Review

Not peer-reviewed version

The Use of Large Language Models in Ophthalmology: A Scoping Review on Current Use-Cases and Considerations for Future Works in This Field

See Ye King Clarence , Lim Khai Shin Alva , Au Wei Yung , Chia Si Yin Charlene , Fan Xiuyi ,
Li Zhenghao Kelvin *

Posted Date: 26 March 2025

doi: 10.20944/preprints202503.1961.v1

Keywords: Large Language Model; Ophthalmology; Artificial intelligence

Review

The use of Large Language Models in Ophthalmology: A Scoping Review on Current Use-Cases and Considerations for Future Works in this Field

See Ye King Clarence [1,2], Lim Khai Shin Alva [3], Au Wei Yung [3], Chia Si Yin Charlene [4], Fan Xiuyi [3,4] and Li Zhenghao Kelvin [1,2,5,*]

[1] Department of Ophthalmology, Tan Tock Seng Hospital, Singapore
[2] National Healthcare Group Eye Institution, Singapore
[3] Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore
[4] School of Computer Science and Engineering, Nanyang Technological University, Singapore
[5] Department of Ophthalmology, Byers Eye Institute, Stanford University School of Medicine, Palo Alto, California
* Correspondence: kelvin_li@ttsh.com.sg

Highlights:

Large Language Models are gaining popularity, and its use has permeated fields of medicine.
The use of LLMs in Ophthalmology is a field of ongoing study.
Newer LLMs like GPT4 appear to have good performance in several clinical areas.
Concerns regarding inaccuracies and harms still exist with LLM use cases.
Standardised frameworks and use of techniques such as prompt engineering are recommended.

Abstract: The advancement of generative artificial intelligence (AI) has resulted in its use permeating many areas of life. Amidst this eruption of scientific output, a wide range of research regarding the usage of LLMs in ophthalmology has emerged. In this study, we aim to map out the landscape of LLM applications in ophthalmology and by consolidating the work done, we aim to produce a point of reference to guide the conduct of future works. 8 databases were searched for articles from 2019 - 2024. 976 studies were screened, and a final 49 were included. The study designs and outcomes of these studies were analysed. The performance of LLMs was further analysed in the areas of exam taking and patient education, diagnostic capability, management capability, administration, inaccuracies, and harm. LLMs performed acceptably in most studies, even surpassing humans in some. Despite the relatively good performance, issues pertaining to study design, grading protocols, hallucinations, inaccuracies, and harm were found to be pervasive. LLMs have received considerable attention through their introduction to the public and have found potential applications in the field of medicine, and in particular, ophthalmology. However, using standardised evaluation frameworks and addressing gaps in current literature when applying LLMs in ophthalmology is recommended through this review.

Keywords: large language model; ophthalmology; artificial intelligence

1. Introduction

The advancement and popularisation of generative artificial intelligence (AI) has resulted in its use permeating many areas of life and scientific research. This has largely been driven by the way

 doi:10.20944/preprints202503.1961.v1

Large Language Models (LLMs) have transformed the use of Natural Language Processing (NLP). Through self-supervised learning, LLMs have been utilised to effectively perform a wide variety of tasks ranging from interpreting and classifying text to generating answers to conversational questions. In November 2022, the release of ChatGPT by OpenAI revolutionised the LLM scene. Through its user-friendly interface and accessibility, ChatGPT has democratised the use of LLMs beyond the realm of computer science researchers, engaging a broad spectrum of users from various fields, sparking unprecedented interest in this field [1]. It took 4 years from the release of the Bidirectional Encoder Representations from Transformers (BERT) language model in October 2018 to develop 8 major LLM applications prior to ChatGPT's release. On the contrary, in the two years since ChatGPT's release, 8 major LLMs - Med-PaLM 1, Google Bard, Glass AI 2.0, GPT-4, Med-PaLM 2, LLaMa, Gemini and Claude were released (Figure 1).

Notably, newer LLMs have superior generalisation capabilities [1] and have been trained to provide more human-like responses, sparking interest in their use within medicine. To date, we have seen encouraging results supporting the use of LLMs in clinical practice, medical education and medical research [2–6].

The field of ophthalmology is no stranger to AI. Machine learning programs have been developed to detect and grade cataracts, while various deep learning programs have demonstrated their utility in detecting glaucomatous optic nerve changes. These applications have allowed ophthalmology to generate a wealth of data, paving the way for LLMs to potentially deliver more streamlined, personalised, and optimised care for ophthalmology patients [7–9].

It is unsurprising therefore, that amid this eruption of scientific output in the realm of AI and LLMs, a wide range of research regarding the usage and efficacy of LLMs in ophthalmology has emerged. Between January and July 2023 alone, a review summarising LLM trends in ophthalmology identified thirty-two articles related to this topic [10]. Inadvertently, this has also resulted in the publication of many isolated studies with overlapping scopes of research resulting in the duplication of efforts. In another review [11] of the usage of LLMs in ophthalmology, a total of 108 studies were identified between January 2016 to June 2023, 55 of which involved overlapping aspects of automated question-answering, while 27 dealt with information screening. Notably, this review did not provide a study-by-study breakdown of the 108 studies but mainly sought to understand general trends of LLM usage in ophthalmology. A literature review of publication in this field suggested that LLM research even in the niche area of ophthalmology appears to have a laissez-faire approach, with each having their own unique design. This potentially complicates the consolidation of research outputs in this field and makes it difficult to compare approaches and results across studies. To tackle such concerns, guidelines such as the SPIRIT-AI and CONSORT-AI initiatives for clinical trials and interventions involving AI have been created [12]. However, the extent to which such protocols are followed is yet to be determined. To our knowledge, to date there has also not been a summarisation of how LLM studies in the field of ophthalmology has been carried out.

In this study, we aim to map out the landscape of LLM applications in ophthalmology. By consolidating the work thus far, we aim to produce a point of reference to guide future research in this field. This review aims to summarise the following points:

1. To identify recent studies (1st January 2019 – 11th February 2024) involving the application of LLMs in ophthalmology. This study period was chosen as it represents the period of LLM breakthroughs after BERT's release [1] (Figure 1).
2. To evaluate how studies of LLM applications in ophthalmology were carried out, in terms of following clinical trial protocols followed, prompt techniques employed, benchmarking methods used, and ethical considerations.
3. To examine how LLMs fared in key areas of healthcare application, including exam taking and patient education, diagnostic and management capability, and clinical administration.
4. To highlight potential issues surrounding the present landscape of LLM applications of ophthalmology, and to discuss directions for future LLM research and development in ophthalmology.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1961.v1

During our literature review, we found that studies utilising LLM in Ophthalmology covered a broad range of applications and had a diverse range of findings and methodologies. Given the broad and diverse nature of works in this field the format of a scoping review was chosen to map out the key trends and findings for this area in recent years, as opposed to a meta-analysis that seeks to draw a conclusion about specific research questions.

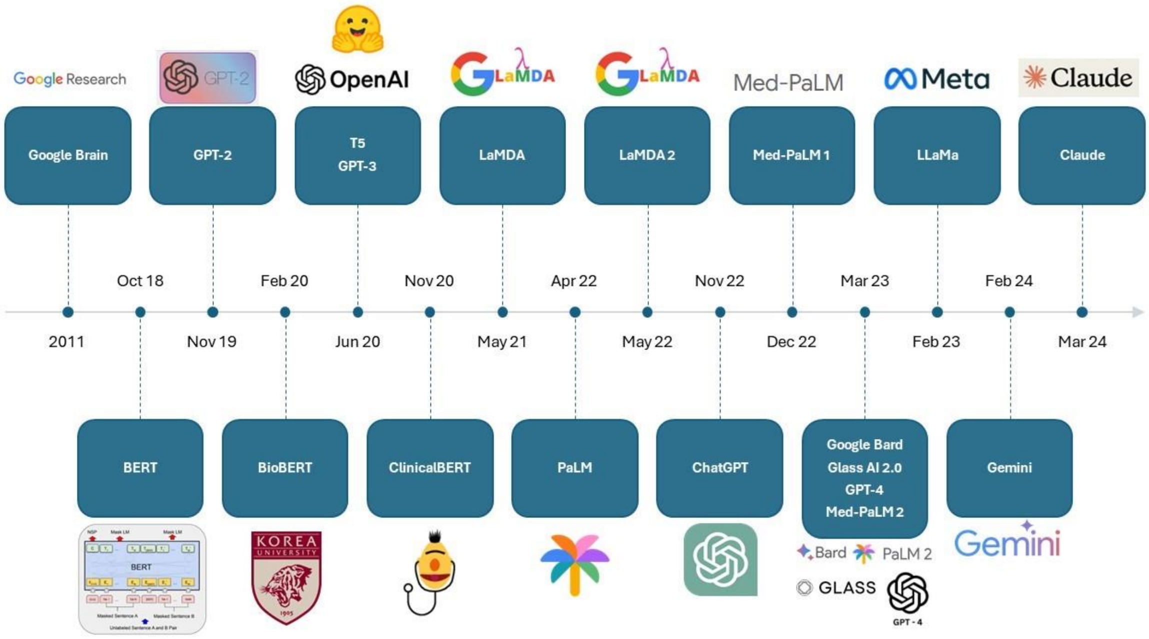

Figure 1. A timeline of major LLMs released since the inception of Google Brain in 2011.

2. Methods

2.1. Search Strategy and Information Sources

A search of PubMed, Embase, SCOPUS, Web of Science, the Institute of Electrical and Electronics Engineers (IEEE) journals, the Association for Computing Machinery (ACM) journals, Google Scholar and DataBase systems and Logic Programming (DBLP) was performed from 1st January 2019 – 11th February 2024. The search strategy can be found in the supplementary information (Appendix Table A1).

The MeSH (medical subject heading) terms included are as follows:

1) For Ophthalmology: Ophthalmology, Ocular Surgery, Eye Disease, Eye Diseases, Eye Disorders.
2) For LLMs: Large Language Model, large language models, large language modelling, Chatbot, ChatGPT, GPT, chatbots, google bard, bing chat, BERT, RoBERTa, distilBERT, BART, MARIAN, llama, palm.

The search strategy was developed in consultation with expert opinion within the research team, which consisted of computer scientists (FXY,CCS) and clinicians [13] (KLZ (ophthalmology), SYKC (ophthalmology)). No additional search filters were applied.

2.2. Selection Process and Eligibility Criteria

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) guideline was utilised [14]. An independent search was conducted by two reviewers (LKA, AWY). Any discrepancies were resolved through discussion between reviewers, with a third (KLZ) included when necessary. This review has not been registered prior.

The inclusion criteria were

1) Peer reviewed primary research studies utilising LLMs.

 doi:10.20944/preprints202503.1961.v1

2) Studies involving ophthalmology.
3) Studies published from January 2019 to March 2024.

The exclusion criteria were

1) Study designs that were reviews, systematic reviews and meta-analyses, case reports, case series, guidelines, letters, correspondences, or protocols.
2) Studies that were not published in English.

2.3. Data Extraction and Analysis

Data on the studies were uploaded into Mendeley and imported into COVIDENCE Systematic Review Software (Veritas Health Innovation, Melbourne, Australia) for screening. As mentioned earlier, differences in screening outcomes were resolved in consultation with a third reviewer.

Data extracted from the papers were analysed on Microsoft Excel (Microsoft, Richmond, Virginia, USA). These included (1) authorship details, (2) LLMs utilised, (3) study methodology, and (4) performance and performance scoring of the LLMs.

In terms of study methodology, we took note of clinical trial protocols used, prompt techniques employed, how benchmarking was done, and ethical considerations in the studies.

The performance of LLMs were also analysed in the following areas: exam taking and patient education, diagnostic capability, management capability, clinical administration, inaccuracies, and harm (glossary A).

The subspecialties studied included (1) Cornea, (2) Glaucoma, (3) Neuro-ophthalmology, (4) Uveitis, (5) Lens and cataract, (6) Paediatrics and strabismus. (7) Retina and Vitreous, (8) Oculoplastics, (9) Optics, (10) Refractive surgery and (11) Pathology. The various LLMs were then assessed on their accuracy and overall completeness of their answers, which were then ranked and compared across the different LLMs employed per study.

3. Results

A total of 976 studies were screened, for which 904 were excluded with 72 being sought for retrieval. A further 22 of these studies did not meet the inclusion criteria, and a final 49 studies [15–63] were included in this study (Figure 2, Table 1).

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1961.v1

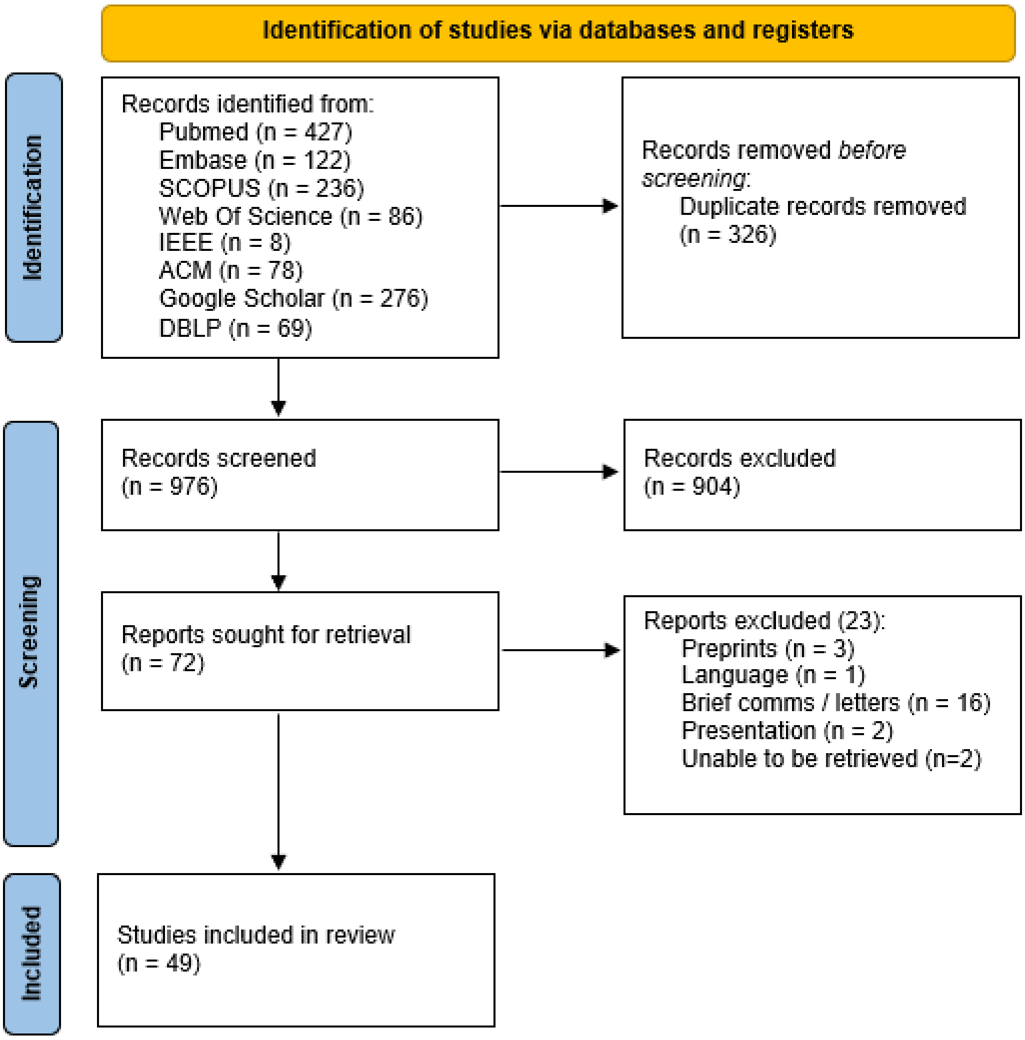

Figure 2. PRISMA flowchart for study screening and selection.

