British Journal of Ophthalmology

Topic Collection: Artificial Intelligence

Capabilities of GPT-4 in Ophthalmology: An Analysis of Model Entropy and Progress Towards Human-Level Medical Question Answering

Fares Antaki, MDCM, FRCSC[1,2,3,4], Daniel Milad, MD[4], Mark A. Chia, MBBS, MMed[1,2], Charles-Édouard Giguère, MSc[5], Samir Touma, MD, CM[4], Jonathan El-Khoury, MD[4], Pearse A. Keane, MD[1,2,*], Renaud Duval, MD[4,*]

1. Institute of Ophthalmology, University College London, London, United Kingdom
2. Moorfields Eye Hospital NHS Foundation Trust, London, United Kingdom
3. The CHUM School of Artificial Intelligence in Healthcare, Montreal, Quebec, Canada
4. Department of Ophthalmology, Université de Montréal, Montreal, Quebec, Canada
5. Research Centre of the Institut Universitaire en Santé Mentale de Montréal, Montreal, Quebec, Canada

*Joint senior authors

ORCID of the authors: Fares Antaki (0000-0001-6679-7276), Daniel Milad (0000-0002-0693-3421), Charles-Édouard Giguère (0000-0002-8507-6412), Samir Touma (0000-0002-6365-0946), Mark A. Chia (0000-0003-0339-5186), Jonathan El-Khoury (0000-0003-3186-2351), Pearse A. Keane (0000-0002-9239-745X), Renaud Duval (0000-0002-3845-3318)

Corresponding authors:
Renaud Duval, MD, CM
Department of Ophthalmology, Université de Montréal, Montreal, Quebec, Canada.
renaud.duval@gmail.com

Pearse A. Keane, MD
Institute of Ophthalmology, University College London, London, United Kingdom
pearse.keane1@nhs.net

Running Head: Performance of GPT-4 in Ophthalmology

Funding Support: None
Competing interests: None

Word Count: 3,238
Number of Figures: 3
Number of Tables: 1
Number of Supplemental Figures: 6
Number of Supplemental Tables: 4

Keywords: artificial intelligence; foundation models; medical education; GPT-4; ChatGPT; ophthalmology

PRECIS

GPT-4, a large language model, demonstrates strong performance in answering ophthalmology questions, surpassing its predecessor GPT-3.5.

ABSTRACT (241/250)

Background Evidence on the performance of Generative Pre-trained Transformer 4 (GPT-4), a large language model, in the ophthalmology question-answering domain is needed.

Methods We tested GPT-4 on two 260-question multiple choice question sets from the Basic and Clinical Science Course (BCSC) Self-Assessment Program and the OphthoQuestions question banks. We compared the accuracy of GPT-4 models with varying temperatures (creativity setting) and evaluated their responses in a subset of questions. We also compared the best-performing GPT-4 model to GPT-3.5 and to historical human performance.

Results GPT-4-0.3 (GPT-4 with a temperature of 0.3) achieved the highest accuracy among GPT-4 models, with 75.8% on the BCSC set and 70.0% on the OphthoQuestions set. The combined accuracy was 72.9%, which represents an 18.3% raw improvement in accuracy compared to GPT-3.5 ($p<0.001$). Human graders preferred responses from models with a temperature higher than 0 (more creative). Exam section, question difficulty and cognitive level were all predictive of GPT-4-0.3 answer accuracy. GPT-4-0.3's performance was numerically superior to human performance on the BCSC (75.8% vs 73.3%) and OphthoQuestions (70.0% vs 63.0%), but the difference was not statistically significant ($p=0.55$ and $p=0.09$).

Conclusion GPT-4, a LLM trained on non-ophthalmology specific data, performs significantly better than its predecessor on simulated ophthalmology board-style exams. Remarkably, its performance tended to be superior to historical human performance, but that difference was not statistically significant in our study.

KEY MESSAGES

What is already known on this topic

Large language models (LLM) are a novel type of artificial intelligence algorithm that can generate text after being trained on large amounts of unlabeled data. GPT-4 is a popular LLM that showed impressive accuracy in answering general medicine questions, but has not yet been extensively evaluated for its test-taking ability in ophthalmology.

What this study adds

Our study reports the accuracy of GPT-4 on questions from the Basic and Clinical Science Course Self-Assessment Program and the OphthoQuestions online question banks. We provide insights on ideal model settings (temperature/ creativity) and compare the best model to GPT-3.5 and historical human performance.

How this study might affect research, practice or policy

Our study provides evidence on the capabilities of LLMs in our specialty. We show that GPT-4, despite being a general purpose model that has not been fine-tuned for ophthalmology, performs better than GPT-3.5 and not significantly different from an average human trainee when answering board-style questions.

INTRODUCTION

Over the past months, natural language processing (NLP) – a specialisation of artificial intelligence (AI) – has gained substantial attention in academia and in the press due to the release of so called 'foundation models'.[1] Foundation models represent a novel paradigm for building AI systems: they are pretrained *at scale* on billions of unannotated multimodal data in a self-supervised manner and then fine-tuned for specific tasks through *transfer learning.*[1,2] Large language models (LLM) are fine-tuned foundation models that are trained on vast text corpora and that can generate responses in natural language.[3] Two prominent examples of such models are OpenAI's Generative Pre-trained Transformer (GPT) and Google's Pathways Language Model (PaLM). Both LLMs were trained on multilingual text data from the internet and can generate human-like text, perform advanced reasoning, and generate code.[4,5]

There has been growing interest in exploring the potential of LLMs in medicine. A first step in evaluating their medical-domain capabilities has been to explore the challenging task of answering medical questions. This task necessitates comprehension of medical context, recall of medical knowledge as well as reasoning – a skill set that requires years of training and hands-on experience to master.[6] In December 2022, Singhal and colleagues demonstrated state-of-the-art performance of Flan-PaLM in responding to US Medical Licensing Examination (USMLE) style questions, reaching 67.6% accuracy.[7] Less than five months later, in May 2023, they reported an accuracy of 86.5% on the same dataset with Med-PaLM 2, marking a 19% improvement over its predecessor.[8] Comparable rapid and substantial improvements in performance were reported by OpenAI when GPT-4 was introduced. GPT-4 performed significantly better than GPT-3.5 on numerous academic benchmarks, exhibiting human-level performance.[9]

In January 2023, we reported the first results on the performance of LLMs in the ophthalmology question-answering space. We showed that ChatGPT (using GPT-3.5) demonstrated improving accuracy in answering questions from the American Academy of Ophthalmology (AAO)'s Basic and Clinical Science Course (BCSC) Self-Assessment Program and the OphthoQuestions online question banks, with accuracy reaching 59.4% and 49.2%, respectively.[10] Since our initial report, numerous studies have expanded on our findings, reporting equivalent or superior performance of various LLMs over GPT-3.5 on a variety of ophthalmic question banks.[11-15]

In this study, we investigate the accuracy of GPT-4 on the BCSC and OphthoQuestions datasets. We generated responses at different 'temperature' settings, controlling the entropy or creativity of GPT-4, with the primary aim of identifying the optimal setting for question-answering within ophthalmology. This included both a quantitative and qualitative analysis of GPT-4 responses through physician rating of answers. Then, we compared the best GPT-4 model to GPT-3.5 and contextualised our findings with historical human performance data.

METHODS

Exploring BCSC and OphthoQuestions

In January 2023, after obtaining written permission from the AAO, we randomly sampled 260 questions from a pool of 4,458 available in the BCSC Self-Assessment Program. Alongside this, we drew an additional 260 questions from a total of 4,539 questions available on OphthoQuestions (www.ophthoquestions.com). We chose to only use questions that did not incorporate visual data such as clinical, radiological, or graphical images, as the GPT-4 model we used was unable to process this kind of data. Although GPT-4 does have image processing capabilities, this feature was not publicly available at the time of writing in July 2023.[16] We produced 20 random questions from each of the 13 ophthalmology subspecialties, as categorized by the BCSC curriculum [cite BCSC curriculum].

