Received: 28 September 2023 | Accepted: 25 January 2024

DOI: 10.1111/opo.13284

OPO THE COLLEGE OF OPTOMETRISTS

REVIEW ARTICLE

# Utility of artificial intelligence-based large language models in ophthalmic care

**Sayantan Biswas** | **Leon N. Davies** | **Amy L. Sheppard** | **Nicola S. Logan** | **James S. Wolffsohn**

School of Optometry, College of Health and Life Sciences, Aston University, Birmingham, UK

**Correspondence**
Sayantan Biswas, School of Optometry, College of Health and Life Sciences, Aston University, Birmingham, UK.
Email: s.biswas2@aston.ac.uk

## Abstract

**Purpose:** With the introduction of ChatGPT, artificial intelligence (AI)-based large language models (LLMs) are rapidly becoming popular within the scientific community. They use natural language processing to generate human-like responses to queries. However, the application of LLMs and comparison of the abilities among different LLMs with their human counterparts in ophthalmic care remain under-reported.

**Recent Findings:** Hitherto, studies in eye care have demonstrated the utility of ChatGPT in generating patient information, clinical diagnosis and passing ophthalmology question-based examinations, among others. LLMs' performance (median accuracy, %) is influenced by factors such as the iteration, prompts utilised and the domain. Human expert (86%) demonstrated the highest proficiency in disease diagnosis, while ChatGPT-4 outperformed others in ophthalmology examinations (75.9%), symptom triaging (98%) and providing information and answering questions (84.6%). LLMs exhibited superior performance in general ophthalmology but reduced accuracy in ophthalmic subspecialties. Although AI-based LLMs like ChatGPT are deemed more efficient than their human counterparts, these AIs are constrained by their nonspecific and outdated training, no access to current knowledge, generation of plausible-sounding 'fake' responses or hallucinations, inability to process images, lack of critical literature analysis and ethical and copyright issues. A comprehensive evaluation of recently published studies is crucial to deepen understanding of LLMs and the potential of these AI-based LLMs.

**Summary:** Ophthalmic care professionals should undertake a conservative approach when using AI, as human judgement remains essential for clinical decision-making and monitoring the accuracy of information. This review identified the ophthalmic applications and potential usages which need further exploration. With the advancement of LLMs, setting standards for benchmarking and promoting best practices is crucial. Potential clinical deployment requires the evaluation of these LLMs to move away from artificial settings, delve into clinical trials and determine their usefulness in the real world.

**KEYWORDS**
artificial intelligence, chatbot, large language model, ophthalmic care, ophthalmology, optometry

## INTRODUCTION

Artificial intelligence (AI)-based chatbots are regularly used in customer service functions in business and marketing.[1] Although the use of AI is still at a relatively early stage in eyecare, it is used by ophthalmologists, optometrists and researchers in screening and diagnosing eye diseases.[2] Current applications of AI in ophthalmology are mostly focused on image-based techniques used for image analysis, recognition and diagnosis of ophthalmic diseases using ophthalmic data from fundus photography and optical coherence tomography images. The two subfields of AI are machine learning and natural language processing (NLP).[3] Machine learning requires 'supervised learning' where experts label and grade individual features and severity from images to develop the AI. A subset of machine learning is deep learning that shows promise in disease screening, diagnosis, risk stratification, treatment monitoring and improved patient care for eyes with myopia,[4] optic disc abnormalities (e.g., glaucoma, papilledema),[5–7] retinal diseases (e.g., age-related macular degeneration, diabetic retinopathy),[2] cataract[2] and corneal disorders.[8,9] Deep learning, referred to as 'unsupervised learning', bypasses this need to label or grade individual features, and instead uses features of the entire image to compare with a diagnosis determined by an expert.[10] The individual predictive features associated with the classification of a disease severity or its diagnosis are 'self-learned' by the AI developed from deep learning. Either way, the performance of deep learning and machine learning is comparable, with decreased error rates, and is better than traditional techniques of screening, diagnosis and management of diseases at a tertiary eyecare level.[10]

However, deep learning is limited by the homogenous training data set, limited data availability for diseases, disagreement and wide interobserver variability in defining disease phenotype. Also, most AI systems have the 'black box' problem where the inputs and operations are unknown to the user. These impenetrable AI systems arrive at a conclusion or decision without providing any reasoning or explanations as to how they were reached; this opaque approach reduces practitioner and patient acceptance of the AI[11] and preference of human expert over AI in decision-making and treatment plan.[12] Even though it is technically possible, AI has not been able to reach its target of converging AI and clinical care so far.[13]

NLP, the other subset of AI, is focused on extracting and processing text data that include written and spoken words. NLP could transform human language (free text) or image into code that computers understand, and has been primarily used to date for information retrieval and text extraction.[3] However, NLP is susceptible to error due to the variable nature of human-generated natural text and limited by the requirement of a huge data set for training NLP models which may or may not utilise deep learning or machine learning. Moreover, NLPs are often trained in specific domains, that impact how word embeddings, which is a method of extracting features out of text, based on the distance between two words, interpret relationships between words in different contexts. Most NLP applications developed so far are in the English language. It is pertinent to develop NLP in non-English languages to promote equity in care, reduce disparity and reach a wider population. Even writer/user presumption about the input (completeness and composition of words, image quality, noise), prior understanding and context are some other variabilities in NLP. Until now, NLP applications have involved text data extraction from clinical, operative and electronic health record notes in the screening of cataract[14] and glaucoma,[15] triaging of outpatient referrals to ophthalmic specialists,[16] prediction of quality-of-life impact of vision loss associated with diabetes[17] and prediction of operative complications related to cataract surgery,[18] to name a few examples.

Large language models (LLMs) are the class of AI primarily succeeding deep learning models that are capable of learning and recognising patterns. They are specific models within NLP that are capable of processing, understanding, generating and manipulating human language.[19] LLMs are trained to predict the sequence of words in natural human language query and in response, generate a novel sequence of words. These models are designed to capture the complexities and nuances of natural language, enabling them to perform a wide range of language-related tasks. They are often based on transformer architectures, which are deep learning models designed to handle sequences of data, such as sentences in natural language. The key innovation in the transformer architecture is the 'self-attention' mechanism, which allows the model to weigh the importance of different words or positions in a sequence while processing each word. They are trained on vast amounts of text data to learn the patterns, semantics and context of

**Key points**

- With the huge interest and popularity of ChatGPT, artificial intelligence-based large language models have a massive role in providing patient information, disease diagnosis, symptom triaging, ophthalmic education and other applications.
- Human experts are the most accurate (86%) in diagnosing disease, whereas ChatGPT-4 tops in responding to text-based ophthalmology examination questions (75.9%) and providing information and answering patient queries (84.6%).
- Large language models perform best in general ophthalmology but worse in ophthalmic subspecialties. Responses are prompt-specific and can often be misleading due to their apparent comprehensiveness to queries and plausible-sounding fake responses.

natural language.[20] These models consist of multiple layers of neural network that process sequential data, such as words or characters, and learn patterns and relationships within the data. The training process of an LLM involves exposing it to large amounts of text data, such as books, articles, websites and other sources of human language. By learning from this vast corpus of text, the model develops a deep understanding of language patterns, grammar, context and even some semantic meaning.[20] As with NLP, an LLM presumes that the input (text), which influences the accuracy of the output, is accurate and up to date.

## RISE OF THE LARGE LANGUAGE MODELS IN HEALTHCARE

LLMs have come into prominence following the recent introduction of the Chat Generative Pre-Trained Transformer, more popularly known as ChatGPT.[21,22] ChatGPT (OpenAI, openai.com) currently relies on GPT-4, a language model that uses deep learning (DL)-based NLP, to produce text with approximately 170 trillion parameters, which is an upgrade from the GPT-3.5 version with approximately 175 billion parameters. Like the neural network of the human brain, these parameters are the weights of connections learned during the training stage of a neural network. This massive network builds the LLMs like ChatGPT and uses supervised and reinforcement-based learning strategies. Using NLP, the LLM can generate responses to queries which can simulate human conversation. Their application in healthcare has seen a widespread use in education, research, practice and electronic health records, among others.[22] In early 2023, ChatGPT gained widespread attention among the medical community worldwide after it performed at or near the passing threshold of 60% accuracy on the United States Medical Licensing Examination (USMLE), primarily because of its ability to respond to an array of natural language queries and human-like interaction.[22,23] ChatGPT-3.5 passed all of the three difficulty levels of the USMLE designed using both multiple choice questions (MCQs) and open-ended questions, while displaying a high level of insight and concordance in its explanations.[23] A slightly better result (67.6% and 67.1% accuracy) was obtained on the MCQs from the USMLE using the instruction-tuned variant of Flan (Fine-tuned Language Net)–PaLM (Pathways Language Model) and Med-PaLM (research.google) which are LLMs with 540 billion parameters.[24] The later versions, GPT-4-base and Med-PaLM 2 (research.google) reached 86.1% and 86.5% accuracy in USMLE, 81.8% and 80.4% in PubMed question answering (PubMedQA), 72.3% and 73.7% in medical domain multiple choice question answering (MedMCQA) and 89.9% and 90.5% in massive multitask language understanding (MMLU) clinical data set, respectively. However, the LLMs remain inferior to human clinician answers.[25] In addition, a team of physicians found that ChatGPT generated written responses to healthcare-related patient questions collected from a public social media forum, that were comparable to those by physicians in quality and empathy, even surpassing some physician responses.[26] However, this does not account for most ophthalmic patient communications being verbalised and being accompanied by expressive body language. With the development of newer LLMs and the exponential progress in technology, newer applications of LLMs have been identified but not all have been assessed comprehensively.

## AIM OF THIS REVIEW

The aim of this review was to provide an overview of the current literature, investigating the recent advent of AI-based LLMs in ophthalmic care. In addition, the most widely used application(s) of LLMs and their future scope in eyecare are described. Finally, the limitations and challenges of implementing LLMs into clinical practice are highlighted.

## METHOD OF LITERATURE SEARCH

A comprehensive review of the literature was performed through a keyword-based and medical subject headings (MeSH)-based search of PubMed/Medline, Web of Science, Scopus, Embase, Google Scholar and pre-print servers until 31 December 2023. The following keywords, their synonyms and combinations were used: "artificial intelligence", "large language model", "LLM", "ChatGPT", "Generative Pre-Trained Transformer", "chatbot", "ophthalmology", "optometry", "ophthalmic", "eye", "health", "vision", "eye disease", "care". Boolean operators "AND," "OR," "NOT" were used to combine all search sets. When a specific application of LLM was identified, the specific factor was also used as a keyword in a second search to identify additional publications with prospective data on the specific use. Relevant articles cited in the reference list of articles obtained through this search were also reviewed. Studies were included if they described original research using LLMs in eyecare. After the selection of articles, all specific applications were grouped, and results were discussed accordingly. The authors manually reviewed each study's title, abstract and manuscript text to validate the relevance of the studies to both eyecare and LLM. Data extracted from each study included: the authors and year of publication, study aim, LLM used, content/query asked, grader/evaluator of the responses, results/outcome, main finding and conclusion. Any disagreements arising were resolved by discussion. All literature reviews and editorials were excluded, resulting in a yield of 70 original reports.

## STATISTICAL ANALYSIS

SPSS Statistics version 29.0 (ibm.com) was used to estimate the median, IQR and range of accuracies. The Kruskal–Wallis

H omnibus test statistic was used to compare the accuracies between the LLMs. The significance level for all statistical tests was set at $p < 0.05$ with Bonferroni correction for post hoc pairwise comparisons. Data are represented as median with full range (%).

## USE OF LLMs IN EYECARE

LLMs are trained to receive large corpora of text data and can interpret natural language inputs and respond with human-like real-time answers. Thus, the focus of LLMs in eyecare is to generate a list of potential diagnoses, information or guidance on management options. There has been a wide application of LLMs in eyecare so far, with studies examining different aspects of them. Figures 1 and 2 show the distribution of research articles published up to 31 December 2023, based on the ophthalmic domain and their subspecialty. Figure 3 illustrates the distribution of accuracy scores (%) across the LLMs and the domain.

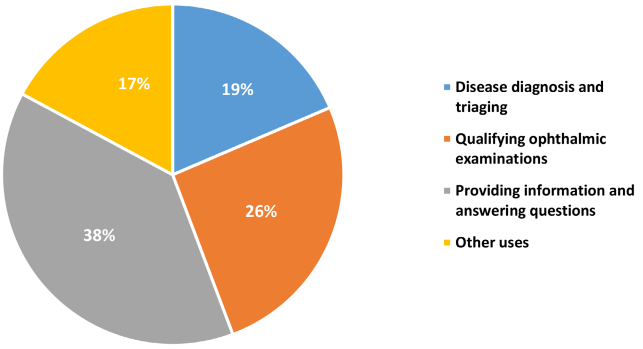


**FIGURE 1** Distribution of published studies using large language models in ophthalmic care.

## Performance in diagnosing ophthalmic diseases and triage accuracy

The accuracy of LLMs for diagnosing common ocular diseases from clinical vignettes and cases was 71.1% (10%–95%).[27–35] ChatGPT-4 (82%, 66%–93%) was superior to other LLMs or clinical decision support systems available, but it was still inferior to trained ophthalmologists (86%, 63.6%–95%) (see Table 1). Not only were the chances of ChatGPT making a correct diagnosis proportionately higher than other LLMs but also ChatGPT was effective in diagnosing written (text based) evidence from Chinese language sources. Additionally, LLMs had a relatively better triage accuracy than ophthalmology trainees (84%–98% vs. 86%).[29,31] A small proportion of misinformation and hallucination in responses existed among LLMs, with Bing Chat and GPT-3 (61.5%–77%) having reduced precision than GPT-4 (80.5%–93%—see Table 1). For image-based diagnosis, the performance of GPT-4V(ision) was poor (30.6%, 26.2%–67.5%)[36–39] with significantly lower median accuracy compared with humans ($p = 0.01$) (Table 2).