3.1. Overall Study Characteristics

A total of 14 different LLMs models (Table 1) were used across the studies with various applications spanning the fields of administration, clinical knowledge, diagnostics, exam taking, manuscript writing, patient education, prognostication, text interpretation, and triage. GPT-3.5 was the most commonly employed LLM, being utilised in 34 studies. GPT-4.0 came in second, appearing in 26 studies. Bard and Bing LLM models were the next most used (Table 1).

Table 1. Study characteristics.

Study	Clinical Application	LLM
Singh 2023 [55]	Administrative	GPT 3.5
Barclay 2023 [20]	Clinical Knowledge	GPT 3.5, GPT 4
Rojas-Carabali (1) 2023 [49]	Diagnostic	GPT 3.5, GPT 4.0. Glass 1.0
Ali 2023 [15]	Diagnostic	GPT 3.5
Shemer 2024 [53]	Diagnostic	GPT 3.5
Rojas-Carabali (2) 2023 [50]	Diagnostic	GPT 3.5, GPT 4
Delsoz 2023 [26]	Diagnostic	GPT 3.5
Sensoy 2023 (1) [27]	Exam Taking	GPT 3.5, Bing, Bard
Moshirfar 2023 [41]	Exam Taking	GPT 3.5, GPT 4
Sensoy 2023 (2) [52]	Exam Taking	GPT 3.5, Bing, Bard

 doi:10.20944/preprints202503.1961.v1

Antaki 2023 (1) [17]	Exam Taking	GPT 3.5, GPT 4
Taloni 2023 [57]	Exam Taking	GPT 3.5, GPT 4
Singer 2023 [54]	Exam Taking	Aeyeconsult, GPT 4
Jiao 2023 [34]	Exam Taking	GPT 3.5, GPT 4
Antaki 2023 (2) [18]	Exam Taking	ChatGPT legacy and ChatGPT Plus
Teebagy 2023 [59]	Exam Taking	GPT 3.5, GPT 4
Fowler 2023 [30]	Exam Taking	GPT 4, Bard
Sakai 2023 [51]	Exam Taking	GPT 3.5, GPT 4
Haddad 2024 [31]	Exam Taking	GPT 3.5, GPT 4
Cai 2023 [23]	Exam Taking	GPT 3.5, GPT 4, Bing Chat
Panthier 2023 [44]	Exam Taking	GPT 4
Hua 2023 [33]	Manuscript Writing	GPT 3.5, GPT 4
Tailor 2024 [56]	Patient Education	GPT 3.5, GPT 4, Claude 2, Bing, Bard
FerroDesideri 2023 [29]	Patient Education	GPT 3.5, Bard, Bing Chat
Potapenko 2023 [46]	Patient Education	GPT 4
Biswas 2023 [22]	Patient Education	GPT 3.5
Nikdel 2023 [42]	Patient Education	GPT 4
Lim 2023 [37]	Patient Education	GPT 3.5, GPT 4, Bard
Kianian 2023 (1) [36]	Patient Education	GPT 3.5
Wu 2023 [61]	Patient Education	GPT 3.5
Bernstein 2023 [21]	Patient Education	GPT 3.5
Balas 2024 [19]	Patient Education	GPT 4
Al-Sharif 2024 [16]	Patient Education	GPT 3.5, Bard
Zandi 2024 [63]	Patient Education	GPT 4, Bard
Eid 2023 [28]	Patient Education	GPT 4.0, Bard
Pushpanathan 2023 [47]	Patient Education	GPT 3.5, GPT 4, Bard
Cappellani 2024 [24]	Patient Education	GPT 3.5
Yılmaz 2024 [62]	Patient Education	GPT 3.5, Bard, Bing AI, AAO website
Patil 2024 [45]	Patient Education	GPT 4, Bard
Kianian 2023 (2) [35]	Patient Education	GPT 4, Bard
Liu 2024 [38]	Patient Education	GPT 3.5
Tao 2024 [58]	Patient Education	GPT 3.5
Wilhelm 2023 [60]	Patient Management	GPT 3.5 Turbo, Command-xlarge-nightly, Claude, Bloomz
Maywood 2024 [40]	Patient Management	GPT 3.5 Turbo
Cirkovic 2023 [25]	Prognostication	GPT 4
Hu 2022 [32]	Prognostication	BERT, RoBerta, DistilBert, BioBERT
Raghu 2023 [48]	Prognostication	GPT 4
Ong 2023 [43]	Text interpretation	GPT 3.5
Lyons 2023 [39]	Triage	GPT 4, Bing Chat, WebMD

In terms of study design, all studies did not follow a standardised clinical trial protocol for artificial intelligence. All studies employed a zero-shot, one-shot or few-shot prompt engineering technique, apart from one study which only utilised contextual priming. There were three studies who additionally used prompt chaining, iterative prompting, and chain-of-thought prompt techniques to supplement their work. Most studies (37 of 49 studies) shared full examples of their prompts (Table 2). Across the studies, the grading of the output generated by the LLMs was heterogeneous with little standardisation, resulting in difficulty in data analyses. 24 studies employed human assessors to benchmark LLM performance in terms of “correctness” of output, two

 doi:10.20944/preprints202503.1961.v1

of which were assisted by automated benchmarking assessments. The remaining studies utilised automated benchmarking to assess "correctness" of output. Nine studies considered harm in their study protocol, all of which were assessed by humans. Only one study by Wilhelm et. al. also utilised an automated form of harm assessment in the form of GPT-4.0 Twelve studies delved into the ethical implications relating to their work, while thirteen only touched very briefly on patient safety without going further into an ethical discussion (Table 2).

Table 2. Summary of study methodologies.

Study	Usage of a research protocol for AI	Ethical / Safety Safeguards considered in methodology	Ethics in Discussion	Prompt Techniques Employed	Prompt examples shared	Benchmarks on Correctness	Benchmarks on Harm
Tailor 2024	No	Yes	Yes	zero-shot (no prior context)	Yes	Human	Human
Sensoy 2023 (1)	No	No	No	Zero-shot	No	Automated (Exact match)	Nil
FerroDesideri 2023	No	No	No	Zero-shot	Yes	Human	Nil
Ong 2023	No	No	Yes	Zero-shot	Yes	Automated (Exact match)	Nil
Lyons 2023	No	No	Yes	Zero-shot	Yes	Human	Nil
Moshirfar 2023	No	No	No	Zero-shot	Yes	Automated (Exact match)	Nil
Potapenko 2023	No	No	No	Zero-shot	Yes	Human	Nil
Sensoy 2023 (2)	No	No	No	Zero-shot	No	Automated (Exact match)	Nil
Biswas 2023	No	No	No	Zero-shot	Yes	Human	Nil
Nikdel 2023	No	No	No	Zero-shot, Prompt Chaining	Yes	Human	Nil
Lim 2023	No	No	Yes	Zero-shot, Iterative Prompting	Yes	Human	Nil
Kianian 2023 (1)	No	No	No (safety but not ethics)	One-shot, Few-shot	Yes	Automated and Human	Nil
Antaki 2023 (1)	No	No	No (safety but not ethics)	Zero-shot	Yes	Human	Nil
Rojas-Carabali (1) 2023	No	No	No (safety but not ethics)	Zero-shot	Yes	Automated (Exact match) and Human	Nil
Ali 2023	No	No	No	Zero-shot	Yes	Human	Nil

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1961.v1

Singh 2023	No	No	No	Contextual Priming	Yes	Human	Nil
Wu 2023	No	No	No	Zero-shot	Yes	Automated (Exact match, Readability)	Nil
Taloni 2023	No	No	No	Zero-shot	No	Automated (Exact match)	Nil
Bernstein 2023	No	Yes	Yes	Zero-shot	Yes	Human	Human
Singer 2023	No	No	No (safety but not ethics)	Zero-shot	No	Automated (Exact match)	Nil
Shemer 2024	No	Yes	No	Zero-shot	Yes	Automated (Exact match)	Nil
Balas 2023	No	No	No	Zero-shot	No	Human	Nil
Al-Sharif 2024	No	No	Yes	Zero-shot	Yes	Human	Nil
Jiao 2023	No	No	Yes	Zero-shot	Yes	Automated (Exact match)	Nil
Rojas-Carabali (2) 2023	No	No	No	Zero-shot	Yes	Automated (Exact match)	Nil
Antaki 2023 (2)	No	No	No (safety but not ethics)	Zero-shot	No	Automated (Exact match)	Nil
Hua 2023	No	No	Yes	Zero-shot	No	Human	Nil
Zandi 2024	No	Yes	No (safety but not ethics)	Zero-shot	No	Human	Human
Cirkovic 2023	No	No?	No	Zero-shot	No	Statistical Analysis including Cohen κ coefficient, a chi-square test, a confusion matrix, accuracy, precision, recall, F1-score, and receiver operating characteristic area under the curve	Nil
Teebagy 2023	No	No	No	Zero-shot	No	Automated (Exact match)	Nil

Wilhelm 2023	No	Yes	No (safety but not ethics)	Zero-shot	No	Automated and Human	Automated and Human
Eid 2023	No	No	No	Zero-shot	Yes	Automated (readability)	Nil
Maywood 2024	No	Yes	No (safety but not ethics)	Zero-shot	Yes	Human	Human
Fowler 2023	No	No	No	Zero-shot	No	Automated (Exact match)	Nil
Sakai 2023	No	No	No	Zero shot, Few-shot	Yes	Automated (Exact match)	Nil
Haddad 2024	No	No	No	Zero-shot	Yes	Automated (Exact match)	Nil
Cai 2023	No	No	No (safety but not ethics)	Zero-shot	Yes	Automated (Exact match)	Nil
Pushpanathan 2023	No	No	No (safety but not ethics)	Zero-shot	Yes	Automated (Exact match)	Nil
Hu 2022	No	No	No	Zero-shot	Yes	Automated (Exact match, F1 score)	Nil
Barclay 2023	No	Yes	No (safety but not ethics)	Zero-shot	Yes	Human	Human
Cappellani 2024	No	Yes	No (safety but not ethics)	Zero-shot	Yes	Human	Human
Panthier 2023	No	No	No	Zero-shot	Yes	Automated (Exact match)	Nil
Yılmaz 2024	No	No	No (safety but not ethics)	Zero-shot	Yes	Automated	Nil
Patil 2024	No	No	Yes	Zero-shot	Yes	Human	Human
Delsoz 2023	No	No	No	Zero-shot	Yes	Human	Nil
Kianian 2023 (2)	No	No	Yes	Zero-shot	Yes	Automated (readability)	Nil
Raghu 2023	No	No	Yes	Zero-shot	Yes	Human	Nil
Liu 2024	No	No	No	Zero-shot, Chain-of-thought (inspired)	Yes	Automated	Nil
Tao 2024	No	Yes	Yes	Zero-shot	Yes	Human	Human

3.2. Breakdown of LLM Benchmarks Studied and General Observations

43 of the 49 included studies [15–27,29–32,34,37–54,56–60,62,63] performed an assessment regarding the "correctness" of the LLM's output in some form - be it in the form of relevance or

accuracy, to name a few. The remaining 6 studies looked at other qualities such as readability [28,35,36], manuscript writing [33], and administration [43,55]. Amongst the 43 studies assessing the "correctness" of the LLM's output, 27 [16–18,20,23,27,29–32,34,37,39,41,45,47,49–52,54,56,57,59,60,62,63] of them compared multiple LLMs against each other (Table 3a), while 16 [15,19,21,22,24–26,38,40,42–44,46,48,53,58] were observational studies using a single LLM (Table 3b). A total of eleven different scoring systems were used to assess for "correctness" (Appendix Table A2). There were 15 studies [20,23,30,31,34,37,41,47,49,51,54,56,57,59,63] which compared GPT-4.0 against humans and/or other LLMs. Among these, GPT-4.0 was the best performer in 10 studies [20,30,31,34,37,41,47,57,59,63]. Amongst the seven studies [16,27,29,37,47,52,56] comparing Bard, Bing, and GPT-3.5, GPT-3.5 had the best performance in 5 [16,29,37,47,56] of them. Amongst the single-armed studies, the LLMs were reported to have largely appropriate responses overall (Table 3b). In the following subsections, we go into further detail regarding LLM performance in specific domains.

