Universal precautions required

Artificial intelligence takes on the Australian Medical Council's trial examination

Oliver Kleinig, Joshua G Kovoor, Aashray K Gupta, Stephen Bacchi

Background and objective
The potential of artificial intelligence in medical practice is increasingly being investigated. This study aimed to examine OpenAI's ChatGPT in answering medical multiple choice questions (MCQ) in an Australian context.

Methods
We provided MCQs from the Australian Medical Council's (AMC) medical licencing practice examination to ChatGPT. The chatbot's responses were graded using AMC's online portal. This experiment was repeated twice.

Results
ChatGPT was moderately accurate in answering the questions, achieving a score of 29/50. It was able to generate answer explanations to most questions (45/50). The chatbot was moderately consistent, providing the same overall answer to 40 of the 50 questions between trial runs.

Discussion
The moderate accuracy of ChatGPT demonstrates potential risks for both patients and physicians using this tool. Further research is required to create more accurate models and to critically appraise such models.

ARTIFICIAL INTELLIGENCE (AI) is hoped to enhance medical practice across many domains.[1] One way in which AI might influence medicine is through chatbots answering questions about diseases and treatments. OpenAI's ChatGPT generates text-based answers in response to text prompts, such as questions (https://chat.openai.com/chat). This chatbot was trained on a variety of texts published up to 2021. The tool has 'gone viral' online,[2] and has even been used to write a journal article.[3] ChatGPT's accuracy in answering multiple choice questions (MCQs) from the US Medical Licencing Examination has been demonstrated to be approximately 60%.[4,5] However, the generalisability of such performance in an Australian context is uncertain.

This study aimed to assess ChatGPT's (GPT-3 version) performance in responding to practice MCQs published by the Australian Medical Council (AMC; https://trial-exam.amc.org.au/). This examination is a barrier requirement for overseas doctors to practice in Australia. Because the examination is standard set yearly,[6] the pass mark is not published.

Methods

On 11 December 2022, 50 MCQs were accessed from the AMC's online practice examination (https://trial-exam.amc.org.au/). An online search of selected questions indicated these MCQs were not indexed by Google. Therefore, it was considered unlikely that ChatGPT was trained on these questions, increasing the validity of this method of performance evaluation.

Each question had five option answers. The questions assessed diagnostic reasoning, interpretation of investigations and management. There was one question requiring the calculation of sensitivity and specificity values. Five questions included images, four of which were radiological scans for interpretation by the test taker, and one was a photograph of a melanoma given only for context.

Question stems with images had these images removed before ChatGPT analysis. This method was used because, at the time of writing, ChatGPT accepts only text input. Therefore, the chatbot was provided only the written portion of questions.

On a computer cleared of cookies to reset the chatbot's memory, test examination questions from AMC's online system were copied into the input field for the chatbot. If ChatGPT gave a clear answer, it was entered into the AMC's online software, which graded each selection at the conclusion of the test. A definitive answer was defined as the chatbot either stating an option answer letter in its response or stating that the text in an option answer was correct. In the event of an unclear or incomplete answer (eg the chatbot declining to answer a question due to insufficient information), the 'regenerate response' button was used to give the

© The Royal Australian College of General Practitioners 2023

chatbot a second attempt. If the response once again failed to provide a definitive answer, the question was automatically scored as 'incorrect'.

To assess consistency, the above procedure was performed twice. In addition, questions were divided into the subgroups 'image-dependent/independent' and 'response provided/absent' for further descriptive analysis. These categories were developed retrospectively. Because only publicly available data and algorithms were used for this study, ethics approval was not required.

Results

The chatbot provided a definitive answer to 48 of the 50 questions. In the two questions it declined to answer, the chatbot stated it had insufficient information. Of the 50 questions, each with five option answers, the chatbot scored 58% (29/50). This performance is notable because random answer selection would result in a score close to 20% (10/50).