## Performance in qualifying ophthalmological board examinations and knowledge assessment

LLMs regularly outperformed the threshold of ophthalmological specialist examinations.[40–57] However, the accuracy of LLMs was 66.9% (22.4%–91%) while the ophthalmology trainees scored 68.4% (33%–75.7%—Table 3). Overall, the performance of GPT-4 (75.8%, 45.8%–91%) and Bing Chat (75.1%, 63.9%–82.9%) was comparable and better than both Google Bard (57.3%, 44.9%–69.4%) and GPT-3.5 (55.5%, 22.4%–69.5%) in passing ophthalmological board examinations. GPT-4 with its better logical reasoning and image processing ability outperformed GPT-3 ($p = 0.005$, Table 2) and GPT-3.5 ($p < 0.001$,

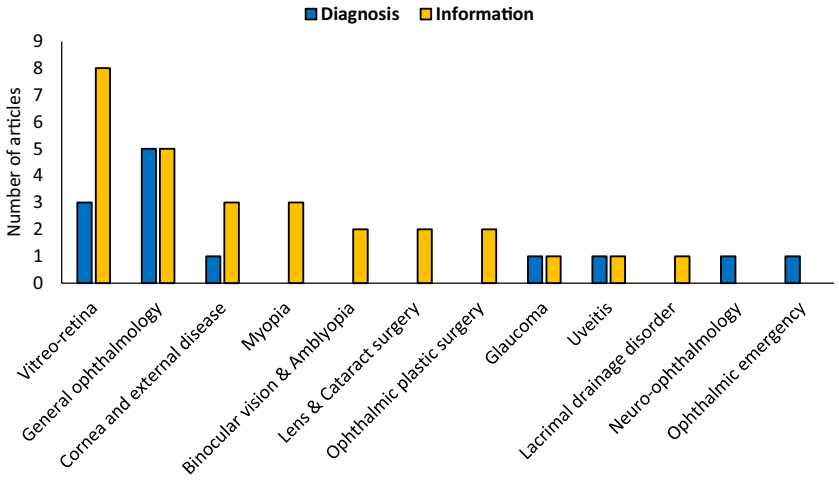


**FIGURE 2** Distribution of published studies based on ophthalmic specialty.

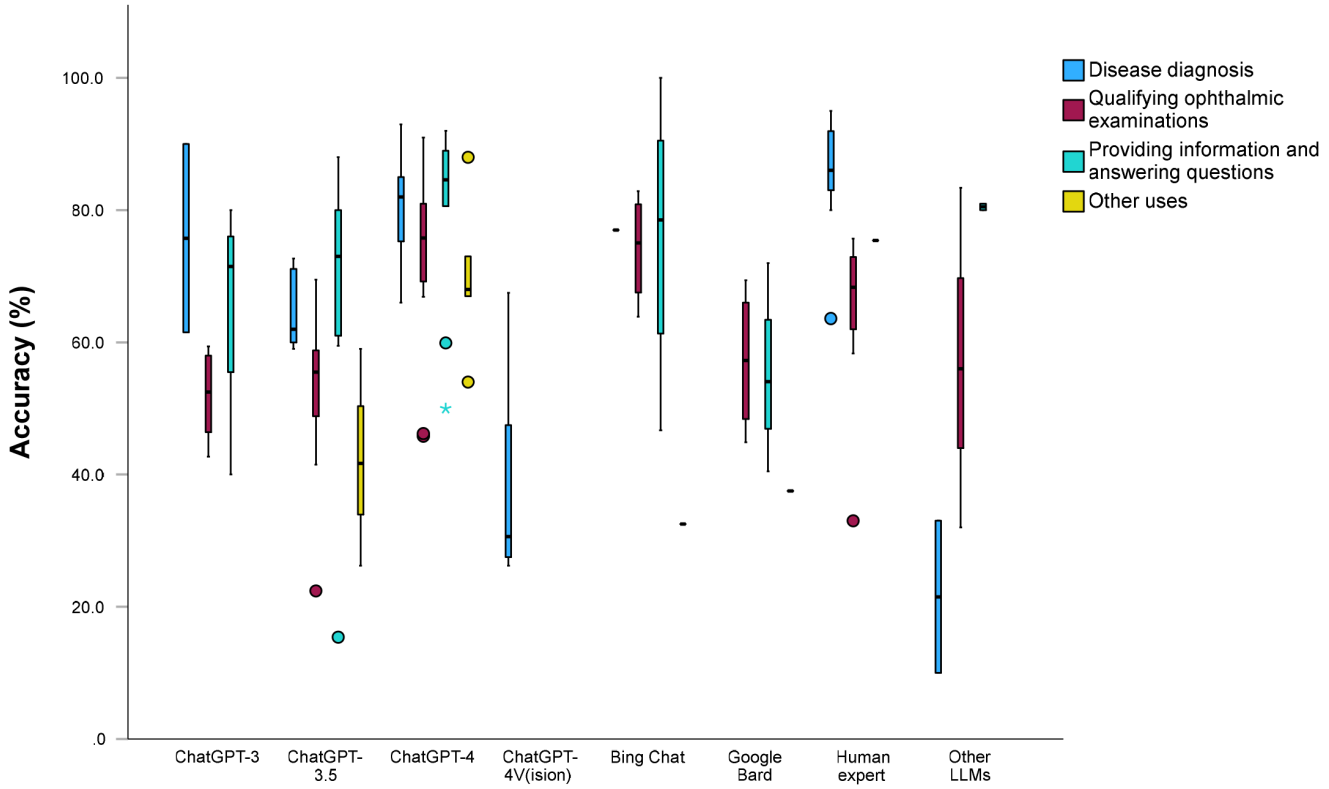


**FIGURE 3** Boxplot illustrating the distribution of accuracy scores (%) across the large language models and their domains of use.

Table 2) in all the examinations. However, the judgement between GPT-4 and its human counterparts (experts) was not yet conclusive and therefore warrants further investigation. Results were similar among the American Academy of Ophthalmology's Basic and Clinical Science Course (46.0%–84.3%),[41,42,46,55] Ophthoquestions (42.7%–84%),[41,47,48] Fellow of The Royal College of Ophthalmologists (FRCOphth) examination questions (32%–88.4%)[51,57] and Statpearls (55.5%–73.2%).[49] In comparison, a lower score was observed in Brazil board examination questions (41.5%)[44] and higher in European board examinations (91%).[50] The performance of LLMs was found to be better for the subspecialties of medicine, cornea, refractive surgery and oncology, and weakest for glaucoma, neuro-ophthalmology, pathology, tumours, optics, oculoplastic and mathematical concepts. The weakness of GPT-3 and GPT-3.5 in answering questions on retina and vitreous (0%–23.1%)[44,48] was overcome in a later version (GPT-4: 100% correct).[47] Both GPT-3.5 and GPT-4's performance was better for first-order questions (recall) and lower for higher order (evaluative/analytical) and image-based questions. Even though ChatGPT has relatively lower hallucinations ('imagines' or 'fabricates' information) and errors in logical reasoning in comparison to other LLMs, ChatGPT provided explanations and additional insights for both its correct and incorrect responses (63%–98%), which can be misleading due to its apparent comprehensiveness (refer to Table 3). Overall, the accuracy and performance of LLMs is improving, and they often surpass the established benchmarks or thresholds of specialised assessments. The ability to pass specialised ophthalmic examinations shows that LLMs can serve as a valuable study aid for board certification examinations, as they can generate practice questions, explanations and feedback to enhance preparation. Besides, it can be a rapid source of specialised knowledge for busy clinicians. This ability to qualify for specialised examinations implies that conventional written evaluations fail to gauge clinical competence. The fact that AI can pass an ophthalmology examination could mean a decrease in the quality of clinicians entering the profession. However, shifting the assessments towards clinical scenarios and decision-making could improve the quality of future graduates. Fortunately, ophthalmic clinical practice relies heavily on physical examination of the eye, an element that cannot be attained solely through text-based interactions with an LLM.[58]

## Performance in providing information and answering questions

In general, the quality of clinical information provided by the interactive LLMs is variable (77.4%, 15.4%–100%) (see Table 4).[21,59–84] The accuracy of the later version of ChatGPT (GPT-4) was higher (84.6%, 50%–92%) than the earlier versions (GPT-3: 71.5%, 40%–80% and GPT-3.5: 73%, 15.4%–88%), Bing Chat (78.5%, 46.7%–100%) and

**TABLE 1** Summary of current studies on the performance of LLM in diagnosing ophthalmic diseases and triage accuracy.

| Author (Year) | Aim/Purpose | LLM used | Content | Grader/Evaluator | Accuracy | Main/Other finding | Author conclusion |
|---|---|---|---|---|---|---|---|
| *Performance in diagnosing ophthalmic diseases and triage accuracy* | | | | | | | |
| Balas and Ing[35] | Disease diagnosis accuracy | ChatGPT-3<br>Isabel Pro | A mix of free text inputs of case descriptions and itemised list of clinical features for 10 ophthalmic emergency cases | Ophthalmologist | 90.0%, median score 1.0 (mean 1.8)<br>10.0%, median score 5.5 (mean 6.1) | ChatGPT generated significantly better provisional and differential diagnoses for common ophthalmic conditions. | ChatGPT has greater accuracy and reliability in correct diagnosis compared to Isabel Pro. |
| Delsoz et al.[34] | Capability in diagnosing corneal eye diseases | ChatGPT-4<br>ChatGPT-3.5<br>Cornea specialist | Details case description of patient's demographics, chief complaint, present illness and major examination findings of various corneal conditions | Cornea specialist/ Ophthalmologist | 85.0%<br>60.0%<br>93.3% | • Interobserver agreement between: GPT-4 and experts were highest (80%), then between GPT-4 and GPT-3.5 (65%), GPT-3.5 and experts (60%).<br>• Specialists took 20–40 min to diagnose cases, while ChatGPT took only a couple of minutes. | • Accuracy of GPT-4.0 in diagnosing patients with various corneal conditions was markedly improved than GPT-3.5.<br>• May enhance corneal diagnostics, improve patient interaction and experience as well as medical education. |
| Delsoz et al.[33] | Capability in the diagnosis of glaucoma | ChatGPT-3.5<br>Ophthalmology resident | Case description of both common and uncommon types of glaucoma | Ophthalmologist | 72.7%<br>54.5%–72.7% | Accuracy of ChatGPT in diagnosing patients with primary and secondary glaucoma, was similar or better than senior ophthalmology residents. | • ChatGPT may enhance glaucoma assessment with potential to be used in clinical care settings, triaging and in eye care clinical practice.<br>• May provide objective and quick diagnoses of glaucoma. |
| Inayat et al.[32] | Assess performance in disease diagnosis and triaging (management urgency) | ChatGPT-3.5<br>Ophthalmology resident<br>Ophthalmologists | Case description with questions on diagnosis and urgency of assessment | Ophthalmologists | 67.2%–75%<br>93.8%<br>87.5% | ChatGPT agreed with 28.1%–38.5% of residents and ophthalmologists on management urgency. | ChatGPT has strong diagnostic performance, but tends to favour more urgent assessment. |
| Knebel et al.[31] | Diagnostic accuracy, triage accuracy, recommended treatment and potential to inflict harm to user/ patient | ChatGPT-3 | Ophthalmological emergency case scenarios | Ophthalmologist | 61.5% | • ChatGPT scored 100% for treatment, 87.2% for triage.<br>• 32% of the recommendations have potential to inflict harm to the patient. | Clinicians should not solely rely on ChatGPT as the primary source of information for acute ophthalmological symptoms. |
| Liu et al.[30] | Effectiveness and reasoning ability in diagnosing retinal vascular diseases in Chinese | ChatGPT-4<br>Ophthalmologist<br>Ophthalmology intern | Fundus fluorescein angiography report and diagnosis of retinal vascular diseases written in Chinese, with prompts in English and Chinese | Ophthalmologist | 80.1% (English) and 70.5% (Chinese)<br>89.4% (Chinese)<br>82.7% (Chinese) | • ChatGPT can derive reasoning process with a low error rate.<br>• Errors such as misinformation (1.96%), hallucination (0.59%) and inconsistency (0.39%) existed. | • ChatGPT can be a helpful medical assistant to provide diagnosis under Chinese clinical environment.<br>• Potential limitations such as language disparity, misinformation and hallucination exists. |
| Lyons et al.[29] | Accuracy of disease diagnosis and triage urgency | ChatGPT-4<br>Bing Chat<br>WebMD symptom checker<br>Ophthalmology trainees | Ophthalmic clinical vignettes | Ophthalmologist | 93.0%<br>77.0%<br>33.0%<br>95.0% | • Triage accuracy was 98%, 84% and 86% for ChatGPT, Bing Chat and physicians.<br>• GPT-4 and physicians had comparably high diagnostic and triage accuracy, whereas Bing Chat had lower accuracy. | • ChatGPT demonstrated high diagnostic and excellent triage performance without any grossly incorrect statements.<br>• Bing Chat had reduced precision, occasional instances of grossly incorrect statements and a propensity to overemphasise triage urgency. |

**TABLE 1** (Continued)

| Author (Year) | Aim/Purpose | LLM used | Content | Grader/Evaluator | Accuracy | Main/Other finding | Author conclusion |
|---|---|---|---|---|---|---|---|
| Madadi et al.[28] | Evaluate the efficiency to assist in diagnosing neuro-ophthalmic diseases | ChatGPT-3.5<br>ChatGPT-4<br>Neuro-ophthalmologists | Case reports of neuro-ophthalmic diseases (both acute and chronic) from a publicly available online database | Neuro-ophthalmologists | 59.0%<br>82.0%<br>86.0% | Diagnosis agreement between: GPT-3.5 vs. GPT-4: 59%, GPT-4 vs. expert: 75%, GPT-3.5 vs. expert: 55%, between experts: 77%. | GPT-4 can diagnose complex neuro-ophthalmic cases, when provided with structured data in case report format. |
| Rojas-Carabali et al.[27] | Test the diagnostic accuracy in various uveitis entities | ChatGPT-3.5<br>ChatGPT-4<br>Ophthalmologist (specialist + fellow) | Prototypical (gold standard) cases of uveitis on most likely diagnosis and differential diagnosis | Uveitis specialist/ Ophthalmologist | 60.0%–64.0%<br>60.0%–72.0%<br>60.0%–100.0% | • AI has potential in uveitis diagnosis and management.<br>• Ophthalmologists excelled in likely diagnosis, exceeding AI. | • AI chatbots may reduce diagnostic errors.<br>• Clinicians should rely on their own clinical judgement and experience while incorporating the information provided by AI models. |
| Sorin et al.[39] | Evaluate the performance in an integrated analysis of ocular pathology images and clinical text | ChatGPT-4V(ision)<br>Non-ophthalmologist physicians | Images without clinical context<br>Images with clinical context<br>Images without clinical context<br>Images with clinical context | Ophthalmologist | 47.5%<br>67.5%<br>58.8%<br>70.0% | • GPT-4 performance on ophthalmology cases was comparable to non-ophthalmology physicians.<br>• GPT-4 and physicians' performance improved with clinical context. | • GPT-4 is not yet suitable for clinical application in ophthalmology.<br>• It can simultaneously analyse and integrate visual and textual data, and arrive at accurate clinical diagnoses. |
| Waisberg et al.[38] | Ability to interpret ophthalmic images | ChatGPT-4 | Fundus photographs, OCT and OCTA images | Ophthalmologist | NA | GPT-4 can correctly identify the image type and describe it, but analysis is inadequate and sometimes extremely inaccurate. | GPT-4 shows promise to the future of ophthalmic imaging analysis, but needs enhancement before clinical use. |
| Xu et al.[37] | Evaluate the capabilities in addressing queries related to ocular multimodal images | ChatGPT-4V(ision) | Slit lamp, SLO, FPP, OCT, FFA, OUS images | Ophthalmologist | 30.6% | • 22.8% of responses were considered highly usable, while 55.6% are in safe (no harm) category.<br>• Slit lamp image identification was excellent but weak for FPP images. | GPT-4V lacks the reliability required for clinical decision-making and patient consultation in ophthalmology. |
| Yu et al.[36] | Evaluate the performance in diagnosing retinal diseases using OCT images | ChatGPT-4<br>Retinal specialist | OCT images of CRVO, DME, CSC and PCV | Retinal specialist/ Ophthalmologist | 26.2%–27.5%<br>93.8% | Efficacy is suboptimal when applied to retinal disease image diagnosis. | GPT-4 is prone to errors when interpreting images involving detailed anatomical and specialised knowledge. |

Abbreviations: AAO's BCSC-SAP, American Academy of Ophthalmology's Basic and Clinical Science Course Self-Assessment Program; AI, artificial intelligence; CRVO, central retinal vein occlusion; CSC, central serous chorioretinopathy; DME, diabetic macular oedema; DOI, digital object identifier; FFA, fundus fluorescein angiography; FPP, fundus photography of the posterior pole; FRCOphtha, Fellowship of Royal College of Ophthalmologists; LLM, Large language model; NA, not available; OCT, optical coherence tomography; OCTA, OCT angiography; OKAP, Ophthalmology Knowledge Assessment Program; OUS, ocular ultrasound; PCV, polypoidal choroidal vasculopathy; SLO, scanning laser ophthalmoscopy.

**TABLE 2** Pairwise comparison of accuracies.

| | | Bonferroni pairwise comparison | |
|---|---|---|---|
| | LLM | Standardised test statistic | $p$ [a] |
| Diagnosis | Human vs. ChatGPT-4V | −3.31 | 0.01 |
| Examination | ChatGPT-3 vs. ChatGPT-4 | −3.66 | 0.005 |
| Examination | ChatGPT-3.5 vs. ChatGPT-4 | −4.59 | <0.001 |
| Information and answer questions | — | — | ns |

Abbreviations: LLM, Large language model; ns, not significant.

[a]Significance values have been adjusted by the Bonferroni correction for multiple tests.

