Assessing the Potential of USMLE-Like Exam Questions Generated by GPT-4

Scott L Fleming, MS*
Keith Morse, MD, MBA
Aswathi Kumar, MD
Chia-Chun Chiang, MD
Birju Patel, MD, MPH
Emma Brunskill, PhD
Nigam Shah, MBBS, PhD

Author Affiliations:
Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA
Department of Computer Science, Stanford University, Stanford, CA, USA
Center for Biomedical Informatics Research, Stanford University School of Medicine, Stanford, CA, USA
Division of Pediatric Hospital Medicine, Department of Pediatrics, Stanford University, Palo Alto, CA
Stanford Healthcare Tri-Valley, Pleasanton, CA
Department of Neurology, Mayo Clinic, Rochester, MN, USA
The Stanford Institute for Human-Centered Artificial Intelligence, Stanford University, Palo Alto, CA
Clinical Excellence Research Center, Stanford University School of Medicine, Stanford, CA, USA
Technology and Digital Services, Stanford Health Care, Palo Alto, CA, USA

*Corresponding author (scottyf@stanford.edu)

Abstract

Prior work has shown that large language models like GPT-4 and Med-PaLM 2 can *answer* sample questions from the USMLE Step 2 Clinical Knowledge (CK) exam with greater than 80% accuracy. But can these generative AI *create* USMLE-like exam questions? This capability could augment humans in writing or preparing for such exams. Here we assess the ability of GPT-4 to generate realistic exam questions by asking licensed physicians to (1) distinguish AI-generated questions from genuine USMLE Step 2 CK questions, and (2) assess the validity of AI-generated questions and answers. We find that GPT-4 can generate question/answer pairs that are largely indistinguishable from human-generated ones, with a majority (64%) deemed “valid” by a panel of licensed physicians.

Introduction

Every year, over 100,000 medical students take a United States Medical Licensing Examination (USMLE) administered by the National Board of Medical Examiners (NBME).[1] Generating exam questions and related preparation materials is costly, and students spend on average over $4,000 on exam preparation.[2] Artificial intelligence (AI) that creates questions similar to those that appear on the USMLE could make preparation material more affordable by redirecting effort from question generation to the faster process of question review. OpenAI's recent AI language model GPT-4 can *answer* sample questions from the USMLE Step 1, Step 2 Clinical Knowledge (CK), and Step 3 exams with greater than 80% accuracy.[3] Such generative technology could potentially also generate USMLE-like questions to facilitate the work of exam content creators and help student learners prepare for the exam. We assess the ability of GPT-4 to generate useful exam questions by asking physician reviewers to (1) distinguish AI-generated questions from genuine USMLE Step 2 CK questions, and (2) assess the validity of AI-generated questions and answers.

Methods

We randomly selected a set of 50 human-generated questions without associated images or abstracts from a bank of 120 publicly available USMLE Step 2 CK test questions.[4] We used a prompt chaining[5] strategy to generate new questions in a similar format using GPT-4 (Figure 1). Fifty human-generated and 50 AI-generated questions were randomly shuffled into a 100-question set. A group of three licensed, practicing physicians were instructed to (1) choose the single best answer to the question without consulting any external reference; and (2) guess whether the question was generated by humans or GPT-4. In a separate task, clinicians evaluated the correctness of the AI-generated exam items using any available external reference, and recorded the type of errors made and also the length of time it took them to make their determination. Data collection and analysis was performed in March and April 2023.

Results

The average accuracy in determining whether a question was generated by humans or GPT-4 (Figure 2) was 50.3% (95% CI 37.8% to 62.8%). All physicians obtained a greater than passing grade on the 100 questions, and on average, individuals' accuracy on AI-generated questions was within 4.6% (absolute percentage points) of their accuracy on human-generated questions.

Of the 50 AI-generated exam items, 32 (64%) were deemed to be "correct" by all assigned reviewers and 18 (36%) were deemed to be "incorrect" by at least one reviewer. Reasons for labeling exam items as "incorrect" included "Multiple correct answer choices" (n=9), "AI-chosen answer is incorrect" (n=6), "No correct answer choice" (n=3). Reviewers spent, on average, 3.21 (95% CI 2.73 to 3.69) minutes reviewing each AI-generated exam item for correctness.

Discussion

We showed that GPT-4, a large language model, can generate USMLE Step 2 CK style questions that are indistinguishable by physicians from questions generated by experts. However, the validity of AI-generated content with GPT-4 was unreliable. It may be possible that having this tool generate candidate items that are later reviewed for quality and correctness is altogether more efficient than having an expert produce an item de novo. We demonstrated that such review for correctness can be fast, though we do not assess other aspects of question quality, such as whether the questions assess integration of knowledge vs. fact-based recall [6].

This study has limitations. First, the physicians who assessed the question and answers were not trained in generating USMLE exam content. Second, we did not assess if GPT-4 questions tested clinical concepts prioritized by the exam blueprint. Finally, GPT-4 is privately owned and its performance changes over time, so these results may not be reproducible.

