Large language models approach expert-level clinical knowledge and reasoning in ophthalmology: A head-to-head cross-sectional study

A.J. Thirunavukarasu[1,2], S. Mahmood[1], A. Malem[3], W.P. Foster[1,4], R. Sanghera[1], R. Hassan[1], S. Zhou[5], S.W. Wong[6], Y.L. Wong[6], Y.J. Chong[7], A. Shakeel[1], Y Chang[8], B.K.J. Tan[9], N. Jain[10], T.F. Tan[11], S. Rauz[7,12], D.S.W. Ting[11,13,14], D.S.J. Ting[7,12,15]

The ROYAL COLLEGE of OPHTHALMOLOGISTS
Annual Congress 2024

Background

- Large language models (LLMs) are a form of language-based artificial intelligence.
- LLMs are attracting great interest after exhibiting remarkable abilities to complete tasks without specific training.
- Understanding the abilities of LLMs is critical as patients may already be using them for advice instead of browsing the internet.
- If validated, LLMs may be integrated into clinical workflows to improve the accuracy and efficiency of healthcare.
- Medical experiments have used examinations to gauge the capabilities and limitations of LLMs, but fail to provide essential context.
- LLMs must be naïve to validation text, but other studies use test material which may be represented within LLM training data.

Methods

- LLMs (GPT-3.5, GPT-4, PaLM 2, LLaMA), expert ophthalmologists, trainee ophthalmologists, and junior doctors without ophthalmology training were tasked with the same 87 FRCOphth Part 2 questions.
- GPT-3.5 and GPT-4 were further tasked with 347 FRCOphth Part 2 questions, and masked expert ophthalmologists graded their responses for accuracy and relevance.
- Performance was analysed with respect to question subject and type (first order recall versus higher order reasoning) to explore the relative strengths and weaknesses of LLMs and human doctors.
- Analysis was conducted in R.

Results

- In the trial of 87 questions, GPT-4 compared well to fully trained (E1-E5) and training (T1-T3) ophthalmologists, and was significantly stronger than untrained doctors (J1-J2; Figure 1).
- GPT-4 was also stronger than the other LLMs (GPT-3.5, LLaMA, PaLM 2), and comparisons with GPT-3.5 over 347 questions suggest that this is due to improved accuracy and relevance of responses (Figure 2; χ^2 test, $p \leq 0.01$ for all comparisons).
- The top performing doctors attained better performance than any of the LLMs (χ^2 test, $p < 0.05$). GPT-4 was equivalent to 3/5 expert and all trainee ophthalmologists (χ^2 test, $p > 0.05$); and superior to both untrained doctors (χ^2 test, $p < 0.05$).
- There was no difference in model or doctor performance when stratified by question complexity (first versus higher order reasoning; χ^2 test, $p > 0.05$) or subject (Fisher's exact test, $p > 0.05$).

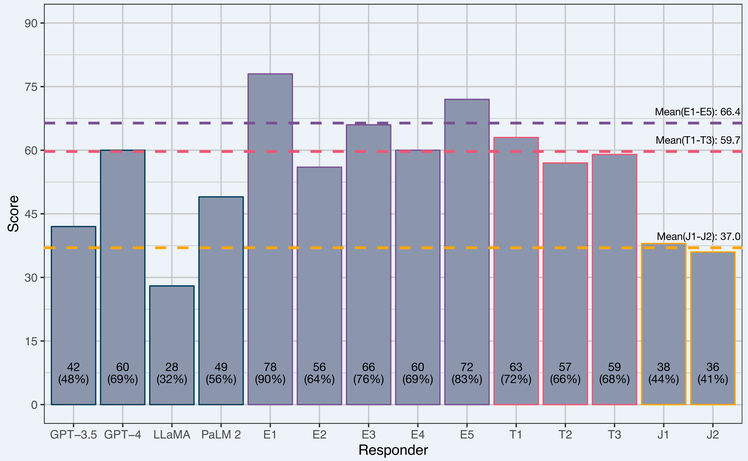


***Figure 1** | State-of-the-art LLMs attain superior performance in ophthalmological questions relative to untrained junior doctors (J1-J2). Moreover, GPT-4 attained comparable performance to training ophthalmologists (T1-T3) and was within the range of scores attained by fully qualified expert ophthalmologists (E1-E5).*