Google Bard (54.1%, 40.5%–72%). Compared to GPT-3.5 and Google Bard, GPT-4 exhibited a reduced proportion of responses that received poor ratings. Of the common questions on myopia, 5.4%–9.7% of the responses were incorrect and 54.8%–87.5% were considered accurate.[61,64] Comparably, for retinal conditions and surgeries, the accuracy of LLM responses varied between 15.4% and 100%.[21,65,67] Nonetheless, GPT-4's responses to common questions on vitreoretinal surgeries were found to be challenging and difficult to comprehend for an average individual without specialised knowledge. Grading of the responses using Flesch–Kincaid grade level and Flesch reading ease scores indicated that it requires at least college graduation to comprehend.[65] The accuracy dropped drastically when it came to responses on lacrimal drainage disorders (40%).[59]

Even though the answers generated by GPT-3.5 had a similar error rate to the responses provided by humans, with comparable likelihood of harm and extent of harm, the presence of incorrect (3.6%–25%) and sometimes fabricated data (due to hallucinations) without supervision/moderation can be harmful, especially if applied in ophthalmic emergencies. Patients and parents should not solely rely on LLMs for their medical guidance, especially on treatment and side effects of medications. The information gathered should serve as a supplement or a basis for engaging in more individualised discussions with a human expert for specialist care and counselling (See Table 4).

## Performance in other potential applications

Further to the applications mentioned above, LLMs have been tested on a diverse range of purposes. Chatbots can generate average quality scientific abstracts (41.7% correct) but remain plagued by fake data and references, when not provided with a data set.[85] GPT-4 scores slightly better than GPT-3.5 with lower fake score and hallucination rates (Table 5). Chatbots can assist people with relatively weak writing or language skills to prepare written assignments both faster and of higher quality. But there is a growing concern that AI chatbots are being abused in writing essays, scientific abstracts and even manuscripts.[85] With the number of factual errors these chatbots generate and their apparently comprehensive response, it is important for authors to know their limitations and pitfalls and for publishers/editors to identify AI-generated text in manuscripts.[86]

GPT-4 can categorise refractive surgery candidates to their ideal procedures (68%–88% correct) with low to moderate agreement (0.399–0.610) with clinicians.[87] However, when it came to recommending ophthalmologists based on their specialty or proximity (location), AI chatbots were unreliable with only 26.2%–37.5% accurate recommendations.[88]

ChatGPT can accurately (59%) generate international classification of disease (ICD) codes from mock retina encounters[89] and even predict the risk of diabetic retinopathy (54%–73%) upon receiving prompts with patient details.[90]

### Applicability to generate novel ideas on future research

ChatGPT-4 was questioned about the 'future research', 'further innovation' and 'technological advancements' in oculoplastic research. It could not come up with any novel idea and displayed convergent thinking in only conveying known ideas for research.[91] ChatGPT's focus is on speed, accuracy, logic and recognising familiar techniques through reapplication of the stored/trained information. It can be viewed as a supplementary research tool, rather than a primary source of original research ideas.

### Ophthalmic operative notes

When asked to generate ophthalmic surgery operative notes, ChatGPT-4 was able to create comprehensive and detailed operative notes across ophthalmic subspecialties.[92,93] However, the response largely depended on the quality of input to GPT. The operative notes were thorough, yet they were deemed to require significant improvement. When appropriately prompted, ChatGPT could integrate specific medications, follow-up directions, consultation timing and location information into discharge summaries.[92]

### Literature review

Two prompts on dry eye disease were used to verify the ability of ChatGPT as a tool for conducting a literature

**TABLE 3** Summary of current studies on the performance of large language model (LLM) in qualifying ophthalmological board examinations.

| Author (Year) | Aim/Purpose | LLM used | Content | Grader/Evaluator | Accuracy | Main/Other finding | Author conclusion |
|---|---|---|---|---|---|---|---|
| *Performance in qualifying ophthalmological board examinations and knowledge assessment* | | | | | | | |
| Antaki et al.[41] | Accuracy in ophthalmology question answering space | ChatGPT-3 | Questions from AAO's BCSC-SAP | Ophthalmologist | 55.8% | • Both ChatGPT and the Plus model had their best performance observed for general medicine, fundamentals and cornea (GPT: 60%–75% vs. GPT Plus: 65%–68.3%).<br>• Weakest performed was neuro-ophthalmology, glaucoma, ophthalmic pathology and intraocular tumours (25%–42.5%) for ChatGPT.<br>• Neuro-ophthalmology, oculoplastics and clinical optics were the weakest for the plus model (40%–45.8%). | • With fixed question section and cognitive level fixed, the difficulty index of questions was predictive of ChatGPT's accuracy.<br>• Improved accuracy was associated with increased difficulty index (which implies an easier question).<br>• ChatGPT's performance and inaccuracies reflects its training, or the lack of it, which aligned well with the mass or collective understanding of its human peers. |
| | | | OphthoQuestions online question bank | | 42.7% | | |
| | | ChatGPT-3 Plus | Questions from AAO's BCSC-SAP | | 59.4% | | |
| | | | OphthoQuestions online question bank | | 49.2% | | |
| Antaki et al.[40] | Performance in ophthalmology question-answering domain | GPT-3.5 | OpthoQuestions and BCSC-SAP test set | Ophthalmologist/ Ophthalmology resident | 50.4%–58.8% | • Human graders preferred (higher ranking) responses from models with a temperature higher than 0 (more creative, less coherent).<br>• GPT-4-0.3 (GPT-4 with 0.3 temperature) achieved the highest accuracy among GPT-4 models.<br>• Simpler, low cognitive level questions (recall tasks) yield better performance than complex, clinical decision-making ones (reasoning). | GPT-4 performs better than its predecessor and historical human performance on simulated ophthalmology board-style exams. |
| | | GPT-4 (Temperature 0) | | | 67.3%–76.2% | | |
| | | GPT-4 (Temperature 0.3) | | | 70.0%–75.8% | | |
| | | GPT-4 (Temperature 0.7) | | | 68.5%-75.8% | | |
| | | GPT-4 (Temperature 1) | | | 66.9%–76.5% | | |
| | | Historical human performance | | | 63.0%–73.3% | | |
| Cai et al.[42] | Ability to answer ophthalmology board style questions | Bing Chat | BCSC-SAP exam questions | Ophthalmologist | 71.2% | • Both Bing Chat and ChatGPT-4 struggled with image interpretation but excelled in workup questions (diagnostic testing).<br>• Hallucinations and errors in logical reasoning were least in GPT-4 (18%) followed by Bing Chat (25.6%) and GPT-3.5 (42.4%). | ChatGPT-4, Bing Chat and human respondents perform similarly with answering questions from the BCSC-SAP. |
| | | ChatGPT-3.5 | | | 58.8% | | |
| | | ChatGPT-4 | | | 71.6% | | |
| | | Human respondent | | | 72.2% | | |

(Continues)

**TABLE 3** (Continued)

| Author (Year) | Aim/Purpose | LLM used | Content | Grader/Evaluator | Accuracy | Main/Other finding | Author conclusion |
|---|---|---|---|---|---|---|---|
| Fowler et al.[43] | Assess performance to answer ophthalmology MCQs | ChatGPT-4<br>Google Bard<br>Historical human performance | Part 1 FRCOphth MCQ examination<br>Book 'MCQs for FRCOpth part 1'<br>Part 1 FRCOphth MCQ examination<br>Part 1 FRCOphth MCQ examination | Mark scheme | 85.7%<br>69.9%<br>44.9%<br>33.0%–62.0% | GPT-4 has the potential to perform well on part 1 FRCOphth MCQ examination when compared with historical human performance. | ChatGPT outperforms Bard, which does not have the same focus on reasoning skills. This highlights the potential application of AI models in medical education and assessment. |
| Gobira et al.[44] | Performance in answering ophthalmology Board questions | ChatGPT-3.5 | Brazilian Council of Ophthalmology Board Examination questions | Ophthalmologist | 41.5% | • 53.7% were wrong, and 4.9% remained undetermined.<br>• Answers to mathematical concepts were poor (23.8% correct), with better responses for theoretical examinations (40.8%–43.2%).<br>• Refractive surgery and oncology had the best performance (100%), while cataract (25%) and retina (23.1%) were the worst. | ChatGPT-3.5 would not succeed in Brazilian ophthalmological board examination as it lacks adequate clinical training data with problems in question formulation. |
| Jiao et al.[45] | Capability in addressing ophthalmic case challenges | ChatGPT-3.5<br>ChatGPT-4 | MCQs from AAO'S 'Diagnose This', focussed on clinical decision-making from different subspecialties | Answers from AAO website | 46.0%<br>75.0% | GPT-4 did better in neuro-ophthalmology, paediatric, image-related questions and generated more concise answers than GPT-3.5. | GPT-4 significantly outperforms GPT-3.5 in addressing ophthalmic case challenges. |
| Lin et al.[46] | Performance in answering practice questions for board certification | ChatGPT-3.5<br>ChatGPT-4<br>Ophthalmology resident and practicing ophthalmologists | AAO's BCSC-SAP image and non-image-based questions | Ophthalmologist | 63.1% and 69.5%<br>76.9% and 84.3%<br>72.6% and 72.9% | • ChatGPT-3.5 and ChatGPT-4 did not answer 48.6% and 51.4% of the image-based questions (non-answers marked incorrect).<br>• On exclusion of the image-based questions, ChatGPT-4 outperformed both its human and tech counterpart.<br>• GPT's performance was worse on higher order questions. | ChatGPT-4's performance is better than its previous ChatGPT-3.5 version. |

**TABLE 3** (Continued)

| Author (Year) | Aim/Purpose | LLM used | Content | Grader/Evaluator | Accuracy | Main/Other finding | Author conclusion |
|---|---|---|---|---|---|---|---|
| Mihalache et al.[47] | Performance in answering practice questions for board certification<br>Repeat assessment after 2 months | ChatGPT-3<br>ChatGPT-3 | OphthoQuestions for the OKAP and Written Qualifying Examination for board certification in ophthalmology | Ophthalmologist | 46.4%<br>58% | • Best performance observed in general medicine (79%) and poorest in retina and vitreous (0%, all incorrect).<br>• Explanations and additional insight were provided for 63% of questions.<br>• However, equal proportion of both correctly and incorrectly answered questions had the explanation and additional insight.<br>• Mean length of both the questions and the responses was similar among questions answered correctly and incorrectly. | ChatGPT as used in this investigation did not answer sufficient multiple-choice questions correctly for it to provide substantial assistance in preparing for board certification. |
| Mihalache et al.[48] | Performance in answering practice questions for board certification | ChatGPT-4<br>Ophthalmology trainees | OphthoQuestions for the OKAP and Written Qualifying Examination for board certification in ophthalmology | Ophthalmologist | 84%<br>71% | • Chatbot responded accurately to general medicine, retina, vitreous and uveitis questions (100%), while lower performance in clinical optics (62%).<br>• Explanations and additional insight were provided for 98% of the questions. | ChatGPT-4's performance appeared to improve compared to the previous version of ChatGPT-3. |
| Moshirfar et al.[49] | Performance in answering ophthalmology questions | ChatGPT-4<br>ChatGPT-3.5<br>Human expert | StatPearls question bank, difficulty levels (rated 1–4) | Ophthalmologist | 73.2%<br>55.5%<br>58.3% | GPT-4's performance was better than human in all categories except 'lens and cataract' where human performed slightly better than GPT-4 (57% vs. 52%, $p = 0.57$). | GPT-4's performance is significantly improved over GPT-3.5 and human professionals. |
| Panthier and Gatinel[50] | Performance in successfully completing ophthalmology examination in French | ChatGPT-4 | European Board of Ophthalmology examination | Ophthalmologist | 91% | • Success rate demonstrates a high level of competency in ophthalmology knowledge and application.<br>• Excelled across all question categories, indicating a strong understanding of basic sciences, clinical knowledge and clinical management. | • AI-LLMs have the potential to positively impact medical education and practice.<br>• With 91% success rate, ChatGPT has shown its potential as a valuable resource for ophthalmologists and other medical professionals seeking guidance on complex cases. |

(Continues)

**TABLE 3** (Continued)

| Author (Year) | Aim/Purpose | LLM used | Content | Grader/Evaluator | Accuracy | Main/Other finding | Author conclusion |
|---|---|---|---|---|---|---|---|
| Raimondi et al.[51] | Ability in answering ophthalmology examinations | ChatGPT-3.5<br>ChatGPT-4<br>ChatGPT-4 prompt<br>Google Bard<br>Bing Chat | Part 1 and Part 2 of the FRCOphth examination (non-image text-based MCQs) | Ophthalmologist | 55.1% and 49.6%<br>NA and 79.1%<br>NA and 88.4%<br>62.6% and 51.9%<br>78.9% and 82.9% | • Bing Chat surpassed ChatGPT-3.5 (Odds ratio, OR 6.37) and Google Bard (OR 3.73) in part 1 questions.<br>• ChatGPT-4 and Bing Chat outperformed ChatGPT-3.5 in part 2. | • Accuracy of the LLMs was notably higher for questions related to the cornea and external eye.<br>• LLMs outperform the set standard or threshold of these specialist examinations.<br>• Traditional assessment does not measure clinical competence. |
| Sakai et al.[52] | Potential in answering ophthalmology specialist examination in Japanese language | ChatGPT-3.5 zero-shot prompting<br>ChatGPT-4 zero-shot prompting<br>ChatGPT-4 few-shot prompting<br>Human examinees | Text-based general questions, either as MCQs with five options or multiple-response questions | Ophthalmologist | 22.4%<br>45.8%<br>46.2%<br>65.7% | • Correct answer rates of GPT-3.5 was 2–3 times lower and GPT-4 was close to 70% of examinees.<br>• Worst performance in paediatric and best in blepharoplasty and orbit. | ChatGPT'S performance was satisfactory, and the results serve as a fundamental basis for considering its practical application using non-English language. |
| Sensoy et al.[53] | Determine the level of knowledge | ChatGPT-3.5<br>Bing Chat<br>Google Bard | Questions from AAO's BCSC on ophthalmic pathology and intraocular tumours | Answer key (2022–2023) available in book | 58.3%<br>63.9%<br>69.4% | The rates of correct answers were similar among the 3 LLMs. | AI chatbots can provide information related to ophthalmic pathologies and intraocular tumours, but not all are correct. |
| Singer et al.[54] | Create and test a new chatbot for ophthalmology | Aeyeconsult (GPT-4 + Langchain + Pinecone)<br>ChatGPT-4 | OKAP style questions from OphthoQuestions | Textbook | 83.4%<br>69.2% | Aeyeconsult's weakest performance was in Clinical Optics (68.1%), but it still outperformed ChatGPT-4 (45.5%). | Aeyeconsultant is more accurate than ChatGPT-4 in answering OKAP style questions. |
| Taloni et al.[56] | Compare the overall performance on answering questions | Human (past examination records)<br>ChatGPT-4<br>ChatGPT-3.5 | MCQs on ophthalmology subspecialties under the practice area of diagnostics/clinics, medical treatment and surgery | Answers from AAO website | 75.7%<br>82.4%<br>65.9% | • GPT-4 has substantial improvement and lower word count over GPT-3.5; both GPTs outperformed human (53.5% and 47.8% vs. 40.1%) in difficult questions.<br>• Worst results were in surgery-related questions. | ChatGPT is achieving higher performance compared to humans, but still limited by inconsistency, especially when it comes to surgery. |

