

## Letter

# Performance of leading large language models in MRCPsych-style examination questions: cross-sectional survey

Richard C. Armitage

**Keywords**

Education and training; medical technology; general adult psychiatry; child and adolescent psychiatry; old age psychiatry.

**Copyright and usage**

© The Author(s), 2025. Published by Cambridge University Press on behalf of Royal College of Psychiatrists.

The potential for ChatGPT and other large language models (LLMs) to assist and improve the delivery of healthcare has been widely discussed. As the capabilities of these technologies have rapidly increased, their competence in answering medical examination questions has been assessed in various specialties.<sup>1</sup> For example, GPT-4 has been shown to perform to varying standards in specialty examination questions relevant to ophthalmology,<sup>2</sup> orthopaedics,<sup>3</sup> internal medicine,<sup>4</sup> general surgery,<sup>5</sup> radiology,<sup>6</sup> paediatrics<sup>7</sup> and dermatology.<sup>8</sup>

However, beyond the predecessors of GPT-4 (such as ChatGPT-3.5), a substantially smaller number of other LLMs have been subjected to medical specialty examination questions.<sup>1</sup> Furthermore, the current leading publicly available LLMs, including the latest iteration of ChatGPT, have not been subjected to such questions.

As of January 2025, ChatGPT o1 (OpenAI, California, USA; <https://openai.com/o1/>), Gemini 1.5 Pro (Google, California, USA; <https://deepmind.google/models/gemini/pro/>), Claude 3.5 Sonnet (Anthropic, California, USA; <https://www.anthropic.com/claude/sonnet>) and Grok-2 (xAI, California, USA; <https://x.ai/news/grok-2>) are among the leading LLMs that are non-experimental, publicly available and accessible via user interface. All four of these LLMs can analyse text, images and documents, and all display the model's reasoning within the answers they provide. All lie behind a paywall, other than Grok-2, for which initial prompts are free but subsequent prompts require payment. Although ChatGPT o1 and Claude 3.5 Sonnet do not have live internet access (which restricts their responses to the data they were trained on), Gemini 1.5 Pro and Grok-2 do have real-time internet access. To test the capabilities of these leading LLMs in the domain of psychiatry education, they were each subjected to a series of questions of the style used in Member of the Royal College of Psychiatrist (MRCPsych) examinations on 18 January 2025.

**Method**

Sample questions from MRCPsych Papers A and B (both written papers) are available to College members via their online portal (<https://www.rcpsych.ac.uk/training/exams/preparing-for-exams>) and to non-members via the Examinations team.

ChatGPT o1, Gemini 1.5 Pro, Claude 3.5 Sonnet and Grok-2 were each prompted with the instruction 'Answer the following questions as if you were a psychiatrist in the UK' and subsequently tasked to answer the same 21 sample questions (five from Paper A, 16 from Paper B). Each question's textual information was copied into each model's context window, and any images that formed part of a question (such as graphs) were attached to the context window

so that each model had access to the full information for each question. No further prompts were given to any model. Questions were attempted sequentially. Each model's answers were collected and marked according to the answer provided with the sample questions at their source. Each model's total score was calculated as a percentage.

**Results**

Of the 21 questions, 19 contained textual information only, and two questions required interpretation of a graph. All models provided answers for all 21 questions. The total scores achieved by Claude 3.5 Sonnet, ChatGPT o1, Gemini 1.5 Pro and Grok-2 were 90.5%, 85.7%, 85.7% and 85.7%, respectively. Although all the questions were multiple choice or extended matching item questions, all models provided comprehensive explanations of the reasoning that supported their answers. Gemini 1.5 Pro and Claude 3.5 Sonnet generally provided the longest reasoning, whereas ChatGPT o1 and Grok-2 provided more concise answers.

For all models, incorrect answers were largely due to factual errors which were incorporated into the models' reasoning rather than errors in the reasoning itself. All models answered incorrectly a question that required interpretation of a forest plot, but all answered correctly the subsequent question that required interpretation of the same forest plot. Incorrect answers were stated by all models with equal confidence as correct answers.