Table 3. a. Overall performance of LLM responses - Multiple LLMs studied. **b. Overall performance of LLM responses - One LLMs studied**

(a)			
Study	**Setting**	**Scoring system**	**Result**
Barclay 2023 [20]	Clinical Knowledge	5 Point Scale	GPT 4 > GPT 3.5
Rojas-Carabali (1) 2023	Diagnostic	Correct or Incorrect	Experts > GPT 4 = GPT 3.5 > Glass 1.0
Rojas-Carabali (2) 2023	Diagnostic	Correct or Incorrect	Ophthalmologist > AI
Singer 2023	Exam Taking	Correct or Incorrect	Aeyeconsult > GPT 4
Antaki 2023 (2)	Exam Taking	Correct or Incorrect	Plus > Legacy
Sensoy 2023 (1)	Exam Taking	Correct or Incorrect	Bard > Bing > GPT 3.5
Sensoy 2023 (2)	Exam Taking	Correct, Incorrect or Unable to Answer	Bard > Bing > GPT 3.5
Moshirfar 2023	Exam Taking	Correct or Incorrect	GPT 4 > humans > GPT 3.5
Antaki 2023 (1)	Exam Taking	Correct or Incorrect	GPT 4-0.3 > GPT 4-0.7 > GPT 4-1 = GPT 4-0 > GPT 3.5
Taloni 2023	Exam Taking	Correct or Incorrect	GPT 4 > Humans > GPT 3.5
Jiao 2023	Exam Taking	Correct or Incorrect	GPT 4 > GPT 3.5
Teebagy 2023	Exam Taking	Correct or Incorrect	GPT 4 > GPT 3.5
Sakai 2023	Exam Taking	Correct or Incorrect	Humans > GPT 4 > GPT 3.5
Haddad 2024	Exam Taking	Correct or Incorrect	GPT 4 > GPT 3.5
Cai 2023	Exam Taking	Correct or Incorrect	Humans > GPT 4 = Bing > GPT 3.5
Fowler 2023	Exam Taking	Correct or Incorrect	GPT 4 > Bard
Yılmaz 2024	Patient Education	SOLO score	ChatGPT > Bard > Bing > AAO
Pushpanathan 2023	Patient Education	5 Point Scale	GPT 4 > GPT 3.5 > Bard
Al-Sharif 2024	Patient Education	4 Point Scale	GPT 3.5 > Bard
FerroDesideri 2023	Patient Education	3 Point Scale	GPT 3.5 > Bard = Bing
Tailor 2024	Patient Education	5 Point Scale	Expert + AI > GPT 3.5 > GPT 4> Expert only > Claude > Bard > Bing
Lim 2023	Patient Education	3 Point Scale	GPT 4 > GPT 3.5 > Bard
Zandi 2024	Patient Education	Correct or Incorrect	GPT 4 > Bard

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1961.v1

Patil 2024	Patient Education	5 Point Scale	ChatGPT > Bard
Wilhelm 2023	Patient Management	mDISCERN	Claude-instant-v1.0 > GPT 3.5-Turbo > Command-xlarge-nightly > Bloomz
Hu 2022	Prognostication	AUROC, F1	BERT > RoBERTa > DistilBERT > BioBert > Humans
Lyons 2023	Triage	5 Point Scale	Ophthalmologists in training > chatGPT > Bing Chat > WebMD

(b)

Study	LLMs	Setting	Scoring system	Result
Ali 2023	GPT 3.5	Diagnostic	3 Point Scale	40% correct 35% partially correct 25% outright incorrect
Shemer 2024	GPT 3.5	Diagnostic	Correct or Incorrect	Residents > Attendings > GPT 3.5
Delsoz 2023	GPT 3.5	Diagnostic	Correct or Incorrect	ChatGPT performed similarly to 2 of 3 residents and better than 1 resident
Panthier 2023	GPT 4	Exam Taking	Correct or Incorrect	6188 / 6785 correct
Biswas 2023	GPT 3.5	Patient Education	5 Point Scale	66 / 275 responses rated as very good 134 / 275 responses rated as good 60 / 275 acceptable 10 / 275 poor 5 / 275 very poor
Bernstein 2023	GPT 3.5	Patient Education	Comparison to humans	GPT 3.5 = Humans
Cappellani 2024	GPT 3.5	Patient Education	5 Point Scale	93 responses scored >= 1 27 responses scored =< -1 9 responses scored -3
Liu 2024	GPT 3.5	Patient Education	Correct or Incorrect	Ophthalmology Attendings > Ophthalmology Interns > English Prompt > Chinese Prompting of ChatGPT
Tao 2024	GPT 3.5	Patient Education	4 Point Scale	2.43 95% CI 1.21, 3.65
Potapenko 2023	GPT 4	Patient Education	Correct or Incorrect	17 / 100 responses were relevant

				without inaccuracies 78 / 100 relevant with inaccuracies that were not harmful 5 /100 relevant with inaccuracies potentially harmful
Nikdel 2023	GPT 4	Patient Education	3 Point Scale	93 / 110 acceptable
Balas 2024	GPT 4	Patient Education	7 Point Scale	43 / 100 scored 6 53 / 100 scored 5 3 / 100 scored 4 1 / 100 scored 3
Maywood 2024	GPT 3.5 Turbo	Patient Management	Correct or Incorrect	33/40 correct 21/40 comprehensive
Cirkovic 2023	GPT 4	Prognostication	Cohens Kappa	6 categories: k = 0.399 2 categories: k = 0.610
Raghu 2023	GPT 4	Prognostication	Cohens Kappa	With central subfield thickness: k = 0.263 Without central subfield thickness: k = 0.351
Ong 2023	GPT 3.5	Text interpretation	Correct: producing at least one correct ICD code Correct only: only the correct ICD code Incorrect: not generating any	Correct: 137 / 181 Correct only: 106/181 Incorrect: 54/181

Human vs Artificial Intelligence

16 studies [17,21,23,25,26,30,31,38,39,41,49–51,53,56,57] investigated the performance of LLMs against humans (Attendings, Ophthalmologists-in-training) in diagnosis, exam taking, patient education, prognostication, and triage. In terms of diagnostic, prognostic, and triage ability, humans consistently outperformed LLMs in all six of these studies (Table 4). In terms of answering exam questions, there was a more even balance with humans being the best in three studies, while GPT-4.0 superseding humans in four studies. It is also worth noting that humans consistently performed better than GPT-3.5 in exam taking for every subspecialty (Table 5). The same could be said in terms of developing patient education materials, with humans bettering GPT 3.5 in one study, equalling GPT 3.5 in another study, and partnering with AI to supersede GPT-4.0 in the last study of this area. Notably, the latter study found that GPT 3.5 produced superior results to humans in terms of how empathetic their patient education material was. (Table 4)

Table 4. Human v AI.

Study	Setting	Results
Rojas-Carabali (1) 2023	Diagnostic	Humans > GPT 4 > Glass

Shemer 2024	Diagnostic	Humans > GPT 3.5
Rojas-Carabali (2) 2023	Diagnostic	Humans > GPT -3.5 and 4 (collectively)
Delsoz 2023	Diagnostic	Humans = GPT 3.5
Moshirfar 2023	Exam Taking	GPT 4 > Humans > GPT 3.5
Antaki 2023 (1)	Exam Taking	GPT 4 > Humans
Taloni 2023	Exam Taking	GPT 4 > Humans > GPT 3.5
Fowler 2023	Exam Taking	GPT 4 > Humans > Bard
Sakai 2023	Exam Taking	Humans > GPT 4 > GPT 3.5
Haddad 2024	Exam Taking	Humans > GPT 4 > GPT 3.5
Cai 2023	Exam Taking	Humans > GPT 4 > Bing > GPT 3.5
Tailor 2024	Patient Education	Quality: Expert + AI = GPT 3.5 = GPT 4 > Expert > Claude > Bard > Bing Empathy: GPT 3.5 = Expert + AI = GPT 4 > Bard > Claude > Expert > Bing
Bernstein 2023	Patient Education	GPT 3.5 = Humans
Liu 2024	Patient Education	Humans > GPT 3.5
Cirkovic 2023	Prognostication	Humans = GPT 4
Lyons 2023	Triage	Human > GPT 4 > Bing > WebMD Symptom Checker

Table 5. Performance in subspecialties.

Study	Moshirfar 2023	Taloni 2023	Singer 2023	Jiao 2023	Antaki 2023 (1)	Antaki 2023 (2)	Teebagy 2023	Sakai 2023	Haddad 2024	Cai 2023	Patil 2024
Clinical application	Exam Taking	Exam Taking	Exam Taking	Exam Taking	Exam Taking	Exam Taking	Exam Taking	Exam Taking	Exam Taking	Exam Taking	Patient Education
LLMs	GPT 3.5, GPT 4	GPT 3.5, GPT 4	Aeyeconsult, GPT 4	GPT 3.5, GPT 4	GPT 3.5, GPT 4	ChatGPT legacy and ChatGPT Plus	GPT 3.5, GPT 4	GPT 3.5, GPT 4	GPT 3.5, GPT 4	GPT 3.5, GPT 4, Bing Chat	GPT 4, Bard
Overall	GPT 4 (73%) > Humans (58%) > GPT 3.5 (55%)	GPT 4 (82.4%) > Humans (75.7%) > GPT 3.5 (65.9%)	Aeyeconsult (83.4%) > GPT 4 (69.2%)	GPT 4 (75%) > GPT 3.5 (46%)	GPT 4-0.3 (72.9%) > Humans (68.15%) > GPT 3.5 (54.6%)	Plus (54.3%) > Legacy (49.25%)	GPT 4 (81%) > GPT 3.5 (57%)	Humans (65.7%) > GPT 4 (46.2% with prompt, 45.8% without) > GPT 3.5 (22.4%)	Humans (70 - 75%) > GPT 4 (70%) > GPT 3.5 (55%)	Humans (72.2%) > GPT 4 (71.6%) > Bing (71.2%) > GPT 3.5 (58.8%)	-

Cornea	GPT 3.5 = GPT 4 = Human	GPT 4 > Human = GPT 3.5	Aeyeconsult = GPT 4	GPT 4 = GPT 3.5	GPT 4-1 > GPT 4-0.7 = GPT 4-0 > GPT 4-3 > GPT 3.5	Plus > Legacy	GPT 4 > 3.5	GPT 4 (few shot) > GPT 4 > GPT 3.5	GPT 4 = GPT 3.5	Human > Bing > GPT 4.0 > GPT 3.5	GPT 4 > Bard
Glaucoma	GPT 4 > GPT 3.5 = Human	GPT 4 > Human > GPT 3.5	Aeyeconsult > GPT 4	GPT 4 = GPT 3.5	GPT 4-1 > GPT 4-0.7 = GPT 4-0.3 = GPT 4-0 > GPT 3.5	Plus > Legacy	GPT 4 > 3.5	GPT 4 > GPT 4 (few shot) > GPT 3.5	GPT 4 = GPT 3.5	Human > GPT 4.0 = Bing > GPT 3.5	-
NeuroOphth	GPT 4 > GPT 3.5 > Human	GPT 4 = Human > GPT 3.5	Aeyeconsult > GPT 4	GPT 4 > GPT 3.5	GPT 4-1 = GPT 4-0.3 = GPT 4-0 > GPT 4-0.7 > GPT 3.5	Plus > Legacy	GPT 4 > 3.5	GPT 4 (few shot) > GPT 4 > GPT 3.5	GPT 4 = GPT 3.5	Human > Bing > GPT 4.0 > GPT 3.5	-
Uveitis	GPT 4 > Human = GPT 3.5	GPT 4 > Human = GPT 3.5	GPT 4 > Aeyeconsult	GPT 4 > GPT 3.5	GPT 4-1 = GPT 4-0.7 = GPT 4-0.3 = GPT 4-0 > GPT 3.5	Plus > Legacy (BSCS) Legacy > Plus (OphtoQuestions)	GPT 4 > 3.5	GPT 4 (few shot) > GPT 4 > GPT 3.5	-	GPT 4.0 > Human = Bing > GPT 3.5	-
Lens and cataract	GPT 3.5 = GPT 4 = Human	GPT 4 = Human > GPT 3.5	Aeyeconsult > GPT 4	-	GPT 4-0.7 = GPT 4-0.3 > GPT 4-1 = GPT 4-0 > GPT 3.5	Legacy > Plus (BCSC) Plus > Legacy (OphthoQuestions)	GPT 4 > 3.5	GPT 4 > GPT 4 (few shot) > GPT 3.5*	GPT 3.5 > GPT 4	Human = GTP4 > Bing > GPT 3.5	GPT 4 > Bard

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1961.v1

Paediatric and strabs	GPT 4 > GPT 3.5 = Human	GPT 4 = Human > GPT 3.5	Aeyeconsult > GPT 4	GPT 4 = GPT 3.5	GPT 4-1 = GPT 4-0.7 = GPT 4-0.3 = GPT 4-0 > GPT 3.5	Legacy > Plus (BCSC) Plus > Legacy (OphthoQuestions)	GPT 4 > 3.5	GPT 4 > GPT 4 (few shot) > GPT 3.5	GPT 4 = GPT 3.5	GPT 4.0 > Bing > Human > GPT 3.5	GPT 4 > Bard
Retina & Vitreous	GPT 3.5 = GPT 4 = Human	GPT 4 = Human = GPT 3.5	Aeyeconsult > GPT 4	GPT 4 = GPT 3.5	GPT 4-0.7 > GPT 4-0.3 > GPT 4-1 = GPT 4-0 > GPT 3.5	Plus > Legacy	GPT 4 > 3.5	GPT 4 = GPT 4 (few shot) > GPT 3.5	GPT 4 = GPT 3.5-	Bing > GPT 4.0 > Humanss > GPT 3.5	GPT 4 > Bard
Oculoplastics	GPT 4 > GPT 3.5 = Human	GPT 4 > Human = GPT 3.5	Aeyeconsult > GPT 4	GPT 3.5 > GPT 4	GPT 4-0.3 = GPT 4-0 > GPT 4-1 = GPT 4-0.7 > GPT 3.5	Legacy > Plus	GPT 4 > 3.5	GPT 4 > GPT 4 (few shot) > GPT 3.5+	GPT 4 = GPT 3.5	GPT 4.0 > Bing > Human > GPT 3.5	GPT 4 > Bard
Optics	GPT 4 > GPT 3.5 = Human	-	Aeyeconsult > GPT 4	-	GPT 4-0.3 > GPT 4-0.7 > GPT 4-0 > GPT 4-1 > GPT 3.5	Legacy > Plus (BCSC) Plus > Legacy (OphthoQuestions)	GPT 4 > 3.5	-	GPT 4 = GPT 3.5#	Human > GPT 4.5 = Bing > GPT 3.5	-
Refractive Surgery	GPT 4 > GPT 3.5 = Human	GPT 4 > Human > GPT 3.5	Aeyeconsult > GPT 4	GPT 4 > GPT 3.5	GPT 4-0.7 > GPT 4-1 = GPT 4-0 > GPT 4-0.3 > GPT 3.5	ChatGPT Plus = ChatGPT Legacy	GPT 4 > 3.5	GPT 4 > GPT 4 (few shot) > GPT 3.5*	GPT 4 = GPT 3.5#	-	Bard > GPT 4

 doi:10.20944/preprints202503.1961.v1

Pathology	GPT 4 > Human = GPT 3.5	GPT 4 = Human = GPT 3.5	Aeyeconsult > GPT 4	GPT 4 > GPT 3.5	GPT 4-0.7 = GPT 4-0.3 = GPT 4-0 > GPT 4-1 > GPT 3.5	Plus > Legacy	GPT 4 > 3.5	GPT 4 > GPT 4 (few shot) > GPT 3.5+	GPT 4 = GPT 3.5-	-	-

*, +, -, #: Results categorized into same subspecialties in reporting.

3.3. Performance of LLM in Exam-Taking and Patient Education

Of the 14 studies which assessed LLM exam-taking capabilities, ten performed focused analysis of individual ophthalmology subspecialties (Table 4). GPT-4.0 was consistently the top performing LLM in all these studies, also scoring more than 50% of answers correct in all but one study (Table 5). The study where GPT-4.0 scored less than 50% of answers correct was performed using a Japanese question bank, highlighting the possible language barriers inherent to LLMs (Table 5).

