Review

Evaluating Large Language Models in Ophthalmology: Systematic Review

Zili Zhang[1*], BE; Haiyang Zhang[1*], MD; Zhe Pan[2]; Zhangqian Bi[2], BE; Yao Wan[2], PhD; Xuefei Song[1], MD, PhD; Xianqun Fan[1], MD, PhD

[1]State Key Laboratory of Eye Health, Department of Ophthalmology, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[2]School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China

[*]these authors contributed equally

Corresponding Author:
Xianqun Fan, MD, PhD
State Key Laboratory of Eye Health, Department of Ophthalmology
Shanghai Ninth People's Hospital
Shanghai Jiao Tong University School of Medicine
639 Zhizaoju Road
Huangpu District
Shanghai, 200011
China
Phone: 86 21 23271699
Email: fanxq@sjtu.edu.cn

Abstract

Background: Large language models (LLMs) have the potential to revolutionize ophthalmic care, but their evaluation practice remains fragmented. A systematic assessment is crucial to identify gaps and guide future evaluation practices and clinical integration.

Objective: This study aims to map the current landscape of LLM evaluations in ophthalmology and explore whether performance synthesis is feasible for a common task.

Methods: A comprehensive search of PubMed, Web of Science, Embase, and IEEE Xplore was conducted up to November 17, 2024 (no language limits). Eligible publications quantitatively assessed an existing or modified LLM on ophthalmology-related tasks. Studies without full-text availability or those focusing solely on vision-only models were excluded. Two reviewers screened studies and extracted data across 6 dimensions (evaluated LLM, data modality, ophthalmic subspecialty, medical task, evaluation dimension, and clinical alignment), and disagreements were resolved by a third reviewer. Descriptive statistics were analyzed and visualized using Python (with *NumPy*, *Pandas*, *SciPy*, and *Matplotlib* libraries). The Fisher exact test compared open- versus closed-source models. An exploratory random-effects meta-analysis (logit transformation; DerSimonian-Laird τ^2) was performed for the diagnosis-making task; heterogeneity was reported with I^2 and subgrouped by model, modality, and subspecialty.

Results: Of the 817 identified records, 187 studies met the inclusion criteria. Closed-source LLMs dominated: 170 for ChatGPT, 58 for Gemini, and 32 for Copilot. Open-source LLMs appeared in only 25 (13.4%) of studies overall, but they appeared in 17 (77.3%) of evaluation-after-development studies, versus 8 (4.8%) pure-evaluation studies ($P<1\times10^{-5}$). Evaluations were chiefly text-only (n=168); image-text tasks, despite the centrality of imaging, were used in 19 studies. Subspecialty coverage was skewed toward comprehensive ophthalmology (n=72), retina and vitreous (n=32), and glaucoma (n=20). Refractive surgery, ocular pathology and oncology, and ophthalmic pharmacology each appeared in 3 or fewer studies. Medical query (n=86), standardized examination (n=41), and diagnosis making (n=29) emerged as the 3 predominant tasks, while research assistance (n=5), patient triaging (n=3), and disease prediction (n=3) received less attention. Accuracy was reported in most studies (n=176), whereas calibration and uncertainty were almost absent (n=5). Real-world patient data (n=45), human performance comparison (n=63), non-English testing (n=24), and real-world deployment (n=4) were relatively absent. Exploratory meta-analysis pooled 28 diagnostic evaluations from 17 studies: overall accuracy was 0.594 (95% CI 0.488-0.692) with extreme heterogeneity (I^2=94.5%). Subgroups remained heterogeneous (I^2>80%), and findings were inconsistent (eg, pooled GPT-3.5>GPT-4).

Conclusions: Evidence on LLM evaluations in ophthalmology is extensive but heterogeneous. Most studies have tested a few closed-source LLMs on text-based questions, leaving open-source systems, multimodal tasks, non-English contexts, and real-world deployment underexamined. High methodological variability precludes meaningful performance aggregation, as illustrated by the heterogeneous meta-analysis. Standardized, multimodal benchmarks and phased clinical validation pipelines are urgently needed before LLMs can be safely integrated into eye care workflows.

(J Med Internet Res 2025;27:e76947) doi: 10.2196/76947

KEYWORDS

ophthalmology; large language model; systematic review; meta-analysis; clinical evaluation; artificial intelligence

Introduction

Background

Driven by rapid advancements in natural language processing technology powered by artificial intelligence (AI), large language models (LLMs), such as ChatGPT (OpenAI), are revolutionizing health care. LLMs, trained on diverse, high-quality datasets, encode extensive knowledge, and generate humanlike responses [1], showing promise in disease prevention, diagnosis, treatment, caregiving, and education [2].

Ophthalmic diseases significantly impact global health [3,4], yet a shortage of ophthalmologists, exacerbated by aging populations, has widened the gap between health care demand and supply, particularly in low- and middle-income countries [5]. LLMs are seen as a potential solution to this resource shortage [6], with studies demonstrating their capabilities in ophthalmology. For instance, Antaki et al [7] found that ChatGPT with GPT-4 [8] achieved an accuracy of more than 70% on simulated ophthalmology board-style exams, outperforming historical human performance. Similarly, Bernstein et al [9] demonstrated that LLMs could appropriately respond to patients' eye health concerns, with response quality comparable to that of ophthalmologists.

However, LLMs are not always reliable [10]; they may generate convincing yet factually incorrect responses [9], posing risks to patient care. Thus, systematic evaluation of LLMs is essential before their clinical integration [6]. While numerous studies have assessed LLMs in ophthalmology [11-13], a comprehensive statistical analysis and synthesis of these evaluation practices is lacking. This gap hinders the standardization of future model evaluations, thereby postponing clinical deployment.

Aim and Scope of This Review

This study aims to systematically map the current landscape of LLM evaluations in ophthalmology. To achieve this, we first summarized existing LLM evaluation practices in ophthalmology across 6 dimensions: evaluated LLM, data modality, ophthalmic subspecialty, medical task, evaluation dimension, and clinical alignment. Moreover, we performed an additional meta-analysis of diagnostic performance evaluation, providing further evidence for the heterogeneity and fragmentation observed in current evaluation studies. By identifying gaps in current practices, we propose a standardized framework to guide future research and clinical implementation, ensuring the safe and effective integration of LLMs into ophthalmic care.

Methods

Design and Registration

A systematic review was conducted in accordance with the relevant sections of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) reporting guidelines [14], and the completed PRISMA checklist is provided in Multimedia Appendix 1.

Ethical Considerations

Ethics approval was not required for this study because it did not involve the recruitment of patients.

Search Strategy

Peer-reviewed studies and preprints were retrieved on November 17, 2024, from 4 databases, including PubMed, Web of Science, Embase, and IEEE Xplore. The key search strategies were based on a combination of 2 themes: LLM and ophthalmology, and synonyms or related terms of these 2 themes were also included in the search strategy to ensure a comprehensive coverage of related publications. The complete search strategies in 4 databases are listed in Multimedia Appendix 2 [15-33].

Study Selection

All records were screened in 3 sequential phases by 2 independent reviewers (ZZ and ZP). First, an automated title-matching script removed exact duplicates before any screening. Second, during title and abstract screening, reviewers applied the inclusion and exclusion criteria (Textbox 1) and manually discarded near-duplicate records (eg, preprint version of a published study). Third, full texts were assessed against the same criteria, and any additional duplicates were removed when the full-text content was found to be substantially identical despite different titles or abstracts. Disagreements were resolved through discussion, with a third reviewer (HZ) acting as arbiter when necessary.

Textbox 1. Inclusion and exclusion criteria for the systematic review of quantitative large language model (LLM) evaluations in ophthalmology.

Inclusion criteria

- Topic: ophthalmology-related tasks, questions, cases, or images
- Model type: original or modified LLMs (fine-tuned, retrieval-augmented generation enhanced [34], or pipeline integrated)
- Evaluation: 1 or more quantitative metrics (eg, accuracy, Likert scale, and readability score)
- Publication: peer-reviewed studies and preprints; any language

Exclusion criteria

- Topic: nonophthalmic content; no artificial intelligence involved
- Model type: traditional deep learning models, such as a convolutional neural network; vision-only generation models, such as DALL·E
- Evaluation: purely qualitative commentaries
- Publication: full text unavailable; preprint version of published study

In line with recent AI-focused reviews [35], we used domains that are more appropriate for AI model evaluations rather than the traditional patient-centered population, intervention, comparison, outcome, and study design headings.

Data Extraction and Analysis Framework

Overview

A wide range of data was extracted and summarized from included publications, mainly across 6 aspects: evaluated LLM, data modality, ophthalmic subspecialty, medical task, evaluation dimension, and clinical alignment. The data extraction was conducted manually by 2 independent reviewers (ZZ and ZP), and a short Python (Python Software Foundation) script was then used to compare the 2 extraction sheets and generate a list of discrepant entries for joint discussion and consensus. Data extraction outcomes are detailed in Multimedia Appendix 3 [2,7,9,11-13,15,16,36-197].

Evaluated LLM

We aggregated LLMs included in each study by model series (eg, grouping GPT-3.5-Turbo, GPT-4, and GPT-4o under "ChatGPT Series"), and we also consolidated models with shared technical lineages (eg, categorizing Pathways Language Model [198], Google Bard, and Gemini [199] under "Gemini Series"). Such categorization ensures consistent tracking of research attention despite frequent model updates. Furthermore, we classified each model as open-source LLM (eg, Large Language Model Meta AI [LLaMA] [200] and DeepSeek [201]) or closed-source LLM (eg, ChatGPT and Gemini) based on public availability of weights and code.

Data Modality

We categorized the evaluation modalities into 2 types: image-text and text-only. Studies involving any image-related questions were classified under the image-text category.

Ophthalmic Subspecialty

A comprehensive classification framework comprising 12 subspecialties and an additional category, "comprehensive ophthalmology," was established based on multiple authoritative sources, including the American Academy of Ophthalmology, the Royal College of Ophthalmologists, and relevant studies [202,203]. Studies were classified as "comprehensive ophthalmology" if they either did not target a single subspecialty or covered 3 or more distinct subspecialties (eg, cataract, glaucoma, and retina) without a predominant focus on any single domain.

Medical Task

Medical tasks across the included studies were classified into 9 distinct categories based on their objectives, target populations, and methodological approaches. This classification framework integrates insights from previous research [204,205], with empirical patterns observed in our literature corpus (refer to definitions in Table S1 in Multimedia Appendix 2).

Evaluation Dimension

We adopted the 7D framework proposed by Bedi et al [204] for testing LLMs in health care applications:

1. Accuracy—concordance between the model's output and a gold-standard reference
2. Comprehensiveness—the extent to which the output addresses all clinically relevant aspects of the prompt
3. Factuality—alignment with perceived consensus and correctness of any cited sources, such as publications and clinical guidelines
4. Robustness—stability of performance under input perturbations, such as typos and paraphrasing, or reproducibility of the answer under repeated identical queries
5. Fairness, bias, and toxicity—absence of harmful, discriminatory, or toxic content toward both majority and ethnic minority groups
6. Deployment metrics—practical considerations, such as latency, computational cost, and memory footprint
7. Calibration and uncertainty—alignment between predicted confidence and actual correctness

Furthermore, "readability and usability" was added as the eighth category to capture an evaluation item that appeared repeatedly in included studies but was not covered earlier. It reflects the linguistic clarity and clinical usability of model outputs for patients or clinicians. This dimension was often assessed with two complementary approaches: (1) objective formulas, most commonly the Simple Measure of Gobbledygook,

Flesch-Kincaid Grade Level, and Flesch-Kincaid Reading Ease and (2) subjective instruments, such as the Patient Education Materials Assessment Tool and Likert scale ratings completed by clinicians or patients. Any study using 1 or more of these measures (or equivalent methods) was classified under this dimension.

Clinical Alignment

Four dimensions assessed clinical alignment. The "language" dimension recorded the languages used in the evaluation to calculate the proportion of studies that included non-English assessments. The "real patient data included" dimension captured whether studies used real patient data, such as clinical information, examination results, or ophthalmic imaging. The "compared with human performance" dimension indicated whether studies compared the performance of LLMs with that of ophthalmologists. Human responses could either be generated specifically for the study or sourced from existing online content. Finally, the "real-world clinical assessment" dimension evaluated whether studies involved deployment, application, and assessment of LLMs in real clinical settings rather than limiting testing to virtual environments.

A study may contain multiple evaluated LLMs, ophthalmic subspecialties, medical tasks, evaluation dimensions, and languages. Multiple results were separated by a slash during data extraction, and each was included in the statistical results; therefore, the sum of their percentages exceeded 100%. "Uncertain" was used in data extraction when the result could not be determined from the full text of the paper.

Exploratory Meta-Analysis

We performed an exploratory meta-analysis limited to the diagnosis task, the only domain with relatively uniform methodology. Eligible studies were screened using 2 criteria: reporting an exact "correct or incorrect" proportion on open-ended diagnostic questions and avoiding multiple-choice or top-N scoring formats. The proportion of correct diagnoses was the primary effect size. Parallel model arms within a study were treated as independent evaluations. Data, including model, case count, modality, subspecialty, and accuracy, were extracted through dual-independent review with third-reviewer adjudication. Random-effects pooling (logit transformation; DerSimonian-Laird τ^2) and inverse-variance weighting were applied in Python (version 3.8.19) using NumPy (version 1.24.3) and SciPy (version 1.10.1) libraries, and heterogeneity was quantified with I^2. Subgroup analyses by LLM, modality, and subspecialty were performed to explore sources of heterogeneity.

Statistical Analysis and Visualization

Bar plots and stacked bar plots were used to visualize the distribution of relevant studies across different categories, and forest plots were generated for the exploratory meta-analysis. A 2-sided Fisher exact test compared the prevalence of open-source LLMs in "evaluation-after-development" studies versus "pure-evaluation" studies. Data statistics and visualization were conducted in Python (version 3.8.19) using NumPy (version 1.24.3), Pandas (version 2.0.3, SciPy (version 1.10.1), and Matplotlib (version 3.7.3) libraries.

Results

Overview

A total of 817 unique records were identified through our systematic search. After removing duplicates and unrelated papers based on title and abstract, 338 (41.4%) studies remained for full-text screening. Ultimately, a total of 187 (22.9%) studies met the study selection criteria and were included in the analysis (Figure 1). Key features of the included studies are described in Table 1.

