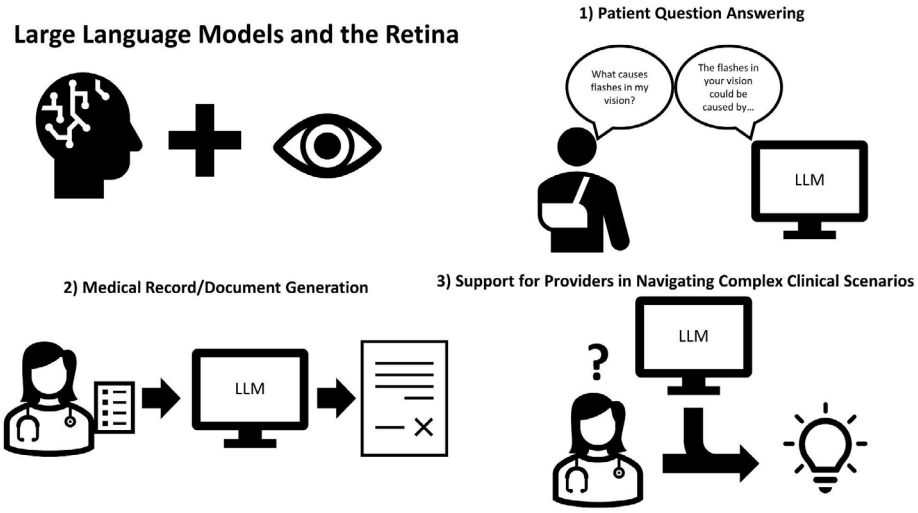
Large Language Models and the Retina: A Review of Current Applications and Future Directions

Aidan Gilson[1], Maxwell Singer[2], Hua Xu[3], Qingyu Chen[3], Ron A Adelman[4]

ABSTRACT

Large Language Models (LLMs) have emerged as a potentially transformative force within retinal healthcare, promising substantial advancements that might be analogous to the impact of anti-VEGF injections. These models signify a shift in patient-provider dynamics and clinical documentation, offering avenues to expedite patient inquiries as well as automate documentation through integration with Electronic Health Records. LLMs may increase direct patient engagement and reduce physician burnout. Simultaneously, provider-centric models may aid in navigating intricate clinical scenarios and rare diseases by assisting in literature review. However, their integration poses unique challenges, including the integration of Protected Health Information, interpreting imaging and other information modalities besides text, and the persistent challenge of generating accurate and verifiable responses. These models mandate rigorous evaluation before integration into clinical workflow. As with all medical interventions, there will always be a possibility of negative outcomes, therefore the critical consideration revolves around the acceptable risk of LLMs vs the substantial benefits they may offer.

Keywords: Artificial Intelligence, Retina, ChatGPT.



INTRODUCTION

Large Language Models have advanced natural language processing (NLP) and have shown great potential in biomedical and health applications.[1–4] They are the latest artificial intelligence (AI) systems for language modeling. The chatbots built from LLMs – such as ChatGPT -- sparked extensive discussions among society, with excitement and concerns alike.[5]

1- BS, Department of Ophthalmology, Yale School of Medicine, New Haven, CT

2- MD, Department of Ophthalmology, Yale School of Medicine, New Haven, CT

3- PhD, Section of Biomedical Informatics and Data Science, Yale School of Medicine, New Haven, CT

4- MD MBA, Department of Ophthalmology, Yale School of Medicine, New Haven, CT

Received: 24.11.2023
Accepted: 27.11.2023

J Ret-Vit 2023; 32: 225-230

DOİ:10.37845/ret.vit.2023.32.38

Correspondence Adress:
Ron A Adelman
Department of Ophthalmology, Yale School of Medicine, New Haven, CT

Phone:

E-mail: ron.adelman@yale.edu

This editorial aims to provide an introduction to LLMs and review its current applications and challenges particularly focusing on its intersection with the retina. We delve into three potential applications: a patient-centric Q&A tool, automated generation of medical records, and support for medical providers in navigating complex clinical scenarios. Throughout each application, we discuss the existing body of literature, prevailing challenges, both present and future, that require attention, and envisage potential end-use scenarios.