20 studies looked at patient education with 16 assessing performance and relevance of output (8 comparative [16,29,37,45,47,56,62,63], 8 non-comparative [19,21,22,24,38,42,46,58]) and the remaining 4 assessing readability [28,35,36,61] (Table 1). Amongst the eight comparative studies, GPT-4.0 was deemed to produce the best patient educational materials in three of the four studies [37,47,63] that it was involved in, while GPT-3.5 performed the best in the remaining four studies [16,29,45,62]. Looking at non-comparative studies, it was found that only GPT-3.5 and 4.0 were used. Both models performed well with the majority of responses being "good", scoring more than 50%, or assessed as "relevant", depending on the scoring systems applied (Table 3b, appendix 1). Regarding readability of patient educational materials, a total of nine different scoring systems were used amongst the 4 studies, showing how varied assessment in this area can be. Results here varied greatly even within individual studies depending on the types of prompts given (Table 6). Both Bard and GPT-4.0 were able to significantly improve the readability scores by varying the types of prompts given [28,35]. GPT-3.5 performed inconsistently with the material produced being beyond the desired reading level in 1 study [35] and being at the desired reading level in another study [36]. In Eid's study [28], GPT-4.0 generated material that was easier to read than Bard. Meanwhile, without prompts, Bard was able to provide educational material that was easier to read in Eid's study [28] but not Kianian's [35].

Table 6. Readibility.

Study	LLMs	Scoring systems	Performance
Eid 2023	GPT 4, Bard	Flesch-Kincaid Reading Ease, Flesch-Kincaid Grade Level, Gunning Fog Index, Coleman-Liau Index, Simple measure of Gobbledygook, Automated readability Index, Linsear write readability score	FKRE: GPT 4 w/ prompt > Bard w/ prompt > Bard > ASOPRS > GPT 4 GFI: GPT 4 > ASOPRS > Bard > Bard w/ prompt > GPT 4 w/ prompt FKGL: GPT 4 > ASOPRS > Bard > Bard w/ prompt > GPT 4 w/ prompt CLI: GPT 4 > ASOPRS > Bard > Bard w/ prompt > GPT 4 w/ prompt SMOG: GPT 4 > ASOPRS > Bard > Bard w/ prompt > GPT 4 w/ prompt ARI: GPT 4 > ASOPRS > Bard > GPT 4 w/ prompt =

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1961.v1

			Bard w/ prompt LWRS: ASOPRS > GPT 4 > GPT 4 w/ prompt > Bard > Bard w/ prompt
Kianian (1) 2023	GPT 3.5	Flesch-Kincaid Reading Ease, Flesch-Kincaid Grade Level, Gunning Fog Index and Simple measure of Gobbledygook	FKRE: GPT > online resources FKGL: GPT < online resources GFI: GPT < online resources SMOG: GPT < online resources
Kianian (2) 2023	GPT 3.5, Bard	Flesch-Kincaid Grade Level	Prompt A: GPT < Bard Prompt B: GPT < Bard
Wu	ChatGPT	Flesch-Kincaid Grade Level, Gunning Fog Index, SMOG index, Dale-Chall-Score	FKGL: ChatGPT > AAO GFI: ChatGPT > AAO SMOG: ChatGPT > AAO Dale-Chall-Score: ChatGPT > AAO

3.4. Diagnostic and Management Capabilities of LLM

11 studies [23,26,37–39,48–50,53,57,63] assessed the diagnostic capabilities of LLM through cases and questions. GPT-4.0 consistently outperformed other LLMs (BingChat, WebMD, Bard) in coming to a diagnosis. (Table 7). As mentioned earlier, humans still performed better than LLMs in this field, nevertheless most studies did not report a great disparity between LLM score versus human scores. It was noted that most LLM outputs included a cautionary line such as "you should seek medical attention from a medical professional".

7 studies [16,22–24,37,50,57] evaluated the management of eye conditions, 5 of which employed multiple LLMs for comparison [16,23,37,50,57]. As noted earlier in LLM diagnostic ability, GPT-4.0 was also superior to GPT-3.5 and Bard in suggesting the appropriate management (Table 8). 2 studies compared the performance of LLMs against humans here [23,57]. The performance of GPT-4 was found to be better or similar to humans in both studies (Table 8). Again, it was noted that most LLMs included a cautionary disclaimer to seek professional medical advice.

Table 7. Diagnostic Capabilities of LLMs.

Study	LLMs	Evaluated Data	# of question / cases	Correct diagnosis
Raghu 2023	GPT 4	Clinical, biochemical and ocular data	111	GPT 4 diagnosis consistent with ophthalmologist in 75/111 cases with CST and 70/111 without CST
Liu 2024	GPT 3.5	FFA reports	1226	Ophthalmologists (89.35%) > Ophthalmologist interns (82.69%) > GPT 3.5-english prompts (80.05%) > ChatGPT 3.5-Chinese prompts (70.47%)
Lyons 2023	GPT 4, Bing Chat, WebMD	History only	44	Ophthalmologists in training (95%) >

Zandi 2024	GPT 4, Bard	History only	80	GPT 4 (93%) > Bing Chat (77%) > WebMD (33%) Correct diagnosis: GPT 4 (53.75%) > Bard (43.75%) Correct diagnosis somewhere in the conversation: GPT 4 (83.75%) > Bard (72.50%)
Shemer 2024	GPT 3.5	History only	126	Residents (75%) > Attendings (71%) > GPT 3.5 (54%)
Lim 2023	GPT 3.5, GPT 4, Bard	History only	2	GPT 4 (100%) = GPT 3.5 (100%) > Bard (50%)
Rojas-Carabali (1) 2023	GPT 3.5, GPT 4.0. Glass 1.0	History and examination findings	25	Completely correct: Uveitis specialist (76% - 92%) > Fellow (76%) > GPT 4 (60%) = GPT 3.5 (60%) Partially correct: Uveitis specialist (4% - 12%) > Fellow (4%) > GPT 4(4%) = GPT 3 (4%)
Delsoz 2023	GPT 3.5	History and examination findings	11	2 Ophthalmologist in training scored 72.7% 1 Ophthalmologist in training scored 54.5% GPT 3.5 scored 72.7%
Rojas-Carabali (2) 2023	GPT 3.5, GPT 4	History, examination findings and Images	6	Experts (100%) > GPT 4 (50%) = GPT 3.5 (50%) > Glass 1.0 (33%)
Taloni 2023	GPT 3.5, GPT 4	Question banks	646	GPT 4 (83.7%) > Humans (75.4% +/- 17.2%) > GPT 3.5 (68.1%)
Cai 2023	GPT 3.5, GPT 4, Bing Chat	Question banks	250	Humans (73.8%) > Bing (60.9%) > GPT 4 (59.4%) > GPT 3.5 (46.4%)

Table 8. Management.

Study	LLMs	Management
Biswas 2023	GPT 3.5	Can myopia be treated? Median 4.0 (Good); IQR 3.0-4.0; Range

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1961.v1

		3.0-4.0 Who can treat myopia? Median 4.0 (Good); IQR 3.0-4.0; Range 3.0-4.0 Which is the single most successful treatment strategy for myopia? Median 4.0 (Good); IQR 3.0-4.0; Range 1.0-5.0 What happens if myopia is left untreated? 4.0 (Good); IQR 4.0-5.0; Range 3.0-5.0
Lim 2023	GPT 3.5, GPT 4, Bard	GPT 3.5: Rating, n (%) Poor 5 (25), Borderline 7 (35), Good 8 (40) GPT 4.0: Rating, n (%) Poor 3 (15), Borderline 3 (15), Good 14 (70) Bard: Rating, n (%) Poor 3 (15), Borderline 8 (40), Good 9 (45)
Taloni 2023	GPT 3.5, GPT 4	GPT 4 Medical Treatment 196 (83.4%); Surgery 106 (74.6%) Humans Medical Treatment 181 ±40 (76.9 ±16.9%); Surgery 106 ± 24 (74.7 ±17.2%) GPT 3.5 Medical Treatment 153 (65.1%); Surgery 81 (57.0%)
Al-Sharif 2024	GPT 3.5, Bard	BARD: X^2 [3, N=25] = 28.0851, p<0.05 GPT 3.5: data not found (only in graph form)
Rojas-Carabali (2) 2023	GPT 3.5, GPT 4	Complete agreement of management and treatment plans: 91.6% of cases Disagreement in 8.3% (1 case)
Cai 2023	GPT 3.5, GPT 4, Bing Chat	GPT 3.5: 58.3% GPT 4.0: 77.0% Bing: 75.4% Humans: 76.1%
Cappellani 2024	GPT 3.5	Overall Median Score for "How is X treated" = 1 (from Likert scale of -3 to 2) General 2; Anterior segment and cornea 2; Glaucoma -1; Neuro-

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1961.v1

			Opth 2; Oncology 1; Paeds 1; Plastics 2; Retina and Uveitis 1

3.5. Clinical Administration Tasks

Only 3 studies evaluated the use of LLM for clinical administration tasks [33,43,55]. In two of the studies which gave LLMs more freedom to write, significant levels of hallucinations were observed.

1 study looked at using LLMs for discharge summary and operative notes writing [55]. It found that the quality of GPT's discharge summaries were affected by the quality of the prompts and tended to be valid but generic. Here, GPT-3.5 hallucinated its own model of the intraocular lenses utilised, but when prompted further, it was able to self-correct to improve the quality of output (Table 9).

Another study [33] evaluated manuscript abstract writing using GPT-3.5 and GPT-4.0. GPT-4.0 outperformed its predecessor on all fronts, including DISCERN score, helpfulness, truthfulness, and harmlessness. However, it was noted that both versions had hallucinated references (Table 9).

The last study was more focused, testing LLM on classifying texts into retina international classification of diseases (ICD) coding [43]. Of the 181 prompts given, 70% of the prompts had at least one correct ICD code generated by the LLM. This accuracy was reduced to 59% when assessed to generate only the correct ICD code (Table 9).

Table 9. Clinical Administration.

Study	LLMs	Clinical Administration	Performance (Qualitative)
Singh 2023	GPT3.5	Discharge Summary and Operative Notes Writing	Discharge Summaries: Divided into different categories including patient details, diagnosis, clinical course, discharge instructions, and case summary. Noted to have valid but very general texts, that upon further prompting, was able to remove generalised texts and provide responses to a greater specificity and detail Operative Notes: Subdivided into categories including patient details, diagnosis, clinical course, discharge instructions, and case summary. Was noted to have levels of inaccuracies and hallucinations which were quickly corrected upon further prompting
Hua 2023	GPT 3.5, GPT 4	Research Manuscript Writing	Mean helpfulness score: GPT4 > GPT3.5

 doi:10.20944/preprints202503.1961.v1

			Mean truthfulness score: GPT4 > GPT3.5 Main harmlessness score: GPT4 > GPT3.5 Modified AI-DISCERN score: GPT4 > GPT3.5 Mean hallucination rate: GPT3.5 > GPT4 Mean GPT-2 Output Detector Fake score: GPT3.5 > GPT4 Mean Sapling AI Detector Fake score: GPT3.5 > GPT4
Ong 2023	GPT 3.5	Retinal ICD Scoring	Correct: 137/181 (70%) Correct only: 106/181 (59%) Incorrect: 54/181 (30%)

3.6. LLM Inaccuracies and Harm

20 studies [15,16,19–24,29,33,37–40,42,46,47,56,60,62]detailed the hallucinations or inaccuracies produced by the LLMs. Bard demonstrated a significant inaccuracy rate, having the most inaccuracies in 4 of the 6 studies it was involved in [16,29,37,47,56,62]. On the other hand, GPT-4.0 had the lowest inaccuracy rate amongst LLMs in all 7 of the studies which included inaccuracy analysis and it [20,23,33,37,39,47,56]. In single LLM studies [15,19,21,22,24,38,40,42,46], we observed that inappropriate responses made up a minority of responses and were at times comparable to the frequency of errors in human answers (Table 10).

Table 10. Hallucination or inaccuracies.

Study	LLMs	Evaluation	Results
Multiple LLMs			
Tailor 2024	GPT 3.5, GPT 4, Claude 2, Bing, Bard	Degree of inaccuracy or correctness with risk of harm	Yes, great significance: Bard > Bing > Claude > GPT 3.5 > Expert + AI > GPT 4 > Expert. No: GPT 4 > Expert > Expert + AI > GPT 3.5 > Claude > Bing > Bard
Al-Sharif 2024	GPT 3.5, Bard	Degree of correctness	Contains both correct and incorrect or outdated: Bard > ChatGPT Contains completely incorrect Bard > ChatGPT
FerroDesideri 2023	GPT 3.5, Bard, Bing Chat	Degree of correctness	Entirely incorrect or contained critical errors: Bard > Bing > GPT 3.5
Yılmaz 2024	GPT 3.5, Bard, Bing AI, AAO website	Degree of inaccuracy	Inaccuracy: AAO > Bing > Bard > ChatGPT
Barclay 2023	GPT 3.5, GPT 4	Degree of inaccuracy	Incorrect facts, little significance: GPT 3.5 >

doi:10.20944/preprints202503.1961.v1

Lim 2023	GPT 3.5, GPT 4, Bard	Degree of inaccuracy	GPT 4 Incorrect facts, great significance: GPT 3.5 > GPT 4 Omission of information, little significance: GPT 3.5 > GPT 4 Omission of information, great significance: GPT 3.5 > GPT 4 Possible factual errors but unlikely to lead to harm: Bard > GPT > GPT 3.5 Inaccuracies that could significantly mislead patients or cause harm: GPT 3.5 = Bard > GPT
Pushpanathan 2023	GPT 3.5, GPT 4, Bard	Degree of inaccuracy	Inaccuracy: Bard > GPT 3.5 > GPT 4
Lyons 2023	GPT 4, Bing Chat, WebMD	Degree of inaccuracy	Grossly inaccurate statements: WebMD > Bing Chat > ChatGPT
Wilhelm 2023	GPT 3.5 Turbo, Command-xlarge-nightly, Claude, Bloomz	Hallucination frequency	Hallucinations: Claude-instant-v1.0 > Command-xlarge-nightly > Bloomz > GPT 3.5 Turbo
Hua 2023	GPT 3.5, GPT 4	Hallucination frequency	Hallucinations: GPT 3.5 > ChatGPT 4
Cai 2023	GPT 3.5, GPT 4, Bing Chat	Hallucination frequency	Hallucinations: GPT 3.5 > Bing > GPT 4
One LLM			
Ali 2023	GPT 3.5	Degree of correctness	Partially correct: 35% Completely factually incorrect: 25%
Cappellani 2024	GPT 3.5	Degree of correctness	27 / 120 graded as =< -1 (incorrect, varying degrees of harm)
Nikdel 2023	GPT 4	Degree of appropriateness	Inappropriate responses: Amblyopia: 5.6% Childhood myopia: 5.4%
Balas 2024	GPT 4	Degree of appropriateness	No inappropriate responses
Bernstein 2023	GPT 3.5	Degree of correctness or inappropriateness	Comparable with human answers (PR 0.92 95% CI [0.77 - 1.10])
Biswas 2023	GPT 3.5	Degree of inaccuracy	Inaccurate: 3.6% Flawed: 1.8%

Potapenko 2023	GPT 4	Degree of inaccuracy	Relevant with inaccuracies: 5 / 100
Liu 2024	GPT 3.5	Hallucination frequency	Hallucination: Step-Chinese > Step-English Misinformation: Step-Chinese > Step-English
Maywood 2024	GPT 3.5 Turbo	Hallucination frequency	12 responses are hallucinations

11 studies [19–21,24,33,37,40,46,56,60,63] evaluated the potential for, the extent of, and the likelihood of harm by the LLMs. In comparative studies [20,24,33,37,40,56,60,63], GPT-4.0 was less likely to generate harmful content when compared to GPT-3.5, Claude 2, Bing and Bard. In some studies, GPT-4.0 did not generate responses that constituted harm [19]. Only two studies compared harm from LLMs against that of humans, both of which found that likelihood of harm by humans and LLM were equivalent [21,56]. Extent of harm was equivalent between humans and chatbots in the study by Bernstein et al [21], while this was lowest in humans in the study by Tailor et al [56] (Table 11).