Figure 1. PRISMA flow diagram of the article screening and identification process. The figure depicts identification, screening, eligibility, and inclusion of 187 peer-reviewed or preprint studies that quantitatively evaluated large language model (LLM) performance for ophthalmic tasks. Searches were conducted in PubMed, Web of Science, Embase, and IEEE Xplore on November 17, 2024, with no language limits. Reasons for exclusion at each stage are shown.

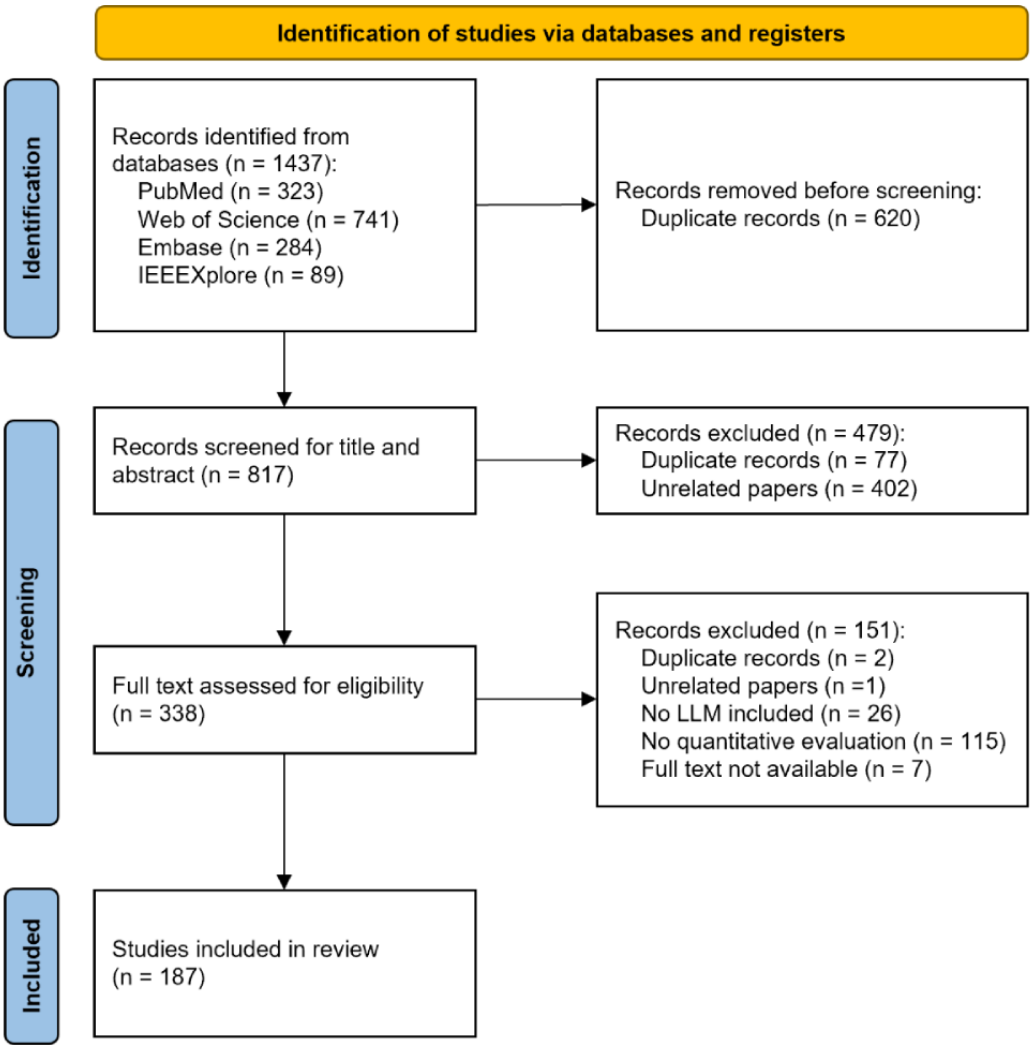


Table 1. Key features of included publications in this study (N=187).

Category and subtype	Studies, n (%)
Publication venue	
Peer-reviewed journal	177 (94.7)
Preprint platform (eg, arXiv and medRxiv)	9 (4.8)
Conference	1 (0.5)
Publication year	
2024	143 (76.5)
2023	44 (23.5)
Evaluation type	
Pure evaluation[a]	165 (88.2)
Evaluation after development[b]	22 (11.8)

[a]Evaluating base large language models without any architectural modification (eg, ChatGPT via its web interface).

[b]Evaluating large language models that were fine-tuned, retrieval-augmented generation enhanced, or integrated into new pipelines within the same study.

Evaluated LLM

Closed-source LLMs overwhelmingly dominated the included studies in this review. In total, 169 (90.4%) of the 187 studies evaluated 1 to 3 models, and more than half (n=100, 53.5%) compared multiple LLMs in parallel (Figure S1 in Multimedia Appendix 2). The 3 most frequently assessed LLM series were all closed-source ones—ChatGPT (n=170, 90.9%), Gemini (n=58, 31%; including Pathways Language Model, Bard, and Gemini), and Copilot (n=32, 17.1%; including Copilot [206] and Bing AI). In contrast, open-source LLMs were only evaluated in 25 (13.4%) studies, for example, LLaMA (n=15, 8%) and ChatGLM [207] (n=6, 3.2%; Table 2).

Table 2. Study characteristics for ophthalmology large language model (LLM) evaluations—evaluated LLM, data modality, and ophthalmic subspecialty (N=187).

Category and subtype	Studies[a], n (%)
LLM series (top 5 most frequently evaluated)	
ChatGPT	170 (90.9)
Gemini	58 (31)
Copilot	32 (17.1)
LLaMA[b]	15 (8)
ChatGLM	6 (3.2)
Data modality	
Text-only	168 (89.8)
Image-text	19 (10.2)
Ophthalmic subspecialty	
Comprehensive ophthalmology	72 (38.5)
Retina and vitreous	32 (17.1)
Glaucoma	20 (10.7)
External disease and cornea	12 (6.4)
Pediatric ophthalmology and strabismus	11 (5.9)
Lens and cataract	8 (4.3)
Oculoplastics and orbit	8 (4.3)
Uveitis and ocular inflammation	7 (3.7)
Neuro-ophthalmology	7 (3.7)
Clinical optics and vision rehabilitation	6 (3.2)
Refractive surgery	3 (1.6)
Ophthalmic pharmacology	2 (1.1)
Ocular pathology and oncology	2 (1.1)

[a]The sum of percentages of evaluated LLM series and ophthalmic subspecialties exceeds 100% because a study may be categorized into more than 1 classification (eg, a study evaluated both ChatGPT and Gemini).

[b]LLaMA: Large Language Model Meta AI.

Notably, open-source models accounted for a substantially larger share of "evaluation-after-development" studies than "pure-evaluation" studies (17/22, 77.3% vs 8/165, 4.8%; $P<1\times10^{-5}$, Fisher exact test; Figure S2 in Multimedia Appendix 2). Detailed definitions of the 2 study types (Table 1) demonstrate superior domain-specific versatility in ophthalmology.

Data Modality

Image-based evaluations were uncommon. Only 19 (10.2%) of the 187 studies assessed the ability of LLMs to process images and text in ophthalmology, whereas most (n=168, 89.8%) focused solely on text-only performance (Table 2). The ophthalmic images involved in these multimodality evaluations include, but are not limited to, slit lamp images, fundus photography of the posterior pole, optical coherence tomography, and ophthalmic ultrasonography. Other modalities such as voice, video, or documents were not included in any of the studies.

Ophthalmic Subspecialty

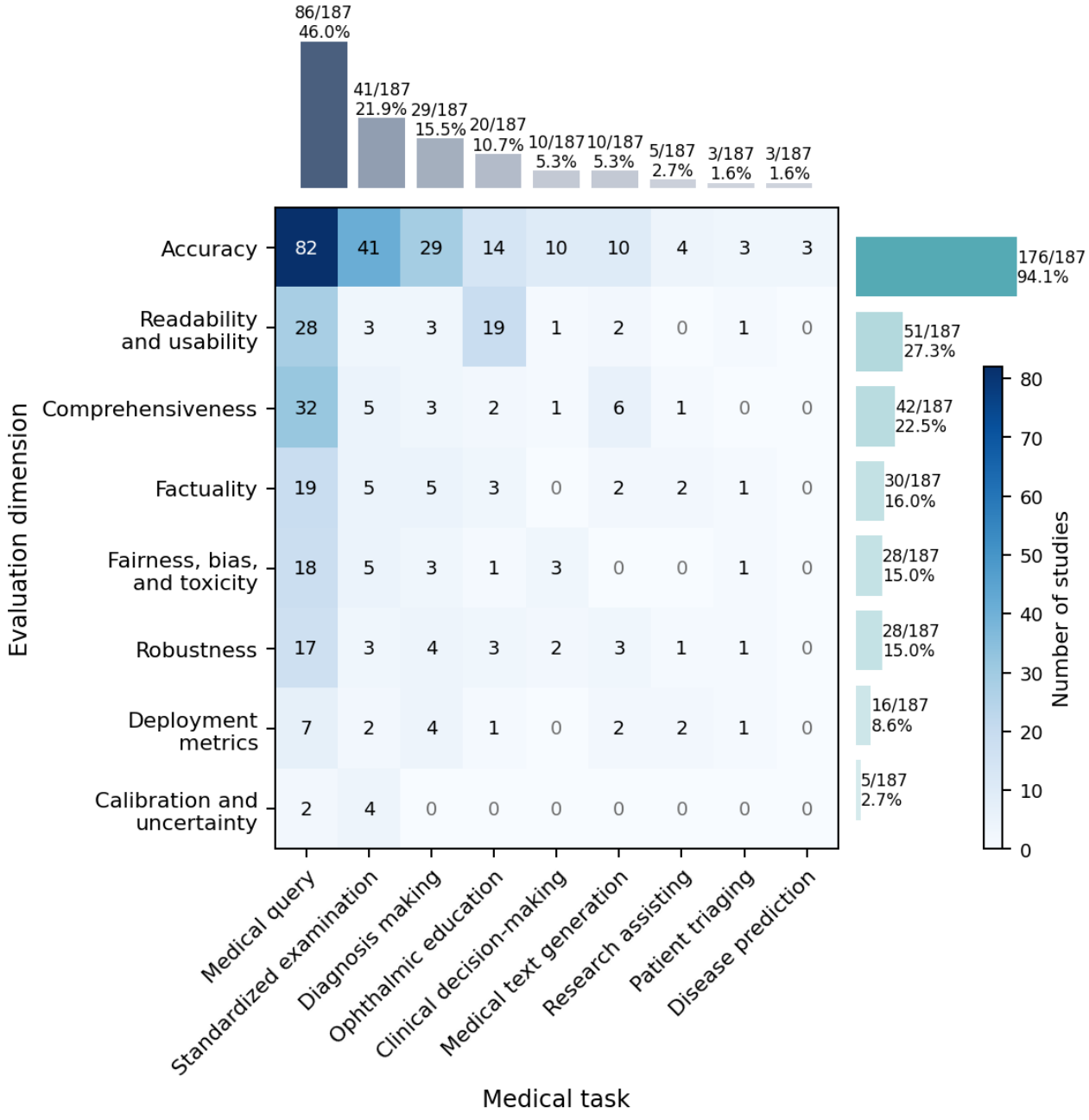
Evaluations clustered in comprehensive ophthalmology (72/187, 38.5%), leaving many subspecialties scarcely explored. Retina and vitreous (32/187, 17.1%) and glaucoma (20/187, 10.7%) followed at a distance, whereas ocular pathology and oncology,

refractive surgery, and ophthalmic pharmacology appeared in no more than 3 studies each (Table 2).

Medical Task

Medical task coverage was unevenly distributed. Most studies concentrated on medical query (86/187, 46.0%), standardized examination (41/187, 21.9%), and diagnosis making (29/187, 15.5%) in ophthalmology, and other common tasks explored (>5%) included ophthalmic education, clinical decision-making, and medical text generation. In contrast, research assisting (5/187, 2.7%), patient triaging (3/187, 1.6%), and disease prediction (3/187, 1.6%) received comparatively less attention (Figure 2).

Figure 2. Heat map of evaluation dimension by medical tasks across 187 ophthalmology large language model studies.



Evaluation Dimension

Evaluation dimensions were concentrated and unevenly distributed. Among 187 included studies, 118 (63.1%) evaluated 2 or more dimensions, yet only 15 (8%) assessed 4 or more dimensions, demonstrating the comprehensiveness of the existing evaluation (Figure S3 in Multimedia Appendix 2). Accuracy (n=176, 94.1%) was the most prevalent evaluation dimension, whereas calibration and uncertainty (n=5, 2.7%) were the least frequently examined. Notably, specific medical tasks showed distinct preferences for evaluation dimensions. For example, among 20 studies on ophthalmic education, 19 (95.0%) assessed readability and usability (Figure 2).

