AMERICAN ACADEMY OF OPHTHALMOLOGY®

Generative Artificial Intelligence Through ChatGPT and Other Large Language Models in Ophthalmology

Clinical Applications and Challenges

Ting Fang Tan, MBBS,[1] Arun James Thirunavukarasu, BA MB, BChir (Cantab),[2,3] J. Peter Campbell, MD, MPH,[4] Pearse A. Keane, MD, FRCOphth,[5] Louis R. Pasquale, MD, FARVO,[6] Michael D. Abramoff, MD, PhD,[7,8,9] Jayashree Kalpathy-Cramer, PhD,[10] Flora Lum, MD,[11] Judy E. Kim, MD, FARVO,[12] Sally L. Baxter, MD, MSc,[13,14] Daniel Shu Wei Ting, MD, PhD[1,15]

The rapid progress of large language models (LLMs) driving generative artificial intelligence applications heralds the potential of opportunities in health care. We conducted a review up to April 2023 on Google Scholar, Embase, MEDLINE, and Scopus using the following terms: "large language models," "generative artificial intelligence," "ophthalmology," "ChatGPT," and "eye," based on relevance to this review. From a clinical viewpoint specific to ophthalmologists, we explore from the different stakeholders' perspectives—including patients, physicians, and policymakers—the potential LLM applications in education, research, and clinical domains specific to ophthalmology. We also highlight the foreseeable challenges of LLM implementation into clinical practice, including the concerns of accuracy, interpretability, perpetuating bias, and data security. As LLMs continue to mature, it is essential for stakeholders to jointly establish standards for best practices to safeguard patient safety.

Financial Disclosure(s): Proprietary or commercial disclosure may be found in the Footnotes and Disclosures at the end of this article. *Ophthalmology Science 2023;3:100394*

The recent hype on large language models (LLMs) has been driven by their capability to leverage deep learning neural networks to learn complex associations between unstructured texts and use these learned patterns to produce useful outputs in response to custom text queries.[1] Generative artificial intelligence (AI) chatbots built with these LLMs facilitate a realistic and interactive user experience through text-based dialogue, which is different from all prior AI applications that have been predominantly single-task based (e.g., classification, segmentation, or prediction) with limited human-AI interaction.[1–3] One of such LLMs would be ChatGPT, built on its backend LLM Generative Pretrained Transformer (GPT)-3.5. Now, GPT-4 has generated great excitement owing to its performance in cognitive tasks including medical problem-solving.[4–7]

Targeted at ophthalmologists, we aim to deepen the understanding of LLMs and their potential opportunities and challenges specific to the field of ophthalmology. We first provide an overview of the development of these LLMs. We then explore potential educational, research, and clinical applications from the different stakeholders' perspectives specific to ophthalmology. Finally, we highlight the challenges of LLM implementation into clinical practice.

Development of LLMs: Evolution of GPT 1 to 4

The rapid development of LLMs is illustrated by considering the evolution of GPT-based models (Table 1).

Generative Pretraining Transformer-1 was first released in 2018. It was engineered through semisupervised training: initial unsupervised language modeling on the BookCorpus dataset with 11 308 books containing 1×10^9 words, followed by supervised fine-tuning to improve performance. Generative Pretraining Transformer-1 achieved decent zero-shot (i.e., no examples of the specified task provided in the input) performance, outperforming bespoke models in most natural language processing (NLP) tasks.[8] Generative Pretraining Transformer-2 was released a year later and trained on 10 times more data from WebText data: over 8 million documents.[9] In addition to its superior performance in general NLP tasks, its performance was maintained even in previously unseen tasks, especially when enhanced with prompting strategies (as described below).

The following year, its successor, GPT-3, was released with 100 times more parameters than GPT-2 and was pretrained with 5 corpora (CommonCrawl, WebText2, Books1, Books2, and Wikipedia), unlocking even higher

https://doi.org/10.1016/j.xops.2023.100394
ISSN 2666-9145/23

performances. Subsequently, GPT-3.5 was developed through fine-tuning of GPT-3 using human-generated text-based input-output pairs, reinforcement learning from human feedback, and further autonomous adversarial training. Reinforcement learning from human feedback involves a reward model trained on human ranking of GPT-3.5-generated outputs, facilitating autonomous reinforcement learning of the LLM based on human feedback.[4] It is important to understand that, fundamentally, the objective function for these (text-based) models is a proxy for linguistic fit and not for objective correctness, which may not be present in the data on which it is trained.

In 2023, GPT-4 has been released.[10] Though model architecture and training datasets remain confidential at the time of writing, GPT-4 incorporates added features inclusive of accommodating multimodal input data types such as images (whereas previous GPT models were limited to only text-based input). Generative Pretraining Transformer-4 outperformed other LLMs with human-level accuracy in professional examinations, which was maintained even in other languages like Welsh and Swahili. Based on human-grading feedback, GPT-4 was found to generate responses that were better aligned with user intent compared to GPT-3.5.[10]

Other generative AI chatbots built on similar LLMs include BlenderBot 3, which uses Open Pretrained Transformer as its backend LLM, and Bard, built on backend LLM Pathways Language Model 2; these also have real-time access to the internet to improve the accuracy and recency of responses. Bing's AI chatbot enables access to a version of GPT-4 without a premium subscription to ChatGPT.[11–13]

Developing LLM Applications for Ophthalmology

In addition to general NLP tasks, foundation LLMs have shown promising results in generalizing to unseen tasks even in medical question-answering requiring scientific expert knowledge.[14–18] These tasks require LLMs to understand the medical context, recall, and interpret relevant medical information in order to formulate an answer. Reported performance in ophthalmology has been mixed, but there appears to be potential to apply LLMs in eye health care applications if important limitations can be addressed.[14–18] Various strategies have been described to develop foundation LLMs with enhanced performance in clinical tasks. These include building domain-specific LLMs by pretraining with curated medical text, fine-tuning foundation LLMs with domain-specific medical data, or using innovative prompting strategies.[14,19–21]