Table 11. Harm.

Study	LLMs	Potentially Harmful	Extent of Harm	Likelihood of harm
Bernstein 2023	GPT 3.5	-	Humans = Chatbots	Humans = Chatbots
Cappellani 2024	GPT 3.5	9 / 120 responses graded as potentially dangerous 27 / 120 responses graded as =< -1 (incorrect, varying degrees of harm)	-	-
Maywood 2024	GPT 3.5 Turbo	3 cases possible harm, 2 cases definitive harm	-	-
Wilhelm 2023	GPT 3.5 Turbo, Command-xlarge-nightly, Claude, Bloomz	Potentially harmful: Claude-instant v1.0 > Bloomz > Command-xlarge-nightly GPT 3.5-turbo no potentially harmful piece	-	-
Hua 2023	GPT 3.5, GPT 4	Updated versions have higher harmlessness scores	-	-
Barclay 2023	GPT 3.5, GPT 4	-	Incorrect facts, little significance: GPT 3.5 > GPT 4 Incorrect facts, great significance: GPT 3.5 > GPT 4	GPT 3.5 more likely than GPT 4

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1961.v1

			Omission of information, little significance: GPT 3.5 > GPT 4 Omission of information, great significance: GPT 3.5 > GPT 4	
Lim 2023	GPT 3.5, GPT 4, Bard	Inaccuracies that could significantly mislead patients or cause harm: GPT 3.5 = Bard > GPT	-	-
Tailor 2024	GPT 3.5, GPT 4, Claude 2, Bing, Bard	-	High risk harm: Bard > Bing > GPT 3.5 > Claude > GPT 4 > Expert + AI > Expert Low risk harm: Bard > Bing > Claude > Expert + AI > GPT 3.5 > GPT 4 > Expert	High likelihood: Bard > Bing > Claude > GPT 3.5 > Expert + AI = Expert = GPT 4 Low likelihood: Expert > GPT 4 > Expert + AI > Claude > Bard > Bing
Potapenko 2023	GPT 4	5/100 responses potentially harmful	-	-
Balas 2024	GPT 4	No responses constituting harm	-	-
Zandi 2024	GPT 4, Bard	Bard potentially more harmful than GPT 4	-	-

4. Discussion

This scoping review identified a total of 49 primary research studies applying LLMs in ophthalmology that were published in the five years and two months' time period of the search. These studies explored a wide range of applications, thereby providing breadth to this nascent field. The results of this scoping review suggest that while state-of-the-art LLMs can exhibit human-level performance, their real-world clinical application still faces several challenges. In the subsequent sections we discuss our results in the context of the study objectives and the implications for evidence and future research. Firstly, we evaluate the conduct of LLM studies in the field of ophthalmology. Thereafter we examine the performance of LLMs in ophthalmology based on current research, according to the major domains of their current applications – namely: patient education and exam taking, ophthalmic diagnostic capabilities, management capabilities, and clinical applications. We then discuss the existing drawbacks and hurdles facing the use of LLMs in ophthalmology. Finally, we discuss directions for future LLM research and development in ophthalmology. To our knowledge, this is the first such review to provide analysis and critique on the conduct of research in the field of LLMs and ophthalmology.

4.1. Evaluation of Past Methodologies

4.1.1. Issues Regarding Standardisation

Amidst the excitement to gather data regarding LLM applications, we have found that recent publications have not been seen to follow suggested frameworks or protocols. Rather, we see diverse

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1961.v1

pockets of data being collected by individual studies over multiple fields. While there is utility in this for widening the breadth of the data pool, the lack of standardized benchmarks leads researchers and experts to use varying benchmarks and implementations, resulting in inconsistent and sometimes incomparable evaluation results. We noted that all included studies also did not follow a fixed AI-related research protocol. This hampers the ability of follow up studies in reproducing these precedents. In the same vein, 12 of the 49 studies did not provide full examples of their prompts, potentially affecting reproducibility of their works. Following protocolised guidelines for AI related clinical trials, the open sharing of specific prompt techniques employed, and the usage of common benchmarks allows for research works in the realm of LLM to be more reproducible and suitable for direct comparison. The SPIRIT-AI and CONSORT-AI initiatives for clinical trials and interventions involving AI are examples of such protocols. Taking the SPIRIT-AI extension for example, interventions are required to specify the procedure for acquiring and selecting the input data for the AI intervention, and to specify the procedure for assessing and handling poor quality or unavailable input data [64]. Such accountability and transparency of steps would benefit future works seeking to build on previous research and allow for better comparison of results.

Beyond issues regarding transparency and standardisation, we noted inconsistencies in terms of the benchmarking of LLM performance. In our study, we encountered significant heterogeneity with respect to the grading systems with some studies grading on a Likert scale with 1 being the worst and 5 being the best [56] and others with 1 being the best and 3 being the worst [29]. Similarly, in evaluating diagnostic capabilities, scoring systems could be binary, meaning whether the responses were correct or incorrect [57], while other studies evaluated agreement with experts [50]. This hinders the inability to perform statistical analysis across studies and hence limits future meta-analysis in this field. Another source of inconsistency was the use of human evaluation. While human evaluation is necessary to grade areas such as harm, many of such evaluations appeared to be arbitrary and not based on evidence-based grading criteria. It is heartening to see open-source frameworks for benchmarking medical machine learning models such as MedPerf gaining traction, but these are yet to be widely adopted [65].

Most studies appeared to take the first output from their LLM platforms. Potential irreproducibility of answers from LLM platforms is a known fact. Answers generated on one occasion may differ from answers generated upon subsequent inputs of the same question. Singer et.al sought to overcome this by considering only the initial answers generated [54], however this runs the risk of missing out better or worse answers subsequently. Future works can seek to overcome this by taking the average of multiple outputs from their LLM platform, such as a best-of-three format.

4.1.2. Harm and Patient Safety

There was a general lack of consideration for patient harm, being evaluated in only nine studies and ethics only being formally discussed in 12 of the included studies (Table 2). As medicine is by nature a practice of non-maleficence, the objective of "doing no harm" has been central to clinical trials throughout medicine. Clinical trials employing AI should be no different, and aids in keeping the patient's welfare at the heart of everything. While many of the included studies were not direct applications on real patients per se, the limited attention paid to ethical safeguards serve as a timely reminder for the future as LLM applications assimilate further with medical practice. Despite being primarily evaluation of technology and not having live patient involvement in most of the studies, the real-world implications of these studies is undeniable. It would be useful for future works to state their findings in relation to patient safety – for instance the certainty to which GPT-4.0 could provide reliable medical advice within a specific field of ophthalmology. The Assessment List for Trustworthy AI (ALTAI) is an example of a self-assessment checklist published by the European Commission as an ethics guideline for trustworthy AI in July 2020 [66,67]. Checklists such as these could be included as supplementary material in AI studies relating to healthcare, serving as an ethical safeguard for patients.

Of the nine studies which evaluated the harm of LLM output, only two studies compared this to harm from human output [21,56]. While data on harm ought to be retrieved from LLMs, it would be insightful when such output is taken in relation to harm from human output. By obtaining human data in the same context for a basis of comparison, we can understand if LLM output is truly more harmful or would a human expert in the same confines of the study be any better. It is worth noting as well that the evaluation of harm involved human assessment in all nine studies in which it was evaluated, showing how LLMs still require a human safety net at this point (Table 2).

4.1.3. Other Issues Relating to Study Design

Amongst the included studies, there also appears to be multiple studies of similar design. Many of these employ zero-shot prompts to test the capabilities of LLMs in a particular area, and then assess their accuracy via exact match benchmarking, or via human assessment. The utility of such repetitive studies overlapping in design is questionable.

There was also an overwhelming bias towards using GPT-3.5 and GPT-4.0, making up the overwhelming majority of LLMs used in the included studies (Table 1). The benefit of doing so is the deeper exploration of GPT models, which are reportedly the most popular LLMs in use in recent history [68]. On the other hand, this runs the risk of under-representing other LLM models. Hence, while current research may provide a good indication of GPT applications in ophthalmology, it may not be representative of LLMs as a whole.

4.2. Evaluation of LLM Performance

Broadly speaking, all the included studies explored two main areas of “correctness” and “inaccuracies”, while a subset also studied the readability and harm of the LLM’s output. In studies where GPT-4.0 was included, it was amongst the best-performing LLMs in all domains of patient education and exam taking, ophthalmic diagnostic capabilities, management capabilities, and clinical applications.

In exam taking ability, LLM could equal and even surpass human scores. Even when faced with inductive subspecialties such as neuro-ophthalmology, GPT-4.0 could perform to the level of or even better than humans [41,57]. It is worth noting however, that all exam questions were text based. Two studies [34,49] attempted to assess the medical image reading ability of LLMs. However, these studies did so by using text-based descriptions of the images as input, rather than a raw image itself. It is known that LLMs have the ability to analyse images, and testing its ability to analyse raw medical image files directly would pave the way for further clinical utility within ophthalmology. Such attempts have already been carried out by non-ophthalmology based studies [69], but results are inconclusive at this point. We also identified an instance where GPT-4.0 was uncharacteristically poor. When a Japanese question bank was utilised, GPT-4.0 performed the worst, scoring less than 50%. Similarly, English prompts fared better than Chinese prompts when reporting fundus fluorescein angiography reports using GPT-3.5 [38]. Possible reasons for this could be the language difference, which GPT-4.0 might not have had exposure to or been trained on and hence fared poorer. This highlights a potentially inherent weakness in LLMs, whereby performance could be hindered by a lack of exposure to the language. While English is a dominant language globally, it is estimated to be spoken by only 20% of the world’s population [70]. The lack of multilingual support is a potential barrier that future works may consider exploring further.

In the area of patient education, it is perhaps unsurprising to find that expert edited LLM responses fared the best in terms of quality [56]. Tailor et. al. reported that human expert-edited LLM responses performed better than purely human expert responses and saved more time when compared to the experts creating a response from scratch [56]. Similarly, Bernstein's study comparing LLMs with those of ophthalmologists found comparable quality in the advice provided [21]. These works demonstrate an interesting direction where more effective human-AI collaboration might be achieved – an area underexplored by most studies which tended to benchmark pure LLM output on its own without human revision. On the other hand, a surprising finding was that LLM output could

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1961.v1

exceed that of humans in terms of empathy scores [56]. This isolated finding was another underexplored yet highly relevant area in this field, as healthcare is not merely a practice of knowledge, but also an art that requires the humanitarian touch. Lastly, the ability of LLMs to personalize the readability of patient education materials to their audience's comprehension levels strengthens its position for future adoption as demonstrated by Bard [28,63].

In terms of diagnostic and management capabilities, LLMs appeared to struggle more when coming to a diagnosis [26,49,50] but fared better when asked for the management plan after the diagnosis had been established. This reflects the higher order thinking that is required for making a diagnosis. In the study by Rojas-Carabali et al., we note that LLMs were possibly disadvantaged in that they were given text descriptions of images, while humans were given the images to assess [50]. It would be useful for future works to assess how LLMs would perform against humans if both were given the same images to come to a diagnosis. It has been shown that a simple combination of patient history and chief complaint could predict an overall diagnostic accuracy of approximately 90% of neuro-ophthalmology cases when read by human assessors [71]. These results seem to suggest that the ability of LLMs to interpret written information falls short of humans despite their potentially greater wealth of knowledge. Also as pointed out earlier, most LLMs included medical disclaimers when posed with diagnostic questions (e.g. "you should seek medical attention from a medical professional"). This drives home the point that while LLMs may close the gap on human accuracy in diagnosis, there is still some way to go before their opinion is taken to be as legitimate as that of a medical professional.

The area of clinical administration tasks was only covered by three studies touching different areas. The dearth of data here calls for more work to explore this area of untapped potential. Singh's research highlights ChatGPT's ability to swiftly generate detailed ophthalmic discharge summaries and operative notes [55], showcasing its potential to streamline administrative processes with tailored content and rapid response times. Similarly, Ong's study demonstrates ChatGPT's capability to interpret text accurately [43], suggesting its potential to ease physician burden in tasks like ICD coding. Moreover, Hua's investigation [33] into manuscript writing reveals that AI-generated ophthalmic scientific abstracts are comparable in quality between different versions of GPT-3.5 and GPT-4.0, though factual errors in references indicate a need for further refinement. Overall, these findings show that LLMs can be helpful for administrative tasks in ophthalmology, but more work is needed to establish them further for practical use in healthcare and ophthalmology.

4.3. Directions for Future Works

4.3.1. Standard Framework for Assessing Accuracy, Validity and Harm

Much like other reviews and commentaries on LLMs in other fields [72,73], this study calls for future works to follow standardised benchmarks and frameworks for assessing the accuracy and validity of LLMs in clinical settings [74]. A robust framework would offer clear guidelines in areas such as providing comprehensive context on diseases, precise wording reporting, incorporating diverse question formats, adopting learning techniques, and using standardised metrics like ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and METEOR (Metric for Evaluation of Translation with Explicit ORdering) [73]. The Standardized Assessment Framework for Evaluations of Large Language Models in Medicine (SAFE-LLM) is one such model which sets out to unify evaluation standards, facilitating the comparison and improvement of LLMs in medical applications [75]. Such frameworks establish a common language and understanding between developers and end-users, fostering collaboration and partnership in the advancement and deployment of LLMs. However, as seen in this review they are yet to be widely adopted. Medical image processing is a growing field within AI, and comprehensive benchmarks such as MedSegBench are already being developed in this area [76]. This is especially relevant to ophthalmology – a field highly dependent on image interpretation. Moving forward it is imperative that standardised benchmarks in this area are employed as well.