Clinical Alignment

The use of real patient data was limited and often undocumented. As Figure 3A shows, of the 187 studies, 45 (24.1%) tested LLMs with real patient data, including clinical information,

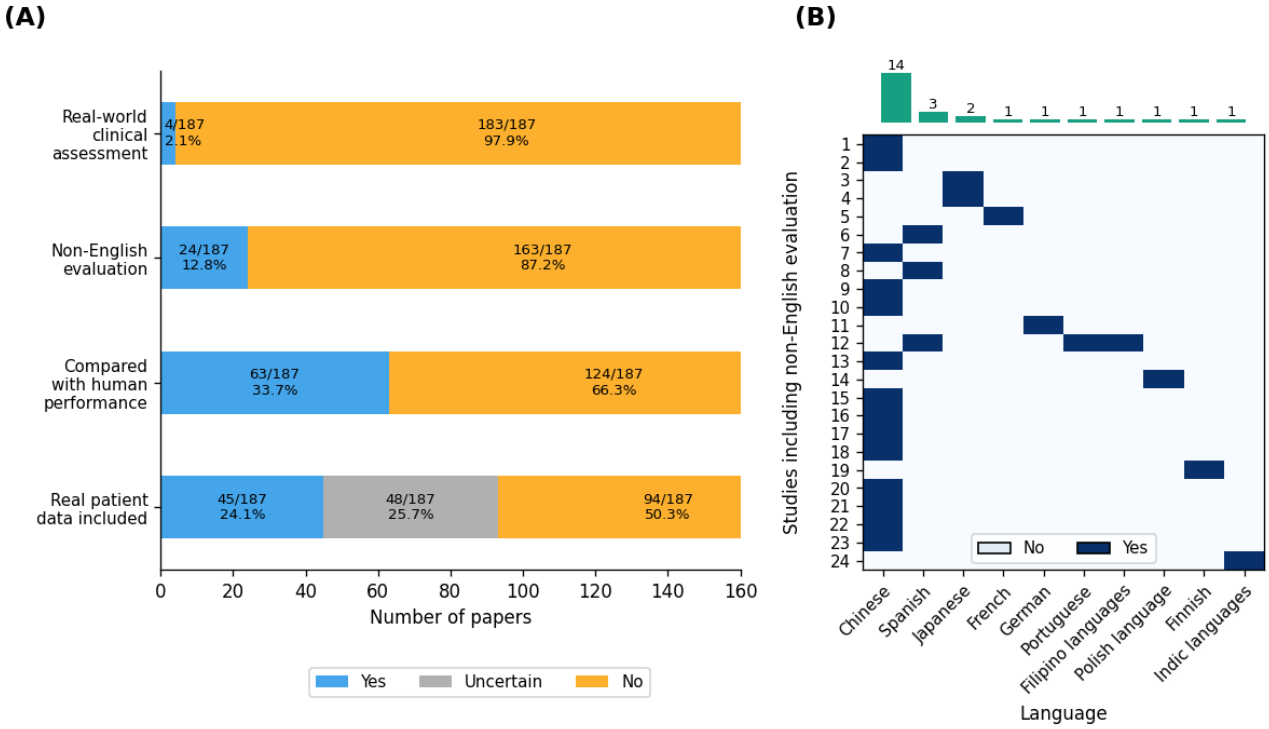
examination results, and imaging; 48 (25.7%) studies did not provide enough methodological detail to ascertain the data source; the remaining 94 (50.3%) studies used purely non–case-based ophthalmic problems or virtual patient data.

Comparisons between LLMs and humans were common but not universal. As Figure 3A shows, of the 187 studies, 63 (33.7%) compared LLM outputs with the performance of ophthalmologists at various training levels, providing a reference point for potential clinical integration.

Multilingual evaluations were scarce. As Figure 3A shows, only 24 (12.8%) of the 187 studies tested LLMs in non-English contexts. As Figure 3B shows, of these 24 studies, Chinese (n=14, 58.3%), Spanish (n=3, 12.5%), and Japanese (n=2, 8.3%) languages dominated these assessments; the remaining 7 languages (eg, French, German, and Finnish) appeared just once each.

Evaluations under real-world deployment were exceedingly rare. As Figure 3A shows, only 2.1% (4/187) of the studies evaluated LLMs in real ophthalmic clinical settings, while the remaining (183/187, 97.9%) studies were all conducted based on simulated clinical scenarios.

Figure 3. Clinical-alignment landscape of ophthalmology large language model (LLM) studies. (A) Shows the proportion of papers that did or did not involve real patient data, comparison with human performance, non-English evaluation, and in-clinic deployment in evaluation practice. (B) Among the 24 papers that conducted non-English evaluation, the heat map shows which of 10 language groups were tested (rows=studies), with the top bar summarizing language frequency.



Exploratory Meta-Analysis of Diagnosis-Making Studies

In total, 17 (58.6%) of the 29 diagnosis-making studies met our stricter inclusion criteria, providing 28 independent LLM evaluations of open-ended diagnostic accuracy (Figure S4 in Multimedia Appendix 2). Data extraction outcomes are detailed in Table S2 in Multimedia Appendix 2. Random-effects pooling yielded an overall correct-diagnosis proportion of 0.594 (95% CI 0.488-0.692) with extreme heterogeneity (I^2=94.5%). Forest plots are provided in Figure S5 in Multimedia Appendix 2. Subgroup pooling did not resolve this variability: I^2 values remained greater than 80% for most models, modalities, and subspecialty strata (Table 3).

MOPH achieved the highest accuracy for diagnosis, though this estimate was based on a single evaluation. Counterintuitively, GPT-4 (pooled accuracy=0.559) underperformed compared to GPT-3.5 (pooled accuracy=0.649). Across modalities, the pooled accuracy for image-text tasks (0.407) was lower than for text-only tasks (0.619). Five subspecialties and “comprehensive ophthalmology” were included in the subspecialty analysis; pooled estimates ranged from 0.483 for retina and vitreous to 0.731 for external disease and cornea, all with wide confidence intervals.

Table 3. Pooled overall and subgroup diagnostic accuracy.

	Pooled accuracy (95% CI)	I^2	Evaluations, n[a]
Overall	0.594 (0.488-0.692)	*94.5* [b]	28
Subgroups: model			
MOPH[c]	0.811 (0.747-0.862)	—[d]	1
GPT-3.5	0.649 (0.503-0.772)	*90.6*	12
GPT-4	0.559 (0.402-0.704)	*93.2*	9
Bing Copilot	0.526 (0.275-0.765)	24.1	2
Bard	0.438 (0.333-0.547)	—	1
Gemini	0.421 (0.226-0.644)	—	2
Glass 1.0	0.333 (0.084-0.732)	—	1
Subgroup: modality			
Text-only	0.619 (0.522-0.708)	*91.9*	25
Image-text	0.407 (0.123-0.770)	*96.9*	3
Subgroup: subspecialty			
External disease and cornea	0.731 (0.427-0.909)	66	2
Glaucoma	0.727 (0.414-0.910)	—	1
Comprehensive ophthalmology	0.601 (0.503-0.692)	*86.2*	12
Uveitis and ocular inflammation	0.589 (0.468-0.701)	—	5
Neuro-ophthalmology	0.547 (0.369-0.715)	51.9	5
Retina and vitreous	0.483 (0.087-0.901)	*99.4*	3

[a]Represents the number of evaluations included in the analysis and not the number of studies.

[b]Italics indicate I^2 values of more than 80%, indicating substantial heterogeneity.

[c]MOPH: LLM of ophthalmology.

[d]For a single line (k=1), I^2 cannot be calculated.

Discussion

Principal Findings

Our synthesis of 187 studies shows a skewed evaluation landscape; certain aspects, such as question-and-answer tasks, comprehensive ophthalmology, and accuracy, dominate, whereas image-based testing, non-English settings, and low-incidence subspecialties (eg, ocular oncology [208]) are rarely explored. Commonly studied areas share 3 traits: broad task generality, easily accessible test data, and high disease prevalence [209-211]. Neglected topics often lack public datasets and have smaller specialist communities [212,213].

Few (19/187, 10.2%) studies included images despite imaging being central to eye care [214]; technical barriers (clinically usable multimodal LLMs only began to emerge in 2023 and are still maturing [8,199,215,216]), privacy concerns, and data-sharing problems all contribute to this low inclusion rate [217]. Open-source LLMs were underrepresented despite their several irreplaceable advantages over closed-source ones in clinical settings (Table 4), yet they appeared much more frequently after new model development, confirming their suitability for customization for specific ophthalmic tasks (eg, fundus image analysis [36-38]). Unlike pure evaluation studies that mainly relied on manual assessment with Likert scales [2,15,39], these postdevelopment evaluations used automated metrics, such as Bilingual Evaluation Understudy [218] and F_1-score, which are standard metrics in AI and natural language processing domains [40,41], enabling rapid testing after LLM customization.

Profound heterogeneity was observed in current evaluation practice and evidence, driven by 3 factors. The first factor was metric heterogeneity. Even within our unified 8D framework, measures under the same heading were not directly comparable; for example, “accuracy” could be reported as an exact-correct ratio, area under the curve, F_1-score, or a Likert scale score with varying anchors. The second factor was data heterogeneity. Assessment sets differed widely; in the “medical query” task, questions spanned multiple difficulty levels, originated from diverse sources, and covered numerous ophthalmic subspecialties. The third factor was model heterogeneity. LLMs are updated frequently (eg, GPT-4o 2024-05-13, 2024-08-06, 2024-11-20), yet most papers note only the major version, obscuring performance shifts between iterations.

Table 4. Comparison of characteristics between open-source large language models (LLMs) and closed-source LLMs.

Key aspect	Open-source LLMs	Closed-source LLMs
Examples	LLaMA[a] and DeepSeek	ChatGPT and Gemini
Costs	Free of charge	Charge occasionally, depending on use
Performance	Remains a performance gap compared to closed-source models, though significant progress has been made in narrowing it [219]	State-of-the-art performance
Data privacy	Enables on-premises data processing, keeping sensitive information internal	Data are transmitted through cloud APIs[b], posing risks of third-party exposure
Accessibility	Available to the public, allowing free downloading, deployment, and use	Only available for use without providing access to their underlying code
Customizability	Tailorable for medical terminology, imaging diagnostics, and hospital workflows [216,220,221]	Limited adaptation to special medical needs (eg, 3D CT[c] or MRI[d] scan analysis)
Interactive interface	Typically provides basic interfaces, with advanced features dependent on community-driven tools or custom development efforts	Well-developed and feature-rich interfaces

[a]LLaMA: Large Language Model Meta AI.

[b]API: application programming interface.

[c]CT: computed tomography.

[d]MRI: magnetic resonance imaging.

To probe whether pooling was feasible despite this variability, we focused on the methodologically most consistent task, diagnosis making, as more than half of these studies (17/29, 59%) used open-ended questions with binary correct or incorrect scoring. However, the pooled accuracy of 0.594 still demonstrated extreme heterogeneity (I^2=94.5%), indicating that this aggregate statistic failed to represent almost any specific clinical context. Subgroup pooling offered little relief. Some findings were merely self-evident (image-text tasks were harder than text-only tasks), others were contradictory (GPT-3.5 pooled accuracy was higher than GPT-4, probably because 1 study of 422 unusually difficult cases tested GPT-4 alone [16]), and most confidence intervals remained wide. Finally, continuous, undocumented model updates, such as minor ChatGPT and Gemini releases, occur every few months, making any version-based synthesis obsolete almost as soon as it is published. Taken together, these observations confirm that high methodological heterogeneity, small task-specific samples, and model drift currently preclude meta-analytic results from informing bedside use or system refinement. Therefore, rigorous, standardized, and multidimensional benchmarks must precede any performance aggregation if future syntheses are to provide actionable clinical guidance in ophthalmology.

Comparison to Prior Work

Previous reviews focused on LLM applications in ophthalmology, but to the best of our knowledge, none systematically analyzed evaluation practice itself [6,222-224]. While cross-disciplinary surveys in medicine [204,205] flagged generic gaps in LLM evaluation, our study is the first to quantify them within ophthalmology.

We deepened the analysis by anchoring it in ophthalmology's distinctive context. Firstly, we gave special weight to multimodal (image-text) evaluation because diagnosis hinges on imaging. Moreover, we paid special attention to non-English evaluation to assess global applicability because eye care resources are unevenly distributed across language regions [225]. Finally, we divided the field into highly granular subspecialties (eg, retina, glaucoma, and ocular oncology), which revealed differences in the degree of attention given to different subspecialties. To capture cutting-edge technical advances, we also expanded our search strategy to include the IEEE Xplore augmentation database, which yielded more than a 5-fold increase in included ophthalmology literature (from 36 to 187 studies) compared to previous reviews [204].

Road to Real-World Deployment and Evaluation

Overview

Only 4 (2.1%) of the 187 ophthalmology studies deployed and evaluated LLMs in real clinical workflows [11,36,42,43], none as a randomized controlled trial (RCT), although RCTs are widely recognized as the gold standard for clinical evidence [226]. We have summarized the obstacles hindering the clinical deployment of LLMs into 10 points, as shown in Table 5. To bridge the LLM implementation-evaluation gap in real-world ophthalmology, we advocate a progressive 3-phase validation road map that keeps pace with rapid LLM iteration while safeguarding patients.

Table 5. Ten key obstacles to clinical large language model (LLM) deployment and evaluation.

Obstacle	Explanation	Illustrative example
Domain and modality limitations [227-231]	General-purpose text models often lack specialized medical knowledge or multimodal integration needed for comprehensive care.	A text-only LLM cannot incorporate fundus images, so it misclassifies macular edema severity compared with an ophthalmic multimodal model.
Lack of clinical validation [204,229,230]	Most LLM prototypes have never been tested in prospective, real-world patient care, so their safety and effectiveness remain unproven.	A hospital pilots an LLM for discharge summaries; after limited laboratory tests, it occasionally omits critical drug - interaction warnings that would be caught in a full clinical trial.
Hallucinations and errors [227-229,231,232]	LLMs can exhibit confidence while generating detailed content that is partially or entirely incorrect.	When asked for glaucoma referral criteria, the model invents a nonexistent "stage 4 angle-closure" classification.
Bias and fairness [227-229,232]	Training data–inherited biases can lead the model to provide systematically different or harmful recommendations for certain groups.	The LLM underrecommends diabetic retinopathy screening intervals for patients in minority groups because those groups were underrepresented in the fine-tuning data.
Data privacy and security [227-232]	Using patient text or images with external AI[a] services risks breaches of confidentiality and regulatory noncompliance.	Ophthalmologists paste private information into a cloud-hosted chatbot that stores inputs for retraining.
Transparency and explainability [227-230,232]	Clinicians and regulators cannot easily trace how or why the model produced a given answer, limiting trust and auditability.	A physician cannot determine which references the LLM used when it advises against a particular antibiotic, so the advice is ignored.
Regulatory gaps [227,229]	Existing medical device rules do not clearly cover general-purpose, continuously updated LLMs.	An updated model version "drifts" after release; no current pathway obliges the vendor to recertify performance each time weights change.
Legal and liability issues [229,231]	It is unclear who is responsible if an LLM's output harms a patient or infringes intellectual property rights.	After following chatbot-generated postoperative instructions, a patient is hospitalized, raising questions of clinician versus vendor liability.
Clinician and patient trust and adoption [232]	Users hesitate to rely on opaque tools that may err, so adoption stalls without clear evidence and oversight.	Nurses stop using an LLM triage assistant after noticing several inappropriate urgency ratings.
Human-AI interaction and usability [228,229]	Poor prompt design, ambiguous outputs, or workflow friction can negate any theoretical performance advantage.	A radiologist must craft complex prompts to extract a usable structured report, making the tool slower than manual dictation.

[a]AI: artificial intelligence.