Given novel human-AI collaboration enabled by language models, we evaluated whether GPT-4 can help create realistic exam content for the USMLE. To determine the actual value of AI models such as GPT-4 in medicine, evaluations based on potential real-world use cases are as important – and perhaps more useful – as those that assess models on tasks intended to evaluate humans.

Figure 1: Prompt Chaining for Question Generation with GPT-4

(A) Prompt 1 (Human): Here is an example question from the USMLE Step 2 CK: ##### A 45-year-old man is brought to the emergency department because of moderate chest pain after a generalized tonic-clonic seizure 30 minutes ago. He has seizure disorder for which he has taken carbamazepine and phenobarbital for the past 20 years. X-ray of the chest shows generalized osteopenia with several rib fractures. Which of the following is the most likely nutritional deficiency? (A) Folic acid (B) Iron (C) Magnesium (D) Vitamin B_2 (riboflavin) (E) Vitamin D (F) Zinc ##### Why is (E) the correct answer and why are the other answers incorrect?	**(C) Prompt 2 (Human):** Here is an example question, answer, and explanation from the USMLE Step 2 CK: ##### Question: A 45-year-old man is brought to the emergency department [...] Which of the following is the most likely nutritional deficiency? (A) Folic acid […] (F) Zinc Answer and Explanation: (E) Vitamin D is the correct answer because [...] In summary, the patient's generalized osteopenia and rib fractures after a seizure point toward compromised bone strength, which is most consistent with a vitamin D deficiency. ##### Generate another question for the USMLE Step 2 CK using a similar format.
(B) Response 1 (GPT-4): (E) Vitamin D is the correct answer because the patient's presentation is consistent with a deficiency in this nutrient. Vitamin D is essential for calcium absorption and bone health. Generalized osteopenia (reduced bone density) and rib fractures in this patient suggest compromised bone strength, which is a characteristic feature of vitamin D deficiency. Here's why the other options are incorrect: […] In summary, the patient's generalized osteopenia and rib fractures after a seizure point toward compromised bone strength, which is most consistent with a vitamin D deficiency.	**(D) Response 2 (GPT-4):** Question: A 28-year-old woman presents to her primary care physician with complaints of fatigue, palpitations, and shortness of breath for the past few weeks. She has a history of Crohn's disease and follows a strict vegan diet. On physical examination, she appears pale and has mild glossitis. A complete blood count reveals a macrocytic anemia. Which of the following is the most likely cause of her anemia? (A) Folic acid deficiency (B) Hemolytic anemia (C) Iron deficiency (D) Sickle cell anemia (E) Vitamin B_{12} (cobalamin) deficiency (F) Thalassemia Answer and Explanation: (E) Vitamin B_{12} (cobalamin) deficiency is the most likely cause of the patient's anemia. Macrocytic anemia, fatigue, palpitations, and shortness of breath are common symptoms of vitamin B_{12} deficiency. A strict vegan diet and Crohn's disease, which can lead to malabsorption, increase the risk of vitamin B_{12} deficiency. Here's why the other options are incorrect: [...]

Caption: The prompt chaining approach used to automatically generate new questions with seed questions from the set of human-generated USMLE Step 2 CK sample questions. First, we use a human-generated sample question (upper left, blue) as part of the prompt to GPT-4 to obtain an AI-generated answer and explanation (lower left, green). Next, both the human-generated question (upper right, blue) and the AI-generated answer and explanation (upper right, green) are passed to the AI model as part of a prompt asking to generate "another question for the USMLE Step 2 CK using a similar format". GPT-4 generates a new question, answer choices, and associated explanations (lower right, red). In this study, we analyze the quality of the AI-generated questions.

Figure 2: USMLE Step 2 CK Questions Generated by GPT-4

(A) AI-generated question, mistaken as human-generated	**(B) AI-generated question, identified as AI-generated**
A 36-year-old woman presents to the clinic with a 2-week history of persistent dry cough, shortness of breath, and mild chest pain. She also reports occasional low-grade fever and night sweats. She is a non-smoker and works as an elementary school teacher. Her medical history is unremarkable, and she takes no medications. On examination, she appears anxious. Lung auscultation reveals decreased breath sounds and dullness to percussion in the left lower lung field. Laboratory studies show: Hematocrit: 41% Mean corpuscular volume: 86 μm³ Leukocyte count: 11,000/mm³ Segmented neutrophils: 68% Lymphocytes: 24% Monocytes: 8% Platelet count: 375,000/mm³ Erythrocyte sedimentation rate: 40 mm/hr A chest x-ray demonstrates a left-sided pleural effusion. Which of the following is the most appropriate next step in management? (A) Bronchoscopy (B) Thoracentesis (C) Computed tomography (CT) scan of the chest (D) Pulmonary function tests (E) Empiric treatment with antibiotics	A randomized controlled trial is conducted to evaluate the effectiveness of a new antihypertensive drug in reducing the incidence of stroke. Ten thousand patients with hypertension are randomly assigned to receive either the new antihypertensive drug or a standard antihypertensive drug. After a follow-up period of 3 years, the results show that the incidence of stroke is 3% in the new antihypertensive drug group and 6% in the standard antihypