ONLINE EXCLUSIVE

ChatGPT cannot pass FRCOphth examinations: implications for ophthalmology and large language model artificial intelligence

BY ARUN JAMES THIRUNAVUKARASU

Large language models are generating a lot of hype for artificial intelligence, but can they assist patients and practitioners in ophthalmology?

Introduction

Deep learning (DL) has emerged in ophthalmology as an exciting form of artificial intelligence (AI) most commonly applied to automated image classification [1,2]. Specifically, DL has been deployed on fundus photography, optical coherence tomography, and visual fields to detect diabetic retinopathy, retinopathy of prematurity, glaucoma, macular oedema, and age-related macular degeneration [1]. DL may also be used to generate natural language processing (NLP) models, although applications in ophthalmology have been limited thus far [2].

Recently, a large language model (LLM), Generative Pre-trained Transformer 3.5 (GPT-3.5), has been trained on a dataset of over 400 billion words derived from books, articles, and webpages on the internet [3]. ChatGPT is a platform which allows users to converse with GPT-3.5 through free text prompts, receiving conversational prose in response.

ChatGPT has generated great excitement and general interest in LLM-based AI, and results across the professions have led to hypotheses that these types of model may complement or even replace doctors [4,5]. Platforms providing accurate, relevant, and reliable information could well be utilised widely in the clinic. Telehealth services could be augmented by automated consultation which could assist triage and even mediate management (where doctors' authorisation is not required) by providing advice to patients. Additionally, advanced chatbots could be used by ophthalmologists, optometrists, and other eye health professionals to promptly resolve confusion, provide guidance, and make evidence-based suggestions. However, tools must be rigorously validated before being applied in the clinic, and it is currently unknown how well ChatGPT's impressive performances generalise to ophthalmology.

The Fellowship of the Royal College of Ophthalmologists (FRCOphth) is an internationally recognised qualification which entitles holders to access the GMC Specialist Register in Ophthalmology. Qualifying by the examination route involves recording sufficient performance in FRCOphth Part 1 and Part 2 written examinations, testing practical and theoretical knowledge through multiple-choice questions, in addition to obtaining a refraction certificate and passing the Part 2 Oral Examination. The Part 1 FRCOphth Written Exam must be based [sat?] by UK trainees before their second year of specialist training ends, whereas the Part 2 FRCOphth Written Exam is sat later; after passing Part 1, and as a prerequisite to sitting the Part 2 Oral Examination. While Part 1 is heavily theoretical, Part 2 requires demonstration of clinical expertise – here, these examinations were used as benchmarks to gauge the domain-specific utility of ChatGPT.

Methods

All available official FRCOphth sample multiple-choice questions were inputted into ChatGPT (9 Jan Version; OpenAI, San Francisco, California, USA) on two separate occasions (27 January 2023 and 30 January 2023), with answers recorded and compared to those provided by the Royal College of Ophthalmologists [6,7]. Questions were omitted from analysis if they included a pictorial figure or table which could not be inputted to ChatGPT. In total, 50 Part 1 questions and 43 Part 2 questions were inputted to ChatGPT, representative of the breadth of difficulty and subject of the written component of both examinations [6,7]. Correct answers were determined by the official answers provided by the Royal