Phase 1: Technical Benchmarking

In the first phase, one should start with an open, version-controlled benchmark of high quality, which could be described as an "OphthoQA" analogue to widely accepted medical LLM benchmarks, such as MedQA [233] or PubMedQA [234], but centered on multimodal ophthalmic tasks. The suite should mix board-style questions, deidentified ophthalmic images, and short clinical vignettes, all autoscored and posted to a public leaderboard so each new LLM update can be retested within hours, mitigating the reproducibility crisis that arises when LLMs change their answers over time [235], whereas traditional evaluation cycles take months to refresh the data.

Phase 2: Retrospective Validation

In the second phase, one should use large, deidentified electronic health record and image archives to compare LLM outputs (diagnoses, summaries, and triage labels) with gold-standard outcomes or expert panels. The assessment conducted by Sarvari et al [236] of LLMs on the MIMIC-IV [237] dataset exemplifies this approach, and ophthalmology already has smaller resources, such as OCTCases for real-world imaging tests [44]. On-premise inference with open-source LLMs allows each institution to assess its own patient data without private information leaving the firewall.

Phase 3: Prospective Trials

In the third phase, one should progressively embed the LLM in clinical workflow as follows: (1) silent-mode logging visible only to investigators; (2) pilot decision-support with mandatory human override; (3) full RCTs measuring diagnostic accuracy, workflow impact, and safety events, with real-time adverse-event monitoring. A recent 2-arm, open-label RCT of a fine-tuned mental health chatbot, featuring continual human review and immediate escalation of unsafe output, demonstrated a viable template [238]. This design offers a valuable reference that future ophthalmology-focused LLM trials can adapt to their specific clinical context. Only after a successful RCT should an LLM be adopted in routine eye care, ensuring maximum benefit and minimum harm to patients.

We hope this road map will provide valuable guidance for developing, evaluating, and refining ophthalmology-specific LLMs, ultimately enhancing ophthalmic clinical workflows and improving patient outcomes.

Future Directions

We propose 4 potential directions for future research. The first direction is the cocreation of a publicly available benchmark for LLMs in ophthalmology, covering all subspecialties, multiple languages, and image-text tasks. The second direction is the

adoption of a standard reporting checklist that requires multidimensional metrics (eg, accuracy, comprehensiveness, and safety) and specifies the exact LLM version. The third direction is continuous evaluation of newly launched LLMs, such as the pioneering reasoning model o1, especially open-source models amenable to ophthalmic fine-tuning (eg, DeepSeek and Qwen [239]). The fourth direction includes clinically driven assessments targeting real-world needs, such as referral triage, postoperative counseling, and low-resource language support, progressing through the 3-phase road map to RCTs.

Limitations

First, the search window closed on November 17, 2024, so publications released afterward, particularly work on the latest LLMs, were not captured. This was especially important given the rapid iteration of recent LLMs. We partially mitigated this by including preprints and 4 multidisciplinary databases, yet progress after the cutoff may have been underestimated.

Second, categorizing medical tasks and evaluation dimensions required subjective judgment; dual-independent extraction with third-reviewer adjudication reduced but did not eliminate misclassification risk.

Third, although we implemented stringent inclusion criteria, the absence of a formal risk-of-bias (RoB) assessment, reflecting the current lack of validated RoB tools specifically for medical AI studies, necessitated equal weighting of all included studies irrespective of methodological rigor. This limitation underscores the need for developing AI-specific RoB frameworks that enable future systematic reviews to appropriately weight evidence by study quality.

Fourth, extreme heterogeneity across studies precluded a definitive meta-analysis, highlighting the need for standardized evaluation protocols and datasets.

Conclusions

This systematic review mapped 187 ophthalmology-focused LLM evaluations and found a landscape that was extensive yet uneven. Research attention was concentrated on a few closed-source models, predominantly text-based tasks, and a limited set of subspecialties, medical tasks, and evaluation dimensions. Open-source LLMs, multimodal assessments, non-English testing, and real-world clinical studies were scarce. Fragmented methods yield heterogeneous evidence, impeding confident clinical adoption. These findings emphasize the necessity of standardized evaluation frameworks and highlight the critical gaps that must be closed before LLMs can be integrated into ophthalmic practice.

Acknowledgments

This study was funded by the Strategic Research and Consulting Project of the Chinese Academy of Engineering (2024-XBZD-18), National Natural Science Foundation of China (grants 82388101, 72293585, and 72293580), the Science and Technology Commission of Shanghai (20DZ2270800), the Shanghai Key Clinical Specialty, and Shanghai Eye Disease Research Center (2022ZZ01003).

Data Availability

All data generated or analyzed during this study are included in this published paper and its multimedia appendix files.

Authors' Contributions

ZZ contributed to conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing the original draft, and reviewing and editing the manuscript. HZ contributed to conceptualization, methodology, project administration, supervision, and reviewing and editing the manuscript. ZP contributed to data curation, investigation, software, and validation. ZB contributed to formal analysis, supervision, writing the original draft, and reviewing and editing the manuscript. YW contributed to supervision and reviewing and editing the manuscript. XS contributed to conceptualization, funding acquisition, resources, and supervision. XF contributed to funding acquisition, resources, and supervision.

Conflicts of Interest

None declared.

Multimedia Appendix 1

PRISMA 2020 checklist.
[XLSX File (Microsoft Excel File), 13 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Search strategies and supplementary figure and table.
[DOCX File , 490 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Data extraction results.
[XLSX File (Microsoft Excel File), 74 KB-Multimedia Appendix 3]

References

1. Han X, Zhang Z, Ding N, Gu Y, Liu X, Huo Y, et al. Pre-trained models: past, present and future. AI Open. 2021;2:225-250. [FREE Full text] [doi: 10.1016/j.aiopen.2021.08.002]
2. Lim ZW, Pushpanathan K, Yew SM, Lai Y, Sun CH, Lam JS, 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. [doi: 10.1016/j.ebiom.2023.104770] [Medline: 37625267]
3. Wittenborn JS, Zhang X, Feagan CW, Crouse WL, Shrestha S, Kemper AR, et al. The economic burden of vision loss and eye disorders among the United States population younger than 40 years. Ophthalmology. Sep 2013;120(9):1728-1735. [FREE Full text] [doi: 10.1016/j.ophtha.2013.01.068] [Medline: 23631946]
4. Wang B, Congdon N, Bourne R, Li Y, Cao K, Zhao A, et al. Burden of vision loss associated with eye disease in China 1990-2020: findings from the Global Burden of Disease Study 2015. Br J Ophthalmol. Feb 12, 2018;102(2):220-224. [doi: 10.1136/bjophthalmol-2017-310333] [Medline: 28607177]
5. Resnikoff S, Felch W, Gauthier TM, Spivey B. The number of ophthalmologists in practice and training worldwide: a growing gap despite more than 200,000 practitioners. Br J Ophthalmol. Jun 26, 2012;96(6):783-787. [doi: 10.1136/bjophthalmol-2011-301378] [Medline: 22452836]
6. Betzler BK, Chen H, Cheng CY, 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. [doi: 10.1016/s2589-7500(23)00201-7]
7. Antaki F, Milad D, Chia MA, Giguère CÉ, 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]
8. OpenAI. Gpt-4 technical report. ArXiv. Preprint posted online on March 15, 2023. [FREE Full text]
9. Bernstein IA, Zhang YV, Govil D, Majid I, Chang RT, Sun Y, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. JAMA Netw Open. Aug 01, 2023;6(8):e2330320. [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.30320] [Medline: 37606922]
10. Cadamuro J, Cabitza F, Debeljak Z, De Bruyne S, Frans G, Perez SM, et al. Potentials and pitfalls of ChatGPT and natural-language artificial intelligence models for the understanding of laboratory medicine test results. An assessment by the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Working Group on Artificial Intelligence (WG-AI). Clin Chem Lab Med. Jun 27, 2023;61(7):1158-1166. [FREE Full text] [doi: 10.1515/cclm-2023-0355] [Medline: 37083166]
11. Wang J, Shi R, Le Q, Shan K, Chen Z, Zhou X, et al. Evaluating the effectiveness of large language models in patient education for conjunctivitis. Br J Ophthalmol. Jan 28, 2025;109(2):185-191. [doi: 10.1136/bjo-2024-325599] [Medline: 39214677]
12. 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]
13. Restrepo D, Nakayama LF, Dychiao RG, Wu C, McCoy LG, Artiaga JC. Seeing beyond borders: evaluating LLMs in multilingual ophthalmological question answering. In: Proceedings of the IEEE 12th International Conference on Healthcare Informatics. 2024. Presented at: ICHI 2024; June.3-6, 2024; Orlando, FL. [doi: 10.1109/ichi61247.2024.00089]
14. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. Mar 29, 2021;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
15. Xu P, Chen X, Zhao Z, Shi D. Unveiling the clinical incapabilities: a benchmarking study of GPT-4V(ision) for ophthalmic multimodal image analysis. Br J Ophthalmol. Sep 20, 2024;108(10):1384-1389. [doi: 10.1136/bjo-2023-325054] [Medline: 38789133]
16. Milad D, Antaki F, Milad J, Farah A, Khairy T, Mikhail D, et al. Assessing the medical reasoning skills of GPT-4 in complex ophthalmology cases. Br J Ophthalmol. Sep 20, 2024;108(10):1398-1405. [doi: 10.1136/bjo-2023-325053] [Medline: 38365427]
17. 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 2023;12(6):3121-3132. [FREE Full text] [doi: 10.1007/s40123-023-00805-x] [Medline: 37707707]
18. 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]
19. 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]

20. Madadi Y, Delsoz M, Lao PA, Fong JW, Hollingsworth TJ, Kahook MY, et al. ChatGPT assisting diagnosis of neuro-ophthalmology diseases based on case reports. J Neuroophthalmol. Oct 10, 2024;45(3):301-306. [doi: 10.1097/WNO.0000000000002274] [Medline: 40830998]
21. 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]
22. Ghalibafan S, Taylor Gonzalez DJ, Cai LZ, 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]
23. 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]
24. Shukla R, Mishra AK, Banerjee N, Verma A. The comparison of ChatGPT 3.5, Microsoft Bing, and Google Gemini for diagnosing cases of neuro-ophthalmology. Cureus. Apr 2024;16(4):e58232. [FREE Full text] [doi: 10.7759/cureus.58232] [Medline: 38745784]
25. Ming S, Yao X, Guo X, Guo Q, Xie K, Chen D, et al. Performance of ChatGPT in ophthalmic registration and clinical diagnosis: cross-sectional study. J Med Internet Res. Nov 14, 2024;26:e60226. [FREE Full text] [doi: 10.2196/60226] [Medline: 39541581]
26. Hu X, Ran AR, Nguyen TX, Szeto S, Yam JC, Chan CK, 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]
27. 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]
28. Mandalos A, Tsouris D. Artificial versus human intelligence in the diagnostic approach of ophthalmic case scenarios: a qualitative evaluation of performance and consistency. Cureus. Jun 2024;16(6):e62471. [doi: 10.7759/cureus.62471] [Medline: 39015855]
29. Liu X, Wu J, Shao A, Shen W, Ye P, Wang Y, et al. Transforming retinal vascular disease classification: a comprehensive analysis of ChatGPT's performance and inference abilities on non-English clinical environment. medRxiv. Preprint posted online on June 29, 2023. [FREE Full text] [doi: 10.1101/2023.06.28.23291931]
30. Sorin V, Kapelushnik N, Hecht I, Zloto O, Glicksberg BS, Bufman H, et al. GPT-4 multimodal analysis on ophthalmology clinical cases including text and images. medRxiv. Preprint posted online on November 27, 2023. [FREE Full text] [doi: 10.1101/2023.11.24.23298953]
31. Mihalache A, Huang RS, Mikhail D, Popovic MM, Shor R, Pereira A, et al. Interpretation of clinical retinal images using an artificial intelligence chatbot. Ophthalmol Sci. Nov 2024;4(6):100556. [FREE Full text] [doi: 10.1016/j.xops.2024.100556] [Medline: 39139542]
32. 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]
33. Gill GS, Blair J, Litinsky S. Evaluating the performance of ChatGPT 3.5 and 4.0 on StatPearls oculoplastic surgery text- and image-based exam questions. Cureus. Nov 2024;16(11):e73812. [doi: 10.7759/cureus.73812] [Medline: 39691123]
34. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020. Presented at: NIPS'20; December 6-12, 2020; Vancouver, BC.
35. Huang SC, Pareek A, Jensen M, Lungren MP, Yeung S, Chaudhari AS. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. NPJ Digit Med. Apr 26, 2023;6(1):74. [FREE Full text] [doi: 10.1038/s41746-023-00811-0] [Medline: 37100953]
36. Li J, Guan Z, Wang J, Cheung CY, Zheng Y, Lim LL, et al. Integrated image-based deep learning and language models for primary diabetes care. Nat Med. Oct 19, 2024;30(10):2886-2896. [doi: 10.1038/s41591-024-03139-8] [Medline: 39030266]
37. Deng Z, Gao W, Chen C, Niu Z, Gong Z, Zhang R, et al. OphGLM: an ophthalmology large language-and-vision assistant. Artif Intell Med. Nov 2024;157:103001. [FREE Full text] [doi: 10.1016/j.artmed.2024.103001] [Medline: 39490063]
38. 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;7:111. [doi: 10.1038/s41746-024-01101-z]
39. 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]