Our prior publication thoroughly outlines the features of the BCSC and OphthoQuestions test sets, including question distribution by examination section, cognitive level, and difficulty.[17] We labelled the questions by cognitive level (high or low) and question difficulty. Low-level questions focused on fact recall, while high-level questions assessed data interpretation and patient management. A difficulty index was derived, indicating the percentage of correct human answers per question bank.[18] Due to the similar distribution of questions in both sets, we combined them for subsequent statistical analyses.

Accessing GPT-4 through the API

ChatGPT (OpenAI, San Francisco) is a chatbot application that was originally based on a fine-tuned model from the GPT-3.5 series called "gpt-3.5-turbo".[19] In March 2023, OpenAI released GPT-4, a new generation LLM exhibiting human-level performance on various academic benchmarks, surpassing GPT-3.5.[9] GPT-4 became available to the public through a limited research preview on the ChatGPT application and through the Application Programming Interface (API). We gained early access to GPT-4 via its API and used it in this research. Using GPT-4 through the API grants unrestricted access to GPT-4 and assures data privacy as the data is not used to enhance the GPT-4 model – a contrast to the research preview available on the ChatGPT application. Moreover, it facilitates integration with other software such as Google Sheets, enabling mass prompting and automation.

Adjusting GPT-4's Temperature

GPT-4 was trained using a vast corpora of text from the internet to reduce the discrepancy between the predicted word and the actual word within the training dataset.[9] Following successful training, the model is capable of creating new text by feeding it an initial prompt, then letting it predict the subsequent word based on statistical patterns learned from its training data. GPT-4 is probabilistic by design, which means it can produce varying responses when given identical prompts. The degree of this variability can be manipulated via the 'temperature' parameter. The ideal temperature setting depends on the specific use case and is often determined *a priori* based on an educated guess.[9] It is generally understood that a temperature of 0 yields coherent and conservative results, while a temperature of 1 fosters high creativity at the expense of coherence. To our knowledge, the ideal temperature for GPT-4 has not yet been defined in the realm of ophthalmology question-answering. Consequently, we decided to identify the optimal temperature for our use case by testing GPT-4 at four distinct temperature settings. For ease of reference in this paper, we will label these as GPT-4-0 for temperature 0, GPT-4-0.3 for 0.3, GPT-4-0.7 for 0.7, and GPT-4-1 for temperature 1.

Human Evaluation of GPT-4 Responses

We carried out human evaluations of long-form responses produced by the GPT-4 models with different temperatures. We randomly sampled 50 questions, with 25 each from BCSC and OphthoQuestions, without controlling for difficulty index, cognitive level, exam section or response accuracy. Our three raters consisted of a recently board-certified ophthalmologist who excelled in OphthoQuestions (Top 10 leaderboards), and two ophthalmology residents from Canada in their third- and a fourth-years of training. In line with the approach proposed by Singhal and colleagues, we directed our raters to rank the model responses based on alignment with medical consensus, knowledge recall, inclusion of irrelevant content and omission of important information.[8] These factors were not judged individually; instead, the raters assigned a comprehensive rank considering all of these domains, with the freedom to weigh them as they deemed appropriate.

Formatting Questions and Zero-Shot Prompting

We maintained the original multiple-choice format of questions with one correct answer and three incorrect options (distractors). We employed a zero-shot approach for the lead-in prompt like in our previous study because this technique is the closest to human test-taking.[7] We used the prompt "Please select the correct answer and provide an explanation" followed by the question and answer options (**Figure 1**).

Sample Question and Response

Question with lead-in prompt
Please select the correct answer and provide an explanation. A 60-year-old patient presents with a homonymous hemianopsia, but with normal visual acuity. Which of the following is the most common cause of this visual field defect in an adult? A. neoplasm, B. trauma, C. ischemic stroke, D. hemorrhage

GPT-4-0.3 response
C. ischemic stroke. Explanation: Homonymous hemianopsia is a condition in which a person sees only the right or left half of the visual field in each eye. This condition is most commonly caused by ischemic stroke, which can damage the optic radiations or the visual cortex in the brain, leading to this type of visual field loss. Neoplasms, trauma, and hemorrhage can also cause homonymous hemianopsia, but these are less common causes than ischemic stroke.

Figure 1. Example of GPT-4-0.3's correct response to this question from the neuro-ophthalmology section of the OphthoQuestions dataset. For reference, this question is considered High cognitive level and of easy difficulty (88% of humans answered correctly).

Historical Human Performance on BCSC and OphthoQuestions

To contextualise the performance of GPT-4, we gathered historical data on human performance for each of the data sets. This information was provided per section and showed average human performance on all the 4,458 BCSC and 4,539 OphthoQuestions questions. However, these figures do not represent the average accuracies for the sample exams of 520 questions that were used to evaluate GPT-4's performance, as this specific data is not available. The BCSC platform offered average peer scores, but these did not include a breakdown by year of training or any historical data. OphthoQuestions provided historical data that matched the user's year of training. The mean accuracies were computed using data from three sequential years of training: the 1st year (2019 – 2020), the 2nd year (2020 – 2021), and the 3rd year (2021 – 2022). Considering this limitation, we decided to adopt a cautious analysis strategy, opting not to establish a non-inferiority threshold. Instead, we assessed whether the performance of GPT-4 differed from that of humans.

Statistical Analysis

We determined accuracy by comparing GPT-4 answers to the answer key provided by the question banks. For each GPT-4 model, accuracy was determined using a single run, as we have previously shown substantial to almost perfect repeatability of GPT-3.5.[17] To compare answer accuracy across different models, we employed a generalized estimating equations using an exchangeable correlation structure and a binomial distribution with a logit link. Since the models were tested on the same questions, we employed *geepack* to allow modelling of correlated data. When we found significant effects, we performed post-hoc analyses and applied Tukey corrections to the p-values. For human evaluation of GPT-4, we measured rater agreement using Kendall's W. We analysed clinician ratings across different GPT-4 models with ANOVA and adjusted using Tukey's method for post-hoc analyses.We used logistic regression to study the influence of exam section, cognitive level, and difficulty on model accuracy. Given that we are dealing with a dichotomous outcome (correct or incorrect answers), we present our results in the form of area under the receiver operating characteristic curve (AUC). Odds ratios could not be used to assess the importance of each variable because we were dealing with both categorical (exam section, cognitive level) and continuous (difficulty index) variables, which makes them non-comparable. We employed Tukey's test to evaluate the effect of each variable while controlling for others. Lastly, we carried out a meta-analysis to compare the best-performing GPT-4 model with historical human data, making adjustments with the *metafor* package. We used R version 4.3.1 for our analyses at a 5% alpha level.

RESULTS

Model temperature does not impact overall accuracy or section performance

Among the GPT-4 models with various temperatures, GPT-4-0.3 achieved the highest combined accuracy. It reached 72.9%, with 75.8% accuracy on the BCSC set and 70.0% on the OphthoQuestions set. Comparatively, the lowest overall accuracy was achieved by GPT-4-0, scoring 71.7%. The maximum difference in overall performance between the best and worst performing models was marginal (1.2%), which is equivalent to 6 questions on the 520 question set. There was no statistically significant difference between the GPT-4 models, (p=0.49). The results are summarised in **Table 1 and Supplemental Table 1.** There were anecdotal performance variations across different exam sections for each GPT-4 model, as seen in **Figure 2**; these differences didn't reach statistical significance (p=0.27) (**Supplemental Figure 1**).

Source	GPT Model				
	GPT-3.5	GPT-4-0	**GPT-4-0.3**	GPT-4-0.7	GPT-4-1
BCSC	58.8	76.2	**75.8**	75.8	76.5
OphthoQuestions	50.4	67.3	**70.0**	68.5	66.9
Combined	54.6	71.7	**72.9**	72.1	71.7

Table 1. Comparison of GPT-4 models results at different temperatures. GPT-4-0.3 had the best overall accuracy. This difference was not statistically significant (Chi-squared= 2.42; p = 0.49).