Introduction to LLMs

Language models leverage extensive text data and capture semantic representations in an unsupervised or self-supervised manner.[6] They can be broadly categorized into (1) non-contextual embeddings models using fully-connected neural networks such word2vec[7] and fasttext,[8] (2) masked language models using the encoder from the transformer architecture such as biomedical bidirectional encoder representations from transformers (BERT),[9] and (3) generative language models using the decoder from the transformer architecture such as Generative pre-trained transformers (GPT). Language models have been considered as the backbone of NLP methods and have been widely adopted in the medical domain.[10–13] LLMs are the latest advancement mostly in the generative language models. They are characterized by the sheer scale of billions or even hundreds of billions of parameters and undergo pre-training on terabytes of text data, employing self-supervised techniques like predicting or generating the next word.[14–16]

The distinguishing feature of LLMs lies in their capacity to adapt and respond to human input.[1,17] First, it undergoes fine-tuning using manually curated pairs of prompts (expressed in natural language to describe specific tasks) and corresponding responses.[15] Second, reinforcement learning is leveraged to enhance their performance, guided by human feedback assessed through rankings provided by human graders for candidate responses generated by LLMs.[18] LLM-powered chatbots, such as ChatGPT, have demonstrated notable capabilities in natural language generation and reasoning tasks where previous language models often encountered challenges such as reading comprehension[19] and long-form question answering.[20] Importantly, studies consistently observe that LLMs exhibit in-context learning abilities: they can effectively interpret and generate text even when provided with minimal prompts (zero-shot learning) or a limited number of example demonstrations (few-shot learning).

Recent studies on LLMs in the medical domains also have reported encouraging outcomes such as question answering[17] and disease diagnosis.[21] However, existing studies on LLMs in the retina are still lacking. In this editorial, we reviewed applications and challenges of applying LLMs to the retina. Specifically, we summarized the following three major use cases and described the main bottlenecks.

Patient Question Answering:

Current applications of Large Language Models (LLMs) in retina-related contexts primarily revolve around foundational question-answering (QA) tasks. These QA tasks serve as a natural initial use case, given that one of the key advantages of LLMs lies in their user-friendly, dialogic interface. Momenaei et al. conducted an assessment of Chat-GPT's performance in addressing QA-based inquiries pertaining to diabetic retinopathy, epiretinal membranes, and macular holes. Impressively, responses were deemed appropriate for 84.6% of the questions, with only 5.1% considered inappropriate.[22] However, it's worth noting that the average Flesch Kincaid Grade Level for these responses was Grade 14, the equivalent of a 2nd year student at the United States university level. This indicates a certain level of college education is needed to fully comprehend the responses.

Translating years of medical education, primary literature, and clinical expertise into a format understandable by patients poses a significant challenge for many physicians. Given the emphasis on patient autonomy in their care, a comprehensive grasp of their clinical condition is essential for making informed decisions. Therefore not only is the accuracy of responses important, given the critical nature of downstream effects on patient safety from inaccurate responses, but the interpretability as well. Some work has already begun at tailoring responses to the knowledge level of patients, even customizing the complexity to individual patients.[23,24] Although the average response complexity is reduced, the responses are still often outside the scope of understanding of patients. Additionally, further work is needed to ensure accuracy does not change as response complexity is reduced.

Providing accurate answers across the scope of all retinal disorders is not necessary to provide benefit to retinal providers. Rosenblatt et al. found that across over 3 million eyes, 5 years, and 58 retinal practices, only four conditions: wet or dry age-related macular degeneration (ARMD), diabetic retinopathy (DR), branch retinal vein occlusion

(BRVO), and central retinal vein occlusion (CRVO) account for 61.0% of the total prevalence of retinal conditions.[25] LLMs that answer questions only around those conditions could provide benefits to providers by reducing clinical tasks they must complete, and patients by reducing wait time before receiving an answer.

Medical Record/Document Generation:

With the now widespread use of electronic health records, the quantity of documentation that physicians must complete has grown, contributing significantly to physician burnout.[26] The use of LLMs in the generation of clinical reports, or encounter notes could potentially reduce physician burnout, and increase the amount of time providers are able to spend on face-to-face interaction with patients. In 2017, the average vitreoretinal physician billed 1,705 clinic visits through Medicare.[27] If LLMs can reduce time spent in the EHR for documentation by three minutes per visit, over 80 hours could be freed to be spent on direct face-to-face patient interaction. This benefit extends internationally to the United Kingdom, where eye services account for over 10% of all outpatient visits in the national healthcare system.[28] As age-related macular degeneration and diabetic retinopathy are two of the most common sight threatening disorders, and each require frequent screening, vitreoretinal specialists operate at maximum patient capacity. Retina specialists may therefore serve to benefit more from LLM integration than other providers.

