Research Letter

Assessment of Large Language Model Performance on Medical School Essay-Style Concept Appraisal Questions: Exploratory Study

Seysha Mehta[1*], BA; Eliot N Haddad[1*], BS; Indira Bhavsar Burke[2], MHPE, MD; Alana K Majors[1], PhD; Rie Maeda[1], BA; Sean M Burke[2], MD; Abhishek Deshpande[1], MD, PhD; Amy S Nowacki[1], PhD; Christina C Lindenmeyer[1], MD; Neil Mehta[1], MBBS

[1]Cleveland Clinic Lerner College of Medicine, School of Medicine, Case Western Reserve University, Cleveland, OH, United States
[2]Department of Internal Medicine, The University of Texas Southwestern Medical Center, Dallas, TX, United States
*these authors contributed equally

Corresponding Author:

Neil Mehta, MBBS
Cleveland Clinic Lerner College of Medicine
School of Medicine, Case Western Reserve University
9500 Euclid Ave, G10
Cleveland, OH, 44195
United States
Phone: 1 2164456512
Fax: 1 2164451007
Email: mehtan@ccf.org

Abstract

Bing Chat (subsequently renamed Microsoft Copilot)—a ChatGPT 4.0–based large language model—demonstrated comparable performance to medical students in answering essay-style concept appraisals, while assessors struggled to differentiate artificial intelligence (AI) responses from human responses. These results highlight the need to prepare students and educators for a future world of AI by fostering reflective learning practices and critical thinking.

JMIR Med Educ 2025;11:e72034; doi: 10.2196/72034

Keywords: essay-type questions; large language models; generative AI; Microsoft Copilot; artificial intelligence

Introduction

Large language models (LLMs) are of growing interest in medical education. LLMs have demonstrated passing scores on the United States Medical Licensing Examination (USMLE), raising questions about their impact on assessment frameworks [1], including whether artificial intelligence (AI) can successfully answer essay-style, reasoning-based questions and whether assessors can distinguish AI-generated and student-written responses. Our medical school's preclinical students complete application-level, essay-type questions—concept appraisals (CAPPs)—every week (Multimedia Appendix 1) [2]. We evaluated LLMs' performance on CAPPs and examined assessors' ability to distinguish AI-generated and human responses.

Methods

Study Design

Ten retired CAPP questions were selected, ensuring representation from multiple preclinical organ-system blocks, including gastroenterology, endocrinology, musculoskeletal science, cardiorespiratory medicine, hematology, renal biology, and immunology. Retired CAPPs were used, so that currently used ones were not exposed to students. Answering these required literature review and application of knowledge to clinical scenarios.

Five student responses from previous classes (before availability of LLMs) were randomly selected and deidentified. Individuals at various medical training levels generated AI responses via Bing Chat (subsequently renamed Microsoft Copilot; Multimedia Appendix 1), which used GPT-4 algorithms and had similar performance on medical tasks

as ChatGPT 4.0—the most advanced LLM at the time of study [3,4]. Users first prompted Bing Chat by using the original CAPP text and then iteratively refined prompts to generate more comprehensive answers and match institutional standards without manual editing (Multimedia Appendix 1).

Ten expert assessors graded responses to 1 CAPP question each. While unaware that any responses had been AI-generated, they graded 5 deidentified student responses and 2 AI-generated responses (presented in random order) for their CAPP question, using a standard rubric (Multimedia Appendix 1). For 2 CAPPs, 4 student responses were used instead of 5 due to lack of consent for inclusion in the registry. Grading each CAPP took approximately 30 minutes; thus, a larger sample size was infeasible for this exploratory study. Afterward, assessors identified whether responses were AI- or student-generated and provided their rationales.

Scoring differences between human- and AI-generated responses and identification accuracy were evaluated, using descriptive statistics. Thematic analysis was conducted on assessors' classification rationales; 2 team members independently analyzed reasons to identify themes, compared findings, and reconciled differences (Multimedia Appendix 1).