As size is a critical component for LLMs exhibiting useful properties, the very limited set of biomedical data makes domain-specific pretraining a difficult challenge.[22] Improved availability of data from electronic patient records, paper-based documentation, and the scientific literature entails overcoming issues of privacy and copyright which may not be feasible for medicine as a whole, let alone individual specialties such as ophthalmology. However, various LLMs have been fine-tuned using curated medical and scientific text, with examples including Med-Pathways Language Model 2, Sci-Bidirectional Encoder Representations from Transformers (BERT), BioBERT, PubMedBERT, Data Augmented Relation Extraction (DARE), ScholarBERT, ClinicalBERT, and BioWordVec.[23–28] These domain-specific LLMs have outperformed foundation LLMs in biomedical NLP tasks.[23–26,29,30] Using available models, prompting strategies requiring minimal computational and economic investment may be used to improve domain-specific performance; these include chain-of-thought (CoT) prompting, where the model is told to provide step-by-step reasoning in deriving a final answer, which may be few-shot (exemplar input-output pairs provided) or zero-shot (no examples provided), and retrieval augmentation, where additional domain-specific context is provided with user requests.[14,31–33] These contextual learning strategies appear to operate via similar mechanisms to domain-specific fine-tuning at a larger scale.[34]

Stakeholders' Perspectives of LLM Integration into Eye Care

Although NLP has been explored in ophthalmology, applications of LLM technology are relatively nascent.[35,36] However, proof-of-concept experiments, validation studies, and directed development have begun to accelerate. While exciting and having the potential to benefit patient and population outcomes, as well as other health care stakeholders (Figure 1), there is currently little evidence for the

Table 1. The Evolution of GPT 1 to 4 and Its Associated Features

	GPT-1	GPT-2	GPT-3	GPT-4
Dataset	1 dataset: BookCorpus (11 380 novels, 1×10^9 words)	1 dataset: Web text (40GB of data, 8 million documents)	5 datasets: CommonCrawl, WebText2, Books1, Books2, Wikipedia (45TB of data)	*
Model architecture	12 layers with 12 attention heads in each self-attention layer	48 layers with 1600 dimensional vectors for word embedding	96 layers with 96 attention heads	*
Parameters	115 million	1.5 billion	175 billion	*

GB = gigabyte; GPT = Generative Pretraining Transformer; TB = terabyte.
*GPT-4 model architecture, pretraining data, and fine-tuning protocols were confidential at the time of writing.

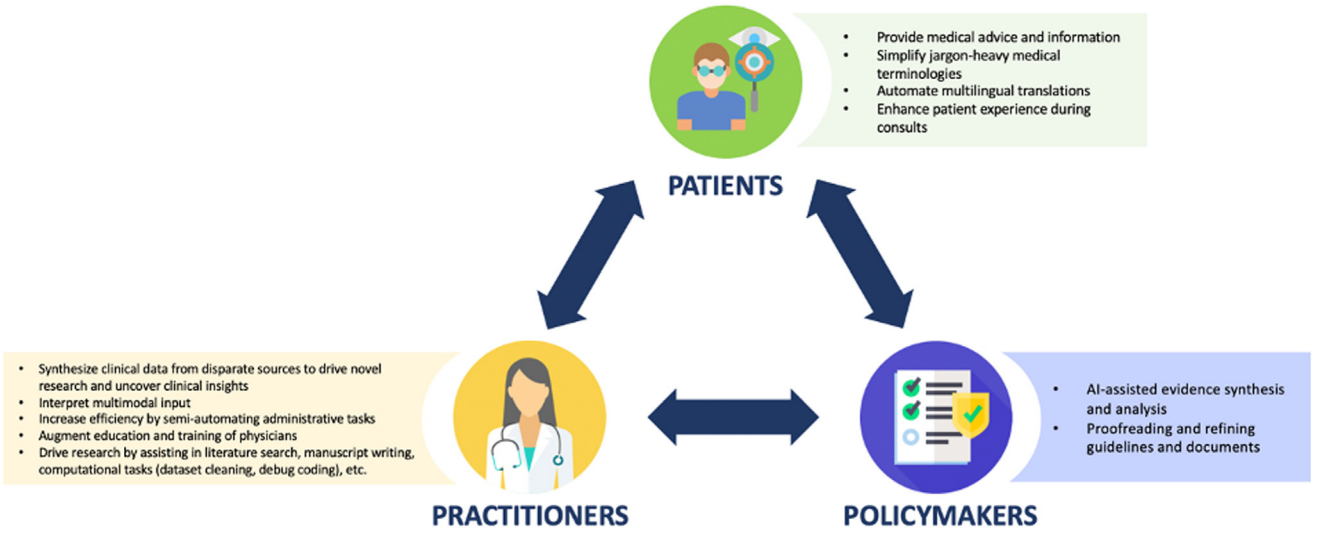

Figure 1. Integration of large language models into eye care from stakeholders' perspectives: patients, practitioners, policymakers. AI = artificial intelligence.

safety, efficacy, ethics, and equity of such LLM applications.