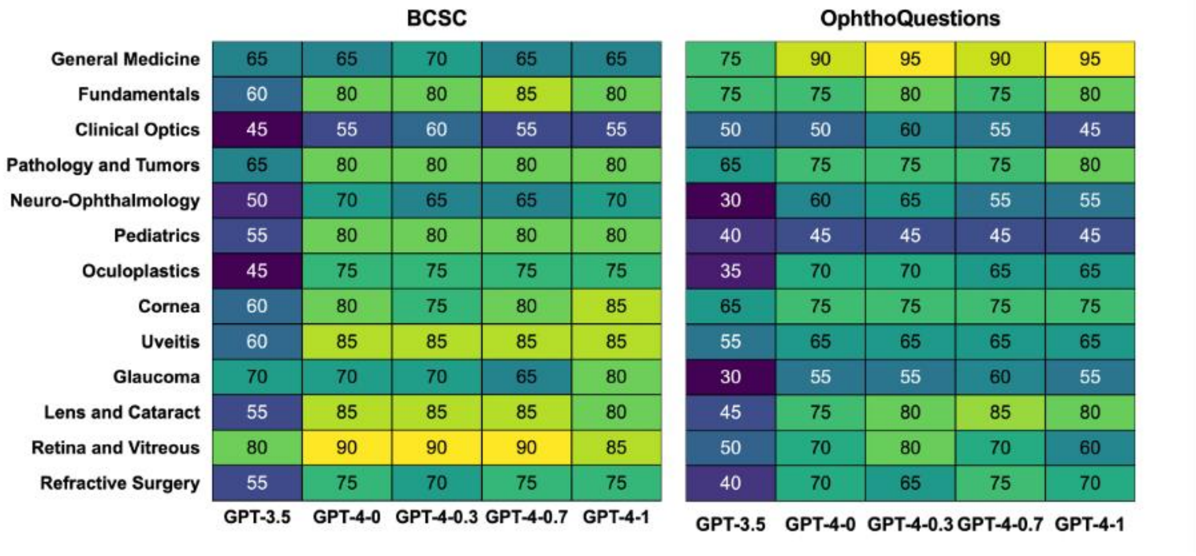


Figure 2. Performance Comparison of GPT Models Across Exam Sections for BCSC and OphthoQuestions. This heatmap provides a colour-coded representation of the performance scores of the various GPT models with varying temperatures across different exam sections and question banks. The scores (percentage) are represented as integers, annotated within each cell and the colours vary from light yellow to dark purple, with lighter colours representing higher performance scores according to the viridis colour palette.

Human raters preferred more probabilistic (creative) models

Our three human raters were in substantial agreement, with a Kendall's W of 0.744 (95% C.I. [0.519, 0.804]), as illustrated in **Supplemental Figure 2**. The mean rankings for the different GPT-4 models were as follows: GPT-4-0 ranked 3.4 (±0.7), GPT-4-0.3 ranked 2.4 (±0.8), and both GPT-4-0.7 and GPT-4-1 ranked 2.1, with a standard deviation of 0.8 and 0.9, respectively. Based on the mean rankings, GPT-4-0 was least preferred compared to all other GPT-4 models ($p<0.001$). There were no statistically significant differences in ranking between the remaining GPT-4-0.3, GPT-4-0.7 and GPT-4-1 models (**Supplemental Table 2**). An example of different GPT-4 responses and the preferred ranking is shown in **Supplemental Figure 3**.

GPT-4-0.3 outperforms GPT-3.5

In all exam sections and across various temperature settings, GPT-4's performance was either on par with or exceeded that of GPT-3.5. The shift from darker to lighter colours on the heatmap in **Figure 2** demonstrates this superior performance. There were two exceptions: the Glaucoma section (BCSC) during the GPT-4-0.7 run and the Clinical Optics section (OphthoQuestions) during the GPT-4-1 run. GPT-4-0.3 outperformed GPT-3.5 by a statistically significant margin, presenting an 18.3% improvement in raw accuracy ($p<0.001$). There were improvements in multiple exam sections, particularly in Lens and Cataract, Oculofacial Plastic and Orbital Surgery, and Retina and Vitreous as seen in **Supplemental Table 3**.

GPT-4-0.3's Accuracy Depends on Exam Section, Cognitive Level and Question Difficulty

Taking the datasets together, GPT-4-0.3 performed best in Retina and Vitreous (85%), General Medicine (82.5%) and Lens and Cataract (82.5%), but not as well in Pediatrics and Strabismus (62.5%), Glaucoma (62.5%) and Clinical Optics (60%). Answer accuracy was most dependent on question difficulty (AUC = 0.69), followed by exam section (AUC = 0.60) and cognitive level (AUC = 0.56), as seen in **Supplemental Figure 4.** We also found that accuracy improved with increased difficulty index (easier questions) while controlling for the examination section and cognitive level. Similar effects were seen for cognitive level (better performance on low cognitive level questions) when controlling for the other two factors. On post-hoc analyses, while controlling for question difficulty and cognitive level, there were significant differences in performance between numerous exam sections as seen in **Supplemental Figure 5**. For example, compared to its strongest performance in Retina and Vitreous, GPT-4-0.3 performed significantly worse in Pediatrics and Strabismus (p=0.017), Oculoplastics (p=0.045) and Glaucoma (p=0.002).

GPT-4-0.3's accuracy is not different from human-level performance

We compared the performance of GPT-4 to historical human performance as reported in the BCSC and OphthoQuestions platforms (**Supplemental Table 6**). The mean accuracy of the GPT-4-0.3 model outperformed historical human averages on both the BCSC (75.8% vs. 73.3%) and OphthoQuestions (70.0% vs. 63.0%) datasets. As depicted in **Figure 3**, GPT-4-0.3 tends to exhibit superior performance compared to human responders, although this varies across different sections. However, an effect size analysis showed no statistically significant difference in performance between GPT-4-0.3 and historical human performance for each of BCSC (p=0.55) and OphthoQuestions (p=0.09). The analyses are shown in **Supplemental Figure 6**.