40. Haghighi T, Gholami S, Sokol JT, Kishnani E, Ahsaniyan A, Rahmanian H, et al. EYE-Llama, an in-domain large language model for ophthalmology. bioRxiv. Preprint posted online on May 22, 2025. [FREE Full text] [doi: 10.1101/2024.04.26.591355] [Medline: 38746183]
41. Chen X, Xu P, Li Y, Zhang W, Song F, He M, et al. ChatFFA: an ophthalmic chat system for unified vision-language understanding and question answering for fundus fluorescein angiography. iScience. Jul 19, 2024;27(7):110021. [FREE Full text] [doi: 10.1016/j.isci.2024.110021] [Medline: 39055931]
42. Shi R, Liu S, Xu X, Ye Z, Yang J, Le Q, et al. Benchmarking four large language models' performance of addressing Chinese patients' inquiries about dry eye disease: a two-phase study. Heliyon. Jul 30, 2024;10(14):e34391. [FREE Full text] [doi: 10.1016/j.heliyon.2024.e34391] [Medline: 39113991]
43. Ramjee P, Sachdeva B, Golechha S, Kulkarni S, Fulari G, Murali K, et al. CataractBot: an LLM-powered expert-in-the-loop chatbot for cataract patients. Proc ACM Interact Mob Wearable Ubiquitous Technol. Jun 18, 2025;9(2):1-31. [doi: 10.1145/3729479]
44. Mihalache A, Huang RS, Cruz-Pimentel M, Patil NS, Popovic MM, Pandya BU, et al. Artificial intelligence chatbot interpretation of ophthalmic multimodal imaging cases. Eye (Lond). Sep 22, 2024;38(13):2491-2493. [doi: 10.1038/s41433-024-03074-5] [Medline: 38649474]
45. 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]
46. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. Ophthalmol Sci. Dec 2023;3(4):100324. [FREE Full text] [doi: 10.1016/j.xops.2023.100324] [Medline: 37334036]
47. 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]
48. 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]
49. Huang X, Raja H, Madadi Y, Delsoz M, Poursoroush A, Kahook MY, et al. Predicting glaucoma before onset using a large language model chatbot. Am J Ophthalmol. Oct 2024;266:289-299. [doi: 10.1016/j.ajo.2024.05.022] [Medline: 38823673]
50. 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]
51. Oca MC, Meller L, Wilson K, Parikh AO, McCoy A, Chang J, et al. Bias and inaccuracy in AI Chatbot ophthalmologist recommendations. Cureus. Sep 2023;15(9):e45911. [FREE Full text] [doi: 10.7759/cureus.45911] [Medline: 37885556]
52. Ali MJ. ChatGPT and lacrimal drainage disorders: performance and scope of improvement. Ophthalmic Plast Reconstr Surg. 2023;39(3):221-225. [FREE Full text] [doi: 10.1097/IOP.0000000000002418] [Medline: 37166289]
53. Ferro Desideri L, Roth J, Zinkernagel M, Anguita R. "Application and accuracy of artificial intelligence-derived large language models in patients with age related macular degeneration". Int J Retina Vitreous. Nov 18, 2023;9(1):71. [FREE Full text] [doi: 10.1186/s40942-023-00511-7] [Medline: 37980501]
54. Dihan Q, Chauhan MZ, Eleiwa TK, Hassan AK, Sallam AB, Khouri AS, et al. Using large language models to generate educational materials on childhood glaucoma. Am J Ophthalmol. Sep 2024;265:28-38. [doi: 10.1016/j.ajo.2024.04.004] [Medline: 38614196]
55. 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]
56. Wu G, Lee DA, Zhao W, Wong A, Sidhu S. ChatGPT: is it good for our glaucoma patients? Front Ophthalmol (Lausanne). 2023;3:1260415. [FREE Full text] [doi: 10.3389/fopht.2023.1260415] [Medline: 38983063]
57. Gupta AS, Sulewski ME, Armenti ST. Performance of ChatGPT in cataract surgery counseling. J Cataract Refract Surg. Apr 01, 2024;50(4):424-425. [doi: 10.1097/j.jcrs.0000000000001345] [Medline: 38523277]
58. Tailor PD, Xu TT, Fortes BH, Iezzi R, Olsen TW, Starr MR, et al. Appropriateness of ophthalmology recommendations from an online chat-based artificial intelligence model. Mayo Clin Proc Digit Health. Mar 2024;2(1):119-128. [FREE Full text] [doi: 10.1016/j.mcpdig.2024.01.003] [Medline: 38577703]
59. Potapenko I, Malmqvist L, Subhi Y, Hamann S. Artificial intelligence-based ChatGPT responses for patient questions on optic disc drusen. Ophthalmol Ther. Dec 2023;12(6):3109-3119. [FREE Full text] [doi: 10.1007/s40123-023-00800-2] [Medline: 37698823]
60. Singh S, Djalilian A, Ali MJ. ChatGPT and ophthalmology: exploring its potential with discharge summaries and operative notes. Semin Ophthalmol. Jul 2023;38(5):503-507. [doi: 10.1080/08820538.2023.2209166] [Medline: 37133418]
61. Cardona G, Argiles M, Pérez-Mañá L. Accuracy of a large language model as a new tool for optometry education. Clin Exp Optom. Apr 2025;108(3):343-346. [doi: 10.1080/08164622.2023.2288174] [Medline: 38044041]

62. 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]
63. Ćirković A, Katz T. Exploring the potential of ChatGPT-4 in predicting refractive surgery categorizations: comparative study. JMIR Form Res. Dec 28, 2023;7:e51798. [FREE Full text] [doi: 10.2196/51798] [Medline: 38153777]
64. Raghu K, S Devishamani C, Rajalakshmi R, Raman R. The utility of ChatGPT in diabetic retinopathy risk assessment: a comparative study with clinical diagnosis. Clin Ophthalmol. 2023;17:4021-4031. [FREE Full text] [doi: 10.2147/OPTH.S435052] [Medline: 38164506]
65. Teebagy S, Colwell L, Wood E, Yaghy A, Faustina M. Improved performance of ChatGPT-4 on the OKAP examination: a comparative study with ChatGPT-3.5. J Acad Ophthalmol (2017). Jul 2023;15(2):e184-e187. [FREE Full text] [doi: 10.1055/s-0043-1774399] [Medline: 37701862]
66. Hua HU, Kaakour AH, Rachitskaya A, Srivastava S, Sharma S, Mammo DA. Evaluation and comparison of ophthalmic scientific abstracts and references by current artificial intelligence chatbots. JAMA Ophthalmol. Sep 01, 2023;141(9):819-824. [doi: 10.1001/jamaophthalmol.2023.3119] [Medline: 37498609]
67. Balas M, Janic A, Daigle P, Nijhawan N, Hussain A, Gill H, et al. Evaluating ChatGPT on orbital and oculofacial disorders: accuracy and readability insights. Ophthalmic Plast Reconstr Surg. 2024;40(2):217-222. [doi: 10.1097/IOP.0000000000002552] [Medline: 37989540]
68. Anguita R, Makuloluwa A, Hind J, Wickham L. Large language models in vitreoretinal surgery. Eye (Lond). Mar 2024;38(4):809-810. [doi: 10.1038/s41433-023-02751-1] [Medline: 37726334]
69. Schumacher I, Bühler VM, Jaggi D, Roth J. Artificial intelligence derived large language model in decision-making process in uveitis. Int J Retina Vitreous. Sep 11, 2024;10(1):63. [FREE Full text] [doi: 10.1186/s40942-024-00581-1] [Medline: 39261870]
70. Cappellani F, Card KR, Shields CL, Pulido JS, Haller JA. Reliability and accuracy of artificial intelligence ChatGPT in providing information on ophthalmic diseases and management to patients. Eye (Lond). May 2024;38(7):1368-1373. [FREE Full text] [doi: 10.1038/s41433-023-02906-0] [Medline: 38245622]
71. Sensoy E, Citirik M. Evaluation of current artificial intelligence programs on the knowledge of glaucoma. Klin Monbl Augenheilkd. Oct 2024;241(10):1140-1144. [FREE Full text] [doi: 10.1055/a-2327-8484] [Medline: 39047763]
72. Caranfa JT, Bommakanti NK, Young BK, Zhao PY. Accuracy of vitreoretinal disease information from an artificial intelligence chatbot. JAMA Ophthalmol. Sep 01, 2023;141(9):906-907. [doi: 10.1001/jamaophthalmol.2023.3314] [Medline: 37535363]
73. Kianian R, Sun D, Crowell EL, Tsui E. The use of large language models to generate education materials about uveitis. Ophthalmol Retina. Feb 2024;8(2):195-201. [FREE Full text] [doi: 10.1016/j.oret.2023.09.008] [Medline: 37716431]
74. Doğan L, Özçakmakcı GB, Yılmaz Ĭ. The performance of chatbots and the AAPOS website as a tool for amblyopia education. J Pediatr Ophthalmol Strabismus. 2024;61(5):325-331. [doi: 10.3928/01913913-20240409-01] [Medline: 38661309]
75. Maywood MJ, 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]
76. Sensoy E, Citirik M. Exploring artificial intelligence programs' understanding of lens, cataract, and refractive surgery information. Middle East Afr J Ophthalmol. 2023;30(3):173-176. [FREE Full text] [doi: 10.4103/meajo.meajo_199_23] [Medline: 39445000]
77. Dihan Q, Chauhan MZ, Eleiwa TK, Brown AD, Hassan AK, Khodeiry MM, et al. Large language models: a new frontier in paediatric cataract patient education. Br J Ophthalmol. Sep 20, 2024;108(10):1470-1476. [doi: 10.1136/bjo-2024-325252] [Medline: 39174290]
78. Rasmussen ML, Larsen AC, Subhi Y, Potapenko I. Artificial intelligence-based ChatGPT chatbot responses for patient and parent questions on vernal keratoconjunctivitis. Graefes Arch Clin Exp Ophthalmol. Oct 2023;261(10):3041-3043. [doi: 10.1007/s00417-023-06078-1] [Medline: 37129631]
79. Balas M, Mandelcorn ED, Yan P, Ing EB, Crawford SA, Arjmand P. ChatGPT and retinal disease: a cross-sectional study on AI comprehension of clinical guidelines. Can J Ophthalmol. Feb 2025;60(1):e117-e123. [FREE Full text] [doi: 10.1016/j.jcjo.2024.06.001] [Medline: 39097289]
80. Jiao C, Edupuganti NR, Patel PA, Bui T, Sheth V. Evaluating the artificial intelligence performance growth in ophthalmic knowledge. Cureus. Sep 2023;15(9):e45700. [FREE Full text] [doi: 10.7759/cureus.45700] [Medline: 37868408]
81. Anguita R, Downie C, Ferro Desideri L, Sagoo MS. Assessing large language models' accuracy in providing patient support for choroidal melanoma. Eye (Lond). Nov 2024;38(16):3113-3117. [doi: 10.1038/s41433-024-03231-w] [Medline: 39003430]
82. Yalla GR, Hyman N, Hock LE, Zhang Q, Shukla AG, Kolomeyer NN. Performance of artificial intelligence chatbots on glaucoma questions adapted from patient brochures. Cureus. Mar 2024;16(3):e56766. [FREE Full text] [doi: 10.7759/cureus.56766] [Medline: 38650824]
83. Inayat H, McDonald HM, Bursztyn LL. Comparison of ChatGPT to ophthalmology resident and staff consultants on an ophthalmological training tool. Can J Ophthalmol (Forthcoming). Oct 26, 2023. [doi: 10.1016/j.jcjo.2023.09.011] [Medline: 39492160]

84. Cheong KX, Zhang C, Tan TE, Fenner BJ, Wong WM, Teo KY, et al. Comparing generative and retrieval-based chatbots in answering patient questions regarding age-related macular degeneration and diabetic retinopathy. Br J Ophthalmol. Sep 20, 2024;108(10):1443-1449. [doi: 10.1136/bjo-2023-324533] [Medline: 38749531]
85. Parikh AO, Oca MC, Conger JR, McCoy A, Chang J, Zhang-Nunes S. Accuracy and bias in artificial intelligence chatbot recommendations for oculoplastic surgeons. Cureus. Apr 2024;16(4):e57611. [FREE Full text] [doi: 10.7759/cureus.57611] [Medline: 38707042]
86. Mihalache A, Grad J, Patil NS, Huang RS, Popovic MM, Mallipatna A, et al. Google Gemini and Bard artificial intelligence chatbot performance in ophthalmology knowledge assessment. Eye (Lond). Sep 2024;38(13):2530-2535. [doi: 10.1038/s41433-024-03067-4] [Medline: 38615098]
87. Ahmed HS, Thrishulamurthy CJ. Evaluating ChatGPT's efficacy and readability to common pediatric ophthalmology and strabismus-related questions. Eur J Ophthalmol. Mar 2025;35(2):466-473. [doi: 10.1177/11206721241272251] [Medline: 39110014]
88. Mohammadi SS, Khatri A, Jain T, Thng ZX, Yoo WS, Yavari N, et al. Evaluation of the appropriateness and readability of ChatGPT-4 responses to patient queries on uveitis. Ophthalmol Sci. 2025;5(1):100594. [FREE Full text] [doi: 10.1016/j.xops.2024.100594] [Medline: 39435137]
89. Tan DN, Tham YC, Koh V, Loon SC, Aquino MC, Lun K, et al. Evaluating Chatbot responses to patient questions in the field of glaucoma. Front Med (Lausanne). 2024;11:1359073. [FREE Full text] [doi: 10.3389/fmed.2024.1359073] [Medline: 39050528]
90. Mihalache A, Huang RS, Patil NS, Popovic MM, Lee WW, Yan P, et al. Chatbot and academy preferred practice pattern guidelines on retinal diseases. Ophthalmol Retina. Jul 2024;8(7):723-725. [FREE Full text] [doi: 10.1016/j.oret.2024.03.013] [Medline: 38499086]
91. Fowler T, Pullen S, Birkett L. Performance of ChatGPT and bard on the official part 1 FRCOphth practice questions. Br J Ophthalmol. Sep 20, 2024;108(10):1379-1383. [doi: 10.1136/bjo-2023-324091] [Medline: 37932006]
92. Potapenko I, Boberg-Ans LC, Stormly Hansen M, Klefter ON, van Dijk EH, Subhi Y. Artificial intelligence-based chatbot patient information on common retinal diseases using ChatGPT. Acta Ophthalmol. Nov 2023;101(7):829-831. [FREE Full text] [doi: 10.1111/aos.15661] [Medline: 36912780]
93. Chen JS, Reddy AJ, Al-Sharif E, Shoji MK, Kalaw FG, Eslani M, et al. Analysis of ChatGPT responses to ophthalmic cases: can ChatGPT think like an ophthalmologist? Ophthalmol Sci. 2025;5(1):100600. [FREE Full text] [doi: 10.1016/j.xops.2024.100600] [Medline: 39346575]
94. 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]
95. Shiraishi M, Tomioka Y, Miyakuni A, Ishii S, Hori A, Park H, et al. Performance of ChatGPT in answering clinical questions on the practical guideline of blepharoptosis. Aesthetic Plast Surg. Jul 2024;48(13):2389-2398. [doi: 10.1007/s00266-024-04005-1] [Medline: 38684536]
96. Kayabaşı M, Köksaldı S, Durmaz Engin C. Evaluating the reliability of the responses of large language models to keratoconus-related questions. Clin Exp Optom. Sep 2025;108(7):784-791. [doi: 10.1080/08164622.2024.2419524] [Medline: 39448387]
97. AlRyalat SA, Musleh AM, Kahook MY. Evaluating the strengths and limitations of multimodal ChatGPT-4 in detecting glaucoma using fundus images. Front Ophthalmol (Lausanne). 2024;4:1387190. [FREE Full text] [doi: 10.3389/fopht.2024.1387190] [Medline: 38984105]
98. Sakai D, Maeda T, Ozaki A, Kanda GN, Kurimoto Y, Takahashi M. Performance of ChatGPT in board examinations for specialists in the Japanese ophthalmology society. Cureus. Dec 2023;15(12):e49903. [FREE Full text] [doi: 10.7759/cureus.49903] [Medline: 38174202]
99. Shaheen A, Afflitto GG, Swaminathan SS. ChatGPT-assisted classification of postoperative bleeding following microinvasive glaucoma surgery using electronic health record data. Ophthalmol Sci. 2025;5(1):100602. [FREE Full text] [doi: 10.1016/j.xops.2024.100602] [Medline: 39380881]
100. Kerci SG, Sahan B. An analysis of ChatGPT4 to respond to glaucoma-related questions. J Glaucoma. Jul 01, 2024;33(7):486-489. [doi: 10.1097/IJG.0000000000002408] [Medline: 38647417]
101. Wu G, Zhao W, Wong A, Lee DA. Patients with floaters: answers from virtual assistants and large language models. Digit Health. 2024;10:20552076241229933. [FREE Full text] [doi: 10.1177/20552076241229933] [Medline: 38362238]
102. Cohen SA, Brant A, Fisher AC, Pershing S, Do D, Pan C. Dr. Google vs. Dr. ChatGPT: exploring the use of artificial intelligence in ophthalmology by comparing the accuracy, safety, and readability of responses to frequently asked patient questions regarding cataracts and cataract surgery. Semin Ophthalmol. Aug 2024;39(6):472-479. [doi: 10.1080/08820538.2024.2326058] [Medline: 38516983]
103. Panthier C, Gatinel D. Success of ChatGPT, an AI language model, in taking the French language version of the European Board of Ophthalmology examination: a novel approach to medical knowledge assessment. J Fr Ophtalmol. Sep 2023;46(7):706-711. [FREE Full text] [doi: 10.1016/j.jfo.2023.05.006] [Medline: 37537126]

