Large language models approach expert-level clinical knowledge and reasoning in ophthalmology

Thirunavukarasu, Arun James; Mahmood, Shathar; Malem, Andrew; Foster, William Paul; Sanghera, Rohan; Hassan, Refaat; Zhou, Sean; Wong, Shiao Wei; Wong, Yee Ling; Chong, Yu Jeat; Shakeel, Abdullah; Chang, Yin-Hsi; Tan, Benjamin Kye Jyn; Jain, Nikhil; Tan, Ting Fang; Rauz, Saaeha; Ting, Daniel Shu Wei; Ting, Darren Shu Jeng

DOI:
10.1371/journal.pdig.0000341

Citation for published version (Harvard):
Thirunavukarasu, AJ, Mahmood, S, Malem, A, Foster, WP, Sanghera, R, Hassan, R, Zhou, S, Wong, SW, Wong, YL, Chong, YJ, Shakeel, A, Chang, Y-H, Tan, BKJ, Jain, N, Tan, TF, Rauz, S, Ting, DSW & Ting, DSJ 2024, 'Large language models approach expert-level clinical knowledge and reasoning in ophthalmology: A head-to-head cross-sectional study', *PLOS digital health*, vol. 3, no. 4, e0000341. https://doi.org/10.1371/journal.pdig.0000341

RESEARCH ARTICLE

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

Arun James Thirunavukarasu[1,2]*, **Shathar Mahmood**[1], **Andrew Malem**[3], **William Paul Foster**[1,4], **Rohan Sanghera**[1], **Refaat Hassan**[1], **Sean Zhou**[5], **Shiao Wei Wong**[6], **Yee Ling Wong**[6], **Yu Jeat Chong**[7], **Abdullah Shakeel**[1], **Yin-Hsi Chang**[8], **Benjamin Kye Jyn Tan**[9], **Nikhil Jain**[10], **Ting Fang Tan**[11], **Saaeha Rauz**[7,12], **Daniel Shu Wei Ting**[11,13,14], **Darren Shu Jeng Ting**[7,12,15]*

1 University of Cambridge School of Clinical Medicine, Cambridge, United Kingdom, **2** Oxford University Clinical Academic Graduate School, University of Oxford, Oxford, United Kingdom, **3** Eye Institute, Cleveland Clinic Abu Dhabi, Abu Dhabi Emirate, United Arab Emirates, **4** Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, United Kingdom, **5** West Suffolk NHS Foundation Trust, Bury St Edmunds, United Kingdom, **6** Manchester Royal Eye Hospital, Manchester University NHS Foundation Trust, Manchester, United Kingdom, **7** Birmingham and Midland Eye Centre, Sandwell and West Birmingham NHS Foundation Trust, Birmingham, United Kingdom, **8** Department of Ophthalmology, Chang Gung Memorial Hospital, Linkou Medical Center, Taoyuan, Taiwan, **9** Yong Loo Lin School of Medicine, National University of Singapore, Singapore, **10** Bedfordshire Hospitals NHS Foundation Trust, Luton and Dunstable, United Kingdom, **11** Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore, **12** Academic Unit of Ophthalmology, Institute of Inflammation and Ageing, University of Birmingham, Birmingham, United Kingdom, **13** Duke-NUS Medical School, Singapore, Singapore, **14** Byers Eye Institute, Stanford University, Palo Alto, California, United States of America, **15** Academic Ophthalmology, School of Medicine, University of Nottingham, Nottingham, United Kingdom

* ajt205@cantab.ac.uk (AJT); ting.darren@gmail.com (DSJT)

Abstract

Large language models (LLMs) underlie remarkable recent advanced in natural language processing, and they are beginning to be applied in clinical contexts. We aimed to evaluate the clinical potential of state-of-the-art LLMs in ophthalmology using a more robust benchmark than raw examination scores. We trialled GPT-3.5 and GPT-4 on 347 ophthalmology questions before GPT-3.5, GPT-4, PaLM 2, LLaMA, expert ophthalmologists, and doctors in training were trialled on a mock examination of 87 questions. Performance was analysed with respect to question subject and type (first order recall and higher order reasoning). Masked ophthalmologists graded the accuracy, relevance, and overall preference of GPT-3.5 and GPT-4 responses to the same questions. The performance of GPT-4 (69%) was superior to GPT-3.5 (48%), LLaMA (32%), and PaLM 2 (56%). GPT-4 compared favourably with expert ophthalmologists (median 76%, range 64–90%), ophthalmology trainees (median 59%, range 57–63%), and unspecialised junior doctors (median 43%, range 41–44%). Low agreement between LLMs and doctors reflected idiosyncratic differences in knowledge and reasoning with overall consistency across subjects and types ($p>0.05$). All ophthalmologists preferred GPT-4 responses over GPT-3.5 and rated the accuracy and relevance of GPT-4 as higher ($p<0.05$). LLMs are approaching expert-level knowledge and

OPEN ACCESS

Citation: Thirunavukarasu AJ, Mahmood S, Malem A, Foster WP, Sanghera R, Hassan R, et al. (2024) Large language models approach expert-level clinical knowledge and reasoning in ophthalmology: A head-to-head cross-sectional study. PLOS Digit Health 3(4): e0000341. https://doi.org/10.1371/journal.pdig.0000341

Editor: Man Luo, Mayo Clinic Scottsdale, UNITED STATES

Received: July 31, 2023

Accepted: February 26, 2024

Published: April 17, 2024

Data Availability Statement: All data are available as supplementary information, excluding copyrighted material from the textbook used for experiments.

Funding: DSWT is supported by the National Medical Research Council, Singapore (NMCR/HSRG/0087/2018; MOH-000655-00; MOH-001014-00), Duke-NUS Medical School (Duke-NUS/RSF/2021/0018; 05/FY2020/EX/15-A58), and Agency for Science, Technology and Research

(A20H4g2141; H20C6a0032). DSJT is supported by a Medical Research Council / Fight for Sight Clinical Research Fellowship (MR/T001674/1). These funders were not involved in the conception, execution, or reporting of this review.

Competing interests: AM is a member of the Panel of Examiners of the Royal College of Ophthalmologists and performs unpaid work as an FRCOphth examiner. DSWT holds a patent on a deep learning system to detect retinal disease. DSJT authored the book used in the study and receives royalty from its sales. The other authors have no competing interests to declare.

reasoning skills in ophthalmology. In view of the comparable or superior performance to trainee-grade ophthalmologists and unspecialised junior doctors, state-of-the-art LLMs such as GPT-4 may provide useful medical advice and assistance where access to expert ophthalmologists is limited. Clinical benchmarks provide useful assays of LLM capabilities in healthcare before clinical trials can be designed and conducted.

Author summary

Large language models (LLMs) are the most sophisticated form of language-based artificial intelligence. LLMs have the potential to improve healthcare, and experiments and trials are ongoing to explore potential avenues for LLMs to improve patient care. Here, we test state-of-the-art LLMs on challenging questions used to assess the aptitude of eye doctors (ophthalmologists) in the United Kingdom before they can be deemed fully qualified. We compare the performance of these LLMs to fully trained ophthalmologists as well as doctors in training to gauge the aptitude of the LLMs for providing advice to patients about eye health. One of the LLMs, GPT-4, exhibits favourable performance when compared with fully qualified and training ophthalmologists; and comparisons with its predecessor model, GPT-3.5, indicate that this superior performance is due to improved accuracy and relevance of model responses. LLMs are approaching expert-level ophthalmological knowledge and reasoning, and may be useful for providing eye-related advice where access to healthcare professionals is limited. Further research is required to explore potential avenues of clinical deployment.

Introduction

Generative Pre-trained Transformer 3.5 (GPT-3.5) and 4 (GPT-4) are large language models (LLMs) trained on datasets containing hundreds of billions of words from articles, books, and other internet sources [1, 2]. ChatGPT is an online chatbot which uses GPT-3.5 or GPT-4 to provide bespoke responses to human users' queries [3]. LLMs have revolutionised the field of natural language processing, and ChatGPT has attracted significant attention in medicine for attaining passing level performance in medical school examinations and providing more accurate and empathetic messages than human doctors in response to patient queries on a social media platform [3,4,5,6]. While GPT-3.5 performance in more specialised examinations has been inadequate, GPT-4 is thought to represent a significant advancement in terms of medical knowledge and reasoning [3,7,8]. Other LLMs in wide use include Pathways Language Model 2 (PaLM 2) and Large Language Model Meta AI 2 (LLaMA 2) [3], [9, p. 2], [10].