The Patient Perspective

Large language model chatbots provide lucid responses to user queries, and patients may use these platforms to obtain medical advice and information. Accuracy may improve by providing LLM platforms with access to real-time information from the internet rather than relying on its nonspecific pretraining corpora and fine-tuning; Google Bard and Bing AI already have this functionality, and ChatGPT is set to follow as it enables plug-in functionality and releases an application programming interface.[13,37–39] Application programming interface access may be especially helpful for developers looking to engineer applications with narrow use cases, such as to provide medical advice to patients. Many patients already self-diagnose using the internet without ever consulting a physician, with consistent search engine activity related to eye disease reported over time.[40] This may have significant benefits, strengthening patient autonomy and even contributing to successful diagnosis.[41–43] However, the safety, presence or lack of bias, and ethical dimensions have not been established, and, thus, there is a risk of patient harm at a large scale, given the potential widespread adoption of such algorithms. Indeed, inaccuracies and "fact fabrication"— often termed hallucination by computer scientists and journalists—where invented, inaccurate statements are presented as lucidly as accurate information—raise a concern that users will be misled and suffer avoidable harm. Until these applications are properly engineered and validated in appropriate settings, they cannot be recommended by clinicians.

As LLM technology is integrated into clinical workflows, patients will be treated by a combination of AI and clinician. While AI applications will likely be subordinate tools used by ophthalmologists, nonhuman contributions to communication and decision-making are significant changes.[44] Change may be positive, as LLM outputs were superior in terms of quality and empathy to doctors replying to medical queries on a social media message board in one study and generally superior to doctors responding to a developer-generated list of patient questions when compared along 9 qualitative parameters in another study.[30,45] Implementation may improve the patient experience; by adopting tools that increase efficiency—particularly in documentation and other administrative tasks—clinicians could have more time to engage with their patients both through conversation and hands-on procedures.[46] This helps facilitate truly patient-centered care, an understudied but important way in which ophthalmology services may be improved, although similar expectations were made for electronic health records, which so far have not materialized.[47–49] However, as patients struggle to differentiate between AI and human text, care must be taken to safeguard them from harm and avoid compromising trust in health care institutions and professionals.[50] It is ophthalmologists' responsibility to ensure that changes to health care systems do not compromise quality of care.[51,52]

The Practitioner Perspective

Multimodal LLMs capable of processing images and text are emerging, with important implications for eye care, which relies heavily on large quantities of nontext-based data.[53] Large language models have already demonstrated that they may encode sufficient information to assist with eye care, and further development and fine-tuning will see this potential improve.[14,17,18] Moreover, the success of deep learning models used to analyze ophthalmic investigations—fundus photographs, OCT, visual fields, and more—suggests that multimodal LLMs will perform to a high standard in this context too.[36] Ophthalmologists may expect LLM applications to rapidly assimilate data

from disparate sources, including clinic notes, correspondence, and investigation results. Validated models may assist with the interpretation of this data and subsequent decision-making. Early examples in general medicine include Foresight, developed by fine-tuning a GPT model with data from approximately 1 million patients' health records.[54] Foresight shows how LLMs could be used as a general risk calculator to triage patients or as a decision aid by facilitating counterfactual simulation of alternative management plans.[54] Other fine-tuned LLMs (BioBERT, BlueBERT, DistilBERT, and ClinicalBERT) exhibited good performance in identifying ophthalmic examinations listed in clinical notes, illustrating the potential of using LLMs to quickly identify and assimilate relevant information from large patient records which would otherwise be daunting.[55] Development of successful tools for ophthalmology may require large quantities of data to fine-tune foundation LLMs, but general medical sources—such as electronic patient records or medical scientific literature—may be sufficient to attain acceptable performance in an eye health context.

Before sophisticated clinical AI assistants are developed and validated, LLM applications may nevertheless have a great impact on clinical practice. Models may already be used as tools provided that ophthalmologists retain responsibility for their patients, and performance is greatest (and most useful) where specialist knowledge is either not required or provided by the user. Large language models can be used to improve the efficiency of administrative work by helping to write letters and notes by accelerating data synthesis and optimizing language on demand.[56] For more straightforward patient queries that nonetheless require consideration of other information (e.g., appointment rescheduling, medication refill requests), responses may be automated using LLMs. As with other clinical applications, clinical utility will increase with multiple modalities. Future models may act as automatic scribes, using transcriptions produced with voice recognition to generate appropriate clinical notes and letters, as well as assisting decision-making. In general, automating cognitive labor should provide ophthalmologists with more time to attend to their patients, which could improve patient and practitioner satisfaction with health care.[57,58]

Large language models may contribute to education in the broadest sense. Ophthalmologists could use LLMs to help explain diagnosis, management, and prognosis to patients and may simultaneously save time and improve communication by providing comprehensive information and tasking an LLM with responding to patient queries autonomously. In addition to simplifying jargon-heavy medical terminology, automating multilingual translation of patient education materials can lower the barrier to accessing information for multiethnic communities. As with clinical applications above, validation, governance, and safeguarding are essential, and ophthalmologists could monitor conversations to mitigate any misunderstandings or inaccuracies.

In addition, LLMs may be used to augment education of doctors. Here, confidence in model outputs is the key to avoid perpetuating misconceptions and inaccurate knowledge. Incremental progress suggests that more basic education will become feasible first, progressing to more advanced teaching as technology improves. The most basic level of ophthalmic training is at medical school, and LLMs may already be appropriate teaching-aids at this level.[7] The next step is ophthalmic teaching for nonspecialists such as general practitioners, and LLMs already exhibit good aptitude in this domain.[7] Large language models currently exhibit greater error rates in response to questions aimed at specialist ophthalmologists, but the significant improvement of ChatGPT using GPT-4 rather than GPT-3.5 suggests that subsequent improvement (facilitating deployment to aid specialist training) is likely.[14,17,18]

Finally, LLMs may contribute to research. Already available models such as GPT-4 are able to improve the quality of text produced for publication.[59] Because LLMs excel in tasks where specialist knowledge is not required or is not provided, other use cases include automatic summarization and synthesis of articles and rewriting and reformatting information for specific purposes such as preparing abstracts for publication or presentation, briefs for the media, or layperson explanations for public engagement. Models fine-tuned with biomedical text, such as BioBERT, Med-Pathways Language Model 2, and PubMedBERT, are likely to perform well in these use cases.[23,24,30] These models may help with the initial writing of perspective pieces and original articles, provided that inputs are carefully curated, outputs are validated to avoid mistakes and plagiarism, and model use is openly disclosed.[34] Authors for the foreseeable future will be responsible for their output, regardless of how much assistance is provided by LLM applications.[34]

