World Academy of Science, Engineering and Technology
International Journal of Educational and Pedagogical Sciences
Vol:18, No:11, 2024

ChatGPT 4.0 Demonstrates Strong Performance in Standardised Medical Licensing Examinations: Insights and Implications for Medical Educators

Authors : K. O'Malley

Abstract : Background: The emergence and rapid evolution of large language models (LLMs) (i.e., models of generative artificial intelligence, or AI) has been unprecedented. ChatGPT is one of the most widely used LLM platforms. Using natural language processing technology, it generates customized responses to user prompts, enabling it to mimic human conversation. Responses are generated using predictive modeling of vast internet text and data swathes and are further refined and reinforced through user feedback. The popularity of LLMs is increasing, with a growing number of students utilizing these platforms for study and revision purposes. Notwithstanding its many novel applications, LLM technology is inherently susceptible to bias and error. This poses a significant challenge in the educational setting, where academic integrity may be undermined. This study aims to evaluate the performance of the latest iteration of ChatGPT (ChatGPT4.0) in standardized state medical licensing examinations. Methods: A considered search strategy was used to interrogate the PubMed electronic database. The keywords ‘ChatGPT’ AND ‘medical education’ OR ‘medical school’ OR ‘medical licensing exam’ were used to identify relevant literature. The search included all peer-reviewed literature published in the past five years. The search was limited to publications in the English language only. Eligibility was ascertained based on the study title and abstract and confirmed by consulting the full-text document. Data was extracted into a Microsoft Excel document for analysis. Results: The search yielded 345 publications that were screened. 225 original articles were identified, of which 11 met the pre-determined criteria for inclusion in a narrative synthesis. These studies included performance assessments in national medical licensing examinations from the United States, United Kingdom, Saudi Arabia, Poland, Taiwan, Japan and Germany. ChatGPT 4.0 achieved scores ranging from 67.1 to 88.6 percent. The mean score across all studies was 82.49 percent (SD= 5.95). In all studies, ChatGPT exceeded the threshold for a passing grade in the corresponding exam. Conclusion: The capabilities of ChatGPT in standardized academic assessment in medicine are robust. While this technology can potentially revolutionize higher education, it also presents several challenges with which educators have not had to contend before. The overall strong performance of ChatGPT, as outlined above, may lend itself to unfair use (such as the plagiarism of deliverable coursework) and pose unforeseen ethical challenges (arising from algorithmic bias). Conversely, it highlights potential pitfalls if users assume LLM-generated content to be entirely accurate. In the aforementioned studies, ChatGPT exhibits a margin of error between 11.4 and 32.9 percent, which resonates strongly with concerns regarding the quality and veracity of LLM-generated content. It is imperative to highlight these limitations, particularly to students in the early stages of their education who are less likely to possess the requisite insight or knowledge to recognize errors, inaccuracies or false information. Educators must inform themselves of these emerging challenges to effectively address them and mitigate potential disruption in academic fora.

Keywords : artificial intelligence, ChatGPT, generative ai, large language models, licensing exam, medical education, medicine, university
Conference Title : ICHME 2024 : International Conference on Healthcare and Medical Education
Conference Location : Paris, France
Conference Dates : November 18-19, 2024