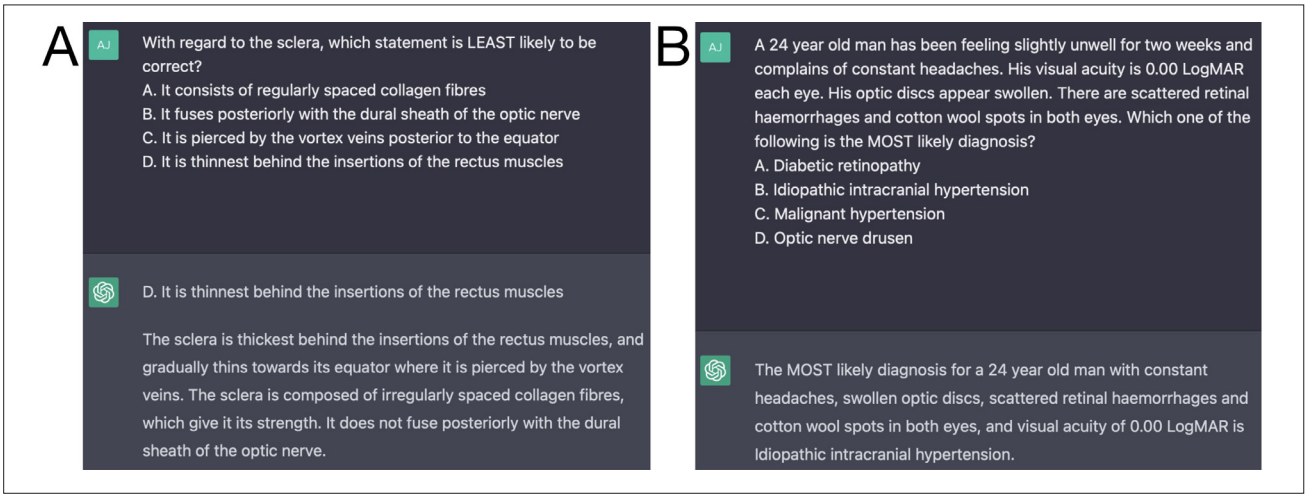

Figure 1: Exemplar question entries and response from ChatGPT: Figure 1A = Part 1 anatomy question, correct answer (A); Figure 1B = Part 2 neuro-ophthalmology question, correct answer (C). Long and plausible (to a non-specialist) explanations are often provided by ChatGPT, inflating confidence in conclusions which are often incorrect. Moreover, even if the provided explanation is factually correct – and sufficient to answer the question correctly – ChatGPT may still respond with the wrong answer, as in Figure 1A.

College of Ophthalmologists. Questions' subjects were defined according to the names of the subsections of the FRCOphth mock papers. Analysis and visualisation of performance was conducted in R (version 4.1.2; R Foundation for Statistical Computing, Vienna, Austria).

Results

For reference, the average FRCOphth multiple-choice question pass mark since 2013 has been 60.2% for Part 1, and 63% for Part 2 [8]. One Part 1 question was omitted due to featuring a table which could not be inputted to the ChatGPT platform. ChatGPT's performance in both examinations was poor: 21/49 (43%) and 26/49 (53%) on the Part 1 questions, 21/43 (49%) and 19/43 (44%) on the Part 2 questions.

Performance was highly variable between subjects (Figure 1). In the Part 1 questions, ChatGPT performed best on questions about investigations (3/3 twice) and pathology (7/10 and 9/10); and worst on questions about anatomy and embryology (0/11 and 3/11). In the Part 2 questions, performance was strongest on questions about ethics (1/1 twice), genetics (1/1 twice), neuro-imaging (1/1 twice), research (1/1 twice), and cornea and the external eye (4/4 and 2/4); and weakest on questions about glaucoma (0/3 twice), ophthalmic investigations (0/2 twice), and statistics (0/1 twice).

ChatGPT exhibited remarkable variability despite its similar performance in two sittings, providing 18/49 (37%) inconsistent Part 1 answers and 11/43 (26%) inconsistent Part 2 answers. The correct answer was reached by both ChatGPT instances in 17/49 (35%) and 16/43 (37%) Part 1 and Part 2 questions, respectively. Ambivalence or uncertainty was expressed by ChatGPT in zero cases. Occasionally, ChatGPT provided longer explanations for its choice of answer (Figure 2), but these did not correlate with accuracy.

Discussion

The inaccuracy and variability of ChatGPT's answers to FRCOphth questions indicates that in their current form, chatbots based on LLMs are not suitable for adoption in clinical settings to provide information to patients or practitioners. In addition,