Large language models may also assist with primary research. Computational ophthalmology work will be enhanced with LLM coding assistants which will semi-automate development, for example, to streamline data cleaning and debug coding.[60,61] Large language models' performance in NLP suits them to new types of research at unprecedented scale using clinical text data. The scalability of LLMs such as ClinicalBERT, GPT, and GatorTron makes the availability of high-quality data the limiting factor.[62–64] Targeted efforts are indicated to curate validated sources of clinical text data: progress notes, investigation reports, referrals, and other letters. This will require collaboration and a commitment to openness to make data available to researchers around the world. Finally, LLMs may assist with nonlanguage-based research, as text data are used to represent other forms of information. AlphaFold represents an example with its ability to deduce protein structures from amino acid sequences represented as text.[65] Other models are emerging for protein and genetic analysis, and potential applications in ophthalmology are diverse: drug development, genetic diagnosis, physiological and pathological research, and more.[66,67]

The Policymaker Perspective

While published trials are beginning to demonstrate the potential of LLMs in medicine, no trials have demonstrated that new applications are safe and effective. Certain use cases may not require a clinical trial to justify adoption, such

as supervised assistance with administrative tasks, though current and proposed regulations in the United States may result in civil rights issues from unconsidered use of such applications.[68] Stakeholders are called upon to ensure new applications are built under an ethical framework, and standards of evidence to justify deployment of more clinical applications must not be compromised.[44]

As with practitioners, LLMs may improve the efficiency and quality of work done by policymakers through implementation of AI-assisted writing, evidence synthesis, and administrative work. General LLMs exhibit promising potential, particularly when integrated with other platforms providing material that requires processing or analysis and enriched with application programming interface "tools" as described earlier from the patient perspective.[37,69,70] There are few documented examples of use in ophthalmology, but LLMs may now feasibly assist with drafting, writing, refining, and proofreading guidelines, regulations, and other documents. The expansion of LLMs' capacity in terms of inputs and outputs increases potential, as does multimodality; GPT-4 accepts or produces up to 25 000 tokens and images compared to 3000 tokens with GPT-3.5 (1 token roughly corresponds to 1 word)—while these limits may currently preclude tasks requiring use of a patient's entire health record, capacity is growing.[10]

Policymakers must contend with a rapidly changing landscape to ensure that innovation works for the benefit of society. This entails overcoming a set of ethical, legal, and safety issues which are discussed at greater length in the following section.

Challenges Impeding Implementation of LLMs

Despite its promising possibilities, there are several challenges of existing LLM applications that limit their maturity for clinical deployment.

First, cautions against ChatGPT and similar applications have been attributed to the lack of accuracy and coherence in its generated responses. Potentially even more concerning for potential patient and population harm is that responses may contain fact fabrication, including made-up but nonexistent peer-reviewed scientific references.[71] Another example would be the trivial guessing heuristics observed in InstructGPT (from backend LLM GPT-3.5) where it often selected choices A and D in multiple-choice question-answering tasks. Closer inspection of the generated CoT explanations showed that this behavior surfaced frequently when the models were not able to answer the question.[33] Poorer performance is observed in tasks that require highly specialized domain-specific knowledge, such as ophthalmology specialty examinations.[17] This is further jeopardized by "falsehood mimicry" observed on occasions when the user input lacked clarity or accuracy, where ChatGPT generated responses to fit the user's incorrect assumptions instead of clarifying the user's intent.[72] Therefore, it is important to build LLM applications that acknowledge doubt and uncertainty rather than outputting unmitigated erroneous responses.[44] This has previously been incorporated in deep learning models, such as by training to flag uncertain cases as "indeterminate" rather than making spurious predictions.[73]

Second, besides Google Bard and Bing AI, many LLM applications do not have real-time internet access. ChatGPT, for example, is trained on data prior to late 2021. This is an important issue, particularly in the health care domain where new breakthroughs and updates in clinical guidelines are constantly evolving. For example, the management of geographic atrophy, a progressive and irreversible blinding retinal condition, has been predominantly limited to low-vision rehabilitation with no approved drug therapies. However, the drug syfovre, a complement inhibitor delivered intravitreally to slow geographic atrophy lesion growth, has been recently approved by the Food and Drug Administration in the United States on February 17, 2023.[74] As a result, patients may be misinformed by medical information that is not up-to-date. More importantly, because these applications are not intended to be deterministic and essentially be "continuously learning," there is currently no framework for determining safety and accuracy, even when established for a previous version.[55]

Third, the "black box" nature of LLMs renders the decision-making process opaque.[75] Unless explicitly asked, generated responses do not contain supporting citations or information sources. This lack of interpretability is compounded by the above observations of fabricated and inaccurate yet plausible-sounding responses. This limits the credibility of generated responses and may be detrimental in the health care domain where patients may be misled by inaccurate medical advice. Possible solutions to enhance interpretability include the use of CoT prompting (an example of a CoT prompt would be "outline a differential diagnosis corresponding to this patient's symptoms using step-by-step reasoning like an expert ophthalmologist") to prompt chatbots to include its reasoning process in addition to the final answer. Human expert annotation of these LLM-generated CoT explanations for medical question-answering tasks revealed that the majority had sound reasoning, thought processes, recall of knowledge, and comprehension of the question and context.[33] Potential additional features that can be explored include uncertainty-aware LLM applications that provide a probability score of generated responses, along with alternative recommendations when the probability score is low, as well as reporting the differential weights of input tokens that contributed to the generated answer.[76]