**TABLE 3** (Continued)

| Author (Year) | Aim/Purpose | LLM used | Content | Grader/Evaluator | Accuracy | Main/Other finding | Author conclusion |
|---|---|---|---|---|---|---|---|
| Teebagy et al.[55] | Performance in answering questions | ChatGPT-3.5<br>ChatGPT-4 | Practise questions for the OKAP examination | Ophthalmologist | 57.0%<br>81.0% | • Improved result for GPT-4 was present even when comparing the different subspecialty in ophthalmology. | • ChatGPT-4 outperformed ChatGPT-3.5 in the OKAP examination, suggesting the capacity to support medical education and practice.<br>• Superior performance of ChatGPT-4 is linked to a refined AI model with wider training and up-to-date data set, which better applied its knowledge to understanding concepts and optimal clinical decision-making. |
| Thirunavukarasu et al.[57] | Gauge the ophthalmological knowledge base and reasoning capability | ChatGPT-4<br>ChatGPT-3.5<br>LLaMA<br>PaLM 2 | FRCOphth Part 2 questions | Ophthalmologist, ophthalmology trainees, junior doctors | 69.0%<br>48.0%<br>32.0%<br>56.0% | • GPT-4 responses compared favourably with expert ophthalmologists (76%), ophthalmology trainees (59%) and unspecialised junior doctors (43%).<br>• Low agreement between LLMs and physicians reflected individual differences in knowledge and reasoning. | LLMs are not capable of replacing ophthalmologists but they are approaching expert-level knowledge and reasoning skills in ophthalmology. |

Abbreviations: AAO, American Academy of Ophthalmology; AAO's BCSC-SAP, American Academy of Ophthalmology's Basic and Clinical Science Course Self-Assessment Program; DOI, digital object identifier; FRCOphth, Fellowship of Royal College of Ophthalmologists; LLM, Large language model; LLaMA, Large Language Model Meta AI; MCQ, multiple choice questions; NA, not available; OKAP, Ophthalmology Knowledge Assessment Program; Optho, Ophthalmologist.

review. ChatGPT was found to provide article titles which were non-existent (60%–70%) with digital object identifiers (DOIs) belonging to different articles. The lack of training to distinguish between valid and invalid sources and their relative importance in a field may be the reason behind the errors encountered. The authors concluded that ChatGPT cannot consistently retrieve appropriate articles reliably and is not recommended for literature reviews on dry eye disease.[94] A recent study reported the use of a 'zero-shot' classification approach, that is, without any previous training or exposure of the LLM in categorising and trend analysis of ophthalmology articles along with identification of emerging scientific trends. The proposed framework had high accuracy with 86% correct classification of articles as well as being time efficient.[95]

### Scientific writing

An evaluation of ChatGPT and DALL-E 2 (openai.com), a prompt-based image generator, in writing a scientific article on the 'Complications of the use of silicone oil in vitreoretinal surgery' showed insufficient accuracy and reliability to produce scientifically rigorous articles. Despite the topic chosen being widely described in the literature, the language generated was uncommon with conceptual errors, the information provided was superficial and references did not represent the existing literature.[96] The image generated by DALL-E 2 was not representative of the topic.

Language models like ChatGPT can serve as research assistants but are limited to aiding in certain stages of research papers, like data analysis, literature review, hypothesis generation and peer review, thereby offering valuable insights. With advancing technology, the potential impact of LLMs in research is expected to expand further. Hence, it is essential to recognise the contributions of LLMs appropriately and acknowledge their involvement. ChatGPT has been considered to meet the first three criteria outlined by the International Committee of Medical Journal Editors (ICMJE) but fails to meet the fourth criterion to qualify as an author.[97] Concerns were raised about the suitability of AI tools as co-authors in research papers, citing ethical considerations and copyright limitations.[98] The existing legal framework does not permit non-human entities, such as AI tools, to possess copyright ownership rights. Furthermore, the authors should take accountability for the integrity of the content, which cannot be effectively applied to LLMs.[99] Most journals agree that LLMs do not qualify for sole authorship, and agree against LLMs as co-authors of research papers.[97]

Most ophthalmology journals, for example, *JAMA Ophthalmology*, discourage the inclusion of AI-generated content, but they do permit its use under the condition that authors acknowledge the AI models' involvement and assume responsibility for the content's integrity. Notably, journals published by Elsevier (Elsevier.com), which encompass prestigious ophthalmology publications including *Ophthalmology*, *Progress in Retinal and Eye Research* and the *American Journal of Ophthalmology*, allow AI tools solely for enhancing readability and clarity during the writing process. Nevertheless, authors are obligated to submit a declaration on AI usage as part of their manuscript submission.[97] Interestingly, Wiley (wiley.com) states that AI cannot be considered capable of initiating an original piece of research without direction by human authors.[100] The World Association of Medical Editors (WAME) recommended that chatbots cannot be an author as they cannot meet ICMJE authorship criteria and do not understand conflict of interest.[101] To our knowledge, there are only two published review articles[102,103] and a pre-print server manuscript[104] which has listed ChatGPT as a co-author in academia. A third review published their corrigendum after Elsevier's Publishing Ethics Policies were revised, removing ChatGPT from the initial cited co-author list.[105]

In summary, LLMs' performance (median accuracy) depended on the iteration, prompts used and the task domain involved. A human expert (86%) was the best performer for disease diagnosis, followed by ChatGPT-4 (82%). However, ChatGPT-4 (75.9%) topped the list when answering text-based ophthalmology examination questions, followed by Bing Chat (75%). Similarly, ChatGPT-4 (84.6%) and Bing Chat (78.5%) were the best options for providing information and answering questions (Figure 3). LLMs performed best in general ophthalmology but worse in the ophthalmic subspecialties. ChatGPT-4 outperformed human experts in symptom triaging (98% vs. 86%). However, LLM can generate fake data and hallucinations based on its training data.

## FUTURE OF LLM AND ITS POTENTIAL USES

### Voice-assisted ChatGPT

Combining the functionality of ChatGPT with voice assistant is used by Farcana (farcana.com) in gaming. It offers gamers a new approach to general account management; for example, the AI-powered voice assistant can teach the bot-specific actions like game strategy by extracting data from previous records of top players.[106] Voice dictation is also helpful in familiarising beginners with the game mechanics, checking account balances, evaluating gaming activity and suggesting improvements.[106] These features enhance the player's level and skill proficiency by allowing gamers to focus better on their game. Thus, chatbots with voice assistants can interact in real time and provide personalised services.[106] This can improve the user experience and further automate the interaction process.

**TABLE 4** Summary of current studies on the performance of LLM in providing information and answering questions.

| Author (Year) | Aim/Purpose | LLM used | Content | Grader/Evaluator | Accuracy | Main/Other finding | Author conclusion |
|---|---|---|---|---|---|---|---|
| *Performance in providing information and answering questions* | | | | | | | |
| Ali[59] | Performance in the context of lacrimal drainage disorders | ChatGPT-3 | Set of prompts through questions and statements spanning common and uncommon aspects of lacrimal drainage disorders | Lacrimal surgeons/ Ophthalmologist | 40.0% | • 40% were graded as correct, 35% partially correct and 25% incorrect.<br>• Responses were detailed but had factual errors in them, especially for prompts related to surgery and not precisely evidence based. | • Performance of ChatGPT was unsatisfactory and can be termed average.<br>• AI chatbot has enormous potential, but needs to be specifically trained and retrained for individual medical subspecialties. |
| Anguita et al.[70] | Assess the information type and accuracy provided in the context of vitreoretinal surgery | ChatGPT-3.5<br>Bing AI<br>Docs-GPT Beta | Common patient queries from daily practice on medical advice and medical conditions/post-operative advice | Vitreo retinal surgeons/ Ophthalmologist | 80%–88%<br>81%–100%<br>80%–81% | ChatGPT offered the most detailed information among the LLMs, while Bing AI provided verifiable references. | • LLMs showed acceptable performance with great potential.<br>• Patients must be educated, and ophthalmologists made aware of its limitations. |
| Barclay et al.[71] | Assess the quality and accuracy of information | ChatGPT-3.5<br>ChatGPT-4 | Commonly asked questions related to endothelial keratoplasty and Fuchs dystrophy | Corneal specialists/ Ophthalmologists | 61.0%<br>89.0% | GPT-4 vs. GPT-3.5 responses:<br>• Improved score (grade): 1.4 (A-) vs. 2.5(B-).<br>• Agree with scientific consensus (5% vs. 35%).<br>• More comprehensible (5% vs. 14%).<br>• Less harmful (4% vs. 12.5%).<br>• Equally biased (4% vs. 3%). | • Quality of responses improved in GPT-4.<br>• Lower odds of giving response against scientific consensus<br>• Capable of making inaccurate statements. |
| Bernstein et al.[60] | Quality of ophthalmological advice compared to an ophthalmologist | ChatGPT-3.5<br>Human (expert) | Patient questions from a single online forum that elaborated on the situation and provided context. Instruction prompt engineering was used to adapt the chatbot to the task of responding to questions | Ophthalmologist | 77.4%<br>75.4% | • Prevalence ratio (PR) of incorrect or inappropriate material was 0.92, harm 0.84 and extent of harm 0.99.<br>• Mean (SD) accuracy for distinguishing between LLM and human responses was 61.3% (9.7%). | • LLM was able to generate responses to long user-written, complex and nuanced medical queries.<br>• Likelihood of chatbot answers containing incorrect or inappropriate material was comparable with human answers.<br>• Responses did not differ significantly from ophthalmologist-written responses. |

(Continues)

**TABLE 4** (Continued)

| Author (Year) | Aim/Purpose | LLM used | Content | Grader/Evaluator | Accuracy | Main/Other finding | Author conclusion |
|---|---|---|---|---|---|---|---|
| Biswas et al.[61] | Assess the utility in providing patient information | ChatGPT-3.5 | Human-like queries on nine categories of disease summary, cause, symptom, onset, prevention, complication, natural history of untreated myopia, treatment and prognosis of myopia | Optometrist | 73.0% | • Of the responses, 24% were rated very good, 48.7% good, 21.8% acceptable, 3.6% poor and 1.8% very poor.<br>• Most errors in ChatGPT responses were in the categories of definition of myopia, its treatment strategy and prevention. | • Overall ChatGPT generated good quality information on myopia.<br>• Further evaluation and awareness concerning its limitations are crucial to avoid potential misinterpretation.<br>• Usefulness in replying to patient queries needs to be tested in clinical setting. |
| Caranfa et al.[72] | Understand whether answers to vitreoretinal queries change over time | ChatGPT-3.5 | Frequently asked questions on vitreoretinal conditions | Vitreoretinal surgeons/ Ophthalmologist | 15.4% | On repeat query, 30.8% answers improved in accuracy, 19.2% showed worse accuracy. | ChatGPT can provide largely inaccurate and inconsistent responses to questions concerning vitreoretinal disease. |
| Cardona et al.[73] | Accuracy of responses and their references evaluated for precision & relevance | ChatGPT-3.5 | Queries on topics of contact lenses, anterior eye, low vision, binocular vision and vision therapy | Expert optometrist<br>Students (UG & PG) | Median score: 6–8<br>Median score: 7.5–9 | References were 24% accurate and 19.3% relevant. Students graded its usefulness between 7 and 8.5. | Useful for optometric education, but expert appraisal of the responses and references are required. |
| Chowdhury et al.[62] | Safety and appropriateness of responses to patient queries | ChatGPT-4 | Questions on cataract surgery postoperative patient questions | Ophthalmologist | 59.9% | • 36.3% responses were rated somewhat helpful.<br>• Majority (92.7%) of the responses were rated having low likelihood of harm, 24.4% had possibility of moderate or mild harm, whereas 9.5% contradicted clinical or scientific consensus.<br>• Symptom information had a higher proportion of incorrect content without clinical reasoning. | Even with no fine-tuning and minimal prompt engineering, LLMs like ChatGPT have the potential to helpfully address routine patient queries from a real-world data set of transcribed questions following cataract surgery. |

**TABLE 4** (Continued)

| Author (Year) | Aim/Purpose | LLM used | Content | Grader/Evaluator | Accuracy | Main/Other finding | Author conclusion |
|---|---|---|---|---|---|---|---|
| Cox et al.[63] | Accuracy and safety of medical information to patients | ChatGPT-4 | Questions on blepharoplasty developed from the American Society of Plastic Surgeons handbook | Ophthalmologist | NA | • ChatGPT-4 delivered prompt and safe medical advice in response to questions from patients seeking blepharoplasty in a language understandable by a lay person.<br>• Answers were broad and detailed as it is possibly incapable of personalised advice. | • ChatGPT-4 holds promise as an adjunctive tool for disseminating medical information to prospective patients.<br>• Since it lacks knowledge updates, utilising GPT-4 to convey medical information requires caution.<br>• Patients should be encouraged to use the information as an adjunct or a foundation for more personalised discussion with surgeons. |
| Eid et al.[74] | Compare the readability of ophthalmic plastic surgery patient education materials (PEM) | ChatGPT-4<br>Google Bard | Prompted to produce PEM at a 6th-grade reading level | Plastic surgeon/ Ophthalmologist | FRES: 36.5<br>FRES: 52.3 | • PEM generated by BARD was easier to read than GPT-4.<br>• When prompted to produce PEM at a 6th-grade reading level, both Chatbots significantly improved their readability scores (easier to read, FRES: 67.9). | Guided by appropriate prompts, chatbots can generate accessible and comprehensible PEMs. |
| Ferro Desideri et al.[75] | Efficacy in addressing age-related macular degeneration patients' questions | ChatGPT-3.5<br>Bing Chat<br>Google Bard | Patient questions related to medical advice and regarding intravitreal injections advice | Retinal specialist/ Ophthalmologist | 80%, Mean (SD): 1.20 (0.41) and 1.07 (0.27)<br>46.7%, 1.60 (0.63) and 1.69 (0.63)<br>53.3%, 1.60 (0.73) and 1.38 (0.63) | Overall agreement between the LLMs is low (Cronbach's $\alpha$ of 0.237). | ChatGPT-3.5 consistently offered the most accurate and satisfactory responses, particularly with technical queries. |
| Hu et al.[69] | Ability to identify rare ophthalmic diseases in simulated patient | ChatGPT-4 | Cases of treatable rare ophthalmic disease with confirmed diagnosis from EyeRounds service | Senior ophthalmologists | 83.3% | • 90% response were graded as 'appropriate' with prompts on chief complaints, history of present illness and other ocular examinations.<br>• Only 50% of prompts with only chief complaint were appropriate. | • ChatGPT can identify rare eye diseases.<br>• Can be a potential consultation assisting tool to obtain referral suggestion.<br>• Junior ophthalmologists may benefit in diagnosing rare eye diseases quickly and accurately in the future. |

(Continues)