Applications and trials of LLMs in ophthalmological settings has been limited despite ChatGPT's performance in questions relating to 'eyes and vision' being superior to other subjects in an examination for general practitioners [7,11]. ChatGPT has been trialled on the North American Ophthalmology Knowledge Assessment Program (OKAP), and Fellowship of the Royal College of Ophthalmologists (FRCOphth) Part 1 and Part 2 examinations. In both cases, relatively poor results have been reported for GPT-3.5, with significant improvement exhibited by GPT-4 [12,13,14,15,16]. However, previous studies are afflicted by two important issues which may affect their validity and interpretability. First, so-called 'contamination', where test material features in the pretraining data used to develop LLMs, may result in inflated performance as models recall previously seen text rather than using clinical reasoning to provide an answer. Second, examination performance in and of itself provides little information regarding the potential of models to contribute to clinical practice as a medical-assistance tool [3]. Clinical benchmarks are required to understanding the meaning and implications of scores in ophthalmological examinations attained by LLMs and are a necessary precursor to clinical trials of LLM-based interventions.

Here, we used FRCOphth Part 2 examination questions to gauge the ophthalmological knowledge base and reasoning capability of LLMs using fully qualified and currently training ophthalmologists as clinical benchmarks. These questions were not freely available online, minimising the risk of contamination. The FRCOphth Part 2 Written Examination tests the clinical knowledge and skills of ophthalmologists in training using multiple choice questions with no negative marking and must be passed to fully qualify as a specialist eye doctor in the United Kingdom.

Methods

Question extraction

FRCOphth Part 2 questions were sourced from a textbook for doctors preparing to take the examination [17]. This textbook is not freely available on the internet, making the possibility of its content being included in LLMs' training datasets unlikely [1]. All 360 multiple-choice questions from the textbook's six chapters were extracted, and a 90-question mock examination from the textbook was segregated for LLM and doctor comparisons. Two researchers matched the subject categories of the practice papers' questions to those defined in the Royal College of Ophthalmologists' documentation concerning the FRCOphth Part 2 written examination. Similarly, two researchers categorised each question as first order recall or higher order reasoning, corresponding to 'remembering' and 'applying' or 'analysing' in Bloom's taxonomy, respectively [18]. Disagreement between classification decisions was resolved by a third researcher casting a deciding vote. Questions containing non-plain text elements such as images were excluded as these could not be inputted to the LLM applications.

Trialling large language models

Every eligible question was inputted into ChatGPT (GPT-3.5 and GPT-4 versions; OpenAI, San Francisco, California, United States of America) between April 29 and May 10, 2023. The answers provided by GPT-3.5 and GPT-4 were recorded and their whole reply to each question was recorded for further analysis. If ChatGPT failed to provide a definitive answer, the question was re-trialled up to three times, after which ChatGPT's answer was recorded as 'null' if no answer was provided. Correct answers ('ground truth') were defined as the answers provided by the textbook and were recorded for every eligible question to facilitate calculation of performance. Upon their release, Bard (Google LLC, Mountain View, California, USA) and HuggingChat (Hugging Face, Inc., New York City, USA) were used to trial PaLM 2 (Google LLC) and LLaMA (Meta, Menlo Park, California, USA) respectively on the portion of the textbook corresponding to a 90-question examination, adhering to the same procedures between June 20 and July 2, 2023.

Clinical benchmarks

To gauge the performance, accuracy, and relevance of LLM outputs, five expert ophthalmologists who had all passed the FRCOphth Part 2 (E1-E5), three trainees (residents) currently in ophthalmology training programmes (T1-T3), and two unspecialised (*i.e*. not in ophthalmology training) junior doctors (J1-J2) first answered the 90-question mock examination independently, without reference to textbooks, the internet, or LLMs' recorded answers. As with the LLMs, doctors' performance was calculated with reference to the correct answers provided by the textbook. After completing the examination, ophthalmologists graded the whole output of GPT-3.5 and GPT-4 on a Likert scale from 1–5 (very bad, bad, neutral, good, very good) to qualitatively appraise accuracy of information provided and relevance of outputs to the question used as an input prompt. For these appraisals, ophthalmologists were blind to the LLM source (which was presented in a randomised order) and to their previous answers to the same questions, but they could refer to the question text and correct answer and explanation provided by the textbook. Procedures are comprehensively described in the protocol issued to the ophthalmologists (S1 Protocol).

Our null hypothesis was that LLMs and doctors would exhibit similar performance, supported by results in a wide range of medical examinations [3, 6]. Prospective power analysis was conducted which indicated that 63 questions were required to identify a 10% superior performance of an LLM to human performance at a 5% significance level (type 1 error rate) with 80% power (20% type 2 error rate). This indicated that the 90-question examination in our experiments was more than sufficient to detect ~10% differences in overall performance. The whole 90-question mock examination was used to avoid over- or under-sampling certain question types with respect to actual FRCOphth papers. To verify that the mock examination was representative of the FRCOphth Part 2 examination, expert ophthalmologists were asked to rate the difficulty of questions used here in comparison to official examinations on a 5-point Likert scale ("much easier", "somewhat easier", "similar", "somewhat more difficult", "much more difficult").

Statistical analysis

Performance of doctors and LLMs were compared using chi-squared (χ^2) tests. Agreement between answers provided by doctors and LLMs was quantified through calculation of Kappa statistics, interpreted in accordance with McHugh's recommendations [19]. To further explore the strengths and weaknesses of the answer providers, performance was stratified by question type (first order fact recall or higher order reasoning) and subject using a chi-squared or Fisher's exact test where appropriate. Likert scale data corresponding to the accuracy and relevance of GPT-3.5 and GPT-4 responses to the same questions were analysed with paired *t*-tests with the Bonferroni correction applied to mitigate the risk of false positive results due to multiple-testing—parametric testing was justified by a sufficient sample size [20]. A chi-squared test was used to quantify the significance of any difference in overall preference of ophthalmologists choosing between GPT-3.5 and GPT-4 responses. Statistical significance was concluded where $p < 0.05$. For additional contextualisation, examination statistics corresponding to FRCOphth Part 2 written examinations taken between July 2017 and December 2022 were collected from Royal College of Ophthalmologists examiners' reports [21]. These statistics facilitated comparisons between human and LLM performance in the mock examination with the performance of actual candidates in recent examinations. Failure cases where all LLMs provided an incorrect answer were appraised qualitatively to explore any specific weaknesses of the technology.

Statistical analysis was conducted in R (version 4.1.2; R Foundation for Statistical Computing, Vienna, Austria), and figures were produced in Affinity Designer (version 1.10.6; Serif Ltd, West Bridgford, Nottinghamshire, United Kingdom).

Results

Questions sources

Of 360 questions in the textbook, 347 questions (including 87 of the 90 questions from the mock examination chapter) were included [17]. Exclusions were all due to non-text elements such as images and tables which could not be inputted into LLM chatbot interfaces. The distribution of question types and subjects within the whole set and mock examination set of questions is summarised in Table 1 and S1 Table alongside performance.

GPT-4 represents a significant advance on GPT-3.5 in ophthalmological knowledge and reasoning. Overall performance over 347 questions was significantly higher for GPT-4 (61.7%) than GPT-3.5 (48.41%; $\chi^2 = 12.32$, $p<0.01$), with results detailed in S1 Fig and S1 Table. ChatGPT performance was consistent across question types and subjects (S1 Table). For GPT-4, no significant variation was observed with respect to first order and higher order questions ($\chi^2 = 0.22$, $p = 0.64$), or subjects defined by the Royal College of Ophthalmologists (Fisher's exact test over 2000 iterations, $p = 0.23$). Similar results were observed for GPT-3.5 with respect to first and second order questions ($\chi^2 = 0.08$, $p = 0.77$), and subjects (Fisher's exact test over 2000 iterations, $p = 0.28$). Performance and variation within the 87-question mock examination was very similar to the overall performance over 347 questions, and subsequent experiments were therefore restricted to that representative set of questions.