Fourth, another limitation of LLMs lies in mirroring the biases that exist in the data they are trained on. Unstructured data such as fundus photographs have been shown to encode factors such as age, sex, and race which could result in LLMs reaching conclusions based on inappropriate assumptions which could perpetuate bias or drive inaccuracy.[77] Therefore, LLMs exhibit a risk of perpetuating socioeconomic stereotypes and negative generalizations against minorities in ethnicity, religion, and gender.[78,79] Other risks to patient safety may arise when LLM applications are misused to spread misinformation or extract confidential patient information. For instance, they can craft unique variants of the same phishing lure and

effectively bypass safety filters that detect possible scams based on identical text sequences. The generated phishing content is also more grammatically accurate and convincing, making it harder to detect. Moreover, in combination with additional tools like text-to-speech, these phishing attempts can potentially take the form of voice calls to imitate realistic and coherent human-like conversation to exploit users.[80] Despite in-built safety nets designed by ChatGPT to mitigate these risks, countermeasures have been devised such as adversarial prompts to exploit ChatGPT to evade these safety features.[81,82]

Further, there is growing concern regarding the security of data inputted into LLM applications such as copyrighted material retained as part of training data, as well as the fact that applications like ChatGPT retains users' conversation content to improve its model performance.[83] For example, employees from Samsung Semiconductor, Inc were sternly warned for using ChatGPT to debug the company's program source code and summarize internal meeting minutes, as highly sensitive company information may be inadvertently disclosed.[84] Also recently, ChatGPT was taken offline temporarily for a bug that resulted in confidential personal information (including payment address, email address, and last 4 digits of a credit card number) and chat history being visible to other active users.[85] Even though OpenAI reassured that it has since rectified the error and established specific actions including system checks to minimize recurrence and introduced an option not to share user conversations with the company, these incidents reinforce the risks to data security. An alternative approach would be to deploy local LLMs for use within clinical centers, but this would entail significant cost (for hardware, software development, and maintenance), difficulty updating decentralized models with new information, and lack of access to state-of-the-art models (currently superior to open-source alternatives) run by for-profit companies.

Finally, because medical records have legal status, the generation, interpretation, and dissemination of such documents without human oversight needs legal analysis and jurisprudence. Regulatory frameworks must be developed to explore how to allocate responsibility for mistakes before issues arise; this is difficult before use cases are decided but necessary to safeguard patients. It seems likely that ophthalmologists will retain complete responsibility for their patients, with LLM applications incorporated as tools under close oversight. As capabilities continue to develop, this issue may have to be revisited accordingly.

Conclusion

The emergence of high-performance LLMs has great potential in ophthalmology through clinical, educational, and research applications. However, caution about deployment in clinical practice is essential as safety, effectiveness, and ethical considerations remain controversial and open areas of enquiry and research. As LLMs continue to mature, it is crucial for all stakeholders to be involved in efforts to establish standards for best practices to promote accuracy, ethical application, and safety—safeguarding patients and striving to improve the quality of eye health care provision.

Acknowledgments

The authors would like to acknowledge the rest of the American Academy of Ophthalmology Committee on Artificial Intelligence including Rishi Singh.

Footnotes and Disclosures

Originally received: April 26, 2023.
Final revision: August 7, 2023.
Accepted: August 30, 2023.
Available online: September 9, 2023. Manuscript no. XOPS-D-23-00082.

[1] Singapore Eye Research Institute, Singapore National Eye Centre, Singapore.

[2] University of Cambridge School of Clinical Medicine, Cambridge, United Kingdom.

[3] Corpus Christi College, University of Cambridge, Cambridge, United Kingdom.

[4] Department of Ophthalmology, Casey Eye Institute, Oregon Health and Science University, Portland, Oregon.

[5] Moorfields Eye Hospital, University of College London, London, United Kingdom.

[6] Department of Ophthalmology, Icahn School of Medicine at Mount Sinai, New York City, New York.

[7] American Medical Association's Digital Medicine Payment Advisory Group (DMPAG) Artificial Intelligence Workgroup, American Medical Association, Chicago, Illinois.

[8] Department of Ophthalmology, University of Iowa, Iowa City, Iowa.

[9] Digital Diagnostics, Inc, Coralville, Iowa.

[10] Department of Ophthalmology, University of Colorado Anschutz Medical Campus, Aurora, Colorado.

[11] American Academy of Ophthalmology, San Francisco, California.

[12] Department of Ophthalmology, Medical College of Wisconsin, Milwaukee, Wisconsin.

[13] Division of Ophthalmology Informatics and Data Science, Viterbi Family Department of Ophthalmology and Shiley Eye Institute, La Jolla, California.

[14] Health Department of Biomedical Informatics, University of California San Diego, La Jolla, California.

[15] Byers Eye Institute, Stanford University, Stanford, California.

Disclosures:

All authors have completed and submitted the ICMJE disclosures form.