Clinical trial protocols such as CONSORT-AI and SPIRIT-AI emphasise the importance of describing the results of any performance errors and how errors were identified. Conducting future LLM studies in line with such protocols would address the critical need for transparency and accountability in assessing the safety and reliability of LLMs, further contributing to building trust among developers and end-users. With standardised AI study protocols in place, stakeholders can communicate their findings with more transparency and uniformity, ensuring the ethical and responsible use of LLMs in various domains, including healthcare. Many of the current works of LLMs in ophthalmology are not yet in the clinical trial stage with live patient testing. This could be due to the ethical considerations mentioned in the next section. Nevertheless, defined protocols even for such studies can build the foundation of transparency and accountability for future real-world patient applications.

4.3.2. Greater Evaluation and Strategies Towards Ethical Considerations

It is encouraging that considerations and suggestions regarding medical ethics were raised in some of the studies in this review. We highlight some of them here.

The inaccuracies of LLM output raises the risk of harm to patients. Al-Sharif et al suggested that LLMs be trained solely on supervised evidence-based ophthalmology datasets, to maintain the "purity" of what the LLM "knows" [16]. Singer [54] and Antaki's [18] studies emphasise the importance of using verified sources to ensure the trustworthiness and accuracy of information provided by LLMs. It has also been shown that fine-tuning medical LLMs significantly improves their safety, reducing their tendency to comply with harmful requests [77]. LLM as a tool has the potential to do good and harm, and it is the responsibility of LLM creators and clinicians to ensure that they are developed to adequate safety standards to limit the harm on patients.

Despite the ability of LLMs to have a reasonably high rate of accuracy, LLM inaccuracies may still be interspersed amongst these facts. Bernstein et al highlighted that these partial truths in LLM outputs may lure patients into a false sense of trust [21]. As shown previously, medical disclaimers are frequently used at the end of LLM medical outputs to mitigate this. Raghu et al also highlighted that equally important is the education of end-users about the capabilities, potential risks, and benefits of this technology [48].

Bernstein et al also noted that patient healthcare information would have to be entered into LLMs to obtain customised and individualised output. OpenAI's privacy policy states that they "may collect Personal Information that is included in the input" [78]. Patient data entering the online domain, or into the servers of private companies are at risk of being hacked and has implications for patient confidentiality and data privacy [79]. In the clinical deployment of LLMs, policies should include strategies to safeguard this. Raghu et al suggested that until such safeguards are in place, only anonymised patient data should be entered into these LLMs [48], while Tao et al suggested to keep personalisation of output offline after online drafts are generated [58].

Regarding information source, there lies the issue of plagiarism as brought up by Tao et al [58]. Data authenticity, data provenance and intellectual property contamination are issues that LLMs are still grappling with [79]. Text generated from LLM output may be taken from copywrite sources illegitimately. We have also seen cases of LLMs hallucinating references [33]. To date, LLM reliability for citation and reference have been found to be inconsistent and occasionally very poor [80]. Further fine-tuning of LLMs in this regard should be a priority moving forward, whilst end-users ought to query the original sources and cite where credit is due.

Jiao et al raised the issue of biases inherent to LLMs, which risks amplifying existing health disparities. LLMs may refer to source material that does not represent all patient populations equally resulting in unequal treatment for specific patient groups [34]. We have also seen how GPT underperformed when tested in a non-English language [51], potentially underserving patients who speak non-English languages. Utilising adversarial testing and bias detection algorithms to identify and remove any discriminatory patterns in the prompts or the AI-generated outputs are possible ways to tackle these biases [81]. While training LLMs on diverse and representative sources are

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1961.v1

possible ways to reduce inequalities associated with LLM use, Kianian et al also argues that improving readability of LLM output can reduce such inequalities too [36]. This is because with poor readability comes poorer health literacy, which have been show to disproportionately affect populations of lower socioeconomic status [82]. Collaboration between prompt engineers, bioethicists and patient advocates may help in designing prompts that are inclusive, diverse, and free from biases based on factors such as race, ethnicity, gender, or socioeconomic status [81].

Finally, Tao et al also questioned how the burden of legal responsibility should be divided between physician and LLM, especially for cases of patient harm or privacy breaches [58]. As AI systems become increasingly autonomous and capable of decision making, it is important to ensure that there is accountability for their actions. This includes ensuring that AI systems are transparent and that there are oversight mechanisms in place to address any errors [66].

4.3.3. Techniques for Improving LLM's Accuracy and Interpretability

In general, prompt engineering, a transformative approach in natural language processing, involves the development of tailored input prompts or instructions to guide LLMs in generating desired outputs or responses. Examples of such methodologies include Retrieval-Augmented Generation (RAG) and fine-tuning. Fine-tuning involves adjusting the model's parameters based on task-specific datasets, essentially operating in a "close-book" manner. Conversely, RAG functions in an "open book" setting, harnessing external information sources to retrieve and integrate relevant data, thereby enhancing the model's comprehension and generative capabilities. For instance, in the domain of healthcare education [83], RAG was chosen due to its capability to provide traceable responses, enhancing trust and explainability, its scalability in accessing vast healthcare knowledge bases, and its flexibility for rapid updates in alignment with evolving clinical guidelines.

Some studies, as seen in patient education use-cases, improved LLM performance by innovative prompt engineering and fine-tuning. This suggests that the limiting factor of output may not only be in the LLM itself, but rather the types of prompts given. Further works exploring the effect of using varying styles of prompts on LLM output would aid in verifying this. Both Eid [28] and Kianian [35] improved the readability of their patient education material output by specifying a level of reading (the 6th-grade reading level in their case). Lim et al. found that even by using a simple prompt "That does not seem quite right. Could you kindly review?", GPT-3.5, GPT-4.0 and Bard were able to able to demonstrate substantial self-correction abilities [37]. Bernstein et al. used instruction prompt engineering to answer patient's questions. This prompt technique uses explicit instructions or cues about the task at hand to adapt the behaviour of the LLM model [21]. With the use of these prompts, they found that human-written and AI-generated answers to patient ophthalmology-related questions were very comparable in terms of accuracy and harm. Notably, assessors could not be "definitely sure" if the responses were AI or human generated in the majority of cases. Another study by Liu et al. [38] utilised chain-of-thought-inspired prompt techniques to elucidate a step-by-step reasoning process from GPT-3.5 for both English and Chinese prompts. Interestingly this study found that English prompts performed better for diagnostic and inference capabilities, as well as providing more complete reasoning steps, suggesting that choice of language affects the quality of output as well.

4.4. Strengths and Limitations

The strengths of this review include the wide search strategy, involving eight bibliographic databases involving both the fields of medical and information technology. The time frame chosen as part of the search criteria (2019 - 2024), gives a reflection of the scene of LLM usage within ophthalmology in this period of LLM breakthroughs since the release of BERT in October 2018. This review also followed best practices in the PRISMA-ScR for conducting a scoping review [14,84]. Expert opinions in the fields of LLM and ophthalmology were also consulted. This was in line with best practice recommendations by the Institute of Medicine (US) Committee on Standards for

 doi:10.20944/preprints202503.1961.v1

Systematic Reviews of Comparative Effectiveness Research [85], as well as Arksey and O'Malley's and Levac et al.'s frameworks for scoping reviews [13,86].

To our knowledge, this is the first scoping review to critique the methodology and conduct of LLM research in ophthalmology. Based on these findings describing the current landscape of LLM research in ophthalmology, this study puts forth key recommendations to strengthen the lack of standardisation and ethical regulation amongst LLM-related studies, and tangible steps to improve the conduct of future works in this field.

Nevertheless, there were shortcomings with regards to the conduct of this review. The search terms chosen aimed to capture all studies relating to LLMs and ophthalmology within the given timeframe. However, due to the rapidly evolving nature of LLMs, newer yet relevant search terms may inadvertently be missed out on. The use of MeSH terms was done with the aim of improving reproducibility of results. However, this ran the risk of missing out on recent articles not yet indexed. The strict exclusion criteria on study design also sought to improve the quality of evidence collected in this review. Nonetheless, this also runs the risk of missing out on novel data, such as from case reports, which is especially possible in the growing field of LLMs. As a trending and growing field, advancements in LLMs are rapid, and recent developments are bound the be missed. For instance, promising and relevant LLM models such as DeepSeek were not covered in any of the included studies. Constant attempts to update the paper to chase each new publication hampers the progress of the paper. Overall, we believe that we have captured a significant portion of time and publications to represent this field at a time when interest in LLMs skyrocketed, while allowing the thoughtful evaluation and discussion of our findings.

The heterogeneity of measures employed in assessing LLMs, and the wide range of study designs made it difficult to compare findings across studies, and to provide firm conclusions. We therefore sought to summarise the assessment of LLMs by the various studies by placing these evaluations into the overarching categories of "exam taking and patient education", "diagnostic capability", "management capability", "clinical administration", and "inaccuracies and harm". Many of the included studies utilised subjective modes of assessment that lacked in strength of evidence, for instance in determining degree of "correctness" or frequency of hallucinations and nonlogical reasoning. Nevertheless, such studies were included as this review did not discriminate against studies based on the strength of their study design, and to reflect the current climate of how LLMs are assessed.

5. Conclusion

LLMs have received considerable attention through their introduction to the general public and have found potential applications in the field of medicine, and in particular, ophthalmology. The main use cases are in exam-taking, patient education, diagnosis and management, and clinical administration. We presented an overview of the landscape of LLM applications in ophthalmology. In our study, we found that the majority of LLMs perform acceptably, with GPT-4.0 having one of the best performances. However, issues pertaining to hallucination, inaccuracies, and harm still exist. We also evaluated how past studies of LLMs in ophthalmology have been carried out and summarized their findings. We have also identified gaps in current literature and have made suggestions for future works to improve on, with hopes that future works can form a more cohesive and clinically useful pool of knowledge that can be applied to patients in a safe and ethical manner. We conclude by advocating for the adoption of standardised frameworks to assess LLMs in healthcare and recommend techniques to improve the performance of LLM in niche fields such as ophthalmology.

6. Glossary

"Exam taking" refers to the ability of the LLMs to answer multiple choice questions set for licensing examinations which are taken by ophthalmology trainees.

 doi:10.20944/preprints202503.1961.v1

"Patient education" refers to the ability of the LLM to produce material appropriate for the laymen to introduce medical conditions, provide guidance on treatment and/or monitoring.

"Diagnostic capability" refers to the ability of the LLM to come to the right diagnosis, or at least relevant differentials when posed with questions describing clinical presentations, findings and/or clinical images. Data was obtained regarding the source of input for the LLM, the number of questions by which the LLM was assessed, and the results of these studies.

"Management capability" refers to the ability of the LLM to manage and treat various eye conditions. Their proposed management plans were graded by trained ophthalmologists and scored accordingly. Scores were extracted and analysed to compare the LLMs in management.

"Clinical administration" refers to utilising the LLM to assist with clinical paperwork, this could be through simplifying clinical notes writing, discharge summaries or optimising clinical scheduling.

"Inaccuracies" refer to the extent of incorrect answers displayed by the LLMs in response to questions. Data was obtained on the form of inaccuracy made, which varied from study to study. These included "Degree of correctness", "Degree of inaccuracy", "Hallucination frequency", and "Degree of appropriateness" as stated by the individual studies based on their grading systems.

"Harm" refers to the possibility of the answers generated by LLMs, often for management purposes, causing potential harm to patients if used clinically. Data was obtained regarding the potential and likelihood of harm, as well as the extent of harm.

Author contributions: CRediT Conceptualisation, S.Y.K.C., F.X. and L.Z.K.; Data Curation, S.Y.K.C., L.K.S.A., A.W.Y., C.S.Y.C., F.X. and L.Z.K.; Formal Analysis, S.Y.K.C., L.K.S.A., A.W.Y., F.X. and L.Z.K.; Investigation, S.Y.K.C., L.K.S.A., A.W.Y., F.X. and L.Z.K.; Methodology, S.Y.K.C., L.K.S.A., A.W.Y., F.X. and L.Z.K.; Resources, S.Y.K.C., L.K.S.A., A.W.Y., C.S.Y.C., F.X. and L.Z.K.; Supervision, S.Y.K.C., F.X. and L.Z.K.; Validation, S.Y.K.C., L.K.S.A., F.X. and L.Z.K.; Writing – original draft, S.Y.K.C., L.K.S.A., A.W.Y., C.S.Y.C., F.X. and L.Z.K.; Writing – review and editing, S.Y.K.C., L.K.S.A., A.W.Y., C.S.Y.C., F.X. and L.Z.K.

Funding: This research received no external funding.

Literature search statement: A search of PubMed, Embase, SCOPUS, Web of Science, the Institute of Electrical and Electronics Engineers (IEEE) journals, the Association for Computing Machinery (ACM) journals, Google Scholar and DataBase systems and Logic Programming (DBLP) was performed from 2019 - 2024. The search strategy can be found in the supplementary information. (Appendix Table A1) MeSH (medical subject heading) terms includes:

1) In ophthalmology included: Ophthalmology, Ocular Surgery, Eye Disease, Eye Diseases, Eye Disorders.
2) For LLMs: Large Language Model, large language models, large language modelling, Chatbot, ChatGPT, GPT, chatbots, google bard, bing chat, BERT, RoBERTa, distilBERT, BART, MARIAN, llama, palm.

No additional search filters were applied.

The inclusion criteria were (1) studies utilising LLMs (2) studies involving ophthalmology (3) studies published from January 2019 to March 2024. The exclusion criteria were (1) study designs that were reviews, systematic reviews and meta-analyses, case reports, case series, guidelines, letters, or protocols (2) non-English studies. Foreign studies that were not in English were excluded. Studies that were not in English were not translated for use in this study.

Conflict of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Search strategy.