Vitreoretinal specialists also utilize imaging in a high percentage of clinic visits. The sheer volume of imaging modalities employed in patient evaluation, including optical coherence tomography, fundus photos, and fluorescein angiograms, presents a specific hurdle.

Shi et al. attempted to combat this limitation using a two part model to "develop an interactive system that harnesses LLMs for report generation and visual question answering in the context of fundus fluorescein angiography (FFA)."[29] The first part, an Image-Text aligning module constructed from an image encoder (ViT) and a language encoder and decoder (BERT), allows for generation of reports from images without physician input. The second part of the model utilized LLaMA-2, an open-source LLM similar to GPT models but with fewer parameters,for QA based on the generated reports. The results showed that 25.7% of reports contained minor errors while 6% contained major errors, while 6.3% of generated responses based on only high-quality reports in the QA task had potential to cause harm. Although promising, this underscores the need for continued refinement and evaluation of LLMs in the context of clinical report generation, particularly when handling visual data. An increase in the number of steps needed in order to generate a document increases the opportunities for errors to be introduced.

Support for Providers in Navigating Complex Clinical Scenarios:

In contrast to patient QA tasks, provider queries to LLMs may be more complex. An example of a patient question may be, "Why could I be seeing flashes and floaters in my vision?" Such questions are commonly encountered by providers every day when they see patients and are reasonably straightforward. Conversely, questions posed by providers to answer complex patient scenarios may be more difficult. Providers need assistance in treating or diagnosing rare diseases, treating complex patients due to disease severity or multimorbidity, or interpreting unusual diagnostic information. Providers are unable to review all primary literature on topics they are less familiar with in the time available for most clinical appointments. A useful response generated by a LLM to physician queries would therefore ideally summarize information from primary literature with citations.

Due to the relatively small number of vitreoretinal providers compared to other ophthalmologic subspecialties, and an increase in subspecialization within the field, it is common for providers to not have other providers within their practice to discuss complex or unique cases. Although large academic centers may have multiple practitioners who can discuss cases, rural and private practice practitioners may not have a colleague with the requisite training to provide assistance.

By mimicking conversations with other providers, the dialogic interface of LLMs provide benefits independent of improvements in information retrieval and summarization.[30] This process is similar to a popular practice known as "rubber ducky debugging" where computer scientists discuss errors in software with an inanimate object, often a rubber duck, to aid in problem solving.[31]

Challenges and limitations:

A critical aspect that remains inadequately addressed is the integration of protected health information (PHI) in both the training and implementation of LLMs. Current models, encompassing various iterations of GPT, PaLM, LLaMA, among others, are either non-compliant with PHI or present logistical challenges for the average clinician to implement

independently.[5,17] ChatGPT is readily accessible and easy to use, however all information which is input into the website is shared with OpenAI meaning it is not compliant with PHI. It is however possible to use GPT models on Microsoft's cloud-computing platform, Azure, and utilize PHI. Additionally, LLaMA is a local model, meaning it can be downloaded directly and information is therefore not shared with a third-party. Both workarounds require significant resources. They may have a significant cost associated with their use or require sufficient computational resources to implement the models locally. Finally, they necessitate some computer science expertise to use in their current form, something that is out of the scope of most retina providers. This is an important hurdle to address, as in order for providers to furnish patients with accurate responses, they must integrate the patient's clinical history into the process of generating an appropriate answer—something that LLMs, due to data safety concerns, struggle to accomplish effectively at this juncture.

As mentioned previously, retinal specialists utilize a large number of imaging modalities in their practice. LLMs must therefore demonstrate proficiency in interpreting this diverse array of information. Although some LLMs, including GPT-4, are capable of receiving multimodal input outside of plain text, their functionality is still below what is necessary for clinical contexts. Task specific models for image classification, segmentation, or interpretation, mostly convolutional neural networks, remain the best performing artificial intelligence tools for image related tasks. Therefore, for optimal performance, LLMs must currently rely on auxiliary tools or physician written reports.