The authors have made the following disclosures: J.P.C.: Supported — grants R01 EY019474, R01 EY031331, and P30 EY10572 from the National Institutes of Health (Bethesda, MD), unrestricted departmental funding and a Career Development Award — Research to Prevent Blindness (New York, New York); Research support — Genentech (San Francisco, California); Consultant — Boston AI Lab (Boston, Massachusetts); Equity owner — Siloam Vision. P.A.K.: Consultant — Google, DeepMind, Roche, Novartis, Apellis, BitFount; Equity owner — Big Picture Medical; Speaker fees — Heidelberg Engineering, Topcon, Allergan, Bayer; Support — Moorfields Eye Charity Career Development Award (R190028A), UK

Research & Innovation Future Leaders Fellowship (MR/T019050/1). L.R.P.: Consultant − Twenty-Twenty, Character Bio; Grant support − National Eye Institute (NEI), Research to Prevent Blindness (RPB), The Glaucoma Foundation (New York). M.D.A.: Investor, director, and consultant − Digital Diagnostics Inc, Coralville, Iowa; Patents and patent applications assigned to the University of Iowa and Digital Diagnostics that are relevant to the subject matter of this manuscript; Chair of Healthcare − AI Coalition, Washington DC, Foundational Principles of AI CCOI Workgroup; Member of the American Academy of Ophthalmology (Academy) Committee on Artificial Intelligence, AI Workgroup Digital Medicine Payment Advisory Group (DMPAG), Collaborative Community for Ophthalmic Imaging (CCOI), Washington DC. S.L.B.: Consulting fees −VoxelCloud; Speaking fees − iVista Medical Education; Equipment support − Optomed, Topcon. D.S.W.T.: Patent − a deep-learning system for the detection of retinal diseases; Supported by grants − National Medical Research Council, Singapore, (NMRC/HSRG/0087/2018; MOH-000655-00; MOH-001014-00), Duke-NUS Medical School, Singapore, (Duke-NUS/RSF/2021/0018; 05/FY2020/EX/15-A58), Agency for Science, Technology and Research, Singapore, (A20H4g2141; H20C6a0032), for research in artificial intelligence.

Daniel Shu Wei Ting, an editor of this journal, was recused from the peer-review process of this article and has no access to information regarding its peer-review.

HUMAN SUBJECTS: No human subjects were included in this study. This review study did not require institutional review board approval.

Author Contributions:

Conception and design: Ting

Analysis and interpretation: N/A

Data collection: Tan, Thirunavukarasu; Obtained funding: N/A, Study was performed as part of regular employment duties at the Singapore National Eye Center. No additional funding was provided.; Overall responsibility: Tan, Thirunavukarasu, Campbell, Keane, Pasquale, Abramof, Kalpathy-Cramer, Kim, Baxter, Ting

Abbreviations and Acronyms:
AI = artifical intelligence; **CoT** = chain-of-thought; **GPT** = Generative Pretrained Transformer; **LLM** = large language model; **NLP** = natural language processing.

Keywords:
Artificial intelligence, Chatbots, ChatGPT, Large language models.

Correspondence:
Daniel Shu Wei Ting, MD (1st Hons), PhD, Duke-NUS Medical School, AI and Digital Innovation, Singapore Eye Research Institute, Singapore Eye Research Institute (SERI), The Academia, 20 College Road, Level 6 Discovery Tower, Singapore 169856. E-mail: daniel.ting@duke-nus.edu.sg; dting45@stanford.edu.

References

1. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Computation and language. http://arxiv.org/abs/2212.13138; 2022. Accessed April 1, 2023.
2. Aggarwal R, Sounderajah V, Martin G, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ Digit Med*. 2021;4:65.
3. Yim J, Chopra R, Spitz T, et al. Predicting conversion to wet age-related macular degeneration using deep learning. *Nat Med*. 2020;26:892−899.
4. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. http://arxiv.org/abs/2203.02155; 2022. Accessed April 1, 2023.
5. Nori H, King N, McKinney SM, et al. Capabilities of GPT-4 on medical challenge problems. http://arxiv.org/abs/2303.13375; 2023. Accessed April 1, 2023.
6. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit Health*. 2023;2:e0000198.
7. Thirunavukarasu AJ, Hassan R, Mahmood S, et al. Trialling a large language model (ChatGPT) in general practice with the applied knowledge test: demonstrating opportunities and limitations in primary care. *JMIR Med Educ*. 2023;9:e46599.
8. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. *Comput Sci*; 2018. Available at: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.
9. Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners. *Comput Sci*. 2019;1(8):9.
10. OpenAI. GPT-4 technical report. https://cdn.openai.com/papers/gpt-4.pdf; 2023. Accessed March 17, 2023.
11. Microsoft Bing. Confirmed: the new Bing runs on OpenAI's GPT-4. https://blogs.bing.com/search/march_2023/Confirmed-the-new-Bing-runs-on-OpenAI's-GPT-4/; 2023. Accessed April 1, 2023.
12. Shuster K, Xu J, Komeili M, et al. BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage. https://doi.org/10.48550/arXiv.2208.03188; 2022. Accessed April 1, 2023.
13. Pichai S. Google AI updates: Bard and new AI features in search. The Keyword. https://blog.google/technology/ai/bard-google-ai-search-updates/; 2023. Accessed March 17, 2023.
14. Thirunavukarasu AJ. ChatGPT cannot pass FRCOphth examinations: implications for ophthalmology and large language model artificial intelligence. Eye News. https://www.eyenews.uk.com/features/ophthalmology/post/chatgpt-cannot-pass-frcophth-examinations-implications-for-ophthalmology-and-large-language-model-artificial-intelligence. Accessed June 1, 2023.
15. Raimondi R, Tzoumas N, Salisbury T, et al. Comparative analysis of large language models in the Royal College of Ophthalmologists fellowship exams. *Eye*. 2023. https://doi.org/10.1038/s41433-023-02563-3.
16. Lin JC, Younessi DN, Kurapati SS, et al. Comparison of GPT-3.5, GPT-4, and human user performance on a practice ophthalmology written examination. *Eye*. 2023. https://doi.org/10.1038/s41433-023-02564-2.
17. Antaki F, Touma S, Milad D, et al. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. *Ophthalmol Sci*. 2023;3:100324.
18. Teebagy S, Colwell L, Wood E, et al. Improved performance of ChatGPT-4 on the OKAP exam: a comparative study with ChatGPT-3.5. medRxiv [Internet]. https://www.medrxiv.org/content/10.1101/2023.04.03.23287957v1; 2023. Accessed April 23, 2023.
19. Taylor R, Kardas M, Cucurull G, et al. Galactica: a large language model for science. http://arxiv.org/abs/2211.09085; 2022. Accessed April 1, 2023.
20. Lehman E, Hernandez E, Mahajan D, et al. Do we still need clinical language models?. http://arxiv.org/abs/2302.08091; 2023. Accessed April 1, 2023.