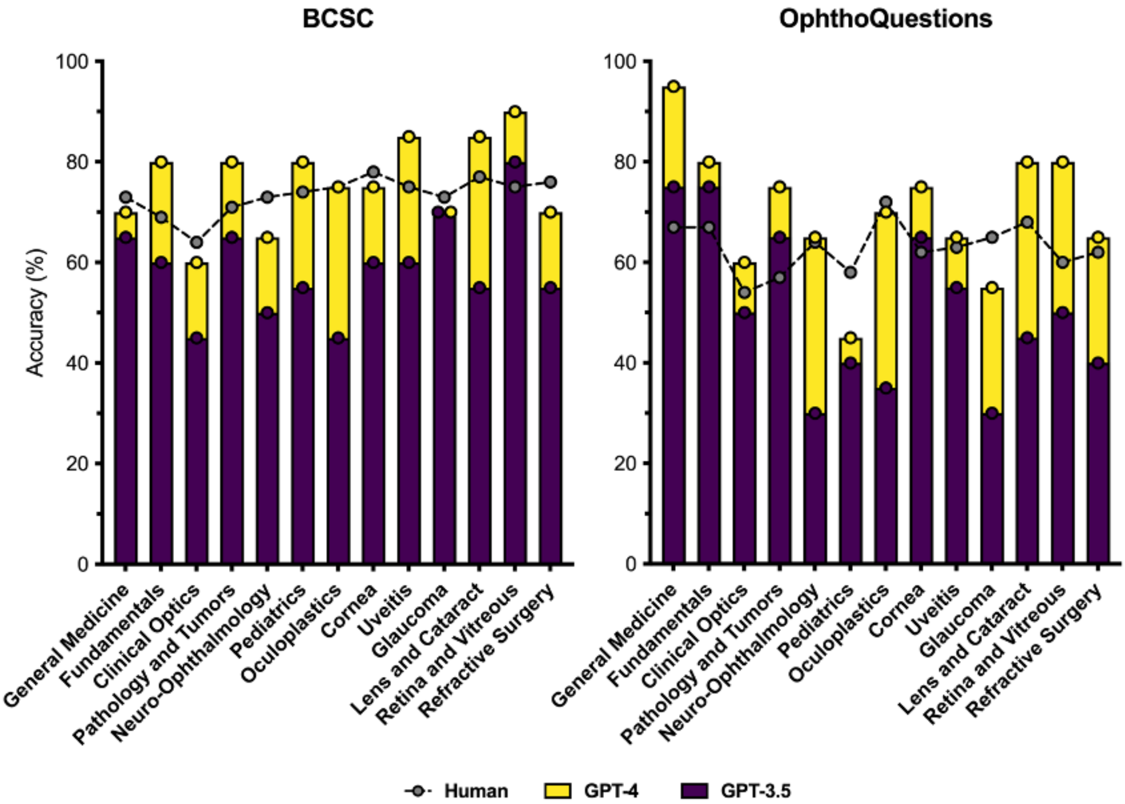


Figure 3. Performance of GPT-3.5 and GPT-4 compared to historical human performance data. GPT-4-0.3 significantly outperforms GPT-3.5 overall ($p<0.001$) and on most exam sections. While it exceeds historical human performance in some sections, there was no significant difference in performance between GPT-4-0.3 and humans for each of BCSC ($p=0.55$) and OphthoQuestions ($p=0.09$), and overall ($p=0.10$).

DISCUSSION

In a previous study, we curated two question datasets from the BCSC and OphthoQuestions to test GPT-3.5 and examine determinants of its performance in the ophthalmology question-answering domain.[10] In this study, we evaluated an updated iteration of GPT, specifically GPT-4, across a spectrum of temperature settings ranging from 0 to 1, which control the creativity of model responses.

The GPT-4 model with a temperature of 0.3 (GPT-4-0.3) had the highest numeric accuracy, but there were no statistically significant differences between GPT-4 models with different temperatures. GPT-4-0.3 had 75.8% accuracy on the BCSC set and 70.0% on the OphthoQuestions set, and a combined overall accuracy of 72.9%. GPT-4-0.3 performed similarly to other GPT-4 models across different exam sections. To our knowledge, the optimal temperature setting for question answering in ophthalmology is not known. The GPT-4 technical report mentions the use of a 0.3 temperature for multiple-choice questions and a 0.6 temperature for free response questions, although the authors clarify that this is merely their best estimation.[9] We found that human raters preferred responses from models operating at temperatures of 0.3, 0.7, or 1, compared to a temperature of 0. We speculate that they were favoured because they are more creative, and possibly pulled from a wider range of knowledge, thereby proving more useful for learning compared to rigid responses. However, creative abilities in models can lead to "hallucinations" or incorrect information, which can be hard to detect. We illustrate an example of such hallucinations in **Supplemental Figure 3.** We believe that a temperature setting ranging from 0.3 to 0.7 is a more secure threshold compared to 1 in a medical question-answering. However, more comprehensive experiments are necessary to confirm this with certainty.

GPT-4-0.3 (72.9%) showed an 18.3% improvement over GPT-3.5 (54.6%) when tested on the same dataset. This improvement was statistically significant. Similar improved accuracy was reported by Med-PaLM2 when tested on USMLE-style general questions, which reported 19% improvement compared to its predecessor MedPaLM.[8] Similar magnitude of improvements were reported in ophthalmology question answering literature. Mihalache and colleagues evaluated the performance of the research preview of GPT-4 (ChatGPT) on a small sample dataset from OphthoQuestions, finding an impressive jump from 58% with GPT-3.5 to 84% with GPT-4.[15] Their reported accuracy exceeds our findings with OphthoQuestions (70%). However, since Mihalache and colleauges used public domain questions from OphthoQuestions free trial, this might have contaminated ChatGPT's training, leading to overestimation of accuracy.[20] Similarly, Teebagy and colleagues saw an improvement of 24%, from 57% to 81%, when assessing the BCSC question set using GPT-4.[13] Meanwhile, Cai and colleagues reported similar results to ours for GPT-3.5 (58.8%) and GPT-4 (71.6%) when using BCSC questions.[12] The discrepancies in reported GPT-4 accuracy could be attributed to differences in the sample datasets (different question difficulties and cognitive level distributions), or even inherent variability in model performance. Indeed, there have been reports of inconsistent behaviour of GPT-4 over time.[21] This raises crucial questions regarding the reproducibility of results from LLMs and issues related to their integration in clinical workflows, particularly if their performance is unpredictable.

We found that GPT-4's answer accuracy depends first on question difficulty, followed by exam section and cognitive level. Simpler, low cognitive level questions—those akin to recall tasks—yield better performance than complex, clinical decision-making ones. While this observation might seem intuitive, its empirical demonstration through our experiments is important. This suggests that GPT-4's current strength lies in memorization-based questions, hinting at limitations in advanced reasoning. Our study also found performance variances across ophthalmic subspecialties like glaucoma and ocular oncology, even after controlling for question difficulty and cognitive level. Such discrepancies reinforce the notion that, although LLMs are trained on a broad corpora of text, their knowledge representation might not uniformly cover all domain subspecialties. In ophthalmology, this could be attributed to factors like the volume of learning materials available online for specific topics (a factor of disease prevalence), the frequency of publications, and other related metrics.

GPT-4-0.3 surpassed historical human scores on the BCSC and OphthoQuestions datasets, individually and combined, with variations observed in different exam sections. This result is significant. Yet, when analysing effect size, the difference was not statistically significant. Such analysis is vital to contextualise GPT-4's performance, especially given its high accuracy. We believe that our analysis provides compelling evidence that GPT-4's performance is on par with human performance in the ophthalmology question answering domain. To determine that, we utilised the aggregate performance of trainees on question banks as a proxy for human-level performance. This method is interesting because the averages we reference come from the experience of thousands of international ophthalmology trainees (residents and fellows), obtained across multiple years and averaged on more than 8,000 questions. Nonetheless, given that these questions are intended as an educational tool, trainees might not perform at their best on those question banks, potentially lowering the average scores. This is balanced out by the likelihood that users, when revisiting questions they have seen before, will answer them correctly.

Despite our encouraging results, we emphasize that this should not be interpreted as suggesting that GPT-4 operates at the same proficiency as a human ophthalmologist overall. Just as with human trainees, performance on multiple-choice examinations (or exam scores) do not dictate overall clinical competency. Such metrics overlook crucial physician competencies, like communication, professionalism and collaboration [cite CanMEDS]. While our study reports on the accuracy of GPT-4 in ophthalmology question-answering, it does not delve into the real-world clinical implications of GPT-4 or LLMs in general. To truly showcase their clinical benefits, we need to use holistic indicators that matter in healthcare like multimodality, applicability, cost-effectiveness, and others. [cite npj paper/ Slack] In ophthalmology, multimodality is crucial, as we frequently depend on imaging data for patient diagnosis and monitoring. LLMs that are integrated into systems equipped for multimodal inputs are poised to be the most beneficial. Over time, building specialised in-domain training for LLMs could prove valuable, and the development of bespoke foundation models for ophthalmology would be optimal.

Contributions:

Acknowledgements: We are deeply grateful to the American Academy of Ophthalmology for generously granting us permission to use the underlying BCSC Self-Assessment Program materials.

Data sharing statement: All data produced in the present study are available upon reasonable request to the authors. The specific question sets are proprietary of the BCSC Self-Assessment Program and OphthoQuestions and cannot be shared.