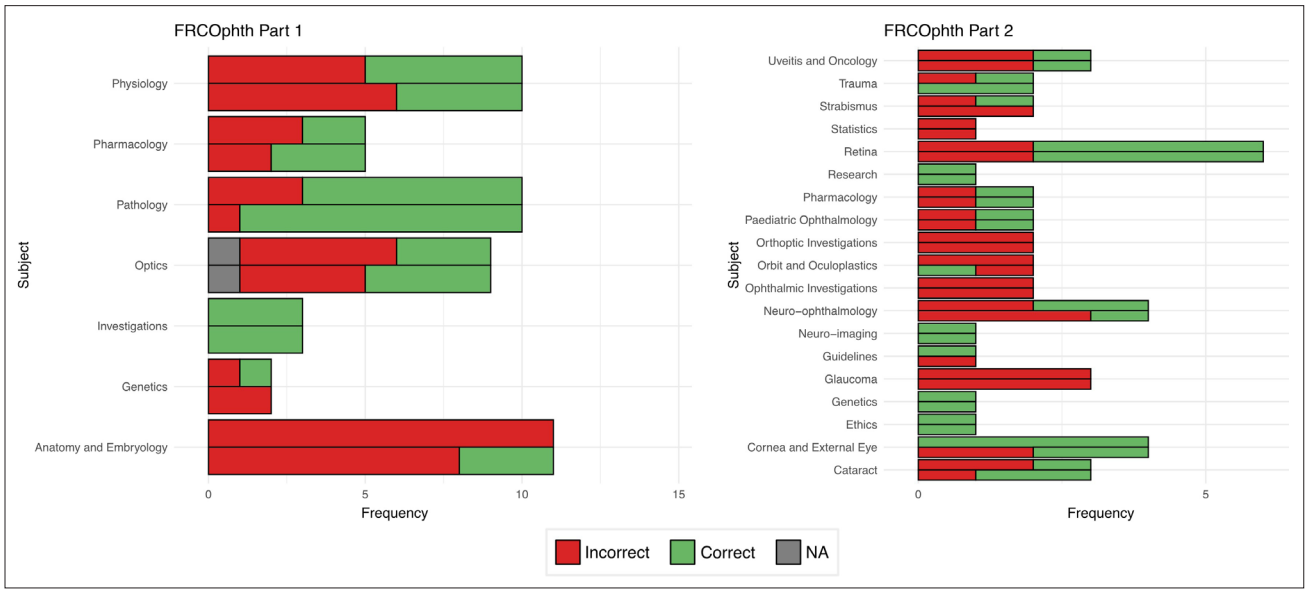

Figure 2: Comparative performance of ChatGPT in FRCOphth Part 1 and 2 questions, categorised by subject. For each subject, the first (higher) column represents Attempt 1 (25 January), and the second (lower) column represents Attempt 2 (30 January). NA = not applicable due to the question featuring an image which could not be inputted to the platform.

the integrity of the FRCOphth written examinations appear not to be threatened specifically by existing LLMs. However, with further technological advances, it is plausible that LLM platforms may contribute in clinical environments, such as by answering clinicians who require rapid responses to queries, or by providing patients with advice regarding their symptoms or ongoing conditions [9].

Although existing training architectures may prove effective means of engineering useful clinical tools, domain-specific training appears necessary for large language models to generate accurate and reliable information to answer ophthalmological queries. Sources of high-yield, domain-specific material may include textbooks, research literature, and clinical notes. In addition, platforms should be designed to provide an indicator of uncertainty rather than blindly 'guessing' – especially with plausible explanatory reasoning – which could serve to confer a false sense of confidence and accuracy.

One schema to gauge uncertainty (which could then be displayed) would be to have queries iterated multiple times by the platform, with greater uncertainty indicated where parallel responses are significantly different from one another semantically. Alternatively, platforms could derive an 'epistemic' uncertainty measure by assaying how well represented queries' words are in its training set (i.e. how familiar the language of the query is) [10].

Clinicians are uniquely placed to advance the development, implementation, and governance of LLMs in ophthalmology (and all of medicine). With their experiential knowledge of the pain points of the specialty, ophthalmologists are the foremost authority on how LLMs may be deployed to improve clinical practice. Ophthalmologists also have access to patients who may provide valuable perspectives on how to improve the delivery and outcomes of eyecare. Ophthalmologists are also responsible for producing and managing a huge quantity of textual information, which (with proper permission and ethical approval) could be leveraged to tune LLMs with ophthalmological material – an opportunity with growing scalability as more hospitals adopt electronic health records [9]. Finally, ophthalmologists will have to be involved in extensive validation projects which must be undertaken to ensure platforms provide accurate information with reliable uncertainty indicators, and to demonstrate that these platforms improve patient care [2]. Validated models have the potential to revolutionise telehealth, triage, patient advice, clinical decision making, and preclinical and clinical education [1,4,9].

TAKE HOME MESSAGES

- Chatbots integrated with large language technology are emerging as exciting platforms to support knowledge acquisition and problem solving.
- Performance answering specialised ophthalmology questions is poor, falling below the standard required to pass FRCOphth examinations.
- In their current form, large language models are not appropriate tools to support practitioners or patients but may improve in the future.
- Improved performance in clinical ophthalmology likely requires domain-specific training, such as from specialist literature and clinical notes.

References

1. Ting DSW, Pasquale LR, Peng, L, et al. Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol* 2019;**103(2)**:167-75.
2. Ting DSW, Peng L, Varadarajan AV, et al. Deep learning in ophthalmology: The technical and clinical considerations. *Progress in Retinal and Eye Research* 2019;**72**:100759.
3. Brown T, Mann B, Ryder N, et al. Language Models are Few-Shot Learners. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf
4. Kung TH, Cheatham M, ChatGPT, et al. Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models. https://doi.org/10.1101/2022.12.19.22283643.
5. Thirunavukarasu, AJ. Large language models will not replace healthcare professionals: curbing popular fears and hype. Journal of the Royal Society of Medicine (2023) doi:10.1177/01410768231173123.
6. Part 1 FRCOphth Exam, The Royal College of Ophthalmologists. https://www.rcophth.ac.uk/examinations/rcophth-exams/part-1-frcophth-exam/
7. Part 2 Written FRCOphth Exam, The Royal College of Ophthalmologists. https://www.rcophth.ac.uk/examinations/rcophth-exams/part-2-written-frcophth-exam/
8. RCOphth Exams, The Royal College of Ophthalmologists. https://www.rcophth.ac.uk/examinations/rcophth-exams/
9. Chen JS, Baxter SL. Applications of natural language processing in ophthalmology: present and future. *Front Med (Lausanne)* https://pubmed.ncbi.nlm.nih.gov/36004369/ 2022. Epub ahead of print.
10. Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning* 2021;**110:** 457-506.

[All links last accessed January-February 2023].

AUTHOR

Arun James Thirunavukarasu,
Final Year Medical Student, University of Cambridge, Cambridge, UK.

Declaration of competing interests: None declared.

Funding statement: This study was not supported by any public or private funders.