Database	Search terms used	Results
Pubmed	(Ophthalmology[MeSH Terms]) OR (Ocular Surgery) OR (Eye Disease) OR (Eye Diseases) OR (Eye Disorders)	427

 doi:10.20944/preprints202503.1961.v1

	AND (Large Language Model) OR (large language models) OR (large language modelling) OR (Chatbot) OR (ChatGPT) OR (GPT) OR (chatbots) OR (google bard) OR (bing chat) OR (BERT) OR (RoBERTa) OR (distilBERT) OR (BART) OR (MARIAN) OR (llama) OR (palm) Limits: 2019 - 2024	Retrieved 11/02/2024
Embase	((Ophthalmology) OR (Ocular Surgery) OR (Eye Disease) OR (Eye Diseases) OR (Eye Disorders)).mp. AND (Large Language Model) OR (large language models) OR (large language modelling) OR (Chatbot) OR (ChatGPT) OR (GPT) OR (chatbots) OR (google bard) OR (bing chat) OR (BERT) OR (RoBERTa) OR (distilBERT) OR (BART) OR (MARIAN) OR (llama) OR (palm).mp. Limits: 2019 - 2024	122 Retrieved 11/02/2024
SCOPUS	TITLE-ABS-KEY ((ophthalmology) OR (ocular AND surgery) OR (eye AND disease) OR (eye AND diseases) OR (eye AND disorders)) AND TITLE-ABS-KEY ((large AND language AND model) OR (large AND language AND models) OR (large AND language AND modelling) OR (chatbot) OR (chatgpt) OR (gpt) OR (chatbots) OR (google AND bard) OR (bing AND chat) OR (bert) OR (roberta) OR (distilbert) OR (bart) OR (marian))) Limits: 2019 - 2024	236 Retrieved 11/02/2024
Web Of Science	(Ophthalmology) OR (Ocular Surgery) OR (Eye Disease) OR (Eye Diseases) OR (Eye Disorders) (Abstract) AND (Large Language Model) OR (large language models) OR (large language modelling) OR (Chatbot) OR (ChatGPT) OR (GPT) OR (chatbots) OR (google bard) OR (bing chat) OR (BERT) OR (RoBERTa) OR (distilBERT) OR (BART) OR (MARIAN) OR (llama) OR (palm) (Abstract) Limits: 2019 - 2024	86 Retrieved 11/02/2024
IEEE	(("All Metadata":Ophthalmology) OR ("All Metadata":"Ocular Surgery") OR ("All Metadata":"Eye Disease") OR ("All Metadata":"Eye Diseases") OR ("All Metadata":"Eye disorders")) AND (("All Metadata":"Large Language Model") OR ("All Metadata":large language models) OR ("All Metadata":ChatGPT)	8 Retrieved 11/02/2024

 doi:10.20944/preprints202503.1961.v1

	OR ("All Metadata":GPT) OR ("All Metadata":chatbots) OR ("All Metadata":Chatbot) OR ("All Metadata":"google bard") OR ("All Metadata":"bing chat") OR ("All Metadata":BERT) OR ("All Metadata":RoBERTa) OR ("All Metadata":distilBERT) OR ("All Metadata":BART) OR ("All Metadata":MARIAN) OR ("All Metadata":llama) OR ("All Metadata":palm)) Limits: 2019 – 2024 and journals	
ACM	[[All: ophthalmology] OR [All: "ocular surgery"] OR [All: "eye disease"] OR [All: "eye diseases"] OR [All: "eye disorders"]] AND [[All: "large language model"] OR [All: or] OR [All: "large language models"] OR [All: "chatgpt"] OR [All: "gpt"] OR [All: "chatbots"] OR [All: "chatbot"] OR [All: "google bard"] OR [All: "bing chat"] OR [All: "bert"] OR [All: "roberta"] OR [All: "distilbert"] OR [All: "bart"] OR [All: "marian"] OR [All: "llama"] OR [All: "palm"]] Limits: 2019 – 2024	78 Retrieved 11/02/2024
Google Scholar	Ophthalmology "Large Language Model" -preprint Limits: 2019 - 2024	276 Retrieved 11/02/2024
DBLP	ophthal* type:Journal_Articles: Limits: 2019 - 2024	69 Retrieved 11/02/2024
Total		1302

Table A2. Scoring systems examples.

Study	Scoring system for responses
Biswas 2023	Likert scale where higher ratings indicated greater quality of information 1: very poor 2: poor 3: acceptable 4: good 5: very good
Nikdel 2023	Acceptable, incomplete or unacceptable
Sharif 2024	comprehensive, correct but inadequate, mixed with correct and incorrect/out-dated data or completely incorrect
Maywood 2024	Correct and comprehensive, correct but inadequate, incorrect
Pushpanathan 2023	Poor. borderline, good
Cappellani 2024	-3: potentially dangerous -2: very poor -1: poor 0: no response 1: good 2: very good 2*: excellent
Patil 2024	Likert scale of 5 from very poor (harmful and incorrect) to excellent (no errors or false claim)

References

1. De Angelis, L.; Baglivo, F.; Arzilli, G.; Privitera, G.P.; Ferragina, P.; Tozzi, A.E.; Rizzo, C. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Front Public Health* **2023**, *11*, 1166120, doi:10.3389/fpubh.2023.1166120.
2. Haupt, C.E.; Marks, M. AI-Generated Medical Advice-GPT and Beyond. *Jama* **2023**, *329*, 1349-1350, doi:10.1001/jama.2023.5321.
3. Kung, T.H.; Cheatham, M.; Medenilla, A.; Sillos, C.; De Leon, L.; Elepaño, C.; Madriaga, M.; Aggabao, R.; Diaz-Candido, G.; Maningo, J.; et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health* **2023**, *2*, e0000198, doi:10.1371/journal.pdig.0000198.
4. Liu, Z.; He, X.; Liu, L.; Liu, T.; Zhai, X. Context Matters: A Strategy to Pre-train Language Model for Science Education. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*; Springer Nature Switzerland: 2023; pp. 666–674.
5. Potapenko, I.; Boberg-Ans, L.C.; Stormly Hansen, M.; Klefter, O.N.; van Dijk, E.H.C.; Subhi, Y. Artificial intelligence-based chatbot patient information on common retinal diseases using ChatGPT. *Acta Ophthalmol* **2023**, *101*, 829-831, doi:10.1111/aos.15661.
6. Thirunavukarasu, A.J.; Hassan, R.; Mahmood, S.; Sanghera, R.; Barzangi, K.; El Mukashfi, M.; Shah, S. Trialling a Large Language Model (ChatGPT) in General Practice With the Applied Knowledge Test: Observational Study Demonstrating Opportunities and Limitations in Primary Care. *JMIR Med Educ* **2023**, *9*, e46599, doi:10.2196/46599.
7. Betzler, B.K.; Chen, H.; Cheng, C.Y.; Lee, C.S.; Ning, G.; Song, S.J.; Lee, A.Y.; Kawasaki, R.; van Wijngaarden, P.; Grzybowski, A.; et al. Large language models and their impact in ophthalmology. *Lancet Digit Health* **2023**, *5*, e917-e924, doi:10.1016/s2589-7500(23)00201-7.
8. Nath, S.; Marie, A.; Ellershaw, S.; Korot, E.; Keane, P.A. New meaning for NLP: the trials and tribulations of natural language processing with GPT-3 in ophthalmology. *Br J Ophthalmol* **2022**, *106*, 889-892, doi:10.1136/bjophthalmol-2022-321141.
9. Soh, Z.D.; Cheng, C.Y. Application of big data in ophthalmology. *Taiwan J Ophthalmol* **2023**, *13*, 123-132, doi:10.4103/tjo.TJO-D-23-00012.
10. Wong, M.; Lim, Z.W.; Pushpanathan, K.; Cheung, C.Y.; Wang, Y.X.; Chen, D.; Tham, Y.C. Review of emerging trends and projection of future developments in large language models research in ophthalmology. *Br J Ophthalmol* **2024**, *108*, 1362-1370, doi:10.1136/bjo-2023-324734.
11. Jin, K.; Yuan, L.; Wu, H.; Grzybowski, A.; Ye, J. Exploring large language model for next generation of artificial intelligence in ophthalmology. *Front Med (Lausanne)* **2023**, *10*, 1291404, doi:10.3389/fmed.2023.1291404.
12. Ibrahim, H.; Liu, X.; Rivera, S.C.; Moher, D.; Chan, A.W.; Sydes, M.R.; Calvert, M.J.; Denniston, A.K. Reporting guidelines for clinical trials of artificial intelligence interventions: the SPIRIT-AI and CONSORT-AI guidelines. *Trials* **2021**, *22*, 11, doi:10.1186/s13063-020-04951-6.
13. Arksey, H.; O'Malley, L. Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology* **2005**, *8*, 19-32, doi:10.1080/1364557032000119616.
14. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Annals of Internal Medicine* **2018**, *169*, 467-473, doi:10.7326/m18-0850 %m 30178033.
15. Ali, M.J. ChatGPT and Lacrimal Drainage Disorders: Performance and Scope of Improvement. *Ophthalmic Plast Reconstr Surg* **2023**, *39*, 221-225, doi:10.1097/iop.0000000000002418.
16. Al-Sharif, E.M.; Penteado, R.C.; Dib El Jalbout, N.; Topilow, N.J.; Shoji, M.K.; Kikkawa, D.O.; Liu, C.Y.; Korn, B.S. 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*, 303-311, doi:10.1097/iop.0000000000002567.
17. Antaki, F.; Milad, D.; Chia, M.A.; Giguère, C.; Touma, S.; El-Khoury, J.; Keane, P.A.; Duval, R. Capabilities of GPT-4 in ophthalmology: an analysis of model entropy and progress towards human-level medical question answering. *Br J Ophthalmol* **2024**, *108*, 1371-1378, doi:10.1136/bjo-2023-324438.

18. 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* **2023**, *3*, 100324, doi:10.1016/j.xops.2023.100324.
19. Balas, M.; Janic, A.; Daigle, P.; Nijhawan, N.; Hussain, A.; Gill, H.; Lahaie, G.L.; Belliveau, M.J.; Crawford, S.A.; Arjmand, P.; et al. Evaluating ChatGPT on Orbital and Oculofacial Disorders: Accuracy and Readability Insights. *Ophthalmic Plast Reconstr Surg* **2024**, *40*, 217-222, doi:10.1097/iop.0000000000002552.
20. Barclay, K.S.; You, J.Y.; Coleman, M.J.; Mathews, P.M.; Ray, V.L.; Riaz, K.M.; De Rojas, J.O.; Wang, A.S.; Watson, S.H.; Koo, E.H.; et al. Quality and Agreement With Scientific Consensus of ChatGPT Information Regarding Corneal Transplantation and Fuchs Dystrophy. *Cornea* **2024**, *43*, 746-750, doi:10.1097/ico.0000000000003439.
21. Bernstein, I.A.; Zhang, Y.V.; Govil, D.; Majid, I.; Chang, R.T.; Sun, Y.; Shue, A.; Chou, J.C.; Schehlein, E.; Christopher, K.L.; et al. Comparison of Ophthalmologist and Large Language Model Chatbot Responses to Online Patient Eye Care Questions. *JAMA Netw Open* **2023**, *6*, e2330320, doi:10.1001/jamanetworkopen.2023.30320.
22. Biswas, S.; Logan, N.S.; Davies, L.N.; Sheppard, A.L.; Wolffsohn, J.S. Assessing the utility of ChatGPT as an artificial intelligence-based large language model for information to answer questions on myopia. *Ophthalmic Physiol Opt* **2023**, *43*, 1562-1570, doi:10.1111/opo.13207.
23. Cai, L.Z.; Shaheen, A.; Jin, A.; Fukui, R.; Yi, J.S.; Yannuzzi, N.; Alabiad, C. Performance of Generative Large Language Models on Ophthalmology Board-Style Questions. *Am J Ophthalmol* **2023**, *254*, 141-149, doi:10.1016/j.ajo.2023.05.024.
24. Cappellani, F.; Card, K.R.; Shields, C.L.; Pulido, J.S.; Haller, J.A. Reliability and accuracy of artificial intelligence ChatGPT in providing information on ophthalmic diseases and management to patients. *Eye (Lond)* **2024**, *38*, 1368-1373, doi:10.1038/s41433-023-02906-0.
25. Ćirković, A.; Katz, T. Exploring the Potential of ChatGPT-4 in Predicting Refractive Surgery Categorizations: Comparative Study. *JMIR Form Res* **2023**, *7*, e51798, doi:10.2196/51798.
26. Delsoz, M.; Raja, H.; Madadi, Y.; Tang, A.A.; Wirostko, B.M.; Kahook, M.Y.; Yousefi, S. The Use of ChatGPT to Assist in Diagnosing Glaucoma Based on Clinical Case Reports. *Ophthalmol Ther* **2023**, *12*, 3121-3132, doi:10.1007/s40123-023-00805-x.
27. 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*, 389-393, doi:10.22336/rjo.2023.61.
28. Eid, K.; Eid, A.; Wang, D.; Raiker, R.S.; 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*, 212-216, doi:10.1097/iop.0000000000002549.
29. 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. *International Journal of Retina and Vitreous* **2023**, *9*, 71, doi:10.1186/s40942-023-00511-7.
30. Fowler, T.; Pullen, S.; Birkett, L. Performance of ChatGPT and Bard on the official part 1 FRCOphth practice questions. *Br J Ophthalmol* **2024**, *108*, 1379-1383, doi:10.1136/bjo-2023-324091.
31. Haddad, F.; Saade, J.S. Performance of ChatGPT on Ophthalmology-Related Questions Across Various Examination Levels: Observational Study. *JMIR Med Educ* **2024**, *10*, e50842, doi:10.2196/50842.
32. Hu, W.; Wang, S.Y. Predicting Glaucoma Progression Requiring Surgery Using Clinical Free-Text Notes and Transfer Learning With Transformers. *Transl Vis Sci Technol* **2022**, *11*, 37, doi:10.1167/tvst.11.3.37.
33. Hua, H.U.; Kaakour, A.H.; Rachitskaya, A.; Srivastava, S.; Sharma, S.; Mammo, D.A. Evaluation and Comparison of Ophthalmic Scientific Abstracts and References by Current Artificial Intelligence Chatbots. *JAMA Ophthalmol* **2023**, *141*, 819-824, doi:10.1001/jamaophthalmol.2023.3119.
34. Jiao, C.; Edupuganti, N.R.; Patel, P.A.; Bui, T.; Sheth, V. Evaluating the Artificial Intelligence Performance Growth in Ophthalmic Knowledge. *Cureus* **2023**, *15*, e45700, doi:10.7759/cureus.45700.
35. Kianian, R.; Sun, D.; Crowell, E.L.; Tsui, E. The Use of Large Language Models to Generate Education Materials about Uveitis. *Ophthalmol Retina* **2024**, *8*, 195-201, doi:10.1016/j.oret.2023.09.008.