21. Lester B, Al-Rfou R, Constant N. *The power of scale for parameter-efficient prompt tuning.arXiv preprint arXiv:2104.08691*. 2021 Apr 18.
22. Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models. arXiv [internet]. http://arxiv.org/abs/2001.08361; 2020. Accessed March 3, 2023.
23. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36:1234–1240.
24. Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*. 2021;3(1):1–23.2007.
25. Papanikolaou Y, Pierleoni A. DARE: data augmented relation extraction with GPT-2. *arXiv preprint arXiv:2004.13845*. 2020 Apr 6.
26. Hong Z, Ajith A, Pauloski G, et al. ScholarBERT: bigger is not always better. https://doi.org/10.48550/arXiv.2205.11342; 2022. Accessed April 1, 2023.
27. Zhang Y, Chen Q, Yang Z, et al. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Sci Data*. 2019;10:52.
28. Alsentzer E, Murphy JR, Boag W, et al. Publicly available clinical BERT embeddings. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics; 2019:72–78. Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323. 2019 Apr 6.
29. Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019:3615–3620. arXiv preprint arXiv:1903.10676. 2019 Mar 26.
30. Singhal K, Tu T, Gottweis J, et al. Towards expert-level medical question answering with large language models. arXiv [internet]. http://arxiv.org/abs/2305.09617; 2023. Accessed May 22, 2023.
31. Wei J, Wang XZ, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. https://doi.org/10.48550/arXiv.2201.11903; 2023. Accessed April 1, 2023.
32. Kojima T, Gu SS, Reid M, et al. Large language models are zero-shot reasoners. https://doi.org/10.48550/arXiv.2205.11916; 2023. Accessed April 1, 2023.
33. Liévin V, Hother CE, Winther O. Can large language models reason about medical questions?. https://doi.org/10.48550/arXiv.2207.08143; 2023. Accessed April 1, 2023.
34. Dai D, Sun Y, Dong L, et al. Why can GPT learn in-context? Language models secretly perform gradient descent as meta-optimizers. arXiv [internet]. http://arxiv.org/abs/2212.10559; 2022. Accessed March 27, 2023.
35. Chen JS, Baxter SL. Applications of natural language processing in ophthalmology: present and future. *Front Med (Lausanne)*. 2022;9:906554.
36. Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol*. 2019;103(2):167–175.
37. OpenAI. *Introducing ChatGPT and whisper APIs*; 2023. https://openai.com/blog/introducing-chatgpt-and-whisper-apis. Accessed April 1, 2023.
38. Microsoft Bing. Introducing the new bing. https://www.bing.com/new; 2023. Accessed April 1, 2023.
39. OpenAI. *ChatGPT plugins*; 2023. https://openai.com/blog/chatgpt-plugins. Accessed April 1, 2023.
40. Thirunavukarasu AJ. Evaluating the mainstream impact of ophthalmological research with Google Trends. *Eye*. 2021;35(11):3165–3167.
41. Van Riel N, Auwerx K, Debbaut P, et al. The effect of Dr Google on doctor-patient encounters in primary care: a quantitative, observational, cross-sectional study. *BJGP Open*. 2017;1:bjgpopen17X100833.
42. Pitt MB, Hendrickson MA. Providing a second opinion to Dr. Google with the WWW framework. *J Gen Intern Med*. 2022;37:222–224.
43. Kuehn BM. More than one-third of US individuals use the Internet to self-diagnose. *JAMA*. 2013;309:756–757.
44. Thirunavukarasu AJ. Large language models will not replace healthcare professionals: curbing popular fears and hype. *J R Soc Med*. 2023;116:181–182.
45. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*. 2023. https://doi.org/10.1001/jamainternmed.2023.1838.
46. Korot E, Wagner SK, Faes L, et al. Will AI replace ophthalmologists? *Transl Vis Sci Technol*. 2020;9:2. Erratum in: Transl Vis Sci Technol. 2021;10:6.
47. Yakar D, Ongena YP, Kwee TC, Haan M. Do people favor artificial intelligence over physicians? A survey among the general population and their view on artificial intelligence in medicine. *Value Health*. 2022;25:374–381.
48. Milne-Ives M, deCock C, Lim E, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. *J Med Internet Res*. 2020;22:e20346.
49. Chow SC, Lam PY, Choy BNK. Patient-centred care in ophthalmology: current practices, effectiveness and challenges. *Graefes Arch Clin Exp Ophthalmol*. 2022;260:3149–3159.
50. Nov O, Singh N, Mann DM. Putting ChatGPT's medical advice to the (Turing) test. medRxiv [internet]. https://www.medrxiv.org/content/10.1101/2023.01.23.23284735v2; 2023. Accessed January 27, 2023.
51. Richardson JP, Smith C, Curtis S, et al. Patient apprehensions about the use of artificial intelligence in healthcare. *NPJ Digit Med*. 2021;4:140.
52. Char DS, Abràmoff MD, Feudtner C. Identifying ethical considerations for machine learning healthcare applications. *Am J Bioeth*. 2020;20:7–17.
53. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. *Nat Med*. 2022;28:1773–1784.
54. Kraljevic Z, Bean D, Shek A, et al. Foresight – generative pretrained transformer (GPT) for modelling of patient timelines using EHRs. https://doi.org/10.48550/arXiv.2212.08072; 2023. Accessed April 1, 2023.
55. Wang SY, Huang J, Hwang H, et al. 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.
56. statMed.org. About statMed.org. https://statmed.org/about; 2023. Accessed April 1, 2023.
57. Patel SB, Lam K. ChatGPT: the future of discharge summaries? *Lancet Digit Health*. 2023;5:e107–e108.
58. Friedberg MW, Chen PG, Van Busum KR, et al. Factors affecting physician professional satisfaction and their implications for patient care, health systems, and health policy. *Rand Health Q*. 2014;3:1.
59. OpenAI. March 20 ChatGPT outage: here's what happened. https://openai.com/blog/march-20-chatgpt-outage; 2023. Accessed April 10, 2023.