REFERENCES

1. Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:210807258.* 2021.
2. Wiggins WF, Tejani AS. On the opportunities and risks of foundation models for natural language processing in radiology. *Radiology: Artificial Intelligence.* 2022;4(4):e220119.
3. 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. *The British Journal of Ophthalmology.* 2022;106(7):889-892.
4. Chowdhery A, Narang S, Devlin J, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:220402311.* 2022.
5. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *Advances in neural information processing systems.* 2020;33:1877-1901.
6. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature.* 2023.
7. Singhal K, Azizi S, Tu T, et al. Large Language Models Encode Clinical Knowledge. *arXiv preprint arXiv:221213138.* 2022.
8. Singhal K, Tu T, Gottweis J, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:230509617.* 2023.
9. OpenAI. GPT-4 Technical Report. *ArXiv.* 2023;abs/2303.08774.
10. 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.*
11. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. *JAMA ophthalmology.* 2023.
12. Cai LZ, Shaheen A, Jin A, et al. Performance of Generative Large Language Models on Ophthalmology Board Style Questions. *American Journal of Ophthalmology.* 2023.
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. *medRxiv.* 2023:2023.2004. 2003.23287957.
14. Raimondi R, Tzoumas N, Salisbury T, et al. Comparative analysis of large language models in the Royal College of Ophthalmologists fellowship exams. *Eye.* 2023.
15. Mihalache A, Huang RS, Popovic MM, Muni RH. Performance of an Upgraded Artificial Intelligence Chatbot for Ophthalmic Knowledge Assessment. *JAMA Ophthalmology.* 2023.
16. OpenAI. GPT-4. 2023.
17. 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.* 2023:100324.
18. Taib F, Yusoff MSB. Difficulty index, discrimination index, sensitivity and specificity of long case and multiple choice questions to predict medical students' examination performance. *Journal of Taibah University Medical Sciences.* 2014;9(2):110-114.

19. OpenAI. ChatGPT: Optimizing Language Models for Dialogue. https://openai.com/blog/chatgpt/. Published 2022. Accessed December 7, 2022.
20. Chia MA, Keane PA. Exploring the Test-Taking Capabilities of Chatbots—From Surgeon to Sommelier. *JAMA ophthalmology.* 2023.
21. Chen L, Zaharia M, Zou J. How is ChatGPT's behavior changing over time? *arXiv preprint arXiv:230709009.* 2023.

Supplemental Material

British Journal of Ophthalmology

Topical Collection: Artificial Intelligence

Capabilities of GPT-4 in Ophthalmology: An Analysis of Model Entropy and Progress Towards Human-Level Medical Question Answering

Fares Antaki, MDCM, FRCSC[1,2,3,4], Daniel Milad, MD[4], Mark A. Chia, MBBS, MMed[1,2], Charles-Édouard Giguère, MSc[5], Samir Touma, MD, CM[4], Jonathan El-Khoury, MD[4], Pearse A. Keane, MD[1,2,*], Renaud Duval, MD[4,*]

1. Institute of Ophthalmology, University College London, London, United Kingdom
2. Moorfields Eye Hospital NHS Foundation Trust, London, United Kingdom
3. The CHUM School of Artificial Intelligence in Healthcare, Montreal, Quebec, Canada
4. Department of Ophthalmology, Université de Montréal, Montreal, Quebec, Canada
5. Research Centre of the Institut Universitaire en Santé Mentale de Montréal, Montreal, Quebec, Canada

*Joint senior authors

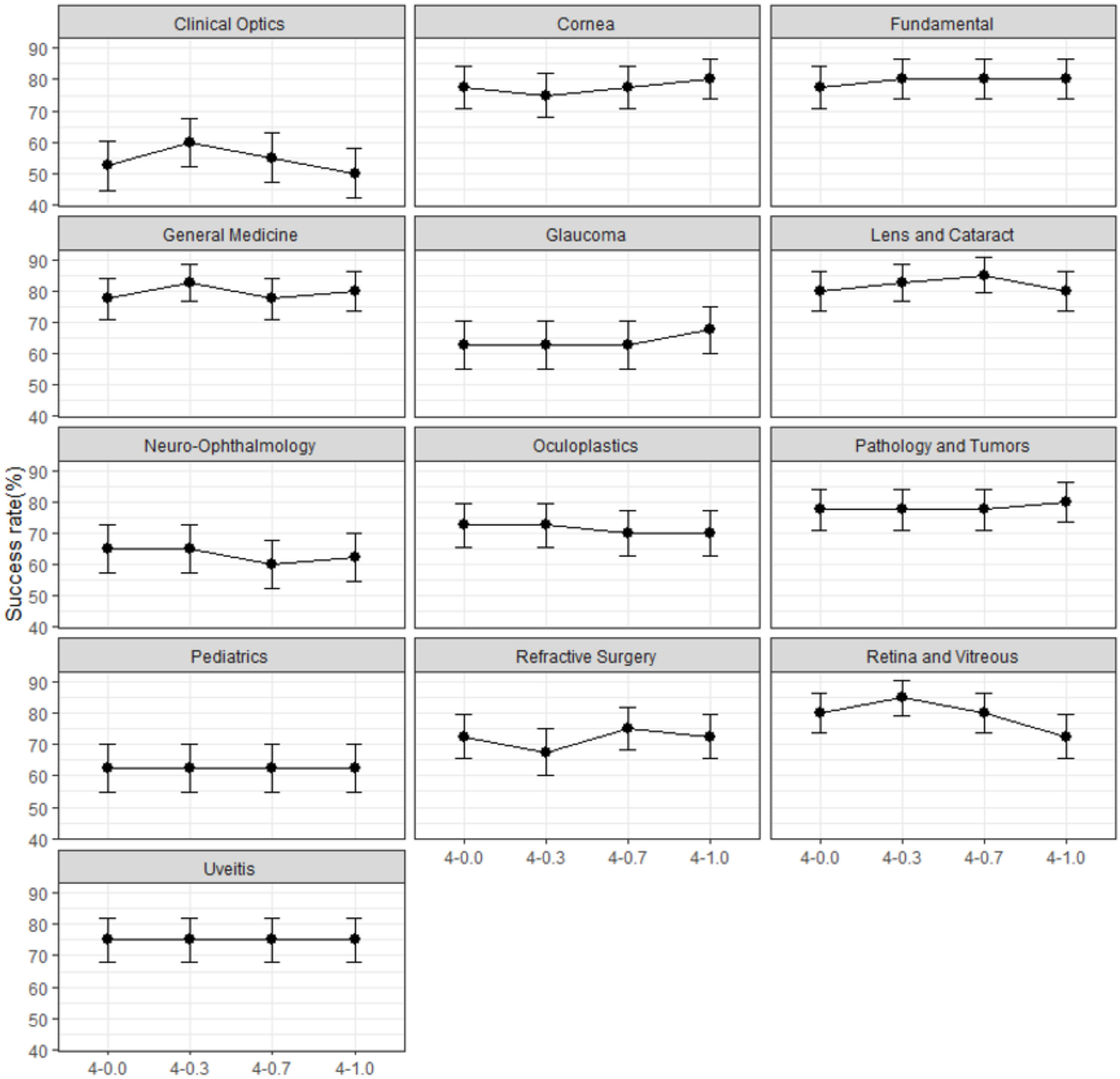
ORCID of the authors: Fares Antaki (0000-0001-6679-7276), Samir Touma (0000-0002-6365-0946), Daniel Milad (0000-0002-0693-3421), Mark A. Chia (0000-0003-0339-5186), Jonathan El-Khoury (0000-0003-3186-2351), Pearse A. Keane (0000-0002-9239-745X), Renaud Duval (0000-0002-3845-3318)

Corresponding authors:
Renaud Duval, MD, CM
Department of Ophthalmology, Université de Montréal, Montreal, Quebec, Canada.
renaud.duval@gmail.com