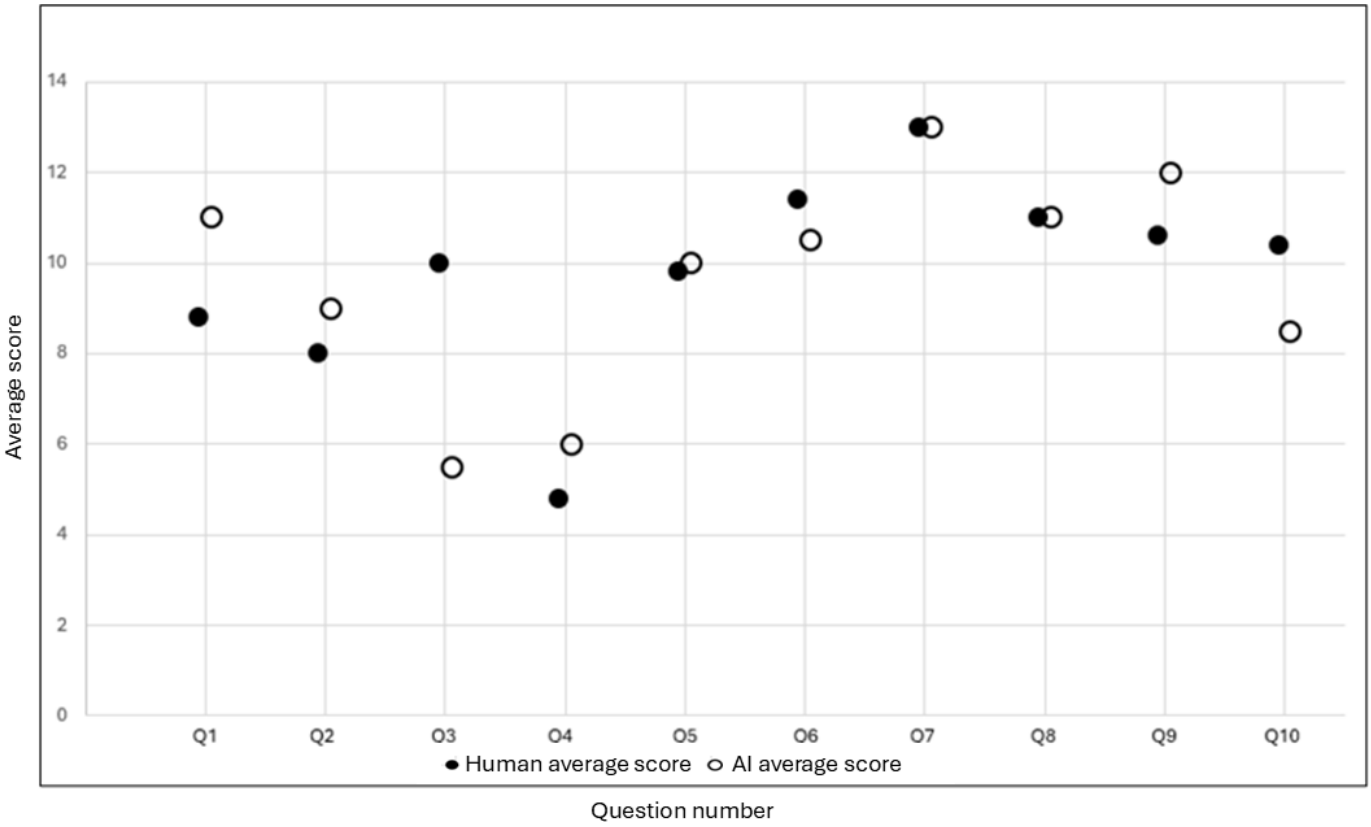
Ethical Considerations

This study used deidentified data from the Cleveland Clinic Institutional Review Board–approved registry #6600. Since this was a registry for which students had already provided informed consent, separate informed consent was not required. Each CAPP reviewer was paid US $100.

Results

AI responses received scores higher than or equal to those for human responses for most questions, with substantial performance variability; AI scored better than, equivalent to, or worse than humans, depending on the CAPP question (Figure 1).

Figure 1. Average of human vs AI scores for each question. CAPP questions were answered either by students (human) or by prompting Microsoft Copilot (AI). Expert graders scored the CAPP questions based on a rubric. The average scores received by humans and AI are shown by question (colored vs open circles, respectively). AI responses received scores higher than or equal to those for human responses for most questions. Each question had a unique maximum score. This figure illustrates the relative scores of humans vs AI. AI: artificial intelligence; CAPP: concept appraisal.



Assessors correctly identified response sources 53% (36/68) of the time (student responses: 27/48, 56%; AI-generated responses: 9/20, 45%). Only 1 assessor correctly classified all responses. Consistent with other studies, 1 assessor who used AI detection tools did not have much success [5] (Table 1).

Table 1. Percentage of responses correctly identified as human or artificial intelligence (AI) responses for each critical appraisal (CAPP) question.[a]

Question number	Correctly identified responses, n/N (%)
Q1	3/6 (50)
Q2	3/7 (43)
Q3	3/7 (43)
Q4	6/7 (86)
Q5	3/6 (50)
Q6	2/7 (29)
Q7[b]	0/7 (0)
Q8	5/7 (71)
Q9	4/7 (58)
Q10[c]	7/7 (100)

[a]Responses for each question were graded by 1 expert. Expert graders were blinded and were not told which responses were generated by humans vs AI.
[b]Despite utilization of AI detection tools, 1 assessor did not correctly classify any of the responses (Q7).
[c]Only 1 assessor correctly classified all responses for their CAPP question (Q10).

Thematic analysis showed that the most cited reason for identification was the perceived "writing style," though many assessors noted an inability to distinguish categories (Multimedia Appendix 1).

Discussion

We demonstrate that AI can provide high-quality answers to essay-style medical education questions requiring detailed research and knowledge application. Content experts struggled to distinguish AI-generated and human-written responses, underscoring the challenges of identifying academic misuse of generative AI.

Iterative prompting of Microsoft Copilot was essential for generating acceptable responses. This process mirrors students' typical workflow for refining drafts through edits; thus, iterative prompting does not necessarily disadvantage AI. Our findings highlight concerns about potential overreliance on AI and its implications for assessment validity, especially as recent survey data suggest that 89% of students use ChatGPT during self-study [6,7].

Given AI responses' similarity to human responses, institutions must consider frameworks for integrating AI into assessments without compromising academic integrity [8]. Potential strategies include structured classroom use of AI during collaborative group work (eg, requiring students to assess AI responses and cite primary evidence to support or refute them) [7,9].

Study limitations include a small sample of AI-generated responses and the research's exploratory nature. Expanding the sample size and including additional questions could provide insights on AI's performance (relative to humans) for specific question types (Multimedia Appendix 1). Additionally, the findings prompt further discussions on ethically integrating generative AI into medical curricula while ensuring students develop critical appraisal and independent reasoning skills [7,10].

AI's performance suggests its potential as a learning enhancement tool. However, medical educators must implement strategies for preventing overreliance on AI, fostering reflective learning practices and critical thinking, and maintaining assessment integrity.

Acknowledgments

The authors would like to thank the following individuals for serving as concept appraisal (CAPP) graders: William Albabish, William Cantrell, Thomas Crilley, Ryan Ellis, Andrew Ford, Emily Frisch, Jeffrey Schwartz, Michael Smith, Mohammad Sohail, and Anirudh Yalamanchali. Financial support was received from The Jones Day Endowment Fund.

Authors' Contributions

IBB and NM contributed to the literature review. NM, AKM, and CCL contributed to the conceptual design. SM, NM, ASN, and AD contributed to data analysis and visualization. IBB and SMB contributed to thematic analysis. SM, ENH, and NM contributed to manuscript writing. All authors contributed to the critical revision of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary materials regarding concept appraisal questions and grading, Bing Chat (subsequently renamed Microsoft Copilot), the iterative prompting used in this study, and the thematic analysis.
[DOCX File (Microsoft Word File), 148 KB-Multimedia Appendix 1]

References

1. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ. Oct 20, 2023;9:e48785. [doi: 10.2196/48785] [Medline: 37862079]
2. Bierer SB, Dannefer EF, Taylor C, Hall P, Hull AL. Methods to assess students' acquisition, application and integration of basic science knowledge in an innovative competency-based curriculum. Med Teach. 2008;30(7):e171-e177. [doi: 10.1080/01421590802139740] [Medline: 18777415]
3. Cai LZ, Shaheen A, Jin A, et al. Performance of generative large language models on ophthalmology board-style questions. Am J Ophthalmol. Oct 2023;254:141-149. [doi: 10.1016/j.ajo.2023.05.024] [Medline: 37339728]
4. Amin KS, Davis MA, Doshi R, Haims AH, Khosla P, Forman HP. Accuracy of ChatGPT, Google Bard, and Microsoft Bing for simplifying radiology reports. Radiology. Nov 2023;309(2):e232561. [doi: 10.1148/radiol.232561] [Medline: 37987662]
5. Elkhatat AM, Elsaid K, Almeer S. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. Int J Educ Integr. Sep 1, 2023;19(1):17. [doi: 10.1007/s40979-023-00140-5]
6. Westfall C. Educators battle plagiarism as 89% of students admit to using OpenAI's ChatGPT for homework. Forbes. Jan 28, 2023. URL: https://www.forbes.com/sites/chriswestfall/2023/01/28/educators-battle-plagiarism-as-89-of-students-admit-to-using-open-ais-chatgpt-for-homework/ [Accessed 2025-04-01]
7. Mehta S, Mehta N. Embracing the illusion of explanatory depth: a strategic framework for using iterative prompting for integrating large language models in healthcare education. Med Teach. Feb 2025;47(2):208-211. [doi: 10.1080/0142159X.2024.2382863] [Medline: 39058399]
8. Silverman JA, Ali SA, Rybak A, van Goudoever JB, Leleiko NS. Generative AI: potential and pitfalls in academic publishing. JPGN Rep. Nov 8, 2023;4(4):e387. [doi: 10.1097/PG9.0000000000000387] [Medline: 38034432]
9. Jowsey T, Stokes-Parish J, Singleton R, Todorovic M. Medical education empowered by generative artificial intelligence large language models. Trends Mol Med. Dec 2023;29(12):971-973. [doi: 10.1016/j.molmed.2023.08.012] [Medline: 37718142]
10. Halkiopoulos C, Gkintoni E. Leveraging AI in e-learning: personalized learning and adaptive assessment through cognitive neuropsychology—a systematic analysis. Electronics (Basel). Sep 22, 2024;13(18):3762. [doi: 10.3390/electronics13183762]

Abbreviations

AI: artificial intelligence
CAPP: concept appraisal
LLM: large language model
USMLE: United States Medical Licensing Examination

Edited by Lorainne Tudor Car; peer-reviewed by David Chartash, Ren Yang; submitted 02.02.2025; final revised version received 11.05.2025; accepted 16.05.2025; published 16.06.2025

Please cite as:
Mehta S, Haddad EN, Burke IB, Majors AK, Maeda R, Burke SM, Deshpande A, Nowacki AS, Lindenmeyer CC, Mehta N
Assessment of Large Language Model Performance on Medical School Essay-Style Concept Appraisal Questions: Exploratory Study
JMIR Med Educ 2025;11:e72034
URL: https://mededu.jmir.org/2025/1/e72034
doi: 10.2196/72034