**Discussion**

To the author's knowledge, this is the first study to assess and compare the capabilities of leading LLMs (as of January 2025) in answering psychiatry examination questions. All models performed impressively well. Claude 3.5 Sonnet scored only one more mark than the other models, suggesting that all the models are capable of comparable performance in answering psychiatry examination questions.


Both correct and incorrect answers were stated with equal confidence by all models. This is concerning, because it implies the models cannot discern uncertainty in their assumptions, which should lower the user's confidence in the validity of all the models' responses (although, in this study, those responses were largely correct). This strengthens previous warnings that such models should not replace but rather augment and strengthen clinician decision-making.<sup>9</sup>

Strengths of the study include the use of multiple LLMs to compare performances, the use of currently leading publicly available LLMs, and the unlikelihood of the examination questions

featuring in the models’ training sets (because they lie within a portal, which is likely to have prevented even those models with live internet access from having ‘seen’ the questions beforehand). The study could have been strengthened by subjecting the models to a larger number of examination questions and by incorporating more questions with non-textual data such as tables of results, graphs and clinical images to more thoroughly assess the broadness of the models’ analytic abilities. These features should form part of future research on this subject.

The performance of all models in this study was undeniably impressive and further strengthens the case that LLMs could be used to assist and improve the delivery of clinical medicine, in this case, in psychiatry. As LLMs continue to improve, those with real-time internet access might develop an advantage over those without it when posed questions about latest clinical guidelines, as such guidelines might not feature in training data and so would be unavailable to those without internet access.

However, despite their rapidly increasing capabilities, LLMs are unlikely to ever replace the psychiatrist. Among other reasons,<sup>9</sup> this is because of the unstructured nature of information presentation in real-world psychiatry practice, which does not reflect the succinct packages of information presented in MRCPsych-style questions. As such, rather than clinicians deferring to LLMs, it is becoming increasingly clear that they can use these technologies to augment and to support their practice, particularly to bolster the continuously evolving knowledge base requirements of clinical practice.

**Richard C. Armitage** , BMBS, BMedSci, MPH, MA, Academic Unit of Population and Lifespan Sciences, School of Medicine, University of Nottingham, Nottingham, UK

Email: [richard.armitage@nhs.net](mailto:richard.armitage@nhs.net)

First received 7 Feb 2025, final revision 1 Apr 2025, accepted 3 Jul 2025

Funding

None.

Declaration of interest

None.

References

- 1 Zong H, Wu R, Cha J, Wang J, Wu E, Li J, et al. Large language models in worldwide medical exams: platform development and comprehensive analysis. *J Med Int Res* 2024; **26**: e66114.
- 2 Thirunavukarasu AJ, Mahmood S, Malem A, Foster WP, Sanghera R, Hassan R, et al. Large language models approach expert-level clinical knowledge and reasoning in ophthalmology: a head-to-head cross-sectional study. *PLOS Digit Health* 2024; **3**: e0000341.
- 3 Saad A, Iyengar KP, Kurisunkal V, Botchu R. Assessing ChatGPTs ability to pass the FRCS orthopaedic part A exam: a critical analysis. *Surgeon* 2023; **21**: 263–6.
- 4 Maitland A, Fowkes R, Maitland S. Can ChatGPT pass the MRCP (UK) written examinations? Analysis of performance and errors using a clinical decision-reasoning framework. *BMJ Open* 2024; **14**: e080558.
- 5 Chan J, Dong T, Angelini GD. The performance of large language models in intercollegiate Membership of the Royal College of Surgeons examination. *Ann R Coll Surg Engl* 2024; **106**: 700–4.
- 6 Ariyaratne S, Jenko N, Davies AM, Iyengar KP, Botchu R. Could ChatGPT pass the UK radiology fellowship examinations? *Acad Radiol* 2024; **31**: 2178–82.
- 7 Armitage R. Performance of GPT-4 in Membership of the Royal College of Paediatrics and Child Health-style examination questions. *BMJ Paediatr Open* 2024; **8**: e002575.
- 8 Passby L, Jenko N, Wernham A. Performance of ChatGPT on Specialty Certificate Examination in Dermatology multiple-choice questions. *Clin Exp Dermatol* 2024; **49**: 722–7.
- 9 Armitage R. Large language models must serve clinicians, not the reverse. *Lancet Infect Dis* 2024; **24**: 453–4.