GPT-4 compares well with other LLMs, junior and trainee doctors and ophthalmology experts. Performance in the mock examination is summarised in Fig 1—GPT-4 (69%) was the top-scoring model, performing to a significantly higher standard than GPT-3.5 (48%; $\chi^2 = 7.33$, $p < 0.01$) and LLaMA (32%; $\chi^2 = 22.77$, $p < 0.01$), but statistically similarly to PaLM 2 (56%) despite a superior score ($\chi^2 = 2.81$, $p = 0.09$). LLaMA exhibited the lowest examination score, significantly weaker than GPT-3.5 ($\chi^2 = 4.58$, $p = 0.03$) and PaLM-2 ($\chi^2 = 10.01$, $p < 0.01$) as well as GPT-4.

The performance of GPT-4 was statistically similar to the mean score attained by expert ophthalmologists (Fig 1; $\chi^2 = 1.18$, $p = 0.28$). Moreover, GPT-4's performance exceeded the mean mark attained across FRCOphth Part 2 written examination candidates between 2017–2022 (66.06%), mean pass mark according to standard setting (61.31%), and the mean official mark required to pass the examination after adjustment (63.75%), as detailed in S2 Table. In individual comparisons with expert ophthalmologists, GPT-4 was equivalent in 3 cases (χ^2 tests, $p > 0.05$, S3 Table), and inferior in 2 cases (χ^2 tests, $p < 0.05$; Table 2). In comparisons with ophthalmology trainees, GPT-4 was equivalent to all three ophthalmology trainees (χ^2 tests, $p > 0.05$; Table 2). GPT-4 was significantly superior to both unspecialised trainee doctors (χ^2 tests, $p < 0.05$; Table 2). Doctors were anonymised in analysis, but their ophthalmological experience is summarised in S3 Table. Unsurprisingly, junior doctors (J1-J2) attained lower scores than expert ophthalmologists (E1-E5; $t = 7.18$, $p < 0.01$), and ophthalmology trainees (T1-T3; $t = 11.18$, $p < 0.01$), illustrated in Fig 1. Ophthalmology trainees approached expert-level scores with no significant difference between the groups ($t = 1.55$, $p = 0.18$). None of the other LLMs matched any of the expert ophthalmologists, mean mark of real examination candidates, or FRCOphth Part 2 pass mark.

Expert ophthalmologists agreed that the mock examination was a faithful representation of actual FRCOphth Part 2 Written Examination papers with a mean and median score of 3/5 (range 2-4/5).

LLM strengths and weaknesses are similar to doctors. Agreement between answers given by LLMs, expert ophthalmologists, and trainee doctors was generally absent ($0 \leq \kappa < 0.2$), minimal ($0.2 \leq \kappa < 0.4$), or weak ($0.4 \leq \kappa < 0.6$), with moderate agreement only recorded for one pairing between the two highest performing ophthalmologists (Fig 2; κ =

Table 1. Examination characteristics and granular performance data. Question subject and type distributions presented alongside scores attained by LLMs (GPT-3.5, GPT-4, LLaMA, and PaLM 2), expert ophthalmologists (E1-E5), ophthalmology trainees (T1-T3), and unspecialised junior doctors (J1-J2). Median scores do not necessarily sum to the overall median score, as fractional scores are impossible.

Question source	Question subset	Large language model performance				Ophthalmologist performance						Trainee performance				Junior performance		
		GPT-3.5	GPT-4	LLaMA	PaLM 2	E1	E2	E3	E4	E5	Median	T1	T2	T3	Median	J1	J2	Median
Mock examination	Overall (N = 87)	42 (48%)	60 (69%)	28 (32%)	49 (56%)	78 (90%)	56 (64%)	66 (76%)	60 (69%)	72 (83%)	66 (76%)	63 (72%)	57 (66%)	59 (68%)	59 (68%)	38 (44%)	36 (41%)	37 (43%)
	First order fact recall (n = 59)	29 (49%)	43 (73%)	19 (32%)	34 (58%)	52 (88%)	40 (68%)	47 (80%)	37 (63%)	46 (78%)	46 (78%)	44 (75%)	41 (69%)	41 (69%)	41 (69%)	24 (42%)	25 (42%)	25 (42%)
	Higher order reasoning (n = 28)	13 (46%)	17 (61%)	9 (32%)	15 (54%)	26 (93%)	16 (57%)	19 (68%)	23 (82%)	26 (93%)	23 (82%)	19 (68%)	16 (57%)	18 (64%)	18 (64%)	14 (50%)	11 (39%)	13 (46%)
	Cataract (n = 3)	0 (0%)	1 (33%)	1 (33%)	1 (33%)	3 (100%)	2 (67%)	3 (100%)	3 (100%)	3 (100%)	3 (100%)	2 (67%)	3 (100%)	2 (67%)	2 (67%)	2 (67%)	1 (33%)	2 (67%)
	Cornea and external eye (n = 8)	3 (38%)	4 (50%)	3 (38%)	1 (13%)	7 (88%)	5 (63%)	5 (63%)	7 (88%)	7 (88%)	7 (88%)	8 (100%)	6 (75%)	7 (88%)	7 (88%)	4 (50%)	4 (50%)	4 (50%)
	Ethics (n = 1)	1 (100%)	0 (0%)	1 (100%)	0 (0%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	0 (0%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	0 (0%)	1 (100%)	1 (100%)
	Genetics (n = 6)	3 (50%)	5 (83%)	0 (0%)	4 (67%)	4 (67%)	4 (67%)	3 (50%)	3 (50%)	3 (50%)	3 (50%)	4 (67%)	1 (17%)	2 (33%)	2 (33%)	0 (0%)	2 (33%)	1 (17%)
	Glaucoma (n = 3)	1 (33%)	2 (67%)	2 (67%)	2 (67%)	3 (100%)	2 (67%)	3 (100%)	3 (100%)	3 (100%)	3 (100%)	3 (100%)	1 (33%)	2 (67%)	2 (67%)	0 (0%)	1 (33%)	1 (33%)
	Guidelines (n = 3)	1 (33%)	2 (67%)	0 (0%)	0 (0%)	2 (67%)	1 (33%)	2 (67%)	2 (67%)	3 (100%)	2 (67%)	3 (100%)	2 (67%)	3 (100%)	3 (100%)	3 (100%)	1 (33%)	2 (67%)
	Neuro-imaging (n = 0)	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)	0 (NA)
	Neuro-ophthalmology (n = 12)	8 (67%)	11 (92%)	4 (33%)	8 (67%)	12 (100%)	7 (58%)	9 (75%)	6 (50%)	10 (83%)	9 (75%)	8 (67%)	8 (67%)	8 (67%)	8 (67%)	8 (67%)	5 (42%)	7 (58%)
	Ophthalmic investigations (n = 7)	3 (43%)	4 (57%)	2 (29%)	3 (43%)	6 (86%)	4 (57%)	7 (100%)	5 (71%)	7 (100%)	6 (86%)	4 (57%)	3 (43%)	6 (86%)	4 (57%)	3 (43%)	2 (29%)	3 (43%)
	Orbit and oculoplastics (n = 5)	3 (60%)	2 (40%)	2 (40%)	2 (40%)	5 (100%)	4 (80%)	4 (80%)	2 (40%)	3 (60%)	4 (80%)	4 (80%)	4 (80%)	3 (60%)	4 (80%)	0 (0%)	2 (40%)	1 (20%)
	Orthoptic investigations (n = 2)	0 (0%)	2 (100%)	1 (50%)	2 (100%)	2 (100%)	1 (50%)	1 (50%)	1 (50%)	2 (100%)	1 (50%)	1 (50%)	2 (100%)	1 (50%)	1 (50%)	2 (100%)	0 (0%)	1 (50%)
	Paediatric ophthalmology (n = 5)	1 (20%)	1 (20%)	2 (40%)	3 (60%)	5 (100%)	2 (40%)	4 (80%)	4 (80%)	5 (100%)	4 (80%)	2 (40%)	3 (60%)	2 (40%)	2 (40%)	1 (20%)	1 (20%)	1 (20%)
	Pharmacology (n = 5)	3 (60%)	3 (60%)	1 (20%)	4 (80%)	4 (80%)	4 (80%)	4 (80%)	4 (80%)	3 (60%)	4 (80%)	2 (40%)	3 (60%)	4 (80%)	3 (60%)	2 (40%)	2 (40%)	2 (40%)
	Research (n = 1)	1 (100%)	1 (100%)	0 (0%)	1 (100%)	1 (100%)	0 (0%)	0 (0%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)	0 (0%)
	Retina (n = 10)	5 (50%)	8 (80%)	3 (30%)	8 (80%)	8 (80%)	8 (80%)	7 (70%)	5 (50%)	7 (70%)	7 (70%)	6 (60%)	8 (80%)	7 (70%)	7 (70%)	4 (40%)	4 (40%)	4 (40%)
	Statistics (n = 2)	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)	1 (50%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)
	Strabismus (n = 2)	1 (50%)	1 (50%)	0 (0%)	1 (50%)	2 (100%)	1 (50%)	1 (50%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)	1 (50%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)
	Trauma (n = 2)	0 (0%)	2 (100%)	0 (0%)	0 (0%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)	1 (50%)	0 (0%)	1 (50%)
	Uveitis and oncology (n = 10)	6 (60%)	9 (90%)	4 (40%)	7 (70%)	9 (90%)	6 (60%)	8 (80%)	8 (80%)	9 (90%)	8 (80%)	8 (80%)	6 (60%)	5 (50%)	6 (60%)	4 (40%)	6 (60%)	5 (50%)