104. Marshall R, Xu H, Dalvin LA, Mishra K, Edalat C, Kirupaharan N, et al. Accuracy and completeness of large language models about antibody-drug conjugates and associated ocular adverse effects. Cornea. Aug 07, 2024;44(7):851-855. [doi: 10.1097/ICO.0000000000003664] [Medline: 39110155]
105. Sensoy E, Citirik M. Assessing the competence of artificial intelligence programs in pediatric ophthalmology and strabismus and comparing their relative advantages. Rom J Ophthalmol. 2023;67(4):389-393. [FREE Full text] [doi: 10.22336/rjo.2023.61] [Medline: 38239420]
106. Moshirfar M, Altaf AW, Stoakes IM, Tuttle JJ, Hoopes PC. Artificial intelligence in ophthalmology: a comparative analysis of GPT-3.5, GPT-4, and human expertise in answering StatPearls questions. Cureus. Jun 2023;15(6):e40822. [FREE Full text] [doi: 10.7759/cureus.40822] [Medline: 37485215]
107. Tao BK, Handzic A, Hua NJ, Vosoughi AR, Margolin EA, Micieli JA. Utility of ChatGPT for automated creation of patient education handouts: an application in neuro-ophthalmology. J Neuroophthalmol. Mar 01, 2024;44(1):119-124. [doi: 10.1097/WNO.0000000000002074] [Medline: 38175720]
108. Alqudah AA, Aleshawi AJ, Baker M, Alnajjar Z, Ayasrah I, Ta'ani Y, et al. Evaluating accuracy and reproducibility of ChatGPT responses to patient-based questions in Ophthalmology: an observational study. Medicine (Baltimore). Aug 09, 2024;103(32):e39120. [FREE Full text] [doi: 10.1097/MD.0000000000039120] [Medline: 39121263]
109. Sudharshan R, Shen A, Gupta S, Zhang-Nunes S. Assessing the utility of ChatGPT in simplifying text complexity of patient educational materials. Cureus. Mar 2024;16(3):e55304. [FREE Full text] [doi: 10.7759/cureus.55304] [Medline: 38559518]
110. Singh S, Watson S. ChatGPT as a tool for conducting literature review for dry eye disease. Clin Exp Ophthalmol. 2023;51(7):731-732. [doi: 10.1111/ceo.14268] [Medline: 37321598]
111. Su Z, Jin K, Wu H, Luo Z, Grzybowski A, Ye J. Assessment of large language models in cataract care information provision: a quantitative comparison. Ophthalmol Ther. Jan 2025;14(1):103-116. [doi: 10.1007/s40123-024-01066-y] [Medline: 39516445]
112. Botross M, Mohammadi SO, Montgomery K, Crawford C. Performance of Google's artificial intelligence chatbot "bard" (now "Gemini") on ophthalmology board exam practice questions. Cureus. Mar 2024;16(3):e57348. [FREE Full text] [doi: 10.7759/cureus.57348] [Medline: 38690460]
113. Thirunavukarasu AJ, Mahmood S, Malem A, Foster WP, Sanghera R, Hassan R, et al. Large language models approach expert-level clinical knowledge and reasoning in ophthalmology: a head-to-head cross-sectional study. PLOS Digit Health. Apr 2024;3(4):e0000341. [FREE Full text] [doi: 10.1371/journal.pdig.0000341] [Medline: 38630683]
114. Raimondi R, Tzoumas N, Salisbury T, Di Simplicio S, Romano MR, North East Trainee Research in Ophthalmology Network (NETRiON). Comparative analysis of large language models in the Royal College of Ophthalmologists fellowship exams. Eye (Lond). Dec 2023;37(17):3530-3533. [FREE Full text] [doi: 10.1038/s41433-023-02563-3] [Medline: 37161074]
115. Biswas S, Logan NS, Davies LN, Sheppard AL, Wolffsohn JS. Assessing the utility of ChatGPT as an artificial intelligence-based large language model for information to answer questions on myopia. Ophthalmic Physiol Opt. Nov 2023;43(6):1562-1570. [doi: 10.1111/opo.13207] [Medline: 37476960]
116. Wu G, Lee DA, Zhao W, Wong A, Jhangiani R, Kurniawan S. ChatGPT and Google Assistant as a source of patient education for patients with amblyopia: content analysis. J Med Internet Res. Aug 15, 2024;26:e52401. [FREE Full text] [doi: 10.2196/52401] [Medline: 39146013]
117. Tao BK, Hua N, Milkovich J, Micieli JA. ChatGPT-3.5 and Bing Chat in ophthalmology: an updated evaluation of performance, readability, and informative sources. Eye (Lond). Jul 2024;38(10):1897-1902. [doi: 10.1038/s41433-024-03037-w] [Medline: 38509182]
118. Wang Y, Liang L, Li R, Wang Y, Hao C. Comparison of the Performance of ChatGPT, Claude and Bard in support of myopia prevention and control. J Multidiscip Healthc. 2024;17:3917-3929. [FREE Full text] [doi: 10.2147/JMDH.S473680] [Medline: 39155977]
119. Barclay KS, You JY, Coleman MJ, Mathews PM, Ray VL, Riaz KM, et al. Quality and agreement with scientific consensus of ChatGPT information regarding corneal transplantation and Fuchs dystrophy. Cornea. Jun 01, 2024;43(6):746-750. [doi: 10.1097/ICO.0000000000003439] [Medline: 38016014]
120. Kianian R, Sun D, Giaconi J. Can ChatGPT aid clinicians in educating patients on the surgical management of glaucoma? J Glaucoma. Feb 01, 2024;33(2):94-100. [doi: 10.1097/IJG.0000000000002338] [Medline: 38031276]
121. Ichhpujani P, Parmar UP, Kumar S. Appropriateness and readability of Google Bard and ChatGPT-3.5 generated responses for surgical treatment of glaucoma. Rom J Ophthalmol. 2024;68(3):243-248. [doi: 10.22336/rjo.2024.45] [Medline: 39464759]
122. Balci AS, Yazar Z, Ozturk BT, Altan C. Performance of Chatgpt in ophthalmology exam; human versus AI. Int Ophthalmol. Nov 06, 2024;44(1):413. [doi: 10.1007/s10792-024-03353-w] [Medline: 39503920]
123. Subramanian B, Rajalakshmi R, Sivaprasad S, Rao C, Raman R. Assessing the appropriateness and completeness of ChatGPT-4's AI-generated responses for queries related to diabetic retinopathy. Indian J Ophthalmol. Jul 01, 2024;72(Suppl 4):S684-S687. [FREE Full text] [doi: 10.4103/IJO.IJO_2510_23] [Medline: 38953134]
124. Edalat C, Kirupaharan N, Dalvin LA, Mishra K, Marshall R, Xu H, et al. Evaluating large language models on their accuracy and completeness: immune checkpoint inhibitors and their ocular toxicities. Retina. Jan 01, 2025;45(1):128-132. [doi: 10.1097/IAE.0000000000004271] [Medline: 39312883]

125. Patil NS, Huang R, Mihalache A, Kisilevsky E, Kwok J, Popovic MM, et al. The ability of artificial intelligence chatbots ChatGPT and Google Bard to accurately convey preoperative information for patients undergoing ophthalmic surgeries. Retina. Jun 01, 2024;44(6):950-953. [doi: 10.1097/IAE.0000000000004044] [Medline: 38215455]
126. Sensoy E, Citirik M. A comparative study on the knowledge levels of artificial intelligence programs in diagnosing ophthalmic pathologies and intraocular tumors evaluated their superiority and potential utility. Int Ophthalmol. Dec 2023;43(12):4905-4909. [doi: 10.1007/s10792-023-02893-x] [Medline: 37880412]
127. Jung H, Oh J, Stephenson KA, Joe AW, Mammo ZN. Prompt engineering with ChatGPT3.5 and GPT4 to improve patient education on retinal diseases. Can J Ophthalmol. Jun 2025;60(3):e375-e381. [FREE Full text] [doi: 10.1016/j.jcjo.2024.08.010] [Medline: 39245293]
128. Nanji K, Yu CW, Wong TY, Sivaprasad S, Steel DH, Wykoff CC, et al. Evaluation of postoperative ophthalmology patient instructions from ChatGPT and Google Search. Can J Ophthalmol. Feb 2024;59(1):e69-e71. [doi: 10.1016/j.jcjo.2023.10.001] [Medline: 37884271]
129. Wu JH, Nishida T, Moghimi S, Weinreb RN. Performance of ChatGPT on responding to common online questions regarding key information gaps in glaucoma. J Glaucoma. Jul 01, 2024;33(7):e54-e56. [doi: 10.1097/IJG.0000000000002409] [Medline: 38647429]
130. Knebel D, Priglinger S, Scherer N, Klaas J, Siedlecki J, Schworm B. Assessment of ChatGPT in the prehospital management of ophthalmological emergencies - an analysis of 10 fictional case vignettes. Klin Monbl Augenheilkd. May 2024;241(5):675-681. [FREE Full text] [doi: 10.1055/a-2149-0447] [Medline: 37890504]
131. Reyhan AH, Mutaf Ç, Uzun İ, Yüksekyayla F. A performance evaluation of large language models in keratoconus: a comparative study of ChatGPT-3.5, ChatGPT-4.0, Gemini, Copilot, Chatsonic, and Perplexity. J Clin Med. Oct 30, 2024;13(21):6512. [FREE Full text] [doi: 10.3390/jcm13216512] [Medline: 39518652]
132. Edhem Yılmaz İ, Berhuni M, Özer Özcan Z, Doğan L. Chatbots talk strabismus: can AI become the new patient educator? Int J Med Inform. Nov 2024;191:105592. [doi: 10.1016/j.ijmedinf.2024.105592] [Medline: 39159506]
133. Muntean GA, Marginean A, Groza A, Damian I, Roman SA, Hapca MC, et al. A qualitative evaluation of ChatGPT4 and PaLM2's response to patient's questions regarding age-related macular degeneration. Diagnostics (Basel). Jul 09, 2024;14(14):1468. [FREE Full text] [doi: 10.3390/diagnostics14141468] [Medline: 39061606]
134. Azzopardi M, Ng B, Logeswaran A, Loizou C, Cheong RC, Gireesh P, et al. Artificial intelligence chatbots as sources of patient education material for cataract surgery: ChatGPT-4 versus Google Bard. BMJ Open Ophthalmol. Oct 17, 2024;9(1):33. [FREE Full text] [doi: 10.1136/bmjophth-2024-001824] [Medline: 39419585]
135. Shiraishi M, Tanigawa K, Tomioka Y, Miyakuni A, Moriwaki Y, Yang R, et al. Blepharoptosis consultation with artificial intelligence: aesthetic surgery advice and counseling from chat generative pre-trained transformer (ChatGPT). Aesthetic Plast Surg. Jun 2024;48(11):2057-2063. [doi: 10.1007/s00266-024-04002-4] [Medline: 38589561]
136. Assayag E, Zadok D, Berkovitz L, Weill Y. Exploring the accuracy and readability of ChatGPT in providing information to patients with keratoconus. J Pediatr Ophthalmol Strabismus. 2024;61(5):e43-e46. [doi: 10.3928/01913913-20240718-01] [Medline: 39301826]
137. Choudhary A, Gopalakrishnan N, Joshi A, Balakrishnan D, Chhablani J, Yadav NK, et al. Recommendations for diabetic macular edema management by retina specialists and large language model-based artificial intelligence platforms. Int J Retina Vitreous. Feb 28, 2024;10(1):22. [FREE Full text] [doi: 10.1186/s40942-024-00544-6] [Medline: 38419083]
138. Momenaei B, Wakabayashi T, Shahlaee A, Durrani AF, Pandit SA, Wang K, et al. Assessing ChatGPT-3.5 versus ChatGPT-4 performance in surgical treatment of retinal diseases: a comparative study. Ophthalmic Surg Lasers Imaging Retina. Aug 2024;55(8):481-482. [FREE Full text] [doi: 10.3928/23258160-20240227-02] [Medline: 38531015]
139. Gill GS, Tsai J, Moxam J, Sanghvi HA, Gupta S. Comparison of Gemini advanced and ChatGPT 4.0's performances on the ophthalmology resident ophthalmic knowledge assessment program (OKAP) examination review question banks. Cureus. Sep 2024;16(9):e69612. [doi: 10.7759/cureus.69612] [Medline: 39421095]
140. 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]
141. Al-Sharif EM, Penteado RC, Dib El Jalbout N, Topilow NJ, Shoji MK, Kikkawa DO, et al. Evaluating the accuracy of ChatGPT and Google BARD in fielding oculoplastic patient queries: a comparative study on artificial versus human intelligence. Ophthalmic Plast Reconstr Surg. 2024;40(3):303-311. [doi: 10.1097/IOP.0000000000002567] [Medline: 38215452]
142. Wang H, Masselos K, Tong J, Connor HR, Scully J, Zhang S, et al. ChatGPT for addressing patient-centered frequently asked questions in glaucoma clinical practice. Ophthalmol Glaucoma. 2025;8(2):157-166. [doi: 10.1016/j.ogla.2024.10.005] [Medline: 39424063]
143. García-Porta N, Vaughan M, Rendo-González S, Gómez-Varela AI, O'Donnell A, de-Moura J, et al. Are artificial intelligence chatbots a reliable source of information about contact lenses? Cont Lens Anterior Eye. Apr 2024;47(2):102130. [FREE Full text] [doi: 10.1016/j.clae.2024.102130] [Medline: 38443210]