36. Kianian, R.; Sun, D.; Giaconi, J. Can ChatGPT Aid Clinicians in Educating Patients on the Surgical Management of Glaucoma? *J Glaucoma* **2024**, *33*, 94-100, doi:10.1097/ijg.0000000000002338.
37. Lim, Z.W.; Pushpanathan, K.; Yew, S.M.E.; Lai, Y.; Sun, C.H.; Lam, J.S.H.; Chen, D.Z.; Goh, J.H.L.; Tan, M.C.J.; Sheng, B.; et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *EBioMedicine* **2023**, *95*, 104770, doi:10.1016/j.ebiom.2023.104770.
38. Liu, X.; Wu, J.; Shao, A.; Shen, W.; Ye, P.; Wang, Y.; Ye, J.; Jin, K.; Yang, J. Uncovering Language Disparity of ChatGPT on Retinal Vascular Disease Classification: Cross-Sectional Study. *J Med Internet Res* **2024**, *26*, e51926, doi:10.2196/51926.
39. Lyons, R.J.; Arepalli, S.R.; Fromal, O.; Choi, J.D.; Jain, N. Artificial intelligence chatbot performance in triage of ophthalmic conditions. *Can J Ophthalmol* **2024**, *59*, e301-e308, doi:10.1016/j.jcjo.2023.07.016.
40. Maywood, M.J.; Parikh, R.; Deobhakta, A.; Begaj, T. PERFORMANCE ASSESSMENT OF AN ARTIFICIAL INTELLIGENCE CHATBOT IN CLINICAL VITREORETINAL SCENARIOS. *Retina* **2024**, *44*, 954-964, doi:10.1097/iae.0000000000004053.
41. Moshirfar, M.; Altaf, A.W.; Stoakes, I.M.; Tuttle, J.J.; Hoopes, P.C. Artificial Intelligence in Ophthalmology: A Comparative Analysis of GPT-3.5, GPT-4, and Human Expertise in Answering StatPearls Questions. *Cureus* **2023**, *15*, e40822, doi:10.7759/cureus.40822.
42. Nikdel, M.; Ghadimi, H.; Tavakoli, M.; Suh, D.W. 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*, 86-89, doi:10.3928/01913913-20231005-02.
43. Ong, J.; Kedia, N.; Harihar, S.; Vupparaboina, S.C.; Singh, S.R.; Venkatesh, R.; Vupparaboina, K.; Bollepalli, S.C.; Chhablani, J. Applying large language model artificial intelligence for retina International Classification of Diseases (ICD) coding. *Journal of Medical Artificial Intelligence* **2023**, *6*.
44. 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* **2023**, *46*, 706-711, doi:10.1016/j.jfo.2023.05.006.
45. Patil, N.S.; Huang, R.; Mihalache, A.; Kisilevsky, E.; Kwok, J.; Popovic, M.M.; Nassrallah, G.; Chan, C.; Mallipatna, A.; Kertes, P.J.; et al. THE ABILITY OF ARTIFICIAL INTELLIGENCE CHATBOTS ChatGPT AND GOOGLE BARD TO ACCURATELY CONVEY PREOPERATIVE INFORMATION FOR PATIENTS UNDERGOING OPHTHALMIC SURGERIES. *Retina* **2024**, *44*, 950-953, doi:10.1097/iae.0000000000004044.
46. Potapenko, I.; Malmqvist, L.; Subhi, Y.; Hamann, S. Artificial Intelligence-Based ChatGPT Responses for Patient Questions on Optic Disc Drusen. *Ophthalmol Ther* **2023**, *12*, 3109-3119, doi:10.1007/s40123-023-00800-2.
47. Pushpanathan, K.; Lim, Z.W.; Er Yew, S.M.; Chen, D.Z.; Hui'En Lin, H.A.; Lin Goh, J.H.; Wong, W.M.; Wang, X.; Jin Tan, M.C.; Chang Koh, V.T.; et al. Popular large language model chatbots' accuracy, comprehensiveness, and self-awareness in answering ocular symptom queries. *iScience* **2023**, *26*, 108163, doi:10.1016/j.isci.2023.108163.
48. Raghu, K.; S, T.; C, S.D.; M, S.; 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, doi:10.2147/opth.S435052.
49. Rojas-Carabali, W.; Sen, A.; Agarwal, A.; Tan, G.; Cheung, C.Y.; Rousselot, A.; Agrawal, R.; Liu, R.; Cifuentes-González, C.; Elze, T.; et al. Chatbots Vs. Human Experts: Evaluating Diagnostic Performance of Chatbots in Uveitis and the Perspectives on AI Adoption in Ophthalmology. *Ocul Immunol Inflamm* **2024**, *32*, 1591-1598, doi:10.1080/09273948.2023.2266730.
50. Rojas-Carabali, W.; Cifuentes-González, C.; Wei, X.; Putera, I.; Sen, A.; Thng, Z.X.; Agrawal, R.; Elze, T.; Sobrin, L.; Kempen, J.H.; et al. Evaluating the Diagnostic Accuracy and Management Recommendations of ChatGPT in Uveitis. *Ocul Immunol Inflamm* **2024**, *32*, 1526-1531, doi:10.1080/09273948.2023.2253471.
51. Sakai, D.; Maeda, T.; Ozaki, A.; Kanda, G.N.; Kurimoto, Y.; Takahashi, M. Performance of ChatGPT in Board Examinations for Specialists in the Japanese Ophthalmology Society. *Cureus* **2023**, *15*, e49903, doi:10.7759/cureus.49903.

52. 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* **2023**, *43*, 4905-4909, doi:10.1007/s10792-023-02893-x.
53. Shemer, A.; Cohen, M.; Altarescu, A.; Atar-Vardi, M.; Hecht, I.; Dubinsky-Pertzov, B.; Shoshany, N.; Zmujack, S.; Or, L.; Einan-Lifshitz, A.; et al. Diagnostic capabilities of ChatGPT in ophthalmology. *Graefes Arch Clin Exp Ophthalmol* **2024**, *262*, 2345-2352, doi:10.1007/s00417-023-06363-z.
54. Singer, M.B.; Fu, J.J.; Chow, J.; Teng, C.C. Development and Evaluation of Aeyeconsult: A Novel Ophthalmology Chatbot Leveraging Verified Textbook Knowledge and GPT-4. *J Surg Educ* **2024**, *81*, 438-443, doi:10.1016/j.jsurg.2023.11.019.
55. Singh, S.; Djalilian, A.; Ali, M.J. ChatGPT and Ophthalmology: Exploring Its Potential with Discharge Summaries and Operative Notes. *Semin Ophthalmol* **2023**, *38*, 503-507, doi:10.1080/08820538.2023.2209166.
56. Tailor, P.D.; Dalvin, L.A.; Chen, J.J.; Iezzi, R.; Olsen, T.W.; Scruggs, B.A.; Barkmeier, A.J.; Bakri, S.J.; Ryan, E.H.; Tang, P.H.; 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*, 100485, doi:10.1016/j.xops.2024.100485.
57. Taloni, A.; Borselli, M.; Scarsi, V.; Rossi, C.; Coco, G.; Scorcia, V.; Giannaccare, G. Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology. *Sci Rep* **2023**, *13*, 18562, doi:10.1038/s41598-023-45837-2.
58. Tao, B.K.; Handzic, A.; Hua, N.J.; Vosoughi, A.R.; Margolin, E.A.; Micieli, J.A. Utility of ChatGPT for Automated Creation of Patient Education Handouts: An Application in Neuro-Ophthalmology. *J Neuroophthalmol* **2024**, *44*, 119-124, doi:10.1097/wno.0000000000002074.
59. 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)* **2023**, *15*, e184-e187, doi:10.1055/s-0043-1774399.
60. Wilhelm, T.I.; Roos, J.; Kaczmarczyk, R. Large Language Models for Therapy Recommendations Across 3 Clinical Specialties: Comparative Study. *J Med Internet Res* **2023**, *25*, e49324, doi:10.2196/49324.
61. Wu, G.; Lee, D.A.; Zhao, W.; Wong, A.; Sidhu, S. ChatGPT: is it good for our glaucoma patients? *Frontiers in Ophthalmology* **2023**, *3*, doi:10.3389/fopht.2023.1260415.
62. Yılmaz, I.B.E.; Doğan, L. Talking technology: exploring chatbots as a tool for cataract patient education. *Clin Exp Optom* **2025**, *108*, 56-64, doi:10.1080/08164622.2023.2298812.
63. Zandi, R.; Fahey, J.D.; Drakopoulos, M.; Bryan, J.M.; Dong, S.; Bryar, P.J.; Bidwell, A.E.; Bowen, R.C.; Lavine, J.A.; Mirza, R.G. Exploring Diagnostic Precision and Triage Proficiency: A Comparative Study of GPT-4 and Bard in Addressing Common Ophthalmic Complaints. *Bioengineering (Basel)* **2024**, *11*, doi:10.3390/bioengineering11020120.
64. Cruz Rivera, S.; Liu, X.; Chan, A.-W.; Denniston, A.K.; Calvert, M.J.; Darzi, A.; Holmes, C.; Yau, C.; Moher, D.; Ashrafian, H.; et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nature Medicine* **2020**, *26*, 1351-1363, doi:10.1038/s41591-020-1037-7.
65. Karargyris, A.; Umeton, R.; Sheller, M.J.; Aristizabal, A.; George, J.; Wuest, A.; Pati, S.; Kassem, H.; Zenk, M.; Baid, U.; et al. Federated benchmarking of medical artificial intelligence with MedPerf. *Nature Machine Intelligence* **2023**, *5*, 799-810, doi:10.1038/s42256-023-00652-2.
66. European Commission: Directorate-General for Communications Networks, C.; Technology. *Ethics guidelines for trustworthy AI*; Publications Office: 2019.
67. Dave, T.; Athaluri, S.A.; Singh, S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Frontiers in Artificial Intelligence* **2023**, *6*, doi:10.3389/frai.2023.1169595.
68. Dam, S.K.; Hong, C.S.; Qiao, Y.; Zhang, C. A Complete Survey on LLM-based AI Chatbots. *arXiv [cs.CL]* **2024**.
69. Waisberg, E.; Ong, J.; Masalkhi, M.; Zaman, N.; Sarker, P.; Lee, A.G.; Tavakkoli, A. GPT-4 and medical image analysis: strengths, weaknesses and future directions. *Journal of Medical Artificial Intelligence* **2023**, *6*.
70. Eberhard, D.M.; Simons, G.F.; Fennig, C.D. Ethnologue: Languages of the World. Twenty-eighth edition. **2025**.

 doi:10.20944/preprints202503.1961.v1

71. Wang, M.Y.; Asanad, S.; Asanad, K.; Karanjia, R.; Sadun, A.A. Value of medical history in ophthalmology: A study of diagnostic accuracy. *J Curr Ophthalmol* **2018**, *30*, 359-364, doi:10.1016/j.joco.2018.09.001.
72. Reddy, S. Evaluating large language models for use in healthcare: A framework for translational value assessment. *Informatics in Medicine Unlocked* **2023**, *41*, 101304, doi:https://doi.org/10.1016/j.imu.2023.101304.
73. Park, Y.-J.; Pillai, A.; Deng, J.; Guo, E.; Gupta, M.; Paget, M.; Naugler, C. Assessing the research landscape and clinical utility of large language models: a scoping review. *BMC Medical Informatics and Decision Making* **2024**, *24*, 72, doi:10.1186/s12911-024-02459-6.
74. Liu, F.; Li, Z.; Zhou, H.; Yin, Q.; Yang, J.; Tang, X.; Luo, C.; Zeng, M.; Jiang, H.; Gao, Y.; et al. Large Language Models in the Clinic: A Comprehensive Benchmark. *arXiv [cs.CL]* **2024**.
75. Mohammadi, I.; Firouzabadi, S.R.; Kohandel Gargari, O.; Habibi, G. Standardized Assessment Framework for Evaluations of Large Language Models in Medicine (SAFE-LLM). *Preprints* **2025**, doi:10.20944/preprints202501.0471.v1.
76. Kuş, Z.; Aydin, M. MedSegBench: A comprehensive benchmark for medical image segmentation in diverse data modalities. *Scientific Data* **2024**, *11*, 1283, doi:10.1038/s41597-024-04159-2.
77. Han, T.; Kumar, A.; Agarwal, C.; Lakkaraju, H. MedSafetyBench: Evaluating and Improving the Medical Safety of Large Language Models. *arXiv [cs.AI]* **2024**.
78. Privacy policy | OpenAI.
79. Ong, J.C.L.; Chang, S.Y.; William, W.; Butte, A.J.; Shah, N.H.; Chew, L.S.T.; Liu, N.; Doshi-Velez, F.; Lu, W.; Savulescu, J.; et al. Ethical and regulatory challenges of large language models in medicine. *Lancet Digit Health* **2024**, *6*, e428-e432, doi:10.1016/s2589-7500(24)00061-x.
80. Mugaanyi, J.; Cai, L.; Cheng, S.; Lu, C.; Huang, J. Evaluation of Large Language Model Performance and Reliability for Citations and References in Scholarly Writing: Cross-Disciplinary Study. *J Med Internet Res* **2024**, *26*, e52935, doi:10.2196/52935.
81. Patil, R.; Heston, T.F.; Bhuse, V. Prompt Engineering in Healthcare. *Electronics* **2024**, *13*, 2961.
82. Schillinger, D. Social Determinants, Health Literacy, and Disparities: Intersections and Controversies. *Health Lit Res Pract* **2021**, *5*, e234-e243, doi:10.3928/24748307-20210712-01.
83. Al Ghadban, Y.; Lu, H.; Adavi, U.; Sharma, A.; Gara, S.; Das, N.; Kumar, B.; John, R.; Devarsetty, P.; Hirst, J.E. Transforming Healthcare Education: Harnessing Large Language Models for Frontline Health Worker Capacity Building using Retrieval-Augmented Generation. *medRxiv* **2023**, 2023.2012.2015.23300009, doi:10.1101/2023.12.15.23300009.
84. Peters, M.D.J.; Marnie, C.; Colquhoun, H.; Garritty, C.M.; Hempel, S.; Horsley, T.; Langlois, E.V.; Lillie, E.; O'Brien, K.K.; Tunçalp, Ö.; et al. Scoping reviews: reinforcing and advancing the methodology and application. *Systematic Reviews* **2021**, *10*, 263, doi:10.1186/s13643-021-01821-3.
85. Institute of Medicine Committee on Standards for Systematic Reviews of Comparative Effectiveness, R. In *Finding What Works in Health Care: Standards for Systematic Reviews*, Eden, J., Levit, L., Berg, A., Morton, S., Eds.; National Academies Press (US)
86. Copyright 2011 by the National Academy of Sciences. All rights reserved.: Washington (DC), 2011.
87. Levac, D.; Colquhoun, H.; O'Brien, K.K. Scoping studies: advancing the methodology. *Implement Sci* **2010**, *5*, 69, doi:10.1186/1748-5908-5-69.