**TABLE 4** (Continued)

| Author (Year) | Aim/Purpose | LLM used | Content | Grader/Evaluator | Accuracy | Main/Other finding | Author conclusion |
|---|---|---|---|---|---|---|---|
| Kianian et al.[77] | Assess the readability of information generated and to evaluate ability to analyse existing information found online | ChatGPT-3.5 | Specific prompts to create patient handouts and evaluate information from existing webpages on glaucoma surgery | Ophthalmologists | FKGL score: 9.4 (ChatGPT) vs. 11.1 (webpages); Discern average score of 50.9 by both human grader and ChatGPT | ChatGPT produced higher (9th) grade reading level although prompted to generate for lower (6th) grade; assessed quality had high precision and same score. | ChatGPT does not generate easy to understand health information for average reader; but can help with quality measurement. |
| Kianian et al.[76] | Assess the ability to generate readable information and improve readability of online health information | ChatGPT-4<br>Google Bard | Specific and non-specific prompts to write patient-targeted health information on uveitis that is easy to understand by an average American | Uveitis specialists/ Ophthalmologists | FKGL score: 6.3–9.2<br>FKGL score: 10.5–11.1 | ChatGPT responses have lower FKGL scores (easier to understand) and contained less complex words than Bard. | • ChatGPT can both create content and rewrite information to patients in an easier-to-digest form if asked with specific prompt.<br>• Can potentially aid patient learning and treat uveitis more adequately. |
| Lim et al.[64] | Performance in delivering accurate responses to common myopia-related queries | ChatGPT-3.5<br>ChatGPT-4<br>Google Bard | Commonly asked myopia care-related questions, which were categorised into six domains—pathogenesis, risk factors, clinical presentation, diagnosis, treatment and prevention and prognosis | Paediatric ophthalmologist | 61.3%<br>80.6%<br>54.8% | • ChatGPT-4 exhibited lower proportions of responses with a 'poor' rating (9.7%), compared with ChatGPT-3.5 (16.1%) and Google Bard (16.1%).<br>• LLM-Chatbots performed consistently across domains, except for 'treatment and prevention' (poor).<br>• ChatGPT-4 still performed superiorly in this domain, receiving 70% 'good' ratings, compared to 40% in ChatGPT-3.5 and 45% in Google Bard. | • ChatGPT-4 outperformed both ChatGPT-3.5 and Google Bard in responding to common myopia-related queries.<br>• This underscores the promising potential of ChatGPT-4 in delivering accurate and comprehensive information regarding myopia care. |
| Momenaei et al.[65] | Appropriateness and readability of medical knowledge | ChatGPT-4 | List of common questions (definition, prevalence, visual impact, diagnostic methods, surgical and nonsurgical treatment options, postoperative information, surgery-related complications and visual prognosis) regarding common vitreoretinal surgeries | Retinal specialist/ Ophthalmologist | 84.6%–92.0% | • Responses were inappropriate for 5.1%–8.3% cases.<br>• Based on the FKGL and Flesch reading ease scores, the answers appear to be difficult or very difficult for the average lay person to comprehend.<br>• Requires a college graduation level of understanding to understand. | • Most of the answers provided by ChatGPT-4 were consistently appropriate.<br>• ChatGPT and other natural LLMs in their current form are not a source of factual information.<br>• Patients, physicians and laypersons should be advised of the limitations of these tools. |

**TABLE 4** (Continued)

| Author (Year) | Aim/Purpose | LLM used | Content | Grader/Evaluator | Accuracy | Main/Other finding | Author conclusion |
|---|---|---|---|---|---|---|---|
| Nanji et al.[78] | Evaluation of postoperative ophthalmology patient instructions | ChatGPT-3.5 | Postoperative patient instructions were asked and examined for seven common ophthalmic procedures | Ophthalmologist | PEMAT-P score (understandability & actionability): 72.0% | PEMAT-P scores for Google, Canada MyHealth and UK NHS were 75%, 76% and 80%, respectively. | Postoperative instructions from ChatGPT, Google Search and two institutions provide procedure specific information at a comparable rate. |
| Nikdel et al.[79] | Assess the responses to frequently asked questions regarding two common paediatric ophthalmologic disorders | ChatGPT-4 | Questions about amblyopia<br>Questions about childhood myopia | Paediatric ophthalmologists | 84.6%<br>87.5% | Most noticeable inaccurate responses were related to the:<br>• Definition of reverse amblyopia.<br>• Threshold of refractive error for prescription of spectacles to myopic children. | ChatGPT has the potential to serve as an adjunct informational tool for paediatric ophthalmology patients and their caregivers by demonstrating a relatively good performance. |
| Nunes et al.[80] | Examine the potential in improving access to eye care in underserved regions | ChatGPT-3 | Common layperson questions on diabetic retinopathy (DR) | Ophthalmologist | NA | • ChatGPT shows promise in counselling DR patients in regions with restricted ophthalmic care.<br>• It can provide educational information and basic counselling. | • Cannot replace expertise of medical professionals.<br>• Can serve as a valuable instrument to improve health education, communication, more effective and efficient healthcare delivery. |
| Potapenko et al.[21] | Accuracy in information for common retinal diseases | ChatGPT-3 | Questions on disease summary, prevention, treatment options and prognosis of common retinal diseases | Retinal specialist/ Ophthalmologist | 71.0% | • Response accuracies were 45% highly accurate, 26% minor non-harmful inaccuracies, 17% potential misinterpretations/ inaccuracies and 12% marked as potentially harmful inaccuracies.<br>• Inaccuracies were mostly related to the treatment. | LLM-based AI-chatbots can improve patient information and reduce the overall costs of eye care. |
| Potapenko et al.[68] | Accuracy of responses to common questions regarding optic disc drusen | ChatGPT-4 | Questions on disease summary, diagnosis, treatment, prevention and prognosis | Ophthalmologists with experience in optic disc drusen and research | 50.0% | • 17% responses were accurate, 33% minor inaccuracies, 45% major inaccuracies without potential harm and 5% marked as potentially harmful.<br>• Inaccuracies were mostly related to the treatment and prognosis. | LLM often provides relevant answers, but inaccuracies become potentially harmful when questions deal with treatment and prognosis. |

(Continues)

**TABLE 4** (Continued)

| Author (Year) | Aim/Purpose | LLM used | Content | Grader/Evaluator | Accuracy | Main/Other finding | Author conclusion |
|---|---|---|---|---|---|---|---|
| Pushpanathan et al.[81] | Proficiency in addressing queries related to ocular symptoms | ChatGPT-3.5<br>ChatGPT-4<br>Google Bard | Queries on common ocular symptoms and visual disturbance | Consultant Ophthalmologist | 59.5%<br>89.2%<br>40.5% | • Comprehensiveness scores of correct responses were optimal (4.6–4.7 out of 5).<br>• Self-awareness or correction capabilities to incorrect responses were subpar to moderate. | ChatGPT-4 has potential to deliver accurate and comprehensive responses to ocular symptom inquiries. |
| Rajabi et al.[82] | Evaluate the accuracy, informativeness and references on TED | ChatGPT-3.5 | Questions that patients commonly ask in ophthalmology clinics | Ophthalmologist | NA | • ChatGPT offered quick and safe medical advice on common TED questions.<br>• Information provided was sufficiently detailed and easily comprehensible but with few fabricated references. | • ChatGPT offers promising medical information with patients seeking knowledge on TED.<br>• Risk of false scientific references. |
| Rasmussen et al.[66] | Accuracy of responses to typical patient-related questions | ChatGPT-3 | Questions on VKC were formulated by two 'experienced clinical experts' related to the aetiology, prognosis, treatment, prevention and allergy | Ophthalmologist | Median score 4.0 (mean 3.3) | Responses were relevant, but sometimes can be inaccurate with misleading and possibly hazard, especially concerning the treatment and potential adverse effects of medications. | • ChatGPT often provides relevant responses to typical patient and parent questions on VKC.<br>• It also provides inaccurate and potentially dangerous statements, particularly regarding treatment and potential side effects of medications.<br>• Patients and parents should exercise caution when relying solely on ChatGPT for medical guidance. |
| Solli et al.[83] | Examine the reproducibility and quality of responses to patient questions commonly asked prior to ophthalmic procedures | ChatGPT-3.5 | Prompted to list possible complications of five common ophthalmic procedures with associated statistics and references | American Academy of Ophthalmology website and consensus statement | NA | • ChatGPT provides moderately reproducible (73%) and relatively accurate answers.<br>• Narrative content is more reproducible than numerical information.<br>• Frequently lists irrelevant or non-existent references. | • ChatGPT holds promise as an informative tool.<br>• It is important to review ChatGPT-generated content cautiously. |

**TABLE 4** (Continued)

| Author (Year) | Aim/Purpose | LLM used | Content | Grader/Evaluator | Accuracy | Main/Other finding | Author conclusion |
|---|---|---|---|---|---|---|---|
| Tsui et al.[67] | Ability to respond to prompts concerning common ocular symptoms | ChatGPT-3 | Scripted prompts reflecting common patient messages about retinal conditions | Retinal specialist/ Ophthalmologist | 80.0% | • 80% of the responses were graded as precise and suitable, whereas two 20% were imprecise and unsuitable.<br>• Unsuitable ones lacked urgency in referral of acute conditions and without detailed inquiry to the level necessary for appropriate triage.<br>• No sources cited or follow-up questions asked. | • Clinicians should be aware of the public health risk from patients using online chatbots.<br>• Chatbots may not be made consistently or appropriately aware of the urgency of symptoms.<br>• Chatbots could auto-draft responses for clinicians to review, edit as needed and send to patients, decreasing typographical burden. |
| Wei et al.[84] | Compare responses when faced with a set of patient questions | ChatGPT-3<br>Bing Chat<br>Google Bard | Questions regarding cataract surgery from patient perspective | Ophthalmologist | 72.0%<br>76.0%<br>72.0% | No AI model provided an answer that was deemed inappropriate with dangerously incorrect information. | Text-based generative AI models can provide mostly accurate information. |

Abbreviations: AAO's BCSC-SAP, American Academy of Ophthalmology's Basic and Clinical Science Course Self-Assessment Program; AI, artificial intelligence; DOI, digital object identifier; FKGL, Flesch–Kincaid Grade Level; FRCOphtha, Fellowship of Royal College of Ophthalmologists; FRES, Flesch Reading Ease Score; LLM, Large language model; NA, not available; NHS, National Health Service; OKAP, Ophthalmology Knowledge Assessment Program; PEM, patient education materials; PEMAT-P, Patient Education Materials Assessment Tool-Printable; PG, postgraduate; TED, thyroid eye disease; UG, undergraduate; VKC, vernal keratoconjunctivitis.

## ChatFFA model

A bilingual (Chinese and English) model utilising ChatGPT-3.5 for answering tasks related to fundus fluorescein angiography (FFA) has been developed.[107] The model achieved 60%–68% accuracy (as graded by an ophthalmologist) in report generation and disease diagnosis. In cases of microaneurysms, diabetic retinopathy and arteriosclerosis, it reached an accuracy of 87%–94%.[107]

## Electronic health records

The way in which large clinical language models, consisting of billions of parameters, can aid medical AI systems in effectively utilising unstructured Electronic Health Records (EHRs) remains uncertain. Fine-tuned LLMs like BioBERT (nvidia.com), BlueBERT (ncbi.nlm.nih.gov), DistilBERT (huggingface.co) and ClinicalBERT (mit.edu) demonstrated impressive capabilities (81.5%–84.3% precision) in accurately detecting ophthalmological examination components such as slit lamp or fundus examination from clinical notes.[108] This highlights the promising prospect of leveraging these language models to extract and comprehend pertinent details efficiently from extensive patient records; a task that would otherwise be quite challenging. GatorTron (nvidia.com), an LLM consisting of >90 billion words of text (including >82 billion words of de-identified clinical text) was evaluated on five clinical NLP tasks including clinical concept extraction, medical relation extraction, semantic textual similarity, natural language inference (NLI) and medical question answering (MQA).[109,110] The GatorTron model was found to increase the scale of the clinical language model significantly, expanding it to 8.9 billion parameters from 110 million (BioBERT or PubMedBERT) and 345 million (ClinicalBERT) parameters.[109] As a result, they enhanced performance across the NLP tasks; the mean GatorTron-large model achieved accuracies of 90% and 93% for NLI and MQA, respectively (which is a 6.8%–9.5% improvement in accuracy over BioBERT and ClinicalBERT).[110] These advancements have the potential to be integrated into medical AI systems, ultimately leading to improved healthcare and/or ophthalmic care delivery.

## Understanding patient satisfaction

Although LLMs are not yet used to determine patient satisfaction, NLP of Healthgrades (Healthgrades.com, a verified physician review website available in the USA) reviews has been used to understand the determinants of patient satisfaction and the sentiment score of ophthalmologists.[111] Since LLMs exhibit remarkable performance across a diverse set of NLP tasks, LLMs are also expected to estimate

**TABLE 5** Summary of current studies on the performance of LLM in other potential applications.

| Author (Year) | Aim/Purpose | LLM used | Content/Query | Grader/Evaluator | Accuracy | Main/Other finding | Author conclusion |
|---|---|---|---|---|---|---|---|
| *Other potential applications* | | | | | | | |
| Ali and Singh[86] | Assessed the ability to generate abstracts | ChatGPT-3.5 | Prompt covered all specialties of ophthalmology, the journal, title, data and outcomes | Ophthalmologist | 41.7% | • Language of the abstract differed in responses to the prompt with significant amount of generic text.<br>• 'Fake data' from unknown sources; could easily be missed as written by human. | ChatGPT, when not provided with a data set, tends to plagiarise from unknown sources in medical literature. |
| Ćirković and Katz[87] | Validate capability in categorising refractive surgery candidates to the ideal procedure | ChatGPT-4 | Prompts with commonly used set of clinical parameters (demographic, lifestyle, ocular) | Ophthalmologist | 68.0%–88.0% | • Minimal to moderate agreement (0.399–0.610). with clinician.<br>• Precision was (75%–88%) and AUC 0.79. | GPT-4 exhibits promising agreement with clinician categorisations but is unstable and variable in response. |
| Hua et al.[85] | Evaluate and compare the quality of ophthalmic scientific abstracts and references generated | ChatGPT-3.5<br>ChatGPT-4 | Scientific abstracts and references for clinical research questions across ophthalmology subspecialties were generated | Physician (author) used modified DISCERN criteria and performance evaluation scores | Score: 35.9<br>Score: 38.1 | GPT-3.5 vs. GPT-4:<br>• Mean fake scores were 65.4%–69.5% and 10.8%–42.7%.<br>• Hallucination rates for references 33% and 29%.<br>• Helpfulness scores 3.36 and 3.79.<br>• Truthfulness scores 3.64 and 3.86.<br>• Harmless scores 3.57 and 3.71. | Chatbots generated average-quality abstracts with high hallucination rate of generating fake references. |
| Oca et al.[88] | Accuracy of and bias in recommendations for ophthalmologists | ChatGPT-3.5<br>Bing AI<br>Google Bard | Recommend ophthalmologists practicing in the 20 most populated cities in the United States | Data from the United States Census Bureau | 26.2%<br>32.5%<br>37.5% | • Tendency against recommending female ophthalmologists and favouring ophthalmologists in academic medicine.<br>• Recommended physicians were often in specialties other than ophthalmology or not in or near the desired city. | AI chatbots are unreliable with substantial bias and inaccuracy present among recommendations for ophthalmologists. |
| Ong et al.[89] | Ability to generate ICD-10 code | ChatGPT-3.5 | Mock retina encounters which included demographics, history of present illness, exam, impression including plan and follow-up | ICD handbook | 59% | • 70% true positive rate.<br>• 59% of response were completely correct.<br>• Remaining 41% have some form of incorrect ICD even if it included the correct diagnosis. | ChatGPT can help in ICD coding but may hallucinate false codes despite understanding the diagnosis. |

**TABLE 5** (Continued)

| Author (Year) | Aim/Purpose | LLM used | Content/Query | Grader/Evaluator | Accuracy | Main/Other finding | Author conclusion |
|---|---|---|---|---|---|---|---|
| Raghu et al.[90] | Ability to predict the risk of DR in patients with diabetes | ChatGPT-4 | Prompt with clinical, biochemical parameters<br>Prompt with clinical, biochemical and ocular parameters | Ophthalmologist | 54.0%–73.0%<br>67.0%–68.0% | Agreement between ChatGPT prediction and the clinical DR status was fair (Cohen's kappa: 0.263–0.351). | ChatGPT has the potential of a preliminary DR screening tool with further optimisation needed for clinical use. |
| Raja et al.[95] | Ability to categorise and trend analysis of ophthalmology articles | Zero shot learning bidirectional and auto-regressive transformers | 1000 ocular disease-related articles | Ophthalmologist | Mean accuracy: 0.86 and mean F1: 0.85 | Accuracy for categories of article type, ocular diseases, study subclasses subranged between 0.82 and 0.93. | The proposed framework achieves high accuracy and efficiency. |
| Seth et al.[91] | Evaluate the capacity, effectiveness and accuracy in generating innovative ideas on designing, implementing and assessing research | ChatGPT-4 | Series of unique questions probing future innovations in oculoplastic surgery were devised | Ophthalmologist | NA | • ChatGPT offered pertinent and precise information.<br>• Some responses lacked depth and provided only a surface-level overview, particularly when addressing more complex queries.<br>• To a lay person without experience in this field, the responses may seem comprehensive. | • ChatGPT displayed convergent thinking by presenting established ideas for future research instead of generating novel insights.<br>• ChatGPT at best should be viewed as an auxiliary research tool, not an original source of research ideas. |
| Singh et al.[92] | Ability to construct ophthalmic discharge summaries and operative notes | ChatGPT-4 | Set of prompts constructed incorporating common ophthalmic surgeries across the subspecialties | Ophthalmologist | NA | • Responses were customised depending on the quality of input (prompt) provided.<br>• Discharge summaries were valid but significantly generic.<br>• Operative notes were comprehensive, but they needed substantial refinement.<br>• ChatGPT consistently acknowledges its errors when confronted with factual inaccuracies.<br>• Mistakes can be prevented in subsequent instances with similar prompts. | • Performance of ChatGPT in the context of ophthalmic discharge summaries and operative notes was encouraging and constructed rapidly in a matter of seconds.<br>• Inclusion of a human verification step is essential to tune the operative notes and avoid error. |