144. Tailor PD, Dalvin LA, Starr MR, Tajfirouz DA, Chodnicki KD, Brodsky MC, et al. A comparative study of large language models, human experts, and expert-edited large language models to neuro-ophthalmology questions. J Neuroophthalmol. Mar 01, 2025;45(1):71-77. [doi: 10.1097/WNO.0000000000002145] [Medline: 38564282]
145. Haddad F, Saade JS. Performance of ChatGPT on ophthalmology-related questions across various examination levels: observational study. JMIR Med Educ. Jan 18, 2024;10:e50842. [FREE Full text] [doi: 10.2196/50842] [Medline: 38236632]
146. Cohen SA, Yadlapalli N, Tijerina JD, Alabiad CR, Chang JR, Kinde B, et al. Comparing the ability of Google and ChatGPT to accurately respond to oculoplastics-related patient questions and generate customized oculoplastics patient education materials. Clin Ophthalmol. 2024;18:2647-2655. [FREE Full text] [doi: 10.2147/OPTH.S480222] [Medline: 39323727]
147. Sensoy E, Citirik M. Assessing the proficiency of artificial intelligence programs in the diagnosis and treatment of cornea, conjunctiva, and eyelid diseases and exploring the advantages of each other benefits. Cont Lens Anterior Eye. Apr 2024;47(2):102125. [doi: 10.1016/j.clae.2024.102125] [Medline: 38443209]
148. Gopalakrishnan N, Joshi A, Chhablani J, Yadav NK, Reddy NG, Rani PK, et al. Recommendations for initial diabetic retinopathy screening of diabetic patients using large language model-based artificial intelligence in real-life case scenarios. Int J Retina Vitreous. Jan 24, 2024;10(1):11. [FREE Full text] [doi: 10.1186/s40942-024-00533-9] [Medline: 38268046]
149. Chang LC, Sun CC, Chen TH, Tsai DC, Lin HL, Liao LL. Evaluation of the quality and readability of ChatGPT responses to frequently asked questions about myopia in traditional Chinese language. Digit Health. 2024;10:20552076241277021. [FREE Full text] [doi: 10.1177/20552076241277021] [Medline: 39229462]
150. Mihalache A, Huang RS, Popovic MM, Muni RH. Artificial intelligence chatbot and Academy Preferred Practice Pattern ® guidelines on cataract and glaucoma. J Cataract Refract Surg. May 01, 2024;50(5):534-535. [doi: 10.1097/j.jcrs.0000000000001317] [Medline: 38468154]
151. Marshall RF, Mallem K, Xu H, Thorne J, Burkholder B, Chaon B, et al. Investigating the accuracy and completeness of an artificial intelligence large language model about uveitis: an evaluation of ChatGPT. Ocul Immunol Inflamm. Nov 2024;32(9):2052-2055. [doi: 10.1080/09273948.2024.2317417] [Medline: 38394625]
152. Strzalkowski P, Strzalkowska A, Chhablani J, Pfau K, Errera MH, Roth M, et al. Evaluation of the accuracy and readability of ChatGPT-4 and Google Gemini in providing information on retinal detachment: a multicenter expert comparative study. Int J Retina Vitreous. Sep 02, 2024;10(1):61. [FREE Full text] [doi: 10.1186/s40942-024-00579-9] [Medline: 39223678]
153. Durmaz Engin C, Karatas E, Ozturk T. Exploring the Role of ChatGPT-4, BingAI, and Gemini as virtual consultants to educate families about retinopathy of prematurity. Children (Basel). Jun 20, 2024;11(6):750. [FREE Full text] [doi: 10.3390/children11060750] [Medline: 38929329]
154. 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. [FREE Full text] [doi: 10.1136/bjo-2023-325143] [Medline: 38448201]
155. Sensoy E, Citirik M. Investigating the comparative superiority of artificial intelligence programs in assessing knowledge levels regarding ocular inflammation, uvea diseases, and treatment modalities. Taiwan J Ophthalmol. 2024;14(3):409-413. [doi: 10.4103/tjo.TJO-D-23-00166] [Medline: 39430359]
156. Bahir D, Zur O, Attal L, Nujeidat Z, Knaanie A, Pikkel J, et al. Gemini AI vs. ChatGPT: a comprehensive examination alongside ophthalmology residents in medical knowledge. Graefes Arch Clin Exp Ophthalmol. Feb 2025;263(2):527-536. [doi: 10.1007/s00417-024-06625-4] [Medline: 39277830]
157. Yaïci R, Cieplucha M, Bock R, Moayed F, Bechrakis NE, Berens P, et al. ChatGPT and the German board examination for ophthalmology: an evaluation. Ophthalmologie. Jul 2024;121(7):554-564. [doi: 10.1007/s00347-024-02046-0] [Medline: 38801461]
158. Nikdel M, Ghadimi H, Tavakoli M, Suh DW. Assessment of the responses of the artificial intelligence-based chatbot ChatGPT-4 to frequently asked questions about amblyopia and childhood myopia. J Pediatr Ophthalmol Strabismus. 2024;61(2):86-89. [doi: 10.3928/01913913-20231005-02] [Medline: 37882183]
159. Gondode P, Duggal S, Garg N, Lohakare P, Jakhar J, Bharti S, et al. Comparative analysis of accuracy, readability, sentiment, and actionability: artificial intelligence chatbots (ChatGPT and Google Gemini) versus traditional patient information leaflets for local anesthesia in eye surgery. Br Ir Orthopt J. 2024;20(1):183-192. [FREE Full text] [doi: 10.22599/bioj.377] [Medline: 39183761]
160. Eid K, Eid A, Wang D, Raiker RS, Chen S, Nguyen J. Optimizing ophthalmology patient education via ChatBot-generated materials: readability analysis of AI-generated patient education materials and the American Society of Ophthalmic Plastic and Reconstructive Surgery patient brochures. Ophthalmic Plast Reconstr Surg. 2024;40(2):212-216. [doi: 10.1097/IOP.0000000000002549] [Medline: 37972974]
161. Raja H, Huang X, Delsoz M, Madadi Y, Poursoroush A, Munawar A, et al. Diagnosing glaucoma based on the ocular hypertension treatment study dataset using chat generative pre-trained transformer as a large language model. Ophthalmol Sci. 2025;5(1):100599. [FREE Full text] [doi: 10.1016/j.xops.2024.100599] [Medline: 39346574]
162. Ermis S, Özal E, Karapapak M, Kumantaş E, Özal SA. Assessing the responses of large language models (ChatGPT-4, Claude 3, Gemini, and Microsoft Copilot) to frequently asked questions in retinopathy of prematurity: a study on readability and appropriateness. J Pediatr Ophthalmol Strabismus. 2025;62(2):84-95. [doi: 10.3928/01913913-20240911-05] [Medline: 39465590]

163. Yılmaz IB, Doğan L. Talking technology: exploring chatbots as a tool for cataract patient education. Clin Exp Optom. Jan 2025;108(1):56-64. [doi: 10.1080/08164622.2023.2298812] [Medline: 38194585]
164. Tuttle JJ, Moshirfar M, Garcia J, Altaf AW, Omidvarnia S, Hoopes PC. Learning the Randleman criteria in refractive surgery: utilizing ChatGPT-3.5 versus internet search engine. Cureus. Jul 2024;16(7):e64768. [doi: 10.7759/cureus.64768] [Medline: 39156271]
165. Aykut A, Sezenoz AS. Exploring the potential of code-free custom GPTs in ophthalmology: an early analysis of GPT store and user-creator guidance. Ophthalmol Ther. Oct 2024;13(10):2697-2713. [doi: 10.1007/s40123-024-01014-w] [Medline: 39141071]
166. Nikdel M, Ghadimi H, Suh DW, Tavakoli M. Accuracy of the image interpretation capability of ChatGPT-4 vision in analysis of Hess screen and visual field abnormalities. J Neuroophthalmol. Nov 07, 2024;45(3):307-315. [doi: 10.1097/WNO.0000000000002267] [Medline: 39508800]
167. Güler MS, Baydemir EE. Evaluation of ChatGPT-4 responses to glaucoma patients' questions: can artificial intelligence become a trusted advisor between doctor and patient? Clin Exp Ophthalmol. Dec 2024;52(9):1016-1019. [doi: 10.1111/ceo.14451] [Medline: 39407089]
168. Gue CC, Rahim ND, Rojas-Carabali W, Agrawal R, Rk P, Abisheganaden J, et al. Evaluating the OpenAI's GPT-3.5 Turbo's performance in extracting information from scientific articles on diabetic retinopathy. Syst Rev. May 16, 2024;13(1):135. [FREE Full text] [doi: 10.1186/s13643-024-02523-2] [Medline: 38755704]
169. Kon MH, Pereira MJ, Molina JA, Yip VC, Abisheganaden JA, Yip W. Unravelling ChatGPT's potential in summarising qualitative in-depth interviews. Eye (Lond). Feb 2025;39(2):354-358. [doi: 10.1038/s41433-024-03419-0] [Medline: 39501005]
170. Alexander AC, Somineni Raghupathy S, Surapaneni KM. An assessment of the capability of ChatGPT in solving clinical cases of ophthalmology using multiple choice and short answer questions. Adv Ophthalmol Pract Res. 2024;4(2):95-97. [FREE Full text] [doi: 10.1016/j.aopr.2024.01.005] [Medline: 38666248]
171. Holmes J, Peng R, Li Y, Hu J, Liu Z, Wu Z, et al. Evaluating multiple large language models in pediatric ophthalmology. arXiv. Preprint posted online on November 7, 2023. [FREE Full text]
172. Parikh RN, Pham C, Carter KD, Shriver EM. What are artificial intelligence-driven large language models saying about ASOPRS surgeons? Implications for scope of practice. Ophthalmic Plast Reconstr Surg. 2024;40(5):586-589. [doi: 10.1097/IOP.0000000000002743] [Medline: 39240205]
173. Tailor PD, Dalvin LA, Chen JJ, Iezzi R, Olsen TW, Scruggs BA, et al. A comparative study of responses to retina questions from either experts, expert-edited large language models, or expert-edited large language models alone. Ophthalmol Sci. 2024;4(4):100485. [FREE Full text] [doi: 10.1016/j.xops.2024.100485] [Medline: 38660460]
174. Tailor PD, D'Souza HS, Castillejo Becerra CM, Dahl HM, Patel NR, Kaplan TM, et al. Utilizing AI-generated plain language summaries to enhance interdisciplinary understanding of ophthalmology notes: a randomized trial. medRxiv. Preprint posted online on September 13, 2024. [FREE Full text]
175. Ruiz-Núñez C, Gismero Rodríguez J, Garcia Ruiz AJ, Gismero Moreno SM, Cañizal Santos MS, Herrera-Peco I. Can generative AI contribute to health literacy? A study in the field of ophthalmology. Multimodal Technol Interact. Sep 04, 2024;8(9):79. [doi: 10.3390/mti8090079]
176. Spina A, Tang J, Picton B, Spiegel S. Using ChatGPT to improve patient accessibility to neuro-ophthalmology research. J Neurol Sci. Dec 2023;455:122104. [doi: 10.1016/j.jns.2023.122104]
177. Solli EM, Tsui E, Mehta N. Analysis of ChatGPT responses to patient-oriented questions on common ophthalmic procedures. Clin Exp Ophthalmol. 2024;52(4):487-491. [doi: 10.1111/ceo.14334] [Medline: 38140836]
178. Ciekalski M, Laskowski M, Koperczak A, Śmierciak M, Sirek S. Performance of ChatGPT and GPT-4 on Polish National Specialty Exam (NSE) in ophthalmology. Postepy Hig Med Dosw. 2024;78(1):111-116. [doi: 10.2478/AHEM-2024-0006]
179. 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. [doi: 10.1001/jamaophthalmol.2023.2754] [Medline: 37440220]
180. Olis M, Dyjak P, Weppelmann TA. Performance of three artificial intelligence chatbots on Ophthalmic Knowledge Assessment Program materials. Can J Ophthalmol. Aug 2024;59(4):e380-e381. [doi: 10.1016/j.jcjo.2024.01.011] [Medline: 38408734]
181. Khake A, Gokhale S, Dindore P, Khake S, Potdar P, Potdar A, et al. Role of ChatGPT 3.5 and bard as self assessment tool for undergraduate level long answer questions in ophthalmology. Int J Acad Med Pharm. 2024;6(2):264-269. [FREE Full text]
182. Wu G, Cheng E, Zhao W, Wong A, Mansubi N, Dorado P, et al. SAT117 diabetes and vision: can Chatgpt educate our diabetic patients? J Endocr Soc. 2023;7(Supplement_1):114-982. [FREE Full text] [doi: 10.1210/jendso/bvad114.982]
183. Luo MJ, Pang J, Bi S, Lai Y, Zhao J, Shang Y, et al. Development and evaluation of a retrieval-augmented large language model framework for ophthalmology. JAMA Ophthalmol. Sep 01, 2024;142(9):798-805. [doi: 10.1001/jamaophthalmol.2024.2513] [Medline: 39023885]