Finally a critical barrier to the ability of LLMs to complete any of the previously mentioned tasks are hallucinations, such as repetitive text in responses, failure to address user queries, and the dissemination of misinformation.[32,33] Broadly, it can be categorized as intrinsic hallucination an extrinsic hallucination.[34]

Intrinsic hallucination, a generated response that contradicts source material, is particularly relevant for rare diseases. LLMs are in their nature generative probabilistic models. The response generated is a result of a stochastic process based on the training data ingested during model construction. Rare diseases or cases, which are likely to be discussed less frequently in training literature than common ones, can be more susceptible to intrinsic error.[35,36] A large language model is likely to have ingested hundreds or thousands of documents pertaining to diabetic retinopathy, its presentation, outcomes, and treatment. But it has likely been trained on far fewer documents related to Bietti's crystalline dystrophy. The combined influence of all documents on a response may lead to incorrect information being presented as an answer to rare diseases.

Extrinsic hallucination, or evidence attribution, is when a response cannot be verified from source material. When there is potential for inaccurate responses, the use of citations to primary literature can both quell concerns, and increases verifiability of the information in the response. However this citation task is currently beyond the scope of large language models to do accurately.[37] If resolved, LLMs can provide a useful adjunct to clinical practice as a first step in performing a comprehensive literature review.

CONCLUSION

Large Language Models hold the potential to revolutionize the clinical practice of retinal specialists, and might even rival the impact of the introduction of anti-VEGF injections. The breadth of potential applications, though not fully expounded upon in this editorial, is staggering. Patient-facing models offer the promise of alleviating the workload of providers while ensuring swift responses to patient inquiries. The integration of LLMs into EHRs has the potential to reduce physician burnout while affording more time for direct patient care. Additionally, provider-facing models can greatly assist in the synthesis and comprehension of complex diseases and presentations. Nevertheless, each of these prospects presents its own set of unique challenges.

In the end, all of these challenges can be distilled into a fundamental question: What level of risk are we willing to accept? Injections entail a risk of endophthalmitis, procedures carry the potential for reduced vision, and LLMs introduce the risk of disseminating inaccurate or potentially harmful information. Therefore, its clinical application necessitates further deliberation to determine the potential risks versus benefits.

REFERENCES

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29:1930-40.
2. Wang DQ, Feng LY, Ye JG, Zou JG, Zheng YF. Accelerating the integration of ChatGPT and other large-scale AI models into biomedical research and healthcare. MedComm–Future Med. 2023;2:e43.