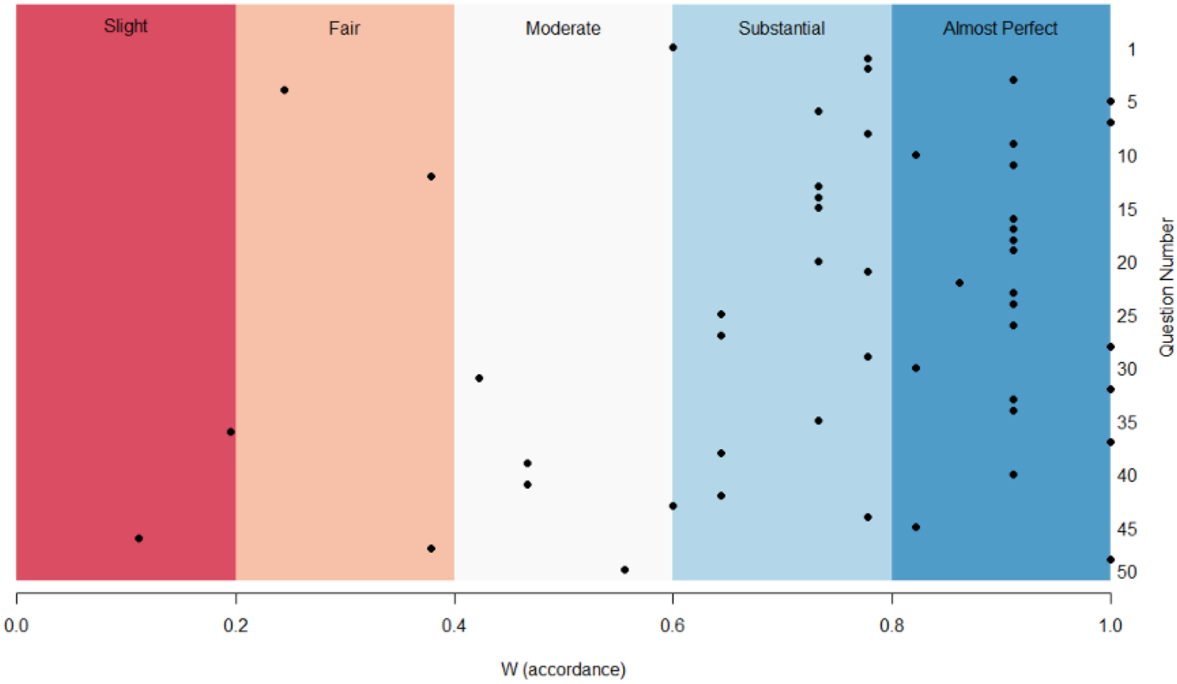
Pearse A. Keane, MD
Institute of Ophthalmology, University College London, London, United Kingdom
pearse.keane1@nhs.net

Section/ Model	GPT-4-0	GPT-4-0.3	GPT-4-0.7	GPT-4-1
Clinical Optics	21 (52.5 %)	33 (82.5 %)	24 (60.0 %)	25 (62.5 %)
Cornea	24 (60.0 %)	31 (77.5 %)	25 (62.5 %)	29 (72.5 %)
Fundamentals	22 (55.0 %)	32 (80.0 %)	29 (72.5 %)	27 (67.5 %)
General Medicine	20 (50.0 %)	25 (62.5 %)	29 (72.5 %)	30 (75.0 %)
Glaucoma	31 (77.5 %)	25 (62.5 %)	28 (70.0 %)	29 (72.5 %)
Lens and Cataract	30 (75.0 %)	25 (62.5 %)	28 (70.0 %)	32 (80.0 %)
Neuro-Ophthalmology	31 (77.5 %)	27 (67.5 %)	31 (77.5 %)	34 (85.0 %)
Oculoplastic	32 (80.0 %)	32 (80.0 %)	31 (77.5 %)	32 (80.0 %)
Pathology and Tumors	31 (77.5 %)	33 (82.5 %)	31 (77.5 %)	29 (72.5 %)
Pediatrics	32 (80.0 %)	34 (85.0 %)	32 (80.0 %)	30 (75.0 %)
Refractive Surgery	32 (80.0 %)	32 (80.0 %)	25 (62.5 %)	30 (75.0 %)
Retina and Vitreous	32 (80.0 %)	26 (65.0 %)	25 (62.5 %)	30 (75.0 %)
Uveitis	31 (77.5 %)	26 (65.0 %)	25 (62.5 %)	30 (75.0 %)

Supplemental Table 1. Comparative Performance Analysis of Various GPT-4 Models Across Exam Sections. This table illustrates the comparative performance of different GPT-4 models across various exam sections. Performance scores are derived from the combined accuracy of each model on the BCSC and OphthoQuestions datasets. Scores are expressed both as raw values (out of a maximum of 40) and as percentages for easier comparison.



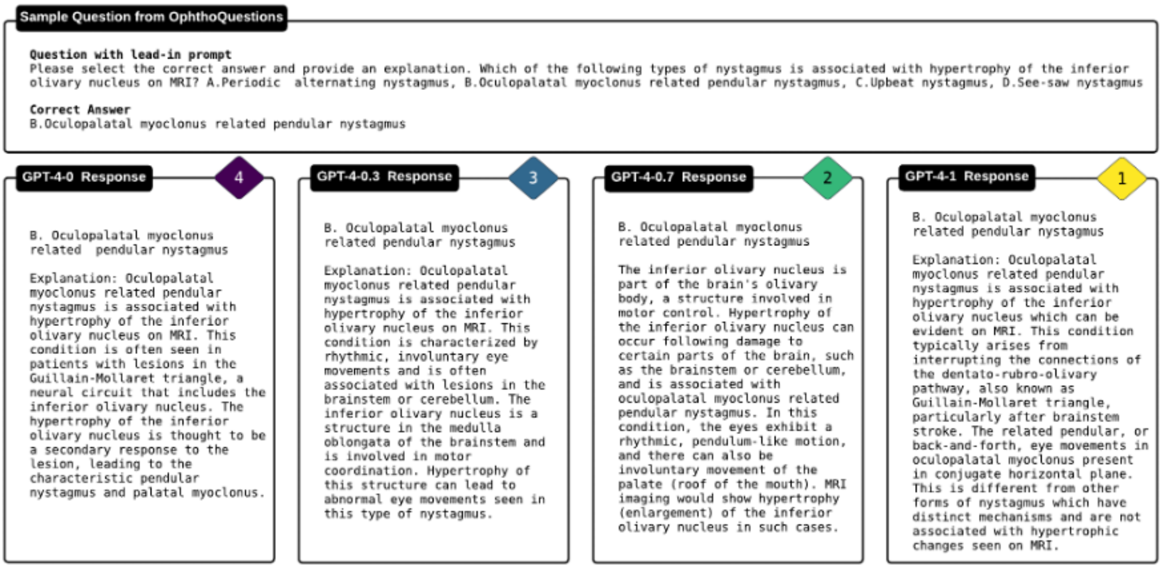
Supplemental Figure 1. Comparative Performance Analysis of GPT-4 Models Across Exam Sections with Standard Errors. This figure illustrates the comparative performance of various GPT-4 models across different exam sections, complete with standard error indications. There were anecdotal variations in performance across exam sections. No statistically significant differences were detected, as evidenced by a Chi-square value of 46.1 and a p-value of 0.27.



Supplemental Figure 2. Kendall's W Representation of Clinician Agreement for 50 Evaluated Questions. This figure displays the Kendall's W representation for the 50 questions independently rated by three clinicians. The figure reveals that for 80% of the questions, the level of agreement between the raters ranges from substantial to perfect, indicating a high degree of inter-rater reliability.

GPT-4 model	**0.0**	**0.3**	**0.7**	**1.0**
Mean (SD) rank	**3.4 (0.7)**	**2.4 (0.8)**	**2.1 (0.8)**	**2.1 (0.9)**
Contrasts	**Difference**	**SE**	**t-value**	**p-value**
0 - 0.3	0.967	0.163	5.930	**<.0001**
0 - 0.7	1.293	0.163	7.930	**<.0001**
0 - 1	1.293	0.163	7.930	**<.0001**
0.3 - 0.7	0.327	0.163	2.000	0.19
0.3 - 1	0.327	0.163	2.000	0.19
0.7 - 1	0.000	0.163	0.000	1.00

Supplemental Table 2. Comparative Analysis of Clinician Ratings for Four Models of GPT-4. This table presents the mean ratings, derived from the evaluations of three clinicians, for each of the four GPT-4 models across 50 questions. While all models demonstrate varying levels of preference, GPT-4-0 was ranked worst compared to the three other models. An ANOVA analysis shows statistically significant differences among the model rankings ($F=28.2$; $p<0.001$). Subsequent contrasts reveal that significant differences exist only between GPT-4-0 and the other three models: GPT-4-0.3, GPT-4-0.7, and GPT-4-1. GPT-4-0 is the least preferred model compared to each of the other models.

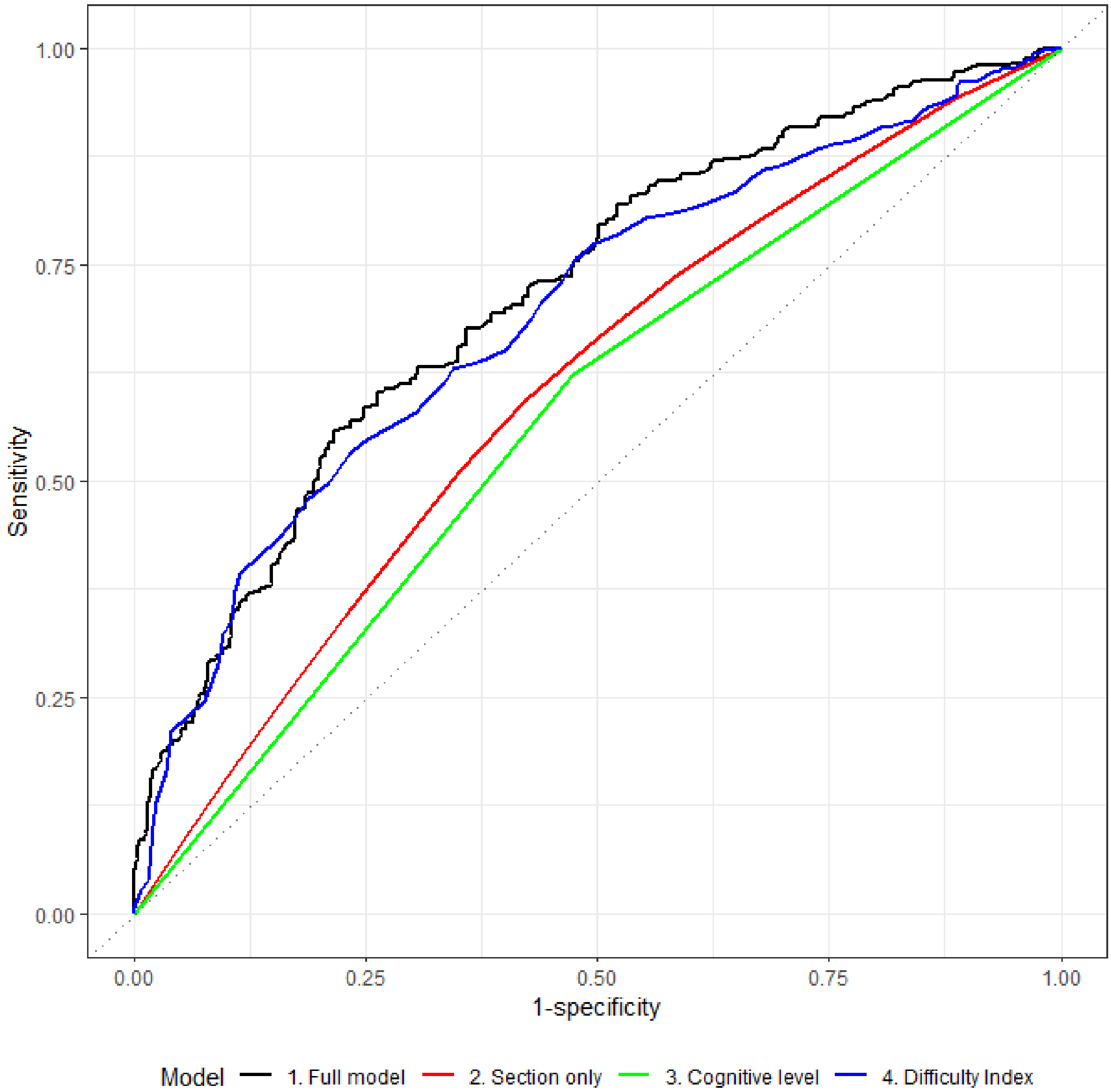


Supplemental Figure 3. Illustrative Example Comparing Responses and Rankings Across Different GPT-4 Models. This figure presents an example of responses generated by the GPT-4 models to a question from the Neuro-Ophthalmology section of OphthoQuestions. This question was of moderate difficulty, with a 67% correct response rate among human responders on OphthoQuestions. All four models provided the correct answer. The response of GPT-4-1 was favoured by clinicians. The GPT-4-1 response is notable for its structured layout: explicitly stating the diagnosis, presenting the pathophysiology (mentioning the important 'Guillain-Mollaret' triangle), before discussing a potential aetiology and clinical findings. However, the explanation is inaccurate (hallucination), as this type of nystagmus is typically described as 'pendular and vertical' rather than horizontal.

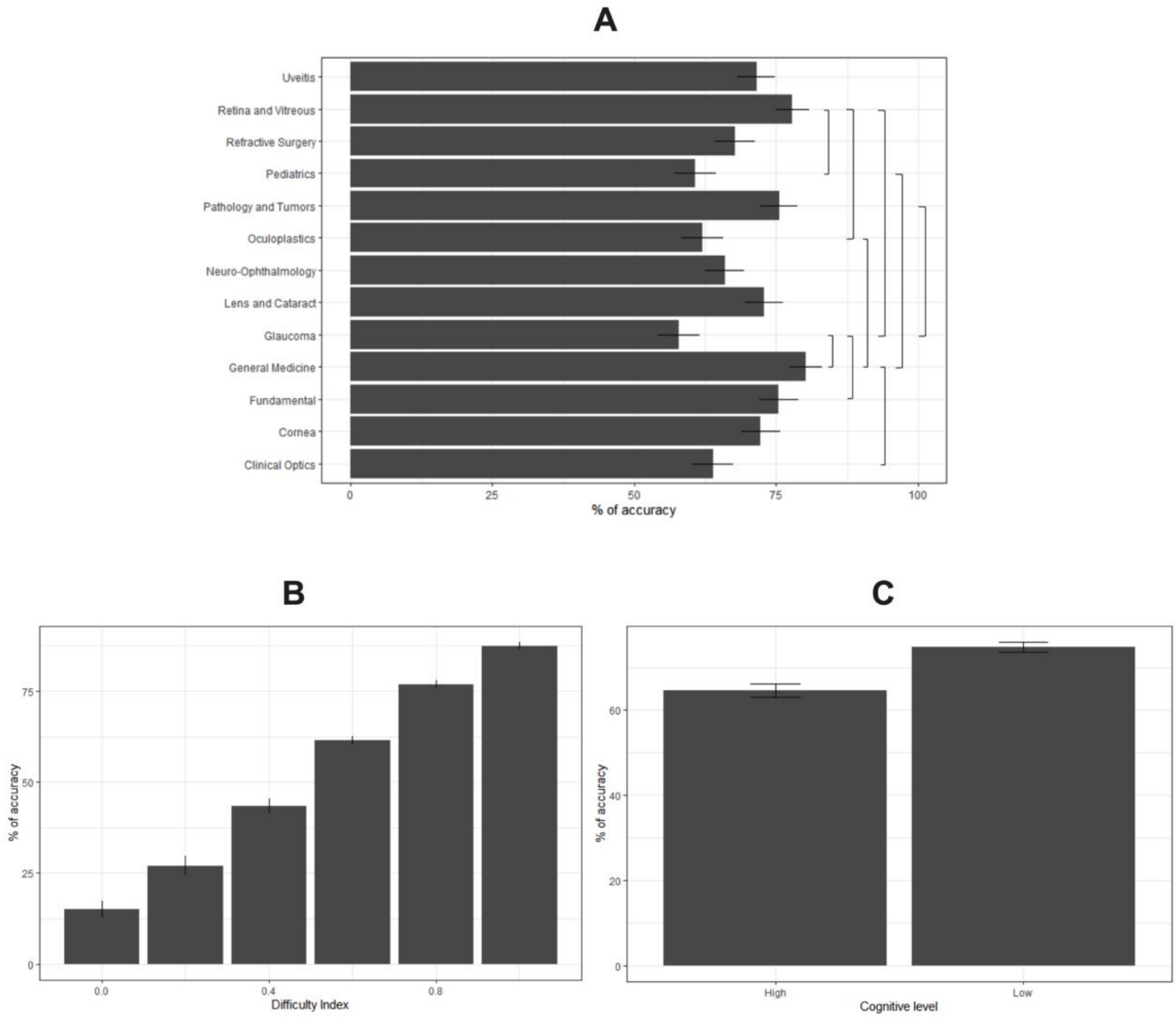
Section/ Model	GPT-3.5	GPT-4-0.3	OR	OR 95% CI