(Continues)

**TABLE 5** (Continued)

| Author (Year) | Aim/Purpose | LLM used | Content/Query | Grader/Evaluator | Accuracy | Main/Other finding | Author conclusion |
|---|---|---|---|---|---|---|---|
| Singh and Watson[94] | Ability to perform a literature review | ChatGPT-3 | Two prompts of listing the seminal articles on dry eye disease (DED) and articles on a specific topic (meibography and DED) with their DOIs | Ophthalmologist | NA | • ChatGPT provided article titles which are non-existent (60%–70%) with DOI belonging to different articles.<br>• Only 30% match between the articles found between the two prompts. | ChatGPT cannot consistently retrieve appropriate articles reliably and is not recommended for literature reviews on dry eye disease. |
| Valentín-Bravo et al.[96] | Potential in writing scientific articles in ophthalmology | ChatGPT-3 and DALL-E 2 (image generator) | Comment on the complications of using silicone oil in vitreo-retinal surgery and expand the findings into an article with specific structure | Vitreo-retinal surgeon and student | NA | • Response showed insufficient accuracy and reliability to write scientifically rigorous articles.<br>• Despite a common chosen topic, the language generated was uncommon with conceptual error.<br>• Information provided was superficial and references were not representing the existing literature.<br>• The image generated was not representative. | • ChatGPT has specific knowledge of ophthalmology.<br>• Scientific accuracy and reliability on specific topics are insufficient to correctly generate a scientifically rigorous article. |
| Waisberg et al.[93] | Potential application to write ophthalmic operative notes | ChatGPT-4 | Asking ChatGPT to generate a cataract surgery operative note | Ophthalmologist | NA | • GPT-4 produced information on patient details, date of surgery, specifics of the surgical procedure, intraocular lens details, steps in surgery, wound closure, postoperative care, complications and section for signatures. | • This can streamline the surgical procedure documentation process.<br>• May lead to improved efficiency in capturing and recording crucial information.<br>• Despite its utility, it is still limited and should be used in conjunction with human expertise and critically evaluated for optimal results. |

Abbreviations: AAO's BCSC-SAP, American Academy of Ophthalmology's Basic and Clinical Science Course Self-Assessment Program; AUC, area under curve; DOI, digital object identifier; DR, Diabetic retinopathy; F1, F1-Score (F-measure is a measure of predictive performance); FRCOphtha, Fellowship of Royal College of Ophthalmologists; ICD, International Classification of Diseases; LLM, Large language model; NA, not available; OKAP, Ophthalmology Knowledge Assessment Program.

patient satisfaction, opinion, sentiment and emotions. Although certain traditional sentiment analysis tasks have approached near-human performance, attaining a thorough understanding of human sentiment, human subjective feelings remain a distant pursuit. The strong capability of LLMs in text comprehension offers sentiment analysis in the era of LLMs.[112]

## Enabling gene discovery

Med-PaLM 2 has shown the ability to identify accurately mouse genes responsible for susceptibility to six biomedical traits including diabetes and cataract. It can also detect a novel causative murine genetic factor for susceptibility to spontaneous hearing loss.[113] This result demonstrates the capability of LLMs in analysing gene–phenotype relationships and facilitating gene discovery.

## Surgical training and education

The role of LLM (ChatGPT-4) has been evaluated as a teaching assistant to plastic surgery residents. A set of eight roles (cognitive, psychomotor and affective domains) were identified with an inter-observer agreement between experts for content output ranging between 30% and 100%.[114] Incorporating LLMs into surgery residency programmes can provide an interactive, dynamic and personalised learning experience.

## Drug discovery, development and other uses

LLMs can be helpful in the discovery and development of drugs. ChatGPT-4 added a retrieval plug-in that searches new documents and assays to update its knowledge base, designed to help drug discovery.[115] Furthermore, LLMs can provide the blueprint of drug compounds with new structures, helping to predict the drug molecule's pharmacodynamics, pharmacokinetics and toxicity. It also assists in understanding a drug's absorption, distribution, metabolism and excretion; thus optimising drug properties, therapeutic target discovery and assessing toxicity.[116]

While taking multiple medications, patients are at increased risk of experiencing adverse events or drug toxicity due to drug–drug interaction (DDI). ChatGPT can predict and explain the DDIs. Although not always correct, ChatGPT is an efficient tool for understanding DDIs, especially for those without access to healthcare facilities.[117]

LLMs have the potential to assist with the complex process of drug repositioning or repurposing, which is finding new therapeutic purposes and targets for existing drugs. This approach saves time and investment required for new drug development without human trials.[116]

Finally, DrugChat (ai.ucsd.edu) is a ChatGPT-like pharmaceutical domain-specific LLM designed to analyse drug compounds, answer questions and generate text descriptions for drugs.[118] It can enable conversational analysis of drug compounds from both text and drug molecule graphs. Thus, LLMs have a role to accelerate drug discovery, find novel therapeutic molecules and perform functions that benefit patients.

## Diagnosing ocular surface disorders

As well as diagnosing ophthalmic diseases of the posterior segment, NLP has also been used for identifying disorders such as herpes zoster ophthalmicus (HZO)[119] and the quantification of microbial keratitis.[120] The sensitivity and specificity of NLP to quantify microbial keratitis from electronic health records ranged between 75%–96% and 91%–96%, respectively.[120] Similarly, NLP used to screen HZO from clinical notes had a high sensitivity (95.6%) and specificity (99.3%).[119] This indicates huge potential for LLMs in detecting ocular surface disorders.

## Newer ophthalmic domain-specific LLM

Although most LLM systems were not built for any specific domain or field of knowledge, the extent of their popularity in health and ophthalmic care is tremendous. However, the text and images used for ophthalmic conditions are different from general web content. Thus, general domain LLMs may have difficulty dealing with professional conversations, resulting in incorrect answers or false facts. Ophthalmology large language-and-vision assistant (OphGLM) is a newly developed LLM (ailab.lv), and uses visual ability (images) specific to ophthalmic conditions.[121] The OphGLM model conducts disease assessment and diagnosis by analysing fundus images and incorporating ophthalmic knowledge data along with real medical conversations. Additionally, this model incorporates visual capabilities and introduces a novel data set for ophthalmic multimodal instruction tracking and dialogue fine-tuning.[121] Although experimental data sets demonstrate outstanding accuracy for diseases such as diabetic retinopathy, glaucoma and other ophthalmic conditions, it will be important to examine its real-life clinical application. The medical and scientific text domain-specific LLMs like Med-PaLM 2, PubMedBERT (microsoft.com), BioBERT, ScholarBERT (Public.Resource.Org), SciBERT (allenai.com), DARE (healx.ai), ClinicalBERT and BioWordVec (ncbi.nlm.nih.gov) are already outperforming the foundation LLMs in biomedical tasks.[25,95,122]

# LIMITATIONS, CHALLENGES AND ADVANCEMENT OF LLM

The clinical deployment of ChatGPT and similar applications has been hindered by various issues and limitations.

First and foremost was the limited training of both GPT-3.5 and GPT-4 up to September 2021. Without continuous updating of the knowledge, the responses may be entirely outdated and even harmful, especially in the field of health and eye care. Second, lack of domain-specific training data leads to the 'garbage in, garbage out' issue. Although the LLM boasts an impressive size with 175 billion parameters, GPT-3.5 utilises a small fraction of the available data (only 570 GB) for its initial training.[123] Third, the absence of any real-time internet access fundamentally limits LLMs like ChatGPT, with the exception of few LLMs like BlenderBot 3 (ai.meta.com)[124] and Sparrow (deepmind.google)[125] which can access the internet while generating responses. Fourth, intensive fine-tuning and training have developed LLMs that can generate responses which sound plausible and coherent, although not necessarily accurate, when presented with queries. These 'hallucinations' or 'fact fabrications' are inaccurate or fake information invented when the information is not represented in the training data set. With the advent of GPT-4 expanded with Advanced Data Analysis (Python, python.org), if prompted, GPT-4 can even create fake data set for research and publication with an author's desired outcome.[126] Moreover, responses are consensus-based, not evidence-based which can be illusional. Other than continuous updating of knowledge from publications, clinicians possess another advantage of accessing data not yet published from conference presentations and workshops. Thus, in the absence of any benchmarking and the presence of fake data, AI may make it difficult to differentiate fact from fiction. However, LLMs can self-improve by employing chain-of-thought prompting and encouraging self-consistency enabled autonomous fine-tuning, leading to a 5%–10% enhancement in reasoning capabilities of an LLM.[127] LLMs have undergone extensive development over the years, resulting in their emergence with 'few-shot' or 'zero-shot' capabilities. This means they can now recognise, interpret and generate text requiring minimal or no fine-tuning. In other words, few-shot and zero-shot are AI developed to complete tasks with or without exposure to initial examples of the task, with accurate generalisation to unseen examples.[123] These impressive few-shot and zero-shot properties become evident and develop when model size, data set size and computational resources reach a significant scale.[128] Fifth, like any AI system, LLM processing has the 'black box' problem, which is the absence or unclear explanations/reasoning behind the model or decision, which reduces confidence, making interpretation and clinical decision-making tougher.[129] Equally, DALL-E 2, which generates images in response to text prompts, risks false-positive diagnosis when reviewed by a lay person (patient), or even a potentially dangerous false-negative diagnosis, which might lead to false reassurance and delayed treatment.[123] The inability of the earlier iterations of ChatGPT to process images may need incorporation of other transfer models such as the Contrastive Language-Image Pretraining (CLIP) model. CLIP generates text description of an image input,[130] which can be used as an adjunct for LLMs without image output. However, CLIP is designed for the domain of medical imaging and not specific for the eye, which might limit its sensitivity to differentiate eye diseases. Furthermore, the fusion of ChatGPT with Argil plugin (argil.ai) might help creation of images from textual prompts.[58] Sixth, LLMs also raise ethical considerations and challenges, such as bias in data and outputs, privacy concerns and the responsible use of AI technology. Addressing these challenges is crucial to ensure the ethical and responsible deployment of LLMs in various health and eye care applications.[131] GPT-4 raises privacy concerns and lack of accountability in containing patient-identifiable and personal data.[123] Seventh, it should be borne in mind that, while LLMs do not possess emotions or consciousness, they can be designed to generate text based on patterns learned from data that appear empathetic and considerate in certain contexts.[132] However, empathy in health care settings is mainly verbal and from body language, so the ethical considerations of using empathetic LLMs should be considered, especially when dealing with sensitive topics or vulnerable users. Finally, most of the existing research is focussed on the qualitative appraisal of LLMs in artificial settings. Real-world clinical interventions tested through randomised controlled trials evaluating the safety, efficacy, morbidity and other parameters are needed for better understanding and deployment of LLMs in clinical care.

Despite the fears and hype, the barriers to implementation of LLMs replacing healthcare professionals in any capacity remain substantial.[133] LLMs continue to be afflicted by mistakes and errors. Fundamentally, LLMs are limited by the quality of information available for training or browsing in response to queries, which remains governed by human activity (e.g., research, policymaking).

To summarise, the existing use and benefits of LLM in eye care are: (i) disease diagnostic support; (ii) symptom assessment and triage; (iii) patient education, information and engagement; (iv) helping prepare for ophthalmic qualifying examinations; (v) literature analysis/review; (vi) preparing ophthalmic operative notes and (vii) drafting editable response to patient queries. The potential future benefits arising from the current trend and speculation based on the ability of LLMs can be predicted to be: (i) remote monitoring and telemedicine, (ii) readily available and accurate information for clinicians in busy clinics, (iii) evidence-based practice and continuing education, (iv) predicting disease progression, (v) vision correction options, (vi) contact lens recommendations, (vii) research assistant, (viii) clinical documentation and report generation, (ix) electronic health records, (x) clinical decision-making support, (xi) image analysis and interpretation, (xii) understanding patient satisfaction, (xiii) pre- and post-operative patient counselling, (xiv) gene discovery for ophthalmic diseases, (xv) surgical training and education, (xvi) drug discovery, development, repurposing, interactions and (xvii) improve the design of

future LLMs. The earlier iterations of ChatGPT (GPT-3 and GPT-3.5) could only process text-based prompts. Although ChatGPT-4 Vision can process and analyse text and image inputs, its performance has been below par. The way forward will be to develop an ophthalmic domain-specific LLM, which will be competent and have both text and image processing capabilities like ChatFFA and OphGLM. Another possibility of improving the diagnostic capability of LLMs could be to synergise them with image-based deep learning algorithms, thereby enhancing the potential for a contextual interpretation of text and image inputs simultaneously.[58] Notwithstanding that the potential benefits of using LLM in eyecare is immense, only some of it has been explored so far. These findings are similar to that previously observed by a systematic review[22] in healthcare where the conversational LLM was found to reduce overall healthcare cost along with a multitude of benefits in research, education and clinical practice.

Unlike the earlier LLMs, the more recent versions can process, analyse and interpret images. If they can be trained for specific tasks, such as grades of diabetic retinopathy or cup–disc ratio, then the opportunities of ChatGPT become infinite at a much lower cost and greater convenience of using a single application to combine and interpret both image and text inputs. It would give access to a plethora of medical knowledge that can be updated and improved. And yet the downsides of LLMs should be addressed before embracing these applications in eyecare. Moreover, clinical decision-making requires human interaction and consideration of the socio-economic–psychological factor of the patient alongside clinical knowledge. LLMs should be used conservatively in the future by identifying the scope and specific areas of use in eyecare while maintaining the human judgement for clinical decisions.

## CONCLUSION

With the introduction of Cerebras GPT (cerebras.net) (a family of seven GPT models ranging from 111 million to 13 billion parameters),[134] GPT-4 (170 trillion parameters) and those already in existence such as GPT-3.5, GPT-3, Chinchilla (deepmind.google), Meta OPT (ai.meta.com), Pythia (eleuther.ai), PubMedGPT (also known as BioMedLM, crfm.stanford.edu), BioGPT (microsoft.com), BioBERT and PaLM, it appears that LLM-based AIs are the tool of the present and the future,[11] capable of undertaking a host of tasks from clinical decision-making to disease diagnosis, raising patient awareness, preparing for examinations and symptom triaging. Although the diagnostic accuracy widely varies based on the LLM iteration, they are more efficient, faster and repeatable than human non-experts and trainees.[22] Primarily, LLMs are used as medical assistants; their use can be broadened into roles which might potentially save consultation time and reduce burden on clinicians and supporting staff like drafting responses that could be edited and sent to patients.[22,26,61] Assessing the practical application of LLMs in a real-world clinical setting is essential before they can be deployed clinically. However, the patient's perspective, attitude and acceptability of LLM in a variety of ethical and minimally harmful clinical contexts must be considered. Although the image processing capability of GPT-4 alongside its text processing is expected to overcome some of the limitations and outdo the existing AI systems in accurate disease screening and diagnosis, developing ophthalmic domain-specialised LLMs combining multimodal ophthalmic data will be a significant step for the future. Nevertheless, given the limitations of LLM, caution should be exercised before embracing these applications. As AI continues to advance, it is essential to ensure that the potential benefits of introducing these applications in eyecare are maximised, while minimising the potential risks of its implementation. This is achievable only through the engagement of the eyecare community and staying up to date with the potential developments.