ChatGPT provided an explanation for 43 of the 48 answered questions. For example, in answering a question about the most appropriate antibiotic therapy, ChatGPT listed all five drug options and gave reasons for and against the use of each (Appendix 1, Answer 1). For five questions, the chatbot did not provide any explanation, giving only a letter in response. The performance on these questions was exactly 20%, consistent with random responses. For question stems without images and where explanations were given, the chatbot scored 61% (25/41).

The chatbot frequently provided responses for image-dependent questions, of which three-quarters of the provided answers were correct. Although the subset was of a small sample size, this result was notable because the chatbot only received the question text and not the accompanying image. For example, on a question including a nuclear medicine scan, ChatGPT generated a response including a description of increased tracer uptake in an image that had not been provided to the algorithm (Appendix 1, Answer 2).

In the second trial, ChatGPT achieved an identical score of 29/50. However, answers between tests were not entirely consistent, with five correct questions in the first trial being answered incorrectly in the repeat examination (and vice versa). For example, a question about multiple subcutaneous lumps was misdiagnosed as neurofibromatosis type 1 during the first trial but was correctly described as subcutaneous lipomas on the second attempt (Appendix 1, Answer 3).

Discussion

This study demonstrates the current level of performance of a popular AI model in answering Australian medical questions. Combined with other studies assessing performance on different examinations,[4,5] it seems that OpenAI's ChatGPT is able to generate answers to a selection of medical MCQs to a moderate level of accuracy. The moderate accuracy of this chatbot highlights both the risks associated with applying current AI technology to answer medical MCQs and the opportunities that natural language processing has in medicine.

This study has demonstrated that ChatGPT, an example of a large language model, might generate answers to medical MCQs for which incomplete information has been provided. Importantly, this finding highlights a significant limitation of ChatGPT. These models are trained to mimic the way a human would express themselves.[7] Although OpenAI's ongoing research is improving the chatbot's accuracy,[8] this and other studies demonstrate the tool can provide incorrect information in a 'confident' manner and project certainty where uncertainty remains. For example, in the instance of the nuclear medicine scan, ChatGPT asserted there was an 'increased uptake in the region of the second metatarsal', without having been provided with the scan. Thus, patients, medical students and clinicians must exercise caution when using this tool because it can be non-obviously misleading.

Furthermore, such models might be less accurate when analysing specialised topics or providing information on rarer topics because there might be fewer human sources for it to analyse and mimic. However, more research is required to determine whether this is a genuine limitation.

In this study, ChatGPT was occasionally inconsistent in its answers to questions. These inconsistencies are likely a reflection of the probabilistic nature of the large language model. In addition, the chatbot might be influenced based on the order in which the examination portal presented the questions. Such chatbots are designed to change responses based on context, thus prior responses might influence subsequent answers. It is unlikely the chatbot learnt from its previous attempts because the answers were not provided to it.

It is plausible that future general practice trainees might attempt to use this or similar tools to generate answers to practice examination questions. This study has shown that this method of learning is fraught and further evidence-based verification of any ChatGPT-generated answers is required. In the short term, the use of closed-book examinations, as already operational for The Royal Australian College of General Practitioners' fellowship examinations,[9] might mitigate the risks ChatGPT poses to the academic integrity of medical licensing exams.

As the public becomes increasingly aware of AI such as ChatGPT, there is a risk of patients using these tools for self-education and self-diagnosis. This research, highlighting the moderate accuracy of ChatGPT, might be useful when counselling patients on the dangers of this approach. Similarly, physician use of this tool to answer questions must recognise ChatGPT's limitations. The use of this AI might provide incorrect diagnoses and incorrect management strategies. This concern is also in the context of ChatGPT's training data containing work only published in or before 2021 (https://chat.openai.com/chat). Therefore, recent changes to practice, such as the 2022 heart failure guidelines,[10] would not be included in ChatGPT's answers. There is the possibility that patients who witness their doctor using this tool to answer questions might have reduced confidence in the medical abilities of their health provider.