184. Jaskari J, Sahlsten J, Summanen P, Moilanen J, Lehtola E, Aho M, et al. DR-GPT: a large language model for medical report analysis of diabetic retinopathy patients. PLoS One. 2024;19(10):e0297706. [FREE Full text] [doi: 10.1371/journal.pone.0297706] [Medline: 39392790]
185. Xue X, Zhang D, Sun C, Shi Y, Wang R, Tan T, et al. Xiaoqing: A Q and A model for glaucoma based on LLMs. Comput Biol Med. May 2024;174:108399. [doi: 10.1016/j.compbiomed.2024.108399] [Medline: 38615461]
186. Upadhyaya DP, Shaikh AG, Cakir GB, Prantzalos K, Golnari P, Ghasia FF, et al. A 360° view for large language models: early detection of amblyopia in children using multi-view eye movement recordings. medRxiv. Preprint posted online on May 7, 2024. [FREE Full text] [doi: 10.1101/2024.05.03.24306688] [Medline: 38765973]
187. Raja H, Munawar A, Mylonas N, Delsoz M, Madadi Y, Elahi M, et al. Automated category and trend analysis of scientific articles on ophthalmology using large language models: development and usability study. JMIR Form Res. Mar 22, 2024;8:e52462. [FREE Full text] [doi: 10.2196/52462] [Medline: 38517457]
188. Chen X, Zhang W, Zhao Z, Xu P, Zheng Y, Shi D, et al. ICGA-GPT: report generation and question answering for indocyanine green angiography images. Br J Ophthalmol. Sep 20, 2024;108(10):1450-1456. [doi: 10.1136/bjo-2023-324446] [Medline: 38508675]
189. Gilson A, Ai X, Arunachalam T, Chen Z, Cheong KX, Dave A, et al. Enhancing large language models with domain-specific retrieval augment generation: a case study on long-form consumer health question answering in ophthalmology. arXiv. Preprint posted online on September 20, 2024. [Medline: 41031070]
190. Gilson A. Bringing large language models to ophthalmology: domain-specific ontologies and evidence attribution. Yale Medicine Thesis Digital Library. URL: https://elischolar.library.yale.edu/ymtdl/4226/ [accessed 2025-05-29]
191. Fu L, Fan B, Du H, Feng Y, Li C, Song H. A role-specific guided large language model for ophthalmic consultation based on stylistic differentiation. arXiv. Preprint posted online on July 26, 2024. [FREE Full text]
192. Tan TF, Elangovan K, Jin L, Jie Y, Yong L, Lim J, et al. Fine-tuning large language model (LLM) artificial intelligence chatbots in ophthalmology and LLM-based evaluation using GPT-4. arXiv. Preprint posted online on February 15, 2024. [FREE Full text]
193. Chen X, Zhao Z, Zhang W, Xu P, Gao L, Xu M, et al. Eyegpt: ophthalmic assistant with large language models. arXiv. Preprint posted online on February 29, 2024. [FREE Full text]
194. Zhao H, Ling Q, Pan Y, Zhong T, Hu JY, Yao J, et al. Ophtha-llama2: a large language model for ophthalmology. arXiv. Preprint posted online on December 8, 2023. [FREE Full text]
195. Yeh CH, Wang J, Graham AD, Liu AJ, Tan B, Chan Y, et al. Insight: a multi-modal diagnostic pipeline using LLMs for ocular surface disease diagnosis. arXiv. Preprint posted online on October 1, 2024. [FREE Full text]
196. Raja H, Munawar A, Delsoz M, Elahi M, Madadi Y, Hassan A. Using large language models to automate category and trend analysis of scientific articles: an application in ophthalmology. arXiv. Preprint posted online on August 31, 2023. [FREE Full text]
197. Gilson A, Ai X, Xie Q, Srinivasan S, Pushpanathan K, Singer MB, et al. Language enhanced model for eye (LEME): an open-source ophthalmology-specific large language model. arXiv. Preprint posted online on October 1, 2024. [Medline: 41031082]
198. Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. PaLM: scaling language modeling with pathways. J Mach Learn Res. 2023;24(1):11324-11436. [FREE Full text] [doi: 10.5555/3648699.3648939]
199. Gemini Team Google. Gemini: a family of highly capable multimodal models. ArXiv. Preprint posted online on December 19, 2023. [FREE Full text]
200. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: open foundation and fine-tuned chat models. ArXiv. Preprint posted online on July 18, 2023. [FREE Full text]
201. DeepSeek-AI. DeepSeek-V3 technical report. ArXiv. Preprint posted online on December 27, 2024. [FREE Full text]
202. Kumar A, Cheeseman R, Durnian JM. Subspecialization of the ophthalmic literature: a review of the publishing trends of the top general, clinical ophthalmic journals. Ophthalmology. Jun 2011;118(6):1211-1214. [doi: 10.1016/j.ophtha.2010.10.023] [Medline: 21269704]
203. Steren BJ, Yee P, Rivera PA, Feng S, Pepple K, Kombo N. Gender distribution and trends of ophthalmology subspecialties, 1992-2020. Am J Ophthalmol. Sep 2023;253:22-28. [doi: 10.1016/j.ajo.2023.04.012] [Medline: 37142172]
204. Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA. Jan 28, 2025;333(4):319-328. [doi: 10.1001/jama.2024.21700] [Medline: 39405325]
205. Tam TY, Sivarajkumar S, Kapoor S, Stolyar AV, Polanska K, McCarthy KR, et al. A framework for human evaluation of large language models in healthcare derived from literature review. NPJ Digit Med. Sep 28, 2024;7(1):258. [FREE Full text] [doi: 10.1038/s41746-024-01258-7] [Medline: 39333376]
206. Perez I, Dedden F, Goodloe A. Copilot 3. NASA. 2020. URL: https://ntrs.nasa.gov/api/citations/20200003164/downloads/20200003164.pdf [accessed 2025-09-24]
207. Team GLM. ChatGLM: a family of large language models from GLM-130B to GLM-4 all tools. ArXiv. Preprint posted online on June 18, 2024. [FREE Full text]

208. Khandekar RB, Al-Towerki AA, Al-Katan H, Al-Mesfer SS, Abboud EB, Al-Hussain HM, et al. Ocular malignant tumors: review of the Tumor Registry at a tertiary eye hospital in central Saudi Arabia. Saudi Med J. 2014;35(4):377-384. [FREE Full text]
209. Jayaram H, Kolko M, Friedman DS, Gazzard G. Glaucoma: now and beyond. The Lancet. Nov 2023;402(10414):1788-1801. [doi: 10.1016/s0140-6736(23)01289-8]
210. Yorston D. Retinal diseases and VISION 2020. Community Eye Health. 2003;16(46):19-20. [FREE Full text] [Medline: 17491850]
211. Camara J, Rezende R, Pires IM, Cunha A. Retinal glaucoma public datasets: what do we have and what is missing? J Clin Med. Jul 02, 2022;11(13):3850. [FREE Full text] [doi: 10.3390/jcm11133850] [Medline: 35807135]
212. Khan SM, Liu X, Nath S, Korot E, Faes L, Wagner SK, et al. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. Lancet Digit Health. Jan 2021;3(1):e51-e66. [doi: 10.1016/s2589-7500(20)30240-5]
213. Bai S, Deng Z, Yang J, Gong Z, Gao W, Shao L, et al. FTSNet: fundus tumor segmentation network on multiple scales guided by classification results and prompts. Bioengineering (Basel). Sep 22, 2024;11(9):950. [FREE Full text] [doi: 10.3390/bioengineering11090950] [Medline: 39329692]
214. Bennett TJ, Barry CJ. Ophthalmic imaging today: an ophthalmic photographer's viewpoint - a review. Clin Exp Ophthalmol. Jan 2009;37(1):2-13. [FREE Full text] [doi: 10.1111/j.1442-9071.2008.01812.x] [Medline: 18947332]
215. Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. In: Proceedings of the Advances in Neural Information Processing Systems 36. 2023. Presented at: NeurIPS 2023; December 10-16, 2023; New Orleans, LA. URL: https://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html
216. Li C, Wong C, Zhang S, Usuyama N, Liu H, Yang J, et al. LLaVA-med: training a large language-and-vision assistant for biomedicine in one day. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. 2023. Presented at: NIPS '23; December 10-16, 2023; New Orleans, LA.
217. Tom E, Keane PA, Blazes M, Pasquale LR, Chiang MF, Lee AY, et al. Protecting data privacy in the age of AI-enabled ophthalmology. Transl Vis Sci Technol. Jul 28, 2020;9(2):36. [FREE Full text] [doi: 10.1167/tvst.9.2.36] [Medline: 32855840]
218. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. 2002. Presented at: ACL '02; July 7-12, 2002; Philadelphia, PA. [doi: 10.3115/1073083.1073135]
219. Manchanda J, Boettcher L, Westphalen M, Jasser J. The open source advantage in large language models (LLMs). ArXiv. Preprint posted online on December 16, 2024. [FREE Full text]
220. Zhang K, Zhou R, Adhikarla E, Yan Z, Liu Y, Yu J, et al. A generalist vision–language foundation model for diverse biomedical tasks. Nat Med. 2024;30:3129-3141. [FREE Full text] [doi: 10.1038/s41591-024-03185-2]
221. Chen Z, Cano AH, Romanou A, Bonnet A, Matoba K, Salvi F, et al. MEDITRON-70B: scaling medical pretraining for large language models. ArXiv. Preprint posted online on November 27, 2023. [FREE Full text]
222. Zhang Q, Wang S, Wang X, Xu C, Liang J, Liu Z. Advancing ophthalmology with large language models: applications, challenges, and future directions. Surv Ophthalmol. 2025;70(5):1019-1028. [doi: 10.1016/j.survophthal.2025.02.009] [Medline: 40032069]
223. Kedia N, Sanjeev S, Ong J, Chhablani J. ChatGPT and beyond: an overview of the growing field of large language models and their use in ophthalmology. Eye (Lond). May 03, 2024;38(7):1252-1261. [doi: 10.1038/s41433-023-02915-z] [Medline: 38172581]
224. Tan TF, Thirunavukarasu AJ, Campbell JP, Keane PA, Pasquale LR, Abramoff MD, et al. Generative artificial intelligence through ChatGPT and other large language models in ophthalmology: clinical applications and challenges. Ophthalmol Sci. Dec 2023;3(4):100394. [FREE Full text] [doi: 10.1016/j.xops.2023.100394] [Medline: 37885755]
225. Gemae MR, Kim P, Sturrock S, Law C. Diversity gaps among practicing ophthalmologists in Canada: a landscape study. Can J Ophthalmol. Apr 2025;60(2):e212-e218. [FREE Full text] [doi: 10.1016/j.jcjo.2024.08.001] [Medline: 39178910]
226. Zhou Q, Chen ZH, Cao YH, Peng S. Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: a systematic review. NPJ Digit Med. Oct 28, 2021;4(1):154. [FREE Full text] [doi: 10.1038/s41746-021-00524-2] [Medline: 34711955]
227. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med. Jul 06, 2023;6(1):120. [FREE Full text] [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]
228. Busch F, Hoffmann L, Rueger C, van Dijk EH, Kader R, Ortiz-Prado E, et al. Current applications and challenges in large language models for patient care: a systematic review. Commun Med (Lond). Jan 21, 2025;5(1):26. [FREE Full text] [doi: 10.1038/s43856-024-00717-2] [Medline: 39838160]
229. Templin T, Perez MW, Sylvia S, Leek J, Sinnott-Armstrong N. Addressing 6 challenges in generative AI for digital health: a scoping review. PLOS Digit Health. May 23, 2024;3(5):e0000503. [doi: 10.1371/journal.pdig.0000503] [Medline: 38781686]
230. Li H, Fu JF, Python A. Implementing large language models in health care: clinician-focused review with interactive guideline. J Med Internet Res. Jul 11, 2025;27:e71916. [FREE Full text] [doi: 10.2196/71916] [Medline: 40644686]

231. Tian S, Jin Q, Yeganova L, Lai PT, Zhu Q, Chen X, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Brief Bioinform. Nov 22, 2023;25(1):bbad493. [FREE Full text] [doi: 10.1093/bib/bbad493] [Medline: 38168838]
232. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on large language models (LLMs). NPJ Digit Med. Jul 08, 2024;7(1):183. [FREE Full text] [doi: 10.1038/s41746-024-01157-x] [Medline: 38977771]
233. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Appl Sci. Jul 12, 2021;11(14):6421. [FREE Full text] [doi: 10.3390/app11146421]
234. Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. PubMedQA: a dataset for biomedical research question answering. ArXiv. Preprint posted online on September 13, 2019. [FREE Full text] [doi: 10.48550/arXiv.1909.06146]
235. Chen L, Zaharia M, Zou J. How is ChatGPT's behavior changing over time? ArXiv. Preprint posted online on July 18, 2023. [FREE Full text]
236. Sarvari P, Al-Fagih Z, Ghuwel A, Al-Fagih O. A systematic evaluation of the performance of GPT-4 and PaLM2 to diagnose comorbidities in MIMIC-IV patients. Health Care Sci. Feb 2024;3(1):3-18. [doi: 10.1002/hcs2.79] [Medline: 38939167]
237. Johnson AE, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. Jan 03, 2023;10(1):1. [FREE Full text] [doi: 10.1038/s41597-022-01899-x] [Medline: 36596836]
238. Heinz MV, Mackin DM, Trudeau BM, Bhattacharya S, Wang Y, Banta HA, et al. Randomized trial of a generative AI chatbot for mental health treatment. NEJM AI. Mar 27, 2025;2(4). [doi: 10.1056/aioa2400802]
239. Bai J, Bai S, Chu Y, Cui Z, Dang K, Deng X, et al. Qwen technical report. ArXiv. Preprint posted online on September 28, 2023. [FREE Full text]

Abbreviations

AI: artificial intelligence
LLaMA: Large Language Model Meta AI
LLM: large language model
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
RoB: risk of bias

Edited by A Coristine; submitted 05.May.2025; peer-reviewed by L Zhu, W Yang, F Li; comments to author 08.Jul.2025; revised version received 03.Aug.2025; accepted 15.Sep.2025; published 27.Oct.2025

Please cite as:
Zhang Z, Zhang H, Pan Z, Bi Z, Wan Y, Song X, Fan X
Evaluating Large Language Models in Ophthalmology: Systematic Review
J Med Internet Res 2025;27:e76947
URL: https://www.jmir.org/2025/1/e76947
doi: 10.2196/76947
PMID: