Abstract 426

IMPACT OF LARGE LANGUAGE MODELS (LLM) AND ARTIFICIAL INTELLIGENCE (AI) ON THE PRACTICE OF RETINA

Adelman R.*[1], Gilson A.[2], Singer M.[3], Chen Q.[2]

[1]Chair of Ophthalmology Mayo Clinic ~ Florida ~ United States of America, [2]Yale School of Medicine ~ New Haven, CT ~ United States of America, [3]Yale School of Medicine ~ New Haven ~ United States of America

Introduction:

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 uderstood, is staggering.

Materials and methods:

Review of current technology and potential future developments.

Results:

The distinguishing feature of LLMs lies in their capacity to adapt and respond to human input. First, it undergoes fine-tuning using manually curated pairs of prompts (expressed in natural language to describe specific tasks) and corresponding responses. 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. 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 comprehension1 and long-form question answering. In this talk we discuss the role of AI in following aspects:
-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.
-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. 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.
-Support for Providers in Navigating Complex Clinical Scenarios: Providers sometimes need assistance in treating or diagnosing very rare diseases, treating complex patients due to disease severity or multi-morbidity, or interpreting unusual diagnostic information. During a short patient encounter , 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.

Conclusions:

Large Language Models have advanced natural language processing (NLP) and have shown great potential in biomedical and health applications. 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.

Sources:

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930-1940.
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(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(1):155-162.
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(1):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(4):1234-1240.
13. Luo R, Sun L, Xia Y, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform. 2022;23(6):bbac409.
14. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877-1901.
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(1):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(1):50-51.
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(8):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(1):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.