Clinical Optics	19 (47.5 %)	24 (60.0 %)	1.66	[0.54, 5.07]
Cornea	25 (62.5 %)	30 (75.0 %)	1.80	[0.54, 6.03]
Fundamentals	27 (67.5 %)	32 (80.0 %)	1.93	[0.41, 9.14]
General medicine	28 (70.0 %)	33 (82.5 %)	2.02	[0.53, 7.66]
Glaucoma	20 (50.0 %)	25 (62.5 %)	1.67	[0.54, 5.15]
Lens & cataract	20 (50.0 %)	33 (82.5 %)	4.71	[1.53, 14.5]
Neuro-Ophthalmology	16 (40.0 %)	26 (65.0 %)	2.79	[0.90, 8.65]
Oculoplastic	16 (40.0 %)	29 (72.5 %)	3.95	[1.22, 12.8]
Pathology & tumors	26 (65.0 %)	31 (77.5 %)	1.85	[0.65, 5.29]
Pediatrics	19 (47.5 %)	25 (62.5 %)	1.84	[0.75, 4.53]
Refractive surgery	19 (47.5 %)	27 (67.5 %)	2.30	[0.71, 7.39]
Retina & vitreous	26 (65.0 %)	34 (85.0 %)	3.05	[1.05, 8.90]
Uveitis	23 (57.5 %)	30 (75.0 %)	2.22	[0.87, 5.67]

Supplemental Table 3. Comparative Performance Analysis of GPT-3.5 and GPT-4-0.3 Across Exam Sections. This table illustrates the comparative performance of different GPT-3.5 and GPT-4 across various exam sections. Performance scores are derived from the combined accuracy of each model on the BCSC and OphthoQuestions datasets. Scores are expressed both as raw values (out of a maximum of 40) and as percentages for easier comparison. Although the odds ratios (OR) are positive and ranging from 1.66 to 4.71, the confidence intervals are too large to be significant. This is due to the small sample size in each section.



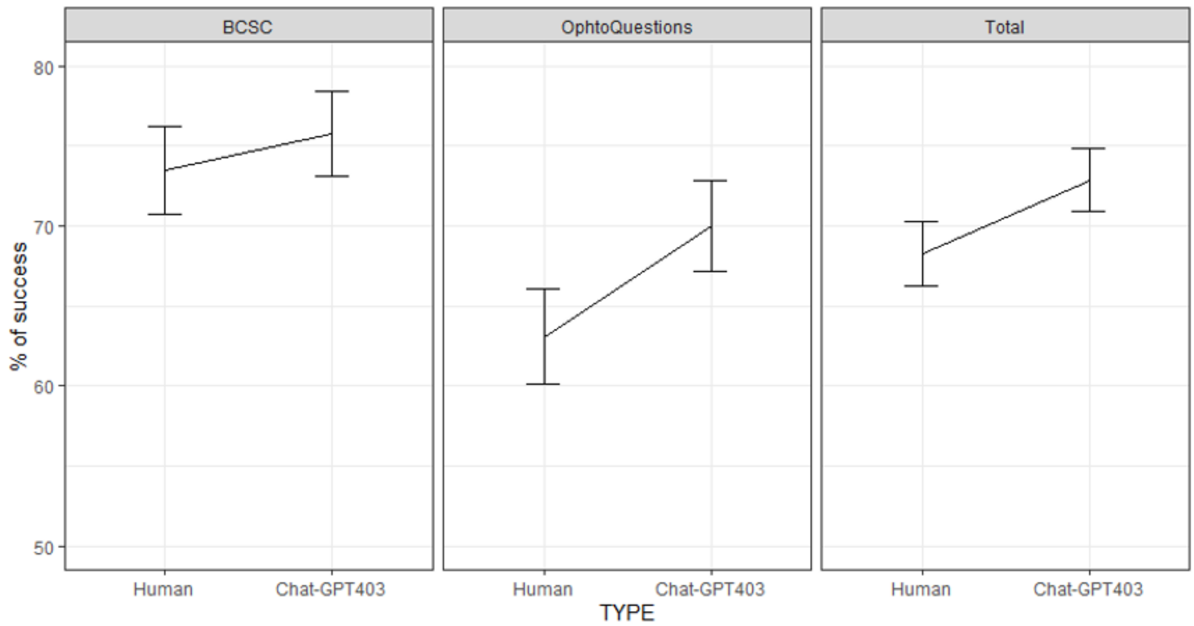
Supplemental Figure 4. ROC Curves for the Logistic Regression Model Considering Various Variables. This figure displays the ROC curves for the logistic regression model. The model predicts answer accuracy using exam section, cognitive level, and difficulty index as variables. The curves represent the overall model (AUC = 0.72), the model using difficulty index (AUC = 0.69), the model considering exam sections only (AUC = 0.60) and the model based on cognitive level (AUC = 0.56).



Supplemental Figure 5. Post hoc analysis using Tukey's test to isolate the effect of exam section, difficulty index and cognitive level using all questions for GPT-4-0.3. (A) Bar plot of the percentage of accuracy by exam section. The blue square brackets identify the significant differences. The significant contrasts were the following: Clinical Optics - General Medicine (estimate -0.164, p=0.021), Fundamentals - Glaucoma (estimate 0.176, p=0.027), General Medicine - Glaucoma (estimate 0.224, p<0.001), General Medicine - Oculoplastics (estimate 0.182, p=0.006), General Medicine - Pediatrics (estimate 0.195, p=0.002), Glaucoma - Pathology and Tumors (estimate -0.176, p=0.021), Glaucoma - Retina and Vitreous (estimate -0.200, p=0.002), Oculoplastics - Retina and Vitreous (estimate -0.158, p=0.045), Pediatrics - Retina and Vitreous (estimate -0.171, p=0.017). **(B)** Predicted percentage of accuracy by difficulty index. Accuracy increased with increasing difficulty index (easier questions). **(C)** Predicted percentage of accuracy by cognitive level. Accuracy increased with low cognitive level questions (recall-style questions).

	BCSC		OphthoQuestions	
	GPT-4-0.3	**Human**	**GPT-4-0.3**	**Human**
General Medicine	70	73	95	67
Fundamentals	80	69	80	67
Clinical Optics	60	64	60	54
Pathology and Tumors	80	71	75	57
Neuro-Ophthalmology	65	73	65	64
Pediatrics	80	74	45	58
Oculoplastics	75	75	70	72
Cornea	75	78	75	62
Uveitis	85	75	65	63
Glaucoma	70	73	55	65
Lens and Cataract	85	77	80	68
Retina and Vitreous	90	75	80	60
Refractive Surgery	70	76	65	62
Mean Accuracy	75.8%	73.3%	70%	63.0%

Supplemental Table 4. GPT-4-0.3 performance per exam section compared to historical human performances from the BCSC and OphthoQuestions sets.



Supplemental Figure 6. Comparative Performance Analysis of GPT-4 and Historical Human Performance with Standard Errors. The performance of the GPT-4-0.3 model consistently exceeds that of historical human performance using both the BCSC (OR 1.13, p=0.55) and OphthoQuestions (OR 1.82, p=0.09) datasets, as well as in the overall analysis (OR 1.25, p=0.10). However, these differences did not reach statistical significance.