60. Cambridge University Press. Authorship and contributorship. Cambridge core. https://www.cambridge.org/core/services/authors/publishing-ethics/research-publishing-ethics-guidelines-for-journals/authorship-and-contributorship; 2023. Accessed April 1, 2023.
61. GitHub Copilot. Your AI pair programmer. GitHub. https://github.com/features/copilot; 2023. Accessed April 1, 2023.
62. Al Madi N. How readable is model-generated code? Examining readability and visual inspection of GitHub copilot. https://doi:10.1145/3551349.3560438; 2022. Accessed April 1, 2023.
63. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. *NPJ Digit Med*. 2022;5:194.
64. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. In: *Association for Computing Machinery Conference on Health, Inference, and Learning 2020 Workshop*. 2020. arXiv preprint arXiv:1904.05342. 2019 Apr 10.
65. Agrawal M, Hegselmann S, Lang H, et al. Large Language models are few-shot clinical information extractors. https://doi.org/10.48550/arXiv.2205.12689; 2022. Accessed April 1, 2023.
66. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596:583−589.
67. Mai DHA, Nguyen LT, Lee EY. TSSNote-CyaPromBERT: development of an integrated platform for highly accurate promoter prediction and visualization of *Synechococcus* sp. and *Synechocystis* sp. through a state-of-the-art natural language processing model BERT. *Front Genet*. 2022;13:1067562.
68. Centers for Medicare and Medicaid Services. Nondiscrimination in health programs and activities. Proposed Rule on 8 April 2022. https://www.federalregister.gov/d/2022-16217; 2022. Accessed April 10, 2023.
69. Spataro J. Introducing Microsoft 365 Copilot − your copilot for work. The Official Microsoft Blog. https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/; 2023. Accessed April 1, 2023.
70. Voolich Wright J. Announcing new generative AI experiences in Google Workspace. Google Workspace Blog. https://workspace.google.com/blog/product-announcements/generative-ai; 2023. Accessed April 1, 2023.
71. Elali FR, Rachid LN. AI-generated research paper fabrication and plagiarism in the scientific community. *Patterns (N Y)*. 2023;4:100706.
72. Ji ZW, Lee N, Frieske R, et al. Survey of hallucination in natural language generation. *ACM Comput Surv*. 2023;55:1−38.
73. Shashikumar SP, Wardi G, Malhotra A, Nemati S. Artificial intelligence sepsis prediction algorithm learns to say "I don't know". *NPJ Digit Med*. 2021;4:134.
74. United States Food and Drug Administration. *NDA approval letter on 17 February 2023*; 2023. https://www.accessdata.fda.gov/drugsatfda_docs/appletter/2023/217171Orig1s000ltr.pdf. Accessed April 14, 2023.
75. Abràmoff MD, Cunningham B, Patel B, et al. Foundational considerations for artificial intelligence using ophthalmic images. *Ophthalmology*. 2022;129:e14−e32.
76. Youssef A, Abramoff M, Char D. Is the algorithm good in a bad world, or has it learned to be bad? The ethical challenges of "locked" versus "continuously learning" artificial intelligence systems, and "Autonomous" Versus "Assistive" AI Tools in Healthcare. *Am J Bioeth*. 2023 May;23(5):43−45. https://doi.org/10.1080/15265161.2023.2191052. PMID: 37130390.In press.
77. Korot E, Pontikos N, Liu X, et al. Predicting sex from retinal fundus photographs using automated deep learning. *Sci Rep*. 2021;11:10286.
78. Guo E, Gupta M, Sinha S, et al. neuroGPT-X: towards an accountable expert opinion tool for vestibular schwannoma. https://doi.org/10.1101/2023.02.25.23286117; 2023. Accessed March 17, 2023.
79. Sun T, Gaut A, Tang S, et al. Mitigating gender bias in natural language processing: literature review. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019:1630−1640. arXiv preprint arXiv:1906.08976. 2019 Jun 21.
80. Au Yeung J, Kraljevic Z, Luintel A, et al. AI chatbots not yet ready for clinical use. https://doi.org/10.1101/2023.03.02.23286705; 2023. Accessed April 1, 2023.
81. Patel A, Sattler J. Creatively malicious prompt engineering. https://labs.withsecure.com/publications/creatively-malicious-prompt-engineering; 2023. Accessed April 10, 2023.
82. Perez F, Ribeiro I. Ignore previous prompt: attack techniques for language models. https://doi.org/10.48550/arXiv.2211.09527; 2022. Accessed March 17, 2023.
83. Taylor J. ChatGPT's alter ego, dan: users jailbreak AI program to get around ethical safeguards. The Guardian. https://www.theguardian.com/technology/2023/mar/08/chatgpt-alter-ego-dan-users-jailbreak-ai-program-to-get-around-ethical-safeguards; 2023. Accessed April 1, 2023.
84. OpenAI. A.P.I. data usage polices. https://openai.com/policies/api-data-usage-policies; 2023. Accessed April 10, 2023.
85. Lewis Maddison. Samsung workers made a major error by using ChatGPT. Techradar. https://www.techradar.com/news/samsung-workers-leaked-company-secrets-by-using-chatgpt; 2023. Accessed April 10, 2023.