## AUTHOR CONTRIBUTIONS

**Sayantan Biswas:** Conceptualization (lead); data curation (lead); formal analysis (lead); investigation (equal); methodology (equal); project administration (equal); resources (equal); software (equal); supervision (equal); validation (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal). **Leon N. Davies:** Data curation (equal); formal analysis (equal); investigation (equal); methodology (equal); validation (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal). **Amy L. Sheppard:** Data curation (equal); formal analysis (equal); investigation (equal); methodology (equal); validation (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal). **Nicola S. Logan:** Data curation (equal); formal analysis (equal); investigation (equal); methodology (equal); validation (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal). **James S. Wolffsohn:** Data curation (equal); formal analysis (equal); investigation (equal); methodology (equal); resources (equal); software (equal); supervision (equal); validation (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal).

## FUNDING INFORMATION

None.

## CONFLICT OF INTEREST STATEMENT

The authors have no proprietary or commercial interest in any materials discussed in this article.

## ORCID

*Sayantan Biswas* https://orcid.org/0000-0001-6011-0365
*Leon N. Davies* https://orcid.org/0000-0002-1554-0566
*Amy L. Sheppard* https://orcid.org/0000-0003-0035-8267
*Nicola S. Logan* https://orcid.org/0000-0002-0538-9516
*James S. Wolffsohn* https://orcid.org/0000-0003-4673-8927

## REFERENCES


1. Misischia CV, Poecze F, Strauss C. Chatbots in customer service: their relevance and impact on service quality. *Procedia Comput Sci*. 2022;201:421–8.
2. Lin WC, Chen JS, Chiang MF, Hribar MR. Applications of artificial intelligence to electronic health record data in ophthalmology. *Transl Vis Sci Technol*. 2020;9:13. https://doi.org/10.1167/tvst.9.2.13
3. Chen JS, Baxter SL. Applications of natural language processing in ophthalmology: present and future. *Front Med*. 2022;9:906554. https://doi.org/10.3389/fmed.2022.906554
4. Foo LL, Lim GYS, Lanca C, Wong CW, Hoang QV, Zhang XJ, et al. Deep learning system to predict the 5-year risk of high myopia using fundus imaging in children. *NPJ Digit Med*. 2023;6:10. https://doi.org/10.1038/s41746-023-00752-8
5. Milea D, Najjar RP, Zhubo J, Ting D, Vasseneix C, Xu X, et al. Artificial intelligence to detect papilledema from ocular fundus photographs. *N Engl J Med*. 2020;382:1687–95.
6. Zhang L, Tang L, Xia M, Cao G. The application of artificial intelligence in glaucoma diagnosis and prediction. *Front Cell Dev Biol*. 2023;11:1173094. https://doi.org/10.3389/fcell.2023.1173094
7. Wang SY, Tseng B, Hernandez-Boussard T. Deep learning approaches for predicting glaucoma progression using electronic health records and natural language processing. *Ophthalmol Sci*. 2022;2:100127. https://doi.org/10.1016/j.xops.2022.100127
8. Gu H, Guo Y, Gu L, Wei A, Xie S, Ye Z, et al. Deep learning for identifying corneal diseases from ocular surface slit-lamp photographs. *Sci Rep*. 2020;10:17851. https://doi.org/10.1038/s41598-020-75027-3
9. Ji Y, Liu S, Hong X, Lu Y, Wu X, Li K, et al. Advances in artificial intelligence applications for ocular surface diseases diagnosis. *Front Cell Dev Biol*. 2022;10:1107689. https://doi.org/10.3389/fcell.2022.1107689
10. Gunasekeran DV, Wong TY. Artificial intelligence in ophthalmology in 2020: a technology on the cusp for translation and implementation. *Asia Pac J Ophthalmol*. 2020;9:61–6.
11. Ting DSW, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, et al. Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol*. 2019;103:167–75.
12. Wu J, Xu L, Yu F, Peng K. Acceptance of medical treatment regimens provided by AI vs human. *Appl Sci*. 2021;12:110. https://doi.org/10.3390/app12010110
13. Murphy TI, Armitage JA, van Wijngaarden P, Abel LA, Douglass AG. A guide to optometrists for appraising and using artificial intelligence in clinical practice. *Clin Exp Optom*. 2023;106:569–79.
14. Peissig PL, Rasmussen LV, Berg RL, Linneman JG, McCarty CA, Waudby C, et al. Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. *J Am Med Inform Assoc*. 2012;19:225–34.
15. Barrows RC Jr, Busuioc M, Friedman C. Limited parsing of notational text visit notes: ad-hoc vs. NLP approaches. *Proc AMIA Symp*. 2000;51–5.
16. Tan Y, Bacchi S, Casson RJ, Selva D, Chan W. Triaging ophthalmology outpatient referrals with machine learning: a pilot study. *Clin Exp Ophthalmol*. 2020;48:169–73.
17. Smith DH, Johnson ES, Russell A, Hazlehurst B, Muraki C, Nichols GA, et al. Lower visual acuity predicts worse utility values among patients with type 2 diabetes. *Qual Life Res*. 2008;17:1277–84.
18. Liu L, Shorstein NH, Amsden LB, Herrinton LJ. Natural language processing to ascertain two key variables from operative reports in ophthalmology. *Pharmacoepidemiol Drug Saf*. 2017;26:378–85.
19. Samant RM, Bachute MR, Gite S, Kotecha K. Framework for deep learning-based language models using multi-task learning in natural language understanding: a systematic literature review and future directions. *IEEE Access*. 2022;10:17078–97.
20. Roumeliotis KI, Tselikas ND. ChatGPT and open-AI models: a preliminary review. *Future Internet*. 2023;15:192. https://doi.org/10.3390/fi15060192
21. Potapenko I, Boberg-Ans LC, Stormly Hansen M, Klefter ON, van Dijk EHC, Subhi Y. Artificial intelligence-based chatbot patient information on common retinal diseases using ChatGPT. *Acta Ophthalmol*. 2023;101:829–31.
22. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare*. 2023;11:887. https://doi.org/10.3390/healthcare11060887
23. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepano C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit Health*. 2023;2:e0000198. https://doi.org/10.1371/journal.pdig.0000198
24. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature*. 2023;620:172–80.
25. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards expert-level medical question answering with large language models. *arxiv* preprint. 2023;2305.090617. https://doi.org/10.48550/arXiv.2305.09617
26. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*. 2023;183:589–96.
27. Rojas-Carabali W, Cifuentes-González C, Wei X, Putera I, Sen A, Thng ZX, et al. Evaluating the diagnostic accuracy and management recommendations of ChatGPT in uveitis. *Ocul Immunol Inflamm*. 2023;1–6. https://doi.org/10.1080/09273948.2023.2253471
28. Madadi Y, Delsoz M, Lao PA, Fong JW, Hollingsworth TJ, Kahook MY, et al. ChatGPT assisting diagnosis of neuro-ophthalmology diseases based on case reports. *medRxiv*. 2023. https://doi.org/10.1101/2023.09.13.23295508
29. Lyons RJ, Arepalli SR, Fromal O, Choi JD, Jain N. Artificial intelligence chatbot performance in triage of ophthalmic conditions. *Can J Ophthalmol*. 2023;S0008-4182(23)00234-X. https://doi.org/10.1016/j.jcjo.2023.07.016
30. Liu X, Wu J, Shao A, Shen W, Ye P, Wang Y, et al. Transforming retinal vascular disease classification: a comprehensive analysis of ChatGPT's performance and inference abilities on non-English clinical environment. *medRxiv*. 2023:2023.06. 28.23291931. https://doi.org/10.1101/2023.06.28.23291931
31. Knebel D, Priglinger S, Scherer N, Klaas J, Siedlecki J, Schworm B. Assessment of ChatGPT in the Prehospital Management of Ophthalmological Emergencies - An Analysis of 10 Fictional Case Vignettes. Klinische Monatsblatter fur Augenheilkunde. 2023. Epub 2023/10/28. ChatGPT in der präklinischen Versorgung augenärztlicher Notfälle – eine Untersuchung von 10 fiktiven Fallvignetten. https://doi.org/10.1055/a-2149-0447
32. Inayat H, McDonald HM, Bursztyn LLCD. Comparison of ChatGPT to ophthalmology resident and staff consultants on an ophthalmological training tool. *Can J Ophthalmol*. 2023;59:e72–e74.
33. Delsoz M, Raja H, Madadi Y, Tang AA, Wirostko BM, Kahook MY, et al. The use of ChatGPT to assist in diagnosing glaucoma based on clinical case reports. *Ophthalmol Therapy*. 2023;12:3121–32.
34. Delsoz M, Madadi Y, Munir WM, Tamm B, Mehravaran S, Soleimani M, et al. Performance of ChatGPT in diagnosis of corneal eye diseases. *medRxiv*. 2023. https://doi.org/10.1101/2023.08.25.23294635
35. Balas M, Ing EB. Conversational AI models for ophthalmic diagnosis: comparison of ChatGPT and the Isabel pro differential diagnosis generator. *JFO Open Ophthalmol*. 2023;1:100005. https://doi.org/10.1016/j.jfop.2023.100005
36. Yu H, Fu Y, Jiang B, Fan P, Shen L, Gao L, et al. Applications of GPT-4 for accurate diagnosis of retinal diseases through optical coherence tomography image recognition. *BMC Ophthalmol*. 2023;1–13. https://doi.org/10.21203/rs.3.rs-3644163/v1
37. Xu P, Chen X, Zhao Z, Zheng Y, Jin G, Shi D, et al. Evaluation of a digital ophthalmologist app built by GPT4-V(ision). *medRxiv*. 2023:2023.11.27.23299056. https://doi.org/10.1101/2023.11.27.23299056
38. Waisberg E, Ong J, Masalkhi M, Kamran SA, Zaman N, Sarker P, et al. Automated ophthalmic imaging analysis in the era of Generative

Pre-Trained Transformer-42023. *Pan Am J Ophthalmol*. 2023;5:50. https://doi.org/10.4103/pajo.pajo_62_23
39. Sorin V, Kapelushnik N, Hecht I, Zloto O, Glicksberg BS, Bufman H, et al. GPT-4 multimodal analysis on ophthalmology clinical cases including text and images. *medRxiv*. 2023:2023.11. 24.23298953. https://doi.org/10.1101/2023.11.24.23298953
40. Antaki F, Milad D, Chia MA, Giguère C, Touma S, El-Khoury J, et al. Capabilities of GPT-4 in ophthalmology: an analysis of model entropy and progress towards human-level medical question answering. *Br J Ophthalmol*. 2023. https://doi.org/10.1136/bjo-2023-324438
41. 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. https://doi.org/10.1016/j.xops.2023.100324
42. Cai LZ, Shaheen A, Jin A, Fukui R, Yi JS, Yannuzzi N, et al. Performance of generative large language models on ophthalmology board-style questions. *Am J Ophthalmol*. 2023;254:141–9.
43. Fowler T, Pullen S, Birkett L. Performance of ChatGPT and Bard on the official part 1 FRCOphth practice questions. *Br J Ophthalmol*. 2023;1–5. https://doi.org/10.1136/bjo-2023-324091
44. Gobira MC, Moreira RC, Nakayama LF, Regatieri CVS, Andrade E, Belfort R Jr. Performance of chatGPT-3.5 answering questions from the Brazilian Council of Ophthalmology Board Examination. *Pan Am J Ophthalmol*. 2023;5:17. https://doi.org/10.4103/pajo.pajo_21_23
45. Jiao C, Edupuganti NR, Patel PA, Bui T, Sheth V. Evaluating the artificial intelligence performance growth in ophthalmic knowledge. *Cureus*. 2023;15:e45700. https://doi.org/10.7759/cureus.45700
46. Lin JC, Younessi DN, Kurapati SS, Tang OY, Scott IU. Comparison of GPT-3.5, GPT-4, and human user performance on a practice ophthalmology written examination. *Eye*. 2023;37:3694–5.
47. Mihalache A, Huang RS, Popovic MM, Muni RH. Performance of an upgraded artificial intelligence chatbot for ophthalmic knowledge assessment. *JAMA Ophthalmol*. 2023;141:798–800.
48. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. *JAMA Ophthalmol*. 2023;141:589–97.
49. Moshirfar M, Altaf AW, Stoakes IM, Tuttle JJ, Hoopes PC. Artificial intelligence in ophthalmology: a comparative analysis of GPT-3.5, GPT-4, and human expertise in answering StatPearls questions. *Cureus*. 2023;15:e40822. https://doi.org/10.7759/cureus.40822
50. 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;46:706–11.
51. Raimondi R, Tzoumas N, Salisbury T, Di Simplicio S, Romano MR. Comparative analysis of large language models in the Royal College of Ophthalmologists fellowship exams. *Eye*. 2023;37:3530–3.
52. Sakai D, Maeda T, Ozaki A, Kanda GN, Kurimoto Y, Takahashi M. Performance of ChatGPT in board examinations for specialists in the Japanese Ophthalmology Society. *Cureus*. 2023;15:e49903. https://doi.org/10.7759/cureus.49903
53. 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;43:4905–9.
54. Singer MB, Fu JJ, Chow J, Teng CC. Development and evaluation of Aeyeconsult: a novel ophthalmology chatbot leveraging verified textbook knowledge and GPT-4. *J Surg Educ*. 2023;S1931-7204(23)00432-4. https://doi.org/10.1016/j.jsurg.2023.11.019
55. 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*. 2023;15:e184–e187.
56. Taloni A, Borselli M, Scarsi V, Rossi C, Coco G, Scorcia V, et al. Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology. *Sci Rep*. 2023;13:18562. https://doi.org/10.1038/s41598-023-45837-2
57. Thirunavukarasu AJ, Mahmood S, Malem A, Foster WP, Sanghera R, Hassan R, et al. Large language models approach expert-level clinical knowledge and reasoning in ophthalmology: a head-to-head cross-sectional study. *medRxiv*. 2023;2023.07. 31.23293474. https://doi.org/10.1101/2023.07.31.23293474
58. Betzler BK, Chen H, Cheng CY, Lee CS, Ning G, Song SJ, et al. Large language models and their impact in ophthalmology. *Lancet Digit Health*. 2023;5:e917–e924.
59. Ali MJ. ChatGPT and lacrimal drainage disorders: performance and scope of improvement. *Ophthal Plast Reconstr Surg*. 2023;39:221–5.
60. Bernstein IA, Zhang YV, Govil D, Majid I, Chang RT, Sun Y, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. *JAMA Netw Open*. 2023;6:e2330320. https://doi.org/10.1001/jamanetworkopen.2023.30320
61. Biswas S, Logan NS, Davies LN, Sheppard AL, Wolffsohn JS. Assessing the utility of ChatGPT as an artificial intelligence-based large language model for information to answer questions on myopia. *Ophthalmic Physiol Opt*. 2023;43:1562–70.
62. Chowdhury M, Lim E, Higham A, McKinnon R, Ventoura N, He Y, et al. Can large language models safely address patient questions following cataract surgery? Proceedings of the 5th Clinical Natural Language Processing Workshop. Toronto: Association for Computational Linguistics; 2023. https://doi.org/10.18653/v1/2023.clinicalnlp-1.17
63. Cox A, Seth I, Xie Y, Hunter-Smith DJ, Rozen WM. Utilizing ChatGPT-4 for providing medical information on blepharoplasties to patients. *Aesthet Surg J*. 2023;43:NP658–P662.
64. Lim ZW, Pushpanathan K, Yew SME, Lai Y, Sun CH, Lam JSH, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *EBioMedicine*. 2023;95:104770. https://doi.org/10.1016/j.ebiom.2023.104770
65. Momenaei B, Wakabayashi T, Shahlaee A, Durrani AF, Pandit SA, Wang K, et al. Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases. *Ophthalmol Retina*. 2023;7:862–8.
66. Rasmussen MLR, Larsen AC, Subhi Y, Potapenko I. Artificial intelligence-based ChatGPT chatbot responses for patient and parent questions on vernal keratoconjunctivitis. *Graefes Arch Clin Exp Ophthalmol*. 2023;261:3041–3.
67. Tsui JC, Wong MB, Kim BJ, Maguire AM, Scoles D, VanderBeek BL, et al. Appropriateness of ophthalmic symptoms triage by a popular online artificial intelligence chatbot. *Eye*. 2023;37:3692–3.
68. Potapenko I, Malmqvist L, Subhi Y, Hamann S. Artificial intelligence-based ChatGPT responses for patient questions on optic disc drusen. *Ophthalmol Therapy*. 2023;12:3109–19.
69. Hu X, Ran AR, Nguyen TX, Szeto S, Yam JC, Chan CKM, et al. What can GPT-4 do for diagnosing rare eye diseases? A pilot study. *Ophthalmol Therapy*. 2023;12:3395–402.
70. Anguita R, Makuloluwa A, Hind J, Wickham L. Large language models in vitreoretinal surgery. *Eye*. 2023. https://doi.org/10.1038/s41433-023-02751-1
71. Barclay KS, You JY, Coleman MJ, Mathews PM, Ray VL, Riaz KM, et al. Quality and agreement with scientific consensus of ChatGPT information regarding corneal transplantation and Fuchs dystrophy. *Cornea*. 2023;1–5. https://doi.org/10.1097/ICO.0000000000003439
72. Caranfa JT, Bommakanti NK, Young BK, Zhao PY. Accuracy of vitreoretinal disease information from an artificial intelligence chatbot. *JAMA Ophthalmol*. 2023;141:906–7.
73. Cardona G, Argiles M, Pérez-Mañá L. Accuracy of a large language model as a new tool for optometry education. *Clin Exp Optom*. 2023;1–4. https://doi.org/10.1080/08164622.2023.2288174
74. Eid K, Eid A, Wang D, Raiker RS, Chen S, Nguyen J. Optimizing ophthalmology patient education via ChatBot-generated materials: readability analysis of AI-generated patient education materials and the American Society of Ophthalmic Plastic and Reconstructive