© The Royal Australian College of General Practitioners 2023

Limitations

A small sample size and a single source of questions are limitations of this study. The AMC examination was selected because it is representative of Australian practice. However, it should be noted that the findings of this study might not generalise to individual Australian subspecialities. The answers provided by the AMC examination were accepted as the gold standard and were not subject to scrutiny in this study. In addition, the categories used for analysis, such as whether questions were image dependent, were developed retrospectively.

Conclusion

This application of ChatGPT demonstrates the moderate performance that might be achieved by such algorithms on Australian medical examinations. Although a future where a machine can explain medical knowledge with expert precision is not immediately foreseeable, the performance of this chatbot in answering medical questions demonstrates the potential for generative AI in medicine. Noting the limitations in ChatGPT's performance is necessary when counselling patients and trainees on their possible use of this AI for education, and the precautions required if doing so.

Authors

Oliver Kleinig MBBS III, Faculty of Health and Medical Sciences, The University of Adelaide, Adelaide, SA; Royal Adelaide Hospital, Adelaide, SA

Joshua Kovoor MS, Surgical Resident Medical Officer, Department of General Surgery, Queen Elizabeth Hospital, Woodville South, SA

Aashray K Gupta MBBS, MS, Cardiothoracic Registrar, Department of Cardiothoracic Surgery, Gold Coast University Hospital, Gold Coast, QLD

Stephen Bacchi MBBS, PhD, Neurology Registrar, Department of Neurology, Royal Adelaide Hospital, Adelaide, SA; Neurology Registrar, Department of Neurology, Flinders University, Bedford Park, SA

Competing interests: None.

Funding: None.

Provenance and peer review: Not commissioned, externally peer reviewed.

Correspondence to:
oliver.kleinig@student.adelaide.edu.au

References

1. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022;28(1):31–38. doi: 10.1038/s41591-021-01614-0.
2. Hern A. TechScape: Meet ChatGPT, the viral AI tool that may be a vision of our weird tech future. The Guardian, 6 December 2022. Available at www.theguardian.com/technology/2022/dec/06/meet-chatgpt-the-viral-ai-tool-that-may-be-a-vision-of-our-weird-tech-future [Accessed 10 July 2023].
3. Perlman AM. The implications of OpenAI's assistant for legal services and society. SSRN, 2022. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4294197 [Accessed 10 July 2023].
4. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLoS Digit Health 2023;2(2):e0000198. doi: 10.1371/journal.pdig.0000198.
5. Gilson A, Safranek CW, Huang T, 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:e45312. doi: 10.2196/45312.
6. Specifications MCQE. Australian Medical Council Limited 2020;2020:12. Australian Medical Council. (2022). Multiple Choice Question Examination Specifications (Version 0.6). Available at: www.amc.org.au/wp-content/uploads/2022/01/2022-01-14-MCQ-Specifications.V0.6.pdf [Accessed 12 July 2023].
7. Floridi L, Chiriatti M. GPT-3: Its nature, scope, limits, and consequences. Minds Mach 2020;30(4):681–94. doi: 10.1007/s11023-020-09548-1.
8. Southern MG. OpenAI's ChatGPT update brings improved accuracy. Search Eng J, 2023. Available at https://www.searchenginejournal.com/openai-chatgpt-update/476116/#close [Accessed 10 July 2023].
9. The Royal Australian College of General Practitioners (RACGP). AKT and KFP guide. RACGP, 2021. Available at www.racgp.org.au/FSDEDEV/media/documents/Education/Registrars/Fellowship%20Pathways/Exams/Examinations-guide.pdf [Accessed 10 July 2023].
10. Sindone AP, De Pasquale C, Amerena J, et al. Consensus statement on the current pharmacological prevention and management of heart failure. Med J Aust 2022;217(4):212–17. doi: 10.5694/mja2.51656.

© The Royal Australian College of General Practitioners 2023