https://doi.org/10.1371/journal.pdig.0000341.t001

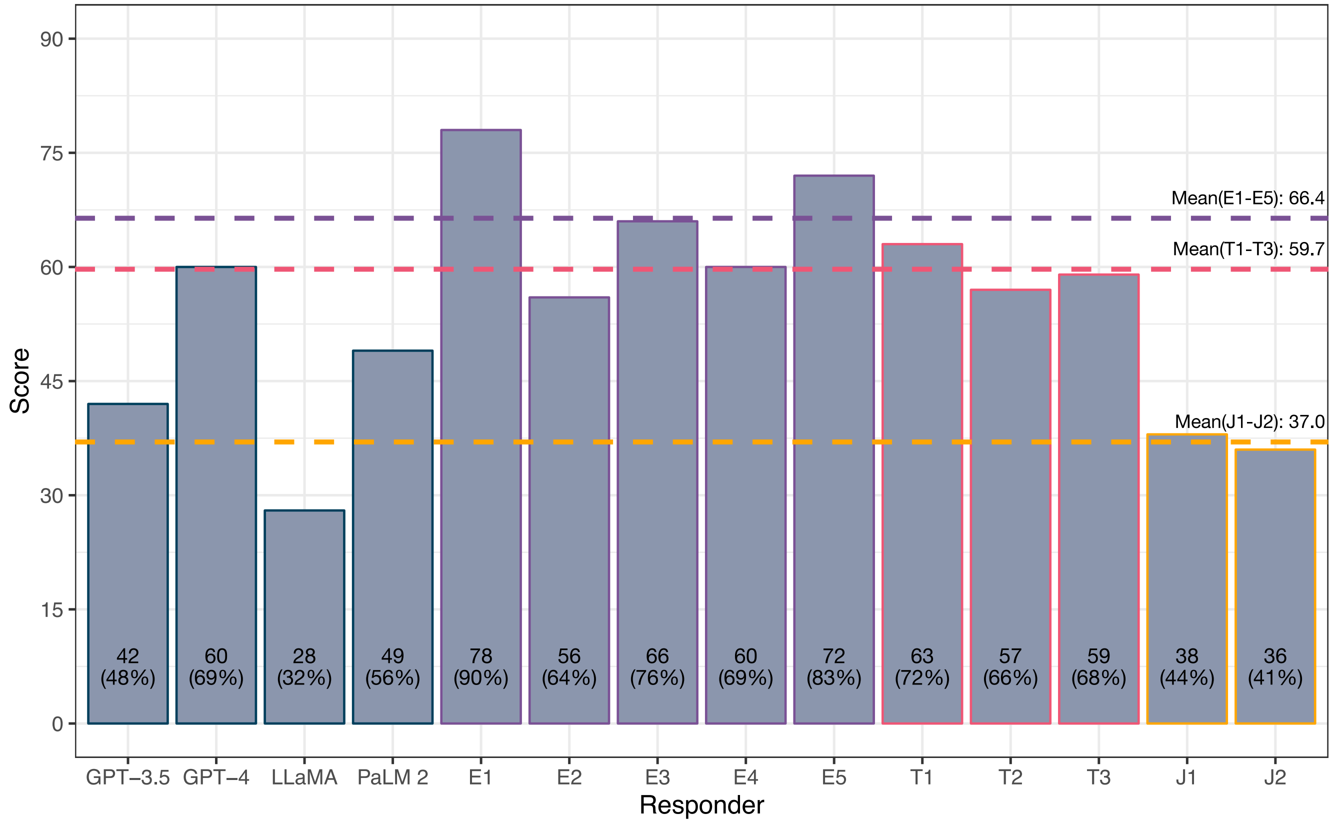


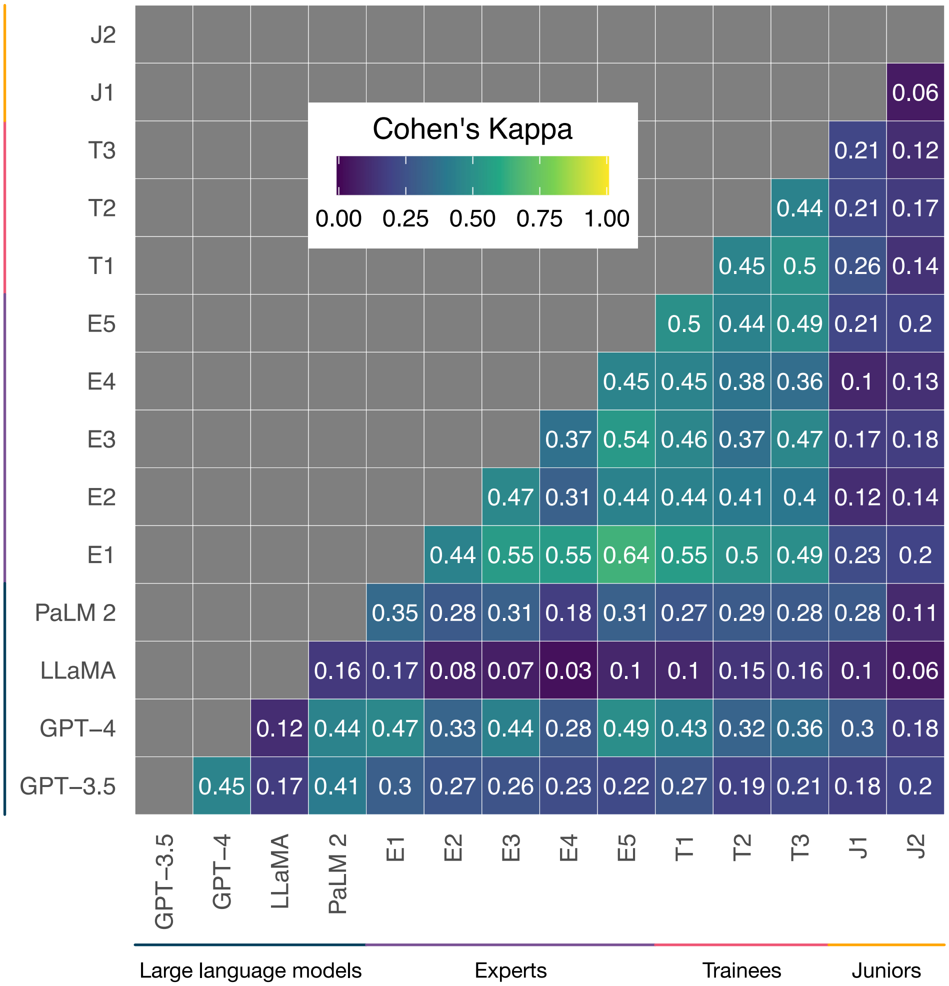
Fig 1. FRCOphth Part 2 performance of LLMs and doctors of variable expertise. Examination performance in the 87-question mock examination used to trial LLMs (GPT-3.5, GPT-4, LLaMA, and PaLM 2), expert ophthalmologists (E1-E5), ophthalmology trainees (T1-T3), and unspecialised junior doctors (J1-J2). Dotted lines depict the mean performance of expert ophthalmologists (66/87; 76%), ophthalmology trainees (60/87; 69%), and unspecialised junior doctors (37/87; 43%). The performance of GPT-4 lay within the range of expert ophthalmologists and ophthalmology trainees.

https://doi.org/10.1371/journal.pdig.0000341.g001

Table 2. GPT-4 compares favourably with LLMs and doctors. Results of pair-wise comparisons of examination performance between GPT-4 and the other answer providers. Significantly greater performance for GPT-4 is highlighted green, significantly inferior performance for GPT-4 is highlighted orange. GPT-4 was superior to all other LLMs and unspecialised junior doctors, and equivalent to most expert ophthalmologists and all ophthalmology trainees.