3. Tian S, Jin Q, Yeganova L, et al. Opportunities and Challenges for ChatGPT and Large Language Models in Biomedicine and Health. ArXiv Prepr ArXiv230610070. Published online 2023.
4. Chen Q, Du J, Hu Y, et al. Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. ArXiv Prepr ArXiv230516326. Published online 2023.
5. Introducing ChatGPT. Accessed September 27, 2023. https://openai.com/blog/chatgpt
6. Qudar M, Mago V. A survey on language models. Assoc Comput Mach. 2020;1.
7. Church KW. Word2Vec. Nat Lang Eng. 2017;23:155-62.
8. Joulin A, Grave E, Bojanowski P, Douze M, Jégou H, Mikolov T. Fasttext. zip: Compressing text classification models. ArXiv Prepr ArXiv161203651. Published online 2016.
9. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv Prepr ArXiv181004805. Published online 2018.
10. Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Sci Data. 2019;6:52. doi:10.1038/s41597-019-0055-0
11. Chen Q, Peng Y, Lu Z. BioSentVec: creating sentence embeddings for biomedical texts. In: 2019 IEEE International Conference on Healthcare Informatics (ICHI). IEEE; 2019:1-5.
12. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36:1234-40.
13. Luo R, Sun L, Xia Y, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform. 2022;23:bbac409.
14. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877-901.
15. Touvron H, Martin L, Stone K, et al. Llama 2: Open foundation and fine-tuned chat models. ArXiv Prepr ArXiv230709288. Published online 2023.
16. Chowdhery A, Narang S, Devlin J, et al. Palm: Scaling language modeling with pathways. ArXiv Prepr ArXiv220402311. Published online 2022.
17. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. ArXiv Prepr ArXiv221213138. Published online 2022.
18. Stiennon N, Ouyang L, Wu J, et al. Learning to summarize with human feedback. Adv Neural Inf Process Syst. 2020;33:3008-3021.
19. Xiao C, Xu SX, Zhang K, Wang Y, Xia L. Evaluating Reading Comprehension Exercises Generated by LLMs: A Showcase of ChatGPT in Education Applications. In: ; 2023:610-625.
20. Bhat MM, Meng R, Liu Y, Zhou Y, Yavuz S. Investigating Answerability of LLMs for Long-Form Question Answering. ArXiv Prepr ArXiv230908210. Published online 2023.
21. Lim ZW, Pushpanathan K, Yew SME, 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.
22. Momenaei B, Wakabayashi T, Shahlaee A, et al. Appropriateness and Readability of ChatGPT-4-Generated Responses for Surgical Treatment of Retinal Diseases. Ophthalmol Retina. Published online June 3, 2023. doi:10.1016/j.oret.2023.05.022
23. Haver HL, Lin CT, Sirajuddin A, Yi PH, Jeudy J. Use of ChatGPT, GPT-4, and Bard to Improve Readability of ChatGPT's Answers to Common Questions About Lung Cancer and Lung Cancer Screening. Am J Roentgenol. Published online 2023:1-4.
24. Campbell DJ, Estephan LE, Mastrolonardo EV, Amin DR, Huntley CT, Boon MS. Evaluating ChatGPT responses on obstructive sleep apnea for patient education. J Clin Sleep Med. Published online 2023:jcsm-10728.
25. Rosenblatt TR, Vail D, Saroj N, Boucher N, Moshfeghi DM, Moshfeghi AA. Increasing incidence and prevalence of common retinal diseases in retina practices across the United States. Ophthalmic Surg Lasers Imaging Retina. 2021;52:29-36.
26. Downing NL, Bates DW, Longhurst CA. Physician burnout in the electronic health record era: are we ignoring the real cause? Ann Intern Med. 2018;169:50-1.
27. Chen E, Feng P, Ahluwalia A, et al. Gender Differences in Procedural Volume of Retina Specialists in the United States. Invest Ophthalmol Vis Sci. 2021;62:2657.
28. Demir E, Southern D, Verner A, Amoaku W. A simulation tool for better management of retinal services. BMC Health Serv Res. 2018;18:759. doi:10.1186/s12913-018-3560-5
29. Shi D, Chen X, Zhang W, et al. FFA-GPT: an Interactive Visual Question Answering System for Fundus Fluorescein Angiography. Published online 2023.
30. Alaboudi A, LaToza TD. Using Hypotheses as a Debugging Aid. In: 2020 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC). ; 2020:1-9. doi:10.1109/VL/HCC50065.2020.9127273
31. Thomas D, Hunt A. The Pragmatic Programmer: Your Journey to Mastery, 20th Anniversary Edition. Addison-Wesley Professional; 2019.
32. Zhang Y, Li Y, Cui L, et al. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. ArXiv Prepr ArXiv230901219. Published online 2023.
33. Guerreiro NM, Alves D, Waldendorf J, et al. Hallucinations in large multilingual translation models. ArXiv Prepr ArXiv230316104. Published online 2023.

34. Du L, Wang Y, Xing X, et al. Quantifying and Attributing the Hallucination of Large Language Models via Association Analysis. ArXiv Prepr ArXiv230905217. Published online 2023.

35. Scao TL, Fan A, Akiki C, et al. Bloom: A 176b-parameter open-access multilingual language model. ArXiv Prepr ArXiv221105100. Published online 2022.

36. Kandpal N, Deng H, Roberts A, Wallace E, Raffel C. Large Language Models Struggle to Learn Long-Tail Knowledge. In: Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J, eds. Proceedings of the 40th International Conference on Machine Learning. Vol 202. Proceedings of Machine Learning Research. PMLR; 2023:15696-15707. https://proceedings.mlr.press/v202/kandpal23a.html

37. Liu NF, Zhang T, Liang P. Evaluating Verifiability in Generative Search Engines. Published online 2023.