once a prompt asked about eye misalignment. While the counseling provided by GPT-4o was often deemed appropriate by a pediatric ophthalmologist and orthoptist, false reassurance in the case of misdiagnoses leaves room for potential future harm. Despite its popularity and accessibility to patients, GPT-4o should not serve as a replacement screening tool at this time. Clinicians, parents, and software developers should be aware of its limitations.

Institutional Review Board Statement

The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of Johns Hopkins School & Medicine (IRB00414185) on 12/06/2023.

Data Sharing Statement

The raw data supporting the conclusions of this article will be made available by the corresponding author on request.

Informed Consent Statement

Informed consent was obtained from all subjects involved in the study. Written informed consent has been obtained from the parent or legal guardian of the patient in Figure 1 to publish this picture.

Acknowledgments

We are grateful to World Pediatric Project (WPP) for their involvement in patient recruitment and attainment of consent. During the preparation of this study, the authors used ChatGPT (GPT-4o) for the purposes of data collection as described in the Materials and Methods section.

Funding

This research project did not receive specific funding from any source.

Disclosure

The authors declare no conflicts of interest.

References

1. Kraus C, Kuwera E. What is strabismus? *JAMA*. 2023;329:856. doi:10.1001/jama.2023.0052
2. Kelly KR, Pang Y, Thompson B, et al. Functional consequences of amblyopia and its impact on health-related quality of life. *Vision Res*. 2025;231:108612. doi:10.1016/j.visres.2025.108612
3. Finney Rutten LJ, Blake KD, Greenberg-Worisek AJ, et al. Online health information seeking among US adults: measuring progress toward a healthy people 2020 objective. *Public Health Rep*. 2019;134:617–625. doi:10.1177/0033354919874074
4. Hello GPT-4o. Available from: https://perma.cc/67PU-L8UK. Accessed October 22, 2025.
5. OpenAI's weekly active users surpass 400 million. *Reuters*. 2025. Available from: https://perma.cc/3KN8-WZ7G. Accessed October 22, 2025.
6. How ChatGPT and our foundation models are developed. OpenAI Help Center. Available from: https://perma.cc/H7EJ-NZLJ. Accessed October 22, 2025.
7. Shu Q, Pang J, Liu Z, et al. Artificial intelligence for early detection of pediatric eye diseases using mobile photos. *JAMA Network Open*. 2024;7:e2425124. doi:10.1001/jamanetworkopen.2024.25124
8. Cardakli N, Wang B, Doyle JJ, Kraus CL. Accuracy of an artificial intelligence chatbot in identifying congenital glaucoma from other ocular etiologies. *Digit J Ophthalmol*. 2025;31. doi:10.5693/djo.01.2025.03.003
9. Hirosawa T, Kawamura R, Harada Y, et al. ChatGPT-generated differential diagnosis lists for complex case-derived clinical vignettes: diagnostic accuracy evaluation. *JMIR Med Inform*. 2023;11:e48808. doi:10.2196/48808
10. Ueda D, Mitsuyama Y, Takita H, et al. ChatGPT's diagnostic performance from patient history and imaging findings on the diagnosis please quizzes. *Radiology*. 2023;308:e231040. doi:10.1148/radiol.231040
11. Horiuchi D, Tatekawa H, Shimono T, et al. Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. *Neuroradiology*. 2024;66:73–79. doi:10.1007/s00234-023-03252-4

12. Shifai N, van Doorn R, Malvehy J, Sangers TE. Can ChatGPT vision diagnose melanoma? An exploratory diagnostic accuracy study. *J Am Acad Dermatol*. 2024;90:1057–1059. doi:10.1016/j.jaad.2023.12.062
13. Balas M, Ing EB. Conversational AI models for ophthalmic diagnosis: comparison of ChatGPT and the Isabel pro differential diagnosis generator. *JFO Open Ophthalmol*. 2023;1:100005. doi:10.1016/j.jfop.2023.100005
14. Chen JS, Reddy AJ, Al-Sharif E, et al. Analysis of ChatGPT responses to ophthalmic cases: can ChatGPT think like an ophthalmologist? *Ophthalmol Sci*. 2025;5:100600. doi:10.1016/j.xops.2024.100600
15. Anagnostidis S, Bulian J. How susceptible are LLMs to influence in prompts? 2024. doi:10.48550/ARXIV.2408.11865.
16. Walsh HL, Parrish A, Hucko L, Sridhar J, Cavuoto KM. Access to pediatric ophthalmological care by geographic distribution and US population demographic characteristics in 2022. *JAMA Ophthalmol*. 2023;141:242–249. doi:10.1001/jamaophthalmol.2022.6010