Answer provider	Score (max = 87)	χ2	*p* value
GPT-4	60	Reference	
GPT-3.5	42	7.68	0.01
LLaMA	28	23.54	<0.01
PaLM 2	49	2.97	0.08
E1	78	11.35	<0.01
E2	56	0.41	0.52
E3	66	1.04	0.31
E4	60	0.00	1.00
E5	72	4.52	0.03
T1	63	0.25	0.62
T2	57	0.23	0.63
T3	59	0.03	0.87
J1	38	11.31	<0.01
J2	36	13.39	<0.01

https://doi.org/10.1371/journal.pdig.0000341.t002



	GPT–3.5	GPT–4	LLaMA	PaLM 2	E1	E2	E3	E4	E5	T1	T2	T3	J1	J2
J2														
J1														0.06
T3													0.21	0.12
T2												0.44	0.21	0.17
T1											0.45	0.5	0.26	0.14
E5										0.5	0.44	0.49	0.21	0.2
E4									0.45	0.45	0.38	0.36	0.1	0.13
E3								0.37	0.54	0.46	0.37	0.47	0.17	0.18
E2							0.47	0.31	0.44	0.44	0.41	0.4	0.12	0.14
E1						0.44	0.55	0.55	0.64	0.55	0.5	0.49	0.23	0.2
PaLM 2					0.35	0.28	0.31	0.18	0.31	0.27	0.29	0.28	0.28	0.11
LLaMA				0.16	0.17	0.08	0.07	0.03	0.1	0.1	0.15	0.16	0.1	0.06
GPT–4			0.12	0.44	0.47	0.33	0.44	0.28	0.49	0.43	0.32	0.36	0.3	0.18
GPT–3.5		0.45	0.17	0.41	0.3	0.27	0.26	0.23	0.22	0.27	0.19	0.21	0.18	0.2

Fig 2. Heat map of Kappa statistics quantifying agreement between answers given by LLMs, expert ophthalmologists, and trainee doctors. Agreement correlates strongly with overall performance and stratification analysis found no particular question type or subject was associated with better performance of LLMs or doctors, indicating that LLM knowledge and reasoning ability is general across ophthalmology rather than restricted to particular subspecialties or question types.

https://doi.org/10.1371/journal.pdig.0000341.g002

0.64) [19]. Disagreement was primarily the result of general differences in knowledge and reasoning ability, illustrated by strong negative correlation between Kappa statistic (quantifying agreement) and difference in examination performance (Pearson's r = -0.63, $p < 0.01$). Answer providers with more similar scores exhibited greater agreement overall irrespective of their category (LLM, expert ophthalmologist, ophthalmology trainee, or junior doctor).

Stratification analysis was undertaken to identify any specific strengths and weaknesses of LLMs with respect to expert ophthalmologists and trainee doctors (Table 1 and S4 Table). No significant difference between performance in first order fact recall and higher order reasoning questions was observed among any of the LLMs, expert ophthalmologists, ophthalmology trainees, or unspecialised junior doctors (S4 Table; χ^2 tests, $p > 0.05$). Similarly, only J1 (junior doctor yet to commence ophthalmology training) exhibited statistically significant variation in performance between subjects (S4 Table; Fisher's exact tests over 2000 iterations, $p = 0.02$); all other doctors and LLMs exhibited no significant variation (Fisher's exact tests over 2000

Table 3. GPT-4 responses are preferred to GPT-3.5 responses by expert ophthalmologists. t-test results with Bonferroni correction applied showing the superior accuracy and relevance of GPT-4 responses relative to GPT-3.5 responses in the opinion of five fully trained ophthalmologists (positive mean differences favour GPT-4), and χ^2 test showing that GPT-4 responses were preferred to GPT-3.5 responses by every ophthalmologist in their blinded qualitative appraisals.

Grader	Accuracy			Relevance			Overall preference		
	Mean difference	*t* statistic	*p* value	Mean difference	*t* statistic	*p* value	GPT-4:GPT-3.5	χ2	*p* value
E1	0.60	3.78	<0.01	0.54	4.38	<0.01	61:26	14.08	<0.01
E2	0.93	4.89	<0.01	0.80	5.16	<0.01	65:22	21.25	<0.01
E3	0.74	5.65	<0.01	1.27	7.53	<0.01	63:24	17.48	<0.01
E4	0.84	4.18	<0.01	0.46	3.15	0.01	65:22	21.25	<0.01
E5	0.59	4.15	<0.01	0.51	5.76	<0.01	72:15	37.35	<0.01

https://doi.org/10.1371/journal.pdig.0000341.t003

iterations, $p > 0.05$). To explore whether consistency was due to an insufficient sample size, similar analyses were run for GPT-3.5 and GPT-4 performance over the larger set of 347 questions (S1 Table; S4 Table). As with the mock examination, no significant differences in performance across question types (S4 Table; χ^2 tests, $p > 0.05$) or subjects (S4 Table; Fisher's exact tests over 2000 iterations, $p > 0.05$) were observed.

LLM examination performance translates to subjective preference indicated by expert ophthalmologists. Ophthalmologists' appraisal of GPT-4 and GPT-3.5 outputs indicated a marked preference for the former over the latter, mirroring objective performance in the mock examination and over the whole textbook. GPT-4 exhibited significantly (*t*-test with Bonferroni correction, $p < 0.05$) higher accuracy and relevance than GPT-3.5 according to all five ophthalmologists' grading (Table 3). Differences were visually obvious, with GPT-4 exhibiting much higher rates of attaining the highest scores for accuracy and relevance than GPT-3.5 (Fig 3). This superiority was reflected in ophthalmologists' qualitative preference indications: GPT-4 responses were preferred to GPT-3.5 responses by every ophthalmologist with statistically significant skew in favour of GPT-4 (χ^2 test, $p < 0.05$; Table 3).

Failure cases exhibit no association with subject, complexity, or human answers. The LLM failure cases—where every LLM provided an incorrect answer—are summarised in Table 4. While errors made by LLMs were occasionally similar to those made by trainee ophthalmologists and junior doctors, this association was not consistent (Table 4). There was no preponderance of ophthalmological subject or first or higher order questions in the failure cases, and questions did not share a common theme, sentence structure, or grammatical construct (Table 4). Examination questions are redacted here to avoid breaching copyright and prevent future LLMs accessing the test data during pretraining but can be provided on request.

Discussion

Here, we present a clinical benchmark to gauge the ophthalmological performance of LLMs, using a source of questions with very low risk of contamination as the utilised textbook is not freely available online [17]. Previous studies have suggested that ChatGPT can provide useful responses to ophthalmological queries, but often use online question sources which may have featured in LLMs' pretraining datasets [7, 12, 15, 22]. In addition, our employment of multiple LLMs as well as fully qualified and training doctors provides novel insight into the potential and limitations of state-of-the-art LLMs through head-to-head comparisons which provide clinical context and quantitative benchmarks of competence in ophthalmology. Subsequent research may leverage our questions and results to gauge the performance of new LLMs and applications as they emerge.

We make three primary observations. First, performance of GPT-4 compares well to expert ophthalmologists and ophthalmology trainees, and exhibits pass-worthy performance in an

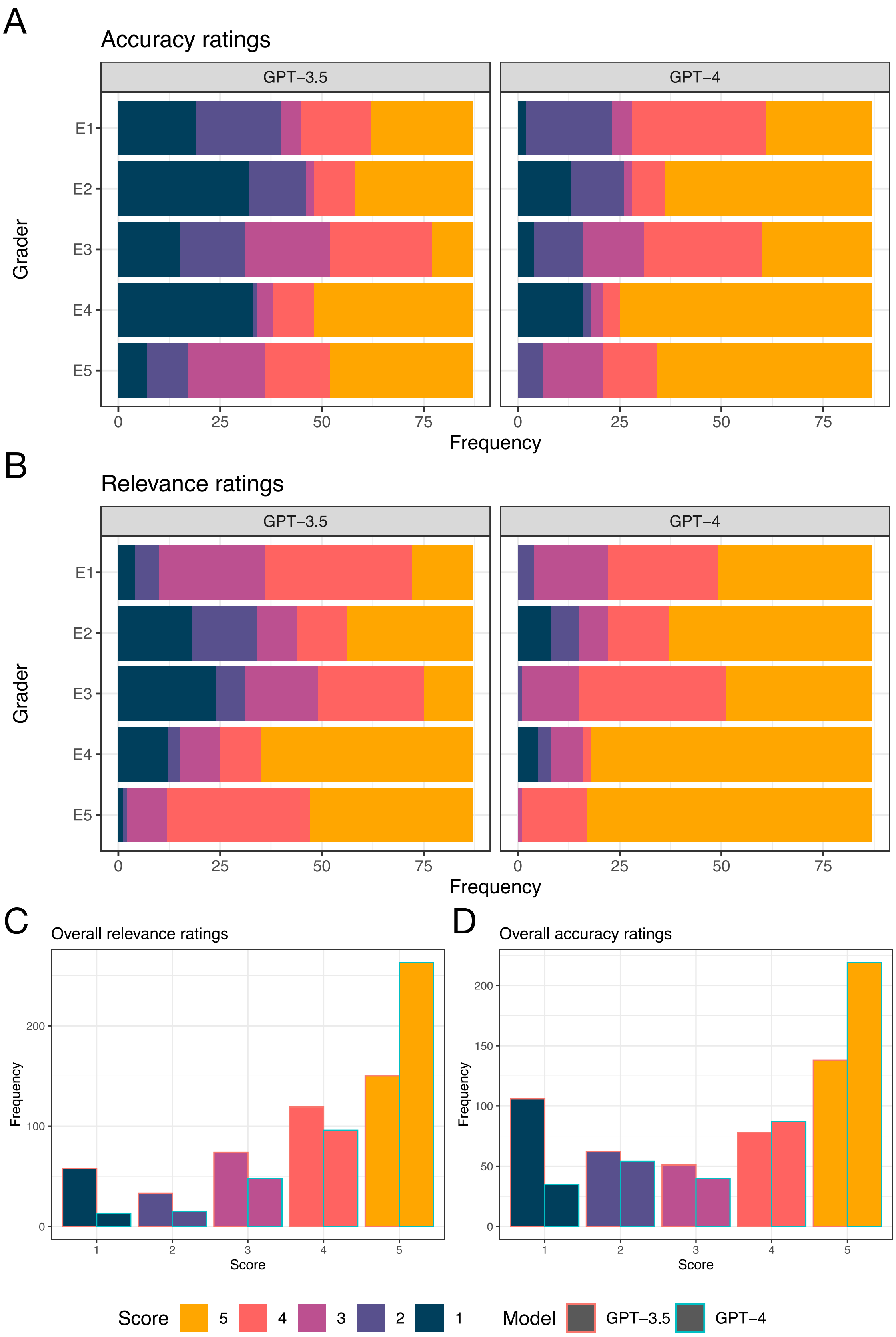

A
Accuracy ratings
GPT–3.5
GPT–4
Grader
E1
E2
E3
E4
E5
0
25
50
75
Frequency
B
Relevance ratings
GPT–3.5
GPT–4
Grader
E1
E2
E3
E4
E5
0
25
50
75
Frequency
C
Overall relevance ratings
Frequency
0
100
200
1
2
3
4
5
Score
D
Overall accuracy ratings
Frequency
0
50
100
150
200
1
2
3
4
5
Score
Score 5 4 3 2 1
Model GPT–3.5 GPT–4

Fig 3. Accuracy and relevance of GPT-3.5 and GPT-4 in response to ophthalmological questions. Accuracy (A) and relevance (B) ratings were provided by five expert ophthalmologists for ChatGPT (powered by GPT-3.5 and GPT-4) responses to 87 FRCOphth Part 2 mock examination questions. In every case, the accuracy and relevance of GPT-4 is significantly superior to GPT-3.5 (t-test with Bonferroni correct applied, $p < 0.05$). Pooled scores for accuracy (C) and relevance (D) from all five raters are presented in the bottom two plots, with GPT-3.5 (left bars) compared directly with GPT-4 (right bars).

https://doi.org/10.1371/journal.pdig.0000341.g003

FRCOphth Part 2 mock examination. PaLM 2 did not attain pass-worthy performance or match expert ophthalmologists' scores but was within the spread of trainee doctors' performance. LLMs are approaching human expert-level knowledge and reasoning in ophthalmology, and significantly exceed the ability of non-specialist clinicians (represented here by unspecialised junior doctors) to answer ophthalmology questions. Second, clinician grading of model outputs suggests that GPT-4 exhibits improved accuracy and relevance when compared with GPT-3.5. Development is producing models which generate better outputs to ophthalmological queries in the opinion of expert human clinicians, which suggests that models are becoming more capable of providing useful assistance in clinical settings. Third, LLM performance was consistent across question subjects and types, distributed similarly to human performance, and exhibited comparable agreement between other LLMs and doctors when corrected for differences in overall performance. Together, this indicates that the ophthalmological knowledge and reasoning capability of LLMs is general rather than limited to certain subspecialties or tasks. LLM-driven natural language processing seems to facilitate similar—although idiosyncratic—clinical knowledge and reasoning to human clinicians, with no obvious blind spots precluding clinical use.

Similarly dramatic improvements in the performance of GPT-4 relative to GPT-3.5 have been reported in the context of the North American Ophthalmology Knowledge Assessment Program (OKAP) [13, 15]. State-of-the-art models exhibit far more clinical promise than their predecessors, and expectations and development should be tailored accordingly. Results from the OKAP also suggest that improvement in performance is due to GPT-4 being more well-rounded than GPT-3.5 [13]. This increases the scope for potential applications of LLMs in ophthalmology, as development is eliminating weaknesses rather than optimising in narrow domains. This study shows that well-rounded LLM performance compares well with expert ophthalmologists, providing clinically relevant evidence that LLMs may be used to provide medical advice and assistance. Further improvement is expected as multimodal foundation models, perhaps based on LLMs such as GPT-4, emerge and facilitate compatibility with image-rich ophthalmological data [3, 23, 24].

Table 4. LLMs do not exhibit consistent weaknesses. Summary of LLM failure cases, where all models provided an incorrect answer to the FRCOphth Part 2 mock examination question. No associations were found with human answers, complexity, subject, theme, sentence structure, or grammatic constructs.

Question	Order	Correct	Category	GPT-3.5	GPT-4	LLaMA	PaLM 2	E1	E2	E3	E4	E5	T1	T2	T3	J1	J2
6	2	A	Cornea and external eye	D	D	C	D	A	A	A	A	A	A	D	A	D	D
15	2	B	Glaucoma	D	C	D	D	B	A	B	B	B	B	C	C	C	D
20	1	C	Cataract	B	B	D	D	C	C	C	C	C	C	C	C	C	D
24	1	D	Uveitis and oncology	C	C	C	C	D	D	D	D	C	D	C	C	A	C
47	1	D	Orbit and oculoplastics	A	A	A	?	D	D	D	A	D	D	D	B	A	D
51	2	A	Strabismus	B	B	B	B	A	B	B	A	A	A	D	A	A	A
55	2	C	Paediatric ophthalmology	D	B	B	B	C	B	C	C	C	B	C	B	B	D
82	1	B	Ophthalmic investigations	A	A	?	D	B	B	B	B	B	D	B	B	D	B
86	2	A	Guidelines	B	B	B	C	A	B	A	A	A	A	A	A	A	D

https://doi.org/10.1371/journal.pdig.0000341.t004

Limitations

This study was limited by three factors. First, examination performance is an unvalidated indicator of clinical aptitude. We sought to ameliorate this limitation by employing expert ophthalmologists, ophthalmology trainees, and unspecialised junior doctors answering the same questions as clinical benchmarks; and compared LLM performance to real cohorts of candidates in recent FRCOphth examinations. However, it remains an issue that comparable performance to clinical experts in an examination does not necessarily demonstrate that an LLM can communicate with patients and practitioners or contribute to clinical decision making accurately and safely. Early trials of LLM chatbots have suggested that LLM responses may be equivalent or even superior to human doctors in terms of accuracy and empathy, and experiments using complicated case studies suggest that LLMs operate well even outside typical presentations and more common medical conditions [4,25,26]. In ophthalmology, GPT-3.5 and GPT-4 have been shown to be capable of providing precise and suitable triage decisions when queried with eye-related symptoms [22,27]. Further work is now warranted in conventional clinical settings.

Second, while the study was sufficiently powered to detect a less than 10% difference in overall performance, the relatively small number of questions in certain categories used for stratification analysis may mask significant differences in performance. Testing LLMs and clinicians with more questions may help establish where LLMs exhibit greater or lesser ability in ophthalmology. Furthermore, researchers using different ways to categorise questions may be able to identify specific strengths and weaknesses of LLMs and doctors which could help guide design of clinical LLM interventions.

Finally, experimental tasks were 'zero-shot' in that LLMs were not provided with any examples of correctly answered questions before it was queried with FRCOphth questions from the textbook. This mode of interrogation entails the maximal level of difficulty for LLMs, so it is conceivable that the ophthalmological knowledge and reasoning encoded within these models is actually even greater than indicated by results here [1]. Future research may seek to fine-tune LLMs by using more domain-specific text during pretraining and fine-tuning, or by providing examples of successfully completed tasks to further improve performance in that clinical task [3].

Future directions

Autonomous deployment of LLMs is currently precluded by inaccuracy and fact fabrication. Our study found that despite meeting expert standards, state-of-the-art LLMs such as GPT-4 do not match top-performing ophthalmologists [28]. Moreover, there remain controversial ethical questions about what roles should and should not be assigned to inanimate AI models, and to what extent human clinicians must remain responsible for their patients [3]. However, the remarkable performance of GPT-4 in ophthalmology examination questions suggests that LLMs may be able to provide useful input in clinical contexts, either to assist clinicians in their day-to-day work or with their education or preparation for examinations [3,13,14,27]. Further improvement in performance may be obtained by specific fine-tuning of models with high quality ophthalmological text data, requiring curation and deidentification [29]. GPT-4 may prove especially useful where access to ophthalmologists is limited: provision of advice, diagnosis, and management suggestions by a model with FRCOphth Part 2-level knowledge and reasoning ability is likely to be superior to non-specialist doctors and allied healthcare professionals working without support, as their exposure to and knowledge of eye care is limited [27,30,31].

However, close monitoring is essential to avoid mistakes caused by inaccuracy or fact fabrication [32]. Clinical applications would also benefit from an uncertainty indicator reducing the risk of erroneous decisions [7]. As LLM performance often correlates with the frequency of query terms' representation in the model's training dataset, a simple indicator of 'familiarity' could be engineered by calculating the relative frequency of query term representation in the training data [7,33]. Users could appraise familiarity to temper their confidence in answers provided by the LLM, perhaps reducing error. Moreover, ophthalmological applications require extensive validation, preferably with high quality randomised controlled trials to conclusively demonstrate benefit (or lack thereof) conferred to patients by LLM interventions [34]. Trials should be pragmatic so as not to inflate effect sizes beyond what may generalise to patients once interventions are implemented at scale [34,35]. In addition to patient outcomes, practitioner-related variables should also be considered: interventions aiming to improve efficiency should be specifically tested to ensure that they reduce rather than increase clinicians' workload [3].

Conclusion

According to comparisons with expert and trainee doctors, state-of-the-art LLMs are approaching expert-level performance in advanced ophthalmology questions. GPT-4 attains pass-worthy performance in FRCOphth Part 2 questions and exceeds the scores of some expert ophthalmologists. As top-performing doctors exhibit superior scores, LLMs do not appear capable of replacing ophthalmologists, but state-of-the-art models could provide useful advice and assistance to non-specialists or patients where access to eye care professionals is limited [27,28]. Further research is required to design LLM-based interventions which may improve eye health outcomes, validate interventions in clinical trials, and engineer governance structures to regulate LLM applications as they begin to be deployed in clinical settings [36].

Supporting information

S1 Fig. ChatGPT performance in questions taken from the whole textbook. Mosaic plot depicting the overall performance of ChatGPT versions powered by GPT-3.5 and GPT-4 in 360 FRCOphth Part 2 written examination questions. Performance was significantly higher for GPT-4 than GPT-3.5, and was close to mean human examination candidate performance and pass mark set by standard setting and after adjustment.
(EPS)

S1 Table. Question characteristics and performance of GPT-3.5 and GPT-4 over the whole textbook. Similar observations were noted here to the smaller mock examination used for subsequent experiments. GPT-4 performs to a significantly higher standard than GPT-3.5
(XLSX)

S2 Table. Examination statistics corresponding to FRCOphth Part 2 written examinations sat between July 2017-December 2022.
(XLSX)

S3 Table. Experience of expert ophthalmologists (E1-E5), ophthalmology trainees (T1-T3), and unspecialised junior doctors (J1-J2) involved in experiments.
(XLSX)

S4 Table. Results of statistical tests of variation in performance between question subjects and types, for each trialled LLM, expert ophthalmologist, and trainee doctor. Statistically significant results are highlighted in green.
(XLSX)

S1 Protocol. Procedures followed by ophthalmologists to grade the output of GPT-3.5 and GPT-4 in terms of accuracy, relevance, and rater-preference of model outputs.
(PDF)

Acknowledgments

The authors extend their thanks to Mr Arunachalam Thirunavukarasu (Betsi Cadwaladr University Health Board) for his advice and assistance with recruitment.

Author Contributions

Conceptualization: Arun James Thirunavukarasu, Darren Shu Jeng Ting.

Data curation: Arun James Thirunavukarasu, Shathar Mahmood, Andrew Malem, William Paul Foster, Rohan Sanghera, Refaat Hassan, Sean Zhou, Shiao Wei Wong, Yee Ling Wong, Yu Jeat Chong, Abdullah Shakeel, Yin-Hsi Chang, Benjamin Kye Jyn Tan, Nikhil Jain, Ting Fang Tan.

Formal analysis: Arun James Thirunavukarasu, Darren Shu Jeng Ting.

Funding acquisition: Arun James Thirunavukarasu, Daniel Shu Wei Ting, Darren Shu Jeng Ting.

Investigation: Arun James Thirunavukarasu, Shathar Mahmood, Andrew Malem, William Paul Foster, Rohan Sanghera, Refaat Hassan, Sean Zhou, Shiao Wei Wong, Yee Ling Wong, Yu Jeat Chong, Abdullah Shakeel, Yin-Hsi Chang, Benjamin Kye Jyn Tan, Nikhil Jain, Ting Fang Tan.

Methodology: Arun James Thirunavukarasu, Darren Shu Jeng Ting.

Project administration: Arun James Thirunavukarasu, Nikhil Jain, Daniel Shu Wei Ting, Darren Shu Jeng Ting.

Resources: Arun James Thirunavukarasu.

Software: Arun James Thirunavukarasu.

Supervision: Arun James Thirunavukarasu, Darren Shu Jeng Ting.

Validation: Arun James Thirunavukarasu.

Visualization: Arun James Thirunavukarasu.

Writing – original draft: Arun James Thirunavukarasu, William Paul Foster, Yin-Hsi Chang, Darren Shu Jeng Ting.

Writing – review & editing: Arun James Thirunavukarasu, Shathar Mahmood, Andrew Malem, William Paul Foster, Rohan Sanghera, Sean Zhou, Shiao Wei Wong, Yee Ling Wong, Yu Jeat Chong, Abdullah Shakeel, Yin-Hsi Chang, Benjamin Kye Jyn Tan, Nikhil Jain, Ting Fang Tan, Saaeha Rauz, Darren Shu Jeng Ting.

References

1. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language Models are Few-Shot Learners. In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2020 [cited 2023 Jan 30]. p. 1877–901. Available from: https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html
2. OpenAI. GPT-4 Technical Report [Internet]. arXiv; 2023 [cited 2023 Apr 11]. Available from: http://arxiv.org/abs/2303.08774

3. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023 Jul 17; 29:1930–40. https://doi.org/10.1038/s41591-023-02448-8 PMID: 37460753

4. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Internal Medicine [Internet]. 2023 Apr 28 [cited 2023 Apr 28]; Available from: https://doi.org/10.1001/jamainternmed.2023.1838 PMID: 37115527

5. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023; 9(101684518):e45312.

