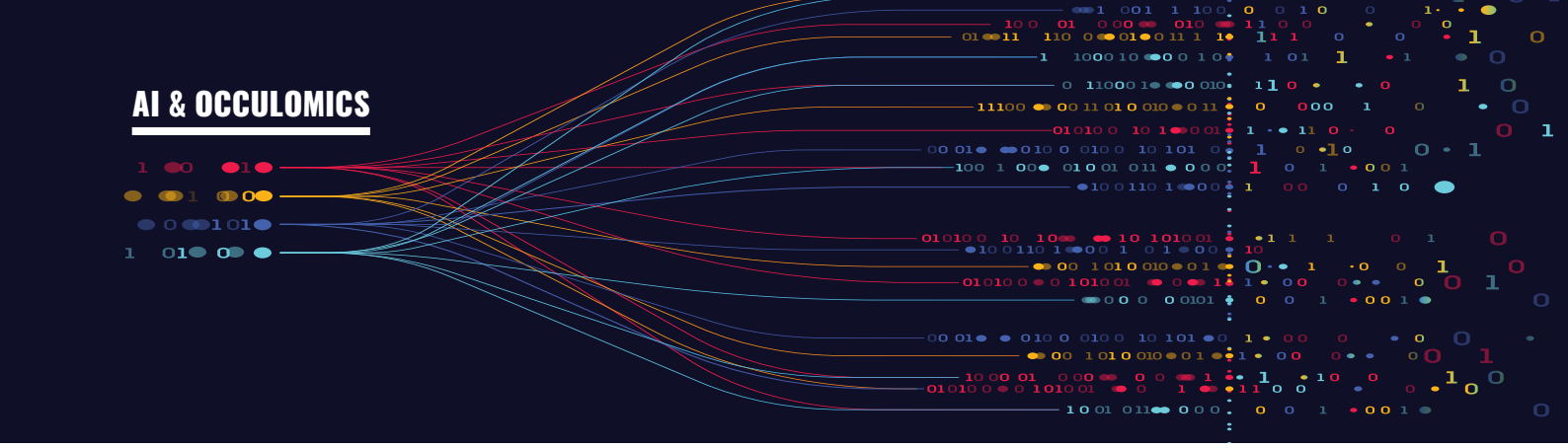

Large language models in ophthalmology

BY FARES ANTAKI

Traditional artificial intelligence (AI) models typically require large amounts of labelled data for training. For example, to develop a model capable of detecting macular pathologies on optical coherence tomography scans, thousands of scans would need to be manually labelled by experts to create a training dataset for the model to learn from. However, a novel paradigm in AI, foundation models, is changing how we train AI models [1].

Foundation models can be trained on unlabelled data—the types of data routinely collected in clinical practice that we do not always label or grade. Through self-supervised learning techniques, these models learn on large amounts of data before we adapt them to various downstream tasks. Thus, a single foundation model serves as a base for multiple applications, reducing the need to create numerous specialised models for narrow tasks [2].

Foundation models have recently gained prominence, particularly with the emergence of large language model (LLM) applications like ChatGPT, which are foundation models designed specifically for text. Large language models are trained on vast amounts of written content, including encyclopaedias, textbooks, and text available on the internet. The training process involves breaking down text data into smaller units called tokens, hiding some of these tokens, and then having the AI model predict the hidden tokens. This process is repeated billions of times, gradually refining the model's ability to understand and generate human-like text [3].

Large language models learn how we write and speak, making them powerful tools in various domains, including medicine and ophthalmology [4]. Most progress in the LLM space has been driven by large technology companies that have spearheaded the development of these models for broader societal benefits. Generally, before releasing these models, companies refine them through various processes, including human grader evaluations, to ensure they align with human goals. It is widely accepted that models should be "helpful, honest, and harmless" [5].

Large language models are generalist by nature, meaning that they are designed to perform a wide range of tasks and understand diverse topics. They can be made more specialised using a variety of techniques, including fine-tuning, which involves adapting the model by training it with specific examples; retrieval-augmented generation, where the model interacts with external sources like textbooks to provide custom information; and prompt engineering, which modifies the model's behaviour by giving specific instructions [6].

In ophthalmology, LLMs have shown promise in medical education, clinical assistance, and workflow improvement [7]. To illustrate their power, several studies have evaluated the performance of LLMs on ophthalmologic knowledge examinations usually taken by humans. For example, using questions from the US, in January 2023, GPT-3.5 achieved 50–60% accuracy on multiple-choice questions [8]. By June 2023, GPT-4 improved to 70–76% accuracy on the same questions [9]. The models have also shown abilities in determining the diagnosis and management steps of complex ophthalmology cases [10]. These findings are noteworthy, because these models, designed to perform well across many domains, have demonstrated impressive capabilities in our specialty despite not being specifically trained for it. Similar results have been observed with FRCOphth questions from the UK, where GPT-4 surpassed its predecessor and compared favourably to expert ophthalmologists [11].

The rise of LLMs opens doors to numerous applications in ophthalmology. Smart chatbots could triage patient queries and provide accurate answers [12]. In medical education, LLMs could generate educational questions from clinical guidelines and create summaries of peer-reviewed research to keep clinicians updated [6]. They could also streamline clinical workflows by reducing the burden of documenting electronic health records or summarising patient histories [13]. Integrating image analysis capabilities with LLMs can further enhance their utility. Foundation models that handle text and images, known as vision-language models (VLMs), hold great potential in our specialty that relies heavily on imaging data. Although generalist models have shown limited abilities in ophthalmology image analysis until now [14], ongoing efforts aim to develop specialised ophthalmology VLMs [15].

While the potential applications of LLMs are exciting, their implementation must be approached with caution. First, since these models may inform clinical decision-making and affect patient care, they will likely need to be regulated after demonstrating their benefits. This can be achieved through randomised or pragmatic clinical trials, or quality improvement projects, though the best approach for demonstrating added benefit will vary with the proposed use case. Second, clinicians must be trained to understand the strengths and limitations of LLMs, recognising when and how to integrate their insights into clinical workflows. They must also be cognisant of the inherent biases of LLMs to ensure these do not adversely influence patient care or perpetuate disparities [16,17].

Finally, once deployed, LLMs must be carefully monitored and governed within existing and emerging frameworks to ensure their use is both safe and effective. This includes implementing robust data governance practices, maintaining transparency in algorithmic decision-making, and continuously updating models to incorporate the latest clinical guidelines and evidence. Overall, we are still in the early stages of understanding how to effectively utilise these models and how to design robust studies to evaluate their performance. Nevertheless, the progress made so far is promising and offers a glimpse into a future where these technologies can help patients and healthcare professionals.

References

1. Chia MA, Antaki F, Zhou Y, et al. Foundation models in ophthalmology. *Br J Ophthalmol* 2024;**10**:325459.
2. Zhou Y, Chia MA, Wagner SK, et al. A foundation model for generalizable disease detection from retinal images. *Nature* 2023;**622**:156–63.
3. Zhao WX, Zhou K, Li J, et al. A Survey of Large Language Models. *arXiv* 2023. doi.org/10.48550/arXiv.2303.18223
4. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. *Nat Med* 2023;**29(8)**:1930–40.
5. Askell A, Bai Y, Chen A, et al. A General Language Assistant as a Laboratory for Alignment. *arXiv* 2021. doi.org/10.48550/arXiv.2112.00861
6. Sevgi M, Antaki F, Keane PA. Medical education with large language models in ophthalmology: custom instructions and enhanced retrieval capabilities. *Br J Ophthalmol* 2024. doi:10.1136/bjo-2023-325046 [ePub ahead of print].
7. Betzler BK, Chen H, Cheng CY, et al. Large language models and their impact in ophthalmology. *Lancet Digit Health* 2023;**5(12)**:e917–24.
8. 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(4)**:100324.
9. Antaki F, Milad D, Chia MA, 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. doi:10.1136/bjo-2023-324438 [ePub ahead of print].
10. Milad D, Antaki F, Milad J, et al. Assessing the medical reasoning skills of GPT-4 in complex ophthalmology cases. *Br J Ophthalmol* 2024. doi:10.1136/bjo-2023-325053 [ePub ahead of print].
11. Thirunavukarasu AJ, Mahmood S, Malem A, et al. Large language models approach expert-level clinical knowledge and reasoning in ophthalmology: A head-to-head cross-sectional study. *PLOS Digit Health* 2024;**3(4)**:e0000341.
12. Lim ZW, Pushpanthan 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**:104770.
13. Raghu Subramanian C, Yang DA, Khanna R. Enhancing Health Care Communication With Large Language Models-The Role, Challenges, and Future Directions. *JAMA Netw Open* 2024;**7(3)**:e240347.
14. Antaki F, Chopra R, Keane PA. Vision-Language Models for Feature Detection of Macular Diseases on Optical Coherence Tomography. *JAMA Ophthalmol* 2024;**142(6)**:573–6.
15. Gao W, Deng Z, Niu Z, et al. OphGLM: Training an Ophthalmology Large Language-and-Vision Assistant based on Instructions and Dialogue. *arXiv* 2023. https://doi.org/10.48550/arXiv.2306.12174
16. Ayoub NF, Balakrishnan K, Ayoub MS, et al. Inherent bias in large language models: A random sampling analysis. Mayo Clinic Proceedings: Digital Health. *Brief Report* 2024;**2(2)**:186–91.
17. Omiye JA, Lester JC, Spichak S, et al. Large language models propagate race-based medicine. *NPJ Digit Med* 2023;**6(1)**:195.

[All links last accessed September 2024]

AUTHOR

Fares Antaki, MDCM, FRCSC,

Vitreoretinal Surgery Fellow, Cleveland Clinic Cole Eye Institute, USA.

SECTION EDITORS

Nima Ghadiri,

Medical Ophthalmology Consultant and Honorary Senior Clinical Lecturer, Liverpool, UK.

nima.ghadiri@liverpoolft.nhs.uk

Arun James Thirunavukarasu, MA, MB, BChir (Cantab),

Academic Foundation Doctor, Oxford University Hospitals NHS Foundation Trust; Clinical Research Fellow, Nuffield Department of Clinical Neurosciences & Big Data Institute, University of Oxford; Rising Leader Fellow, Aspen Institute, UK.

ajt205@cantab.ac.uk

Declaration of competing interests: Arun James Thirunavukarasu has received research funding from HealthSense for work related to machine learning applications in evidence-based medicine.

Want to learn more about AI and oculomics? Check out ***General purpose AI: Transitioning from high performance in highly curated settings to useful applications in ophthalmology***
by Arun James Thirunavukarasu