Surgery Patient Brochures. *Ophthal Plast Reconstr Surg*. 2023;1–4. https://doi.org/10.1097/IOP.0000000000002549
75. Ferro Desideri L, Roth J, Zinkernagel M, Anguita R. Application and accuracy of artificial intelligence-derived large language models in patients with age related macular degeneration. *Int J Retina Vitr*. 2023;9:71. https://doi.org/10.1186/s40942-023-00511-7
76. Kianian R, Sun D, Crowell EL, Tsui E. The use of large language models to generate education materials about uveitis. *Ophthalmol Retina*. 2023;8:195–201.
77. Kianian R, Sun D, Giaconi J. Can ChatGPT aid clinicians in educating patients on the surgical management of glaucoma? *J Glaucoma*. 2023;33:94–100.
78. Nanji K, Yu CW, Wong TY, Sivaprasad S, Steel DH, Wykoff CC, et al. Evaluation of postoperative ophthalmology patient instructions from ChatGPT and Google search. *Can J Ophthalmol*. 2023;59:e69–e71.
79. Nikdel M, Ghadimi H, Tavakoli M, Suh DW. Assessment of the responses of the artificial intelligence-based chatbot ChatGPT-4 to frequently asked questions about amblyopia and childhood myopia. *J Pediatr Ophthalmol Strabismus*. 2023;1–4. https://doi.org/10.3928/01913913-20231005-02
80. Nunes BF, Reis JS, de Souza AD, Soares AN. Exploring the use of ChatGPT for counseling patients with diabetic retinopathy in regions with limited ophthalmic care. *InterSciencePlace*. 2023;18:102–11.
81. Pushpanathan K, Lim ZW, Er Yew SM, Chen DZ, Hui'En Lin HA, Lin Goh JH, et al. Popular large language model chatbots' accuracy, comprehensiveness, and self-awareness in answering ocular symptom queries. *iScience*. 2023;26:108163. https://doi.org/10.1016/j.isci.2023.108163
82. Rajabi MT, Rafizadeh SM, Ghahvehchian H. Exploring the use of ChatGPT in delivering evidence-based information to patients with thyroid eye disease. *Ophthal Plast Reconstr Surg*. 2023;40:113–5.
83. Solli EM, Tsui E, Mehta N. Analysis of ChatGPT responses to patient-oriented questions on common ophthalmic procedures. *Clin Exp Ophthalmol*. 2023;1–4. https://doi.org/10.1111/ceo.14334
84. Wei L, Mohammed ISK, Francomacaro S, Munir WM. Evaluating text-based generative artificial intelligence models for patient information regarding cataract surgery. *J Cataract Refract Surg*. 2024;50:95–6.
85. Hua HU, Kaakour AH, Rachitskaya A, Srivastava S, Sharma S, Mammo DA. Evaluation and comparison of ophthalmic scientific abstracts and references by current artificial intelligence chatbots. *JAMA Ophthalmol*. 2023;141:819–24.
86. Ali MJ, Singh S. ChatGPT and scientific abstract writing: pitfalls and caution. *Graefes Arch Clin Exp Ophthalmol*. 2023;261:3205–6.
87. Ćirković A, Katz T. Exploring the potential of ChatGPT-4 in predicting refractive surgery categorizations: comparative study. *JMIR Form Res*. 2023;7:e51798. https://doi.org/10.2196/51798
88. Oca MC, Meller L, Wilson K, Parikh AO, McCoy A, Chang J, et al. Bias and inaccuracy in AI chatbot ophthalmologist recommendations. *Cureus*. 2023;15:e45911. https://doi.org/10.7759/cureus.45911
89. Ong J, Kedia N, Harihar S, Vupparaboina SC, Singh SR, Venkatesh R, et al. Applying large language model artificial intelligence for retina International Classification of Diseases (ICD) coding. *J Med Artif Intell*. 2023;6:21. https://doi.org/10.21037/jmai-23-106
90. Raghu K, Tamilselvi S, Devishamani CS, Suchetha M, Rajalakshmi R, Raman R. The utility of ChatGPT in diabetic retinopathy risk assessment: a comparative study with clinical diagnosis. *Clin Ophthalmol*. 2023;17:4021–31.
91. Seth I, Bulloch G, Xie Y, Zhu Z. Exploring the potential of ChatGPT for advancing oculoplastic surgical research. *Ann Ophthalmol Vis Sci*. 2023;6:1038–41.
92. Singh S, Djalilian A, Ali MJ. ChatGPT and ophthalmology: exploring its potential with discharge summaries and operative notes. *Semin Ophthalmol*. 2023;38:503–7.
93. Waisberg E, Ong J, Masalkhi M, Kamran SA, Zaman N, Sarker P, et al. GPT-4 and ophthalmology operative notes. *Ann Biomed Eng*. 2023;51:2353–5.
94. Singh S, Watson S. ChatGPT as a tool for conducting literature review for dry eye disease. *Clin Exp Ophthalmol*. 2023;51:731–2.
95. Raja H, Munawar A, Delsoz M, Elahi M, Madadi Y, Hassan A, et al. Using large language models to automate category and trend analysis of scientific articles: an application in ophthalmology. *arXiv* preprint. 2023;2308.16688. https://doi.org/10.48550/arXiv.2308.16688
96. Valentín-Bravo FJ, Mateos-Álvarez E, Usategui-Martín R, Andrés-Iglesias C, Pastor-Jimeno JC, Pastor-Idoate S. Artificial intelligence and new language models in ophthalmology: complications of the use of silicone oil in vitreoretinal surgery. *Arch Soc Esp Oftalmol*. 2023;98:298–303.
97. Salimi A, Saheb H. Large language models in ophthalmology scientific writing: ethical considerations blurred lines or not at all? *Am J Ophthalmol*. 2023;254:177–81.
98. Teixeira da Silva JA. Is ChatGPT a valid author? *Nurse Educ Pract*. 2023;68:103600. https://doi.org/10.1016/j.nepr.2023.103600
99. Flanagin A, Bibbins-Domingo K, Berkwits M, Christiansen SL. Nonhuman "Authors" and implications for the integrity of scientific publication and medical knowledge. *JAMA*. 2023;329:637–9.
100. Aghemo A, Forner A, Valenti L. Should artificial intelligence-based language models be allowed in developing scientific manuscripts? A debate between ChatGPT and the editors of liver international. *Liver Int*. 2023;43:956–7.
101. Zielinski C, Winker M, Aggarwal R, Ferris L, Heinemann M, Lapeña JF, et al. Chatbots, ChatGPT, and scholarly manuscripts: WAME recommendations on ChatGPT and chatbots in relation to scholarly publications. *Natl Med J India*. 2023;36:1–4.
102. Curtis N, ChatGPT. To ChatGPT or not to ChatGPT? The impact of artificial intelligence on academic publishing. *Pediatr Infect Dis J*. 2023;42:275. https://doi.org/10.1097/INF.0000000000003852
103. Transformer CGP-t, Zhavoronkov A. Rapamycin in the context of Pascal's wager: generative pre-trained transformer perspective. *J Oncoscience*. 2022;9:82–4.
104. Transformer GGP, Thunström AO, Steingrimsson S. Can GPT-3 write an academic paper on itself, with minimal human input? In: HAL open science. Lyon: ML Research Press; 2022.
105. O'Connor S. Corrigendum to "Open artificial intelligence platforms in nursing education: tools for academic progress or abuse?". *Nurse Educ Pract*. 2023;66:103537. https://doi.org/10.1016/j.nepr.2022.103537
106. Shafeeg A, Shazhaev I, Mihaylov D, Tularov A, Shazhaev I. Voice assistant integrated with chat GPT. *J Indones J Comput Sci*. 2023;12:22–31.
107. Chen X, Xu P, Li Y, Zhang W, Song F, Zheng Y-F, et al. ChatFFA: interactive visual question answering on fundus fluorescein angiography image using ChatGPT. Available at SSRN 4578568. 2023. http://dx.doi.org/10.2139/ssrn.4578568
108. Wang SY, Huang J, Hwang H, Hu W, Tao S, Hernandez-Boussard T. Leveraging weak supervision to perform named entity recognition in electronic health records progress notes to identify the ophthalmology exam. *Int J Med Inform*. 2022;167:104864. https://doi.org/10.1016/j.ijmedinf.2022.104864
109. Arora A, Arora A. The promise of large language models in health care. *Lancet*. 2023;401:641. https://doi.org/10.1016/S0140-6736(23)00216-7
110. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. *NPJ Digit Med*. 2022;5:194. https://doi.org/10.1038/s41746-022-00742-2
111. Jo JJ, Cheng CP, Ying S, Chelnis JG. Physician review websites: understanding patient satisfaction with ophthalmologists using natural language processing. *J Ophthalmol*. 2023;2023:4762460. https://doi.org/10.1155/2023/4762460
112. Zhang W, Deng Y, Liu B, Pan SJ, Bing L. Sentiment analysis in the era of large language models: a reality check. *arXiv* preprint. 2023;2305.15005. https://doi.org/10.48550/arXiv.2305.15005

113. Tu T, Fang Z, Cheng Z, Spasic S, Palepu A, Stankovic K, et al. Genetic discovery enabled by a large language model. bioRxiv 2023:2023.11.09.566468. https://doi.org/10.1101/2023.11.09.566468
114. Mohapatra DP, Thiruvoth FM, Tripathy S, Rajan S, Vathulya M, Lakshmi P, et al. Leveraging large language models (LLM) for the plastic surgery resident training: do they have a role? *Indian J Plast Surg*. 2023;56:413–20.
115. Savage N. Drug discovery companies are customizing ChatGPT: here's how. *Nat Biotechnol*. 2023;41:585–6.
116. Chakraborty C, Bhattacharya M, Lee SS. Artificial intelligence enabled ChatGPT and large language models in drug target discovery, drug discovery, and development. *Mol Ther Nucleic Acids*. 2023;33:866–8.
117. Juhi A, Pipil N, Santra S, Mondal S, Behera JK, Mondal H. The capability of ChatGPT in predicting and explaining common drug–drug interactions. *Cureus*. 2023;15:e36272. https://doi.org/10.7759/cureus.36272
118. Liang Y, Zhang R, Zhang L, Xie P. DrugChat: towards enabling ChatGPT-like capabilities on drug molecule graphs. *TechRxiv*. 2023:1–14. https://doi.org/10.48550/arXiv.2309.03907
119. Zheng C, Luo Y, Mercado C, Sy L, Jacobsen SJ, Ackerson B, et al. Using natural language processing for identification of herpes zoster ophthalmicus cases to support population-based study. *Clin Exp Ophthalmol*. 2019;47:7–14.
120. Maganti N, Tan H, Niziol LM, Amin S, Hou A, Singh K, et al. Natural language processing to quantify microbial keratitis measurements. *Ophthalmology*. 2019;126:1722–4.
121. Gao W, Deng Z, Niu Z, Rong F, Chen C, Gong Z, et al. OphGLM: training an ophthalmology large language-and-vision assistant based on instructions and dialogue. *arXiv* Preprint. 2023:2306.12174. 2023. https://doi.org/10.48550/arXiv.2306.12174
122. Tan TF, Thirunavukarasu AJ, Campbell JP, Keane PA, Pasquale LR, Abramoff MD, et al. Generative artificial intelligence through ChatGPT and other large language models in ophthalmology: clinical applications and challenges. *Ophthalmol Sci*. 2023;3:100394. https://doi.org/10.1016/j.xops.2023.100394
123. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29:1930–40.
124. Shuster K, Xu J, Komeili M, Ju D, Smith EM, Roller S, et al. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv* preprint. 2022;2208.03188. https://doi.org/10.48550/arXiv.2208.03188
125. Glaese A, McAleese N, Trębacz M, Aslanides J, Firoiu V, Ewalds T, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv* preprint. 2022;2209.14375. https://doi.org/10.48550/arXiv.2209.14375
126. Taloni A, Scorcia V, Giannaccare G. Large language model advanced data analysis abuse to create a fake data set in medical research. *JAMA Ophthalmol*. 2023;141:1174–5.
127. Wang X, Wei J, Schuurmans D, Le Q, Chi E, Narang S, et al. Self-consistency improves chain of thought reasoning in language models. *arXiv* preprint. 2022;2203.11171. https://doi.org/10.48550/arXiv.2203.11171
128. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. *Nature*. 2023;616:259–65.
129. Lin Z, Trivedi S, Sun J. Generating with confidence: uncertainty quantification for black-box large language models. *arXiv* preprint. 2023;2305.19187. https://doi.org/10.48550/arXiv.2305.19187
130. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al., editors. Learning transferable visual models from natural language supervision. International conference on Machine learning. Maastricht: ML Research Press; 2021.
131. Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. *Lancet Digit Health*. 2023;5:e333–e335.
132. Sorin V, Brin D, Barash Y, Konen E, Charney A, Nadkarni G, et al. Large language models (LLMs) and empathy—a systematic review. 2023:2023.08.07.23293769. https://doi.org/10.1101/2023.08.07.23293769
133. Thirunavukarasu AJ. Large language models will not replace healthcare professionals: curbing popular fears and hype. *J R Soc Med*. 2023;116:181–2.
134. Dey N, Gosal G, Khachane H, Marshall W, Pathria R, Tom M, et al. Cerebras-GPT: open compute-optimal language models trained on the Cerebras wafer-scale cluster. *arXiv* preprint. 2023;2304.03208. https://doi.org/10.48550/arXiv.2304.03208

**How to cite this article:** Biswas S, Davies LN, Sheppard AL, Logan NS, Wolffsohn JS. Utility of artificial intelligence-based large language models in ophthalmic care. *Ophthalmic Physiol Opt*. 2024;00:1–31. https://doi.org/10.1111/opo.13284