6. Kung TH, Cheatham M, Medenilla A, Sillos C, Leon LD, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health. 2023 Feb 9; 2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198 PMID: 36812645

7. Thirunavukarasu AJ, Hassan R, Mahmood S, Sanghera R, Barzangi K, Mukashfi ME, et al. Trialling a Large Language Model (ChatGPT) in General Practice With the Applied Knowledge Test: Observational Study Demonstrating Opportunities and Limitations in Primary Care. JMIR Medical Education. 2023 Apr 21; 9(1):e46599. https://doi.org/10.2196/46599 PMID: 37083633

8. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems [Internet]. arXiv; 2023 [cited 2023 Mar 26]. Available from: http://arxiv.org/abs/2303.13375

9. Google. PaLM 2 Technical Report [Internet]. 2023 [cited 2023 May 11]. Available from: https://ai.google/static/documents/palm2techreport.pdf

10. Touvron H, Martin L, Stone K. Llama 2: Open Foundation and Fine-Tuned Chat Models [Internet]. 2023. Available from: https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/

11. Ting DSJ, Tan TF, Ting DSW. ChatGPT in ophthalmology: the dawn of a new era? Eye (Lond). 2023 Jun 27; https://doi.org/10.1038/s41433-023-02619-4 PMID: 37369764

12. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of its Successes and Shortcomings. Ophthalmology Science [Internet]. 2023 May 4 [cited 2023 May 8];0(0). Available from: https://www.ophthalmologyscience.org/article/S2666-9145(23)00056-8/fulltext https://doi.org/10.1016/j.xops.2023.100324 PMID: 37334036

13. Teebagy S, Colwell L, Wood E, Yaghy A, Faustina M. Improved Performance of ChatGPT-4 on the OKAP Exam: A Comparative Study with ChatGPT-3.5 [Internet]. medRxiv; 2023 [cited 2023 Apr 23]. p. 2023.04.03.23287957. Available from: https://www.medrxiv.org/content/10.1101/2023.04.03.23287957v1

14. Raimondi R, Tzoumas N, Salisbury T, Di Simplicio S, Romano MR. Comparative analysis of large language models in the Royal College of Ophthalmologists fellowship exams. Eye. 2023 May 9;1–4. https://doi.org/10.1038/s41433-023-02563-3 PMID: 37161074

15. Mihalache A, Popovic MM, Muni RH. Performance of an Artificial Intelligence Chatbot in Ophthalmic Knowledge Assessment. JAMA Ophthalmology. 2023 Jun 1; 141(6):589–97. https://doi.org/10.1001/jamaophthalmol.2023.1144 PMID: 37103928

16. Thirunavukarasu AJ. ChatGPT cannot pass FRCOphth examinations: implications for ophthalmology and large language model artificial intelligence. Eye News [Internet]. 2023 Apr 26 [cited 2023 Apr 26]; 30(1). Available from: https://www.eyenews.uk.com/features/ophthalmology/post/chatgpt-cannot-pass-frcophth-examinations-implications-for-ophthalmology-and-large-language-model-artificial-intelligence

17. Ting DSJ, Steel D. MCQs for FRCOphth Part 2. Oxford University Press; 2020. 253 p.

18. Adams NE. Bloom's taxonomy of cognitive learning objectives. J Med Libr Assoc. 2015 Jul; 103(3):152–3. https://doi.org/10.3163/1536-5050.103.3.010 PMID: 26213509

19. McHugh ML. Interrater reliability: the kappa statistic. Biochemia Medica. 2012 Oct 15; 22(3):276–82. https://doi.org/10.1016/j.jocd.2012.03.005 PMID: 23092060

20. Sullivan GM, Artino AR. Analyzing and Interpreting Data From Likert-Type Scales. J Grad Med Educ. 2013 Dec; 5(4):541–2. https://doi.org/10.4300/JGME-5-4-18 PMID: 24454995

21. Part 2 Written FRCOphth Exam [Internet]. The Royal College of Ophthalmologists. [cited 2023 Jan 30]. Available from: https://www.rcophth.ac.uk/examinations/rcophth-exams/part-2-written-frcophth-exam/

22. Tsui JC, Wong MB, Kim BJ, Maguire AM, Scoles D, VanderBeek BL, et al. Appropriateness of ophthalmic symptoms triage by a popular online artificial intelligence chatbot. Eye. 2023 Apr 29;1–2. https://doi.org/10.1038/s41433-023-02556-2 PMID: 37120656

23. Nath S, Marie A, Ellershaw S, Korot E, Keane PA. New meaning for NLP: the trials and tribulations of natural language processing with GPT-3 in ophthalmology. British Journal of Ophthalmology. 2022 Jul 1; 106(7):889–92. https://doi.org/10.1136/bjophthalmol-2022-321141 PMID: 35523534

24. Kline A, Wang H, Li Y, Dennis S, Hutch M, Xu Z, et al. Multimodal machine learning in precision health: A scoping review. npj Digit Med. 2022 Nov 7; 5(1):1–14.

25. Kulkarni PA, Singh H. Artificial Intelligence in Clinical Diagnosis: Opportunities, Challenges, and Hype. JAMA [Internet]. 2023 Jul 6 [cited 2023 Jul 7]; Available from: https://doi.org/10.1001/jama.2023.11440

26. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023 Jul 12;1–9.

27. Waisberg E, Ong J, Zaman N, Kamran SA, Sarker P, Tavakkoli A, et al. GPT-4 for triaging ophthalmic symptoms. Eye. 2023 May 25;1–2. https://doi.org/10.1038/s41433-023-02595-9 PMID: 37231187

28. Thirunavukarasu AJ. Large language models will not replace healthcare professionals: curbing popular fears and hype. J R Soc Med. 2023; 116(5):181–2. https://doi.org/10.1177/01410768231173123 PMID: 37199678

29. Tan TF, Thirunavukarasu AJ, Campbell JP, Keane PA, Pasquale LR, Abramoff MD, et al. Generative Artificial Intelligence through ChatGPT and Other Large Language Models in Ophthalmology: Clinical Applications and Challenges. Ophthalmology Science. 2023 Sep 5; 3(4):100394. https://doi.org/10.1016/j.xops.2023.100394 PMID: 37885755

30. Alsaedi MG, Alhujaili HO, Fairaq GS, Alwdaan SA, Alwadan RA. Emergent Ophthalmic Disease Knowledge among Non-Ophthalmologist Healthcare Professionals in the Western Region of Saudi Arabia: Cross-Sectional Study. The Open Ophthalmology Journal [Internet]. 2022 Mar 25 [cited 2023 Jul 12]; 16(1). Available from: https://openophthalmologyjournal.com/VOLUME/16/ELOCATOR/e187436412203160/FULLTEXT/

31. Tan TF, Thirunavukarasu AJ, Jin L, Lim J, Poh S, Teo ZL, et al. Artificial intelligence and digital health in global eye health: opportunities and challenges. The Lancet Global Health. 2023 Sep 1; 11(9):e1432–43. https://doi.org/10.1016/S2214-109X(23)00323-6 PMID: 37591589

32. Bakken S. AI in health: keeping the human in the loop. Journal of the American Medical Informatics Association. 2023 Jul 1; 30(7):1225–6. https://doi.org/10.1093/jamia/ocad091 PMID: 37337923

33. Biderman S, Schoelkopf H, Anthony Q, Bradley H, O'Brien K, Hallahan E, et al. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling [Internet]. arXiv; 2023 [cited 2023 May 6]. Available from: http://arxiv.org/abs/2304.01373

34. Thirunavukarasu AJ. How Can the Clinical Aptitude of AI Assistants Be Assayed? Journal of Medical Internet Research. 2023 Dec 5; 25(1):e51603. https://doi.org/10.2196/51603 PMID: 38051572

35. Tossaint-Schoenmakers R, Versluis A, Chavannes N, Talboom-Kamp E, Kasteleyn M. The Challenge of Integrating eHealth Into Health Care: Systematic Literature Review of the Donabedian Model of Structure, Process, and Outcome. J Med Internet Res. 2021 May 10; 23(5):e27180. https://doi.org/10.2196/27180 PMID: 33970123

36. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. npj Digit Med. 2023 Jul 6; 6(1):1–6.