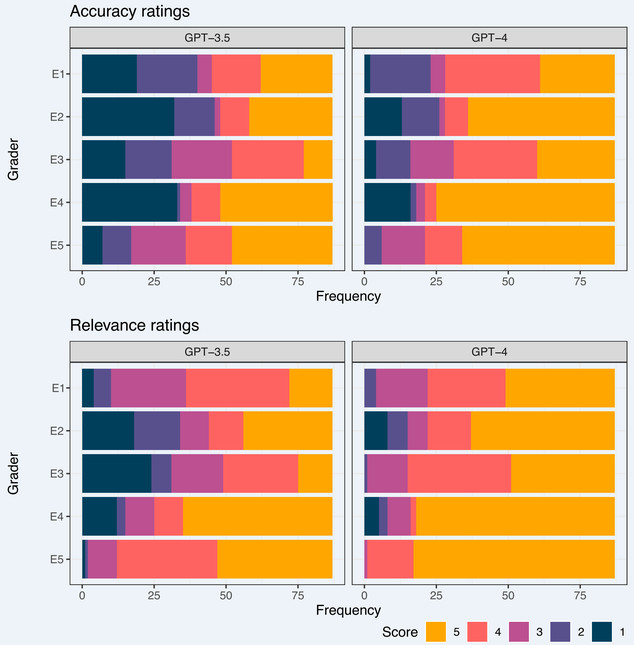


***Figure 2** | GPT-4 responses to ophthalmological questions are more accurate and relevant than GPT-3.5. Responses from both LLMs to the same 347 questions were assessed by full trained ophthalmologists (E1-E5) on a Likert scale ranging from 1 (very poor) to 5 (very good). Every ophthalmologist found that GPT-4 responses were more accurate and more relevant to the question than GPT-3.5 responses (t-test with Bonferroni correction applied, $p < 0.05$). All ophthalmologists preferred GPT-4 responses to GPT-3.5 responses (χ^2 test, $p < 0.05$).*

Conclusions

- State-of-the-art LLMs exhibit comparable ophthalmology knowledge and reasoning ability to training and expert ophthalmologists.
- LLMs exhibit superior knowledge and reasoning when compared to doctors without specialty-specific ophthalmology training.
- The best LLMs exhibit superior accuracy and relevance to their predecessors.
- LLMs may provide useful advice and guidance where access to ophthalmology expertise is limited, such as in lower income countries and in geographically remote locations.
- Patients may already be leveraging LLM tools; with comparable effects to 'Dr Google' phenomena.
- Clinical applications require further validation before they are deployed at scale, ideally through randomised-control trials of LLM interventions pitted against usual care in pragmatic settings.
- LLM tools are unlikely to replace ophthalmologists, but may augment their work in the face of growing demand for expertise and care around the world.

Peer-Reviewed Report

Thirunavukarasu AJ, et al. (2024) Large language models approach expert-level clinical knowledge and reasoning in ophthalmology: A head-to-head cross-sectional study. PLOS Digital Health 3(4): e0000341

1) University of Cambridge School of Clinical Medicine; 2) Oxford University Clinical Academic Graduate School; 3) Eye Institute, Cleveland Clinic Abu Dhabi; 4) Department of Physiology, Development and Neuroscience, University of Cambridge; 5) West Suffolk NHS Foundation Trust; 6) Manchester Royal Eye Hospital; 7) Birmingham and Midland Eye Centre; 8) Department of Ophthalmology, Chang Gung Memorial Hospital; 9) Yong Loo Lin School of Medicine, National University of Singapore; 10) Bedfordshire Hospitals NHS Foundation Trust; 11) Singapore Eye Research Institute, Singapore National Eye Centre; 12) Academic Unit of Ophthalmology, University of Birmingham; 13) Duke-NUS Medical School; 14) Byers Eye Institute, Stanford University; 15) School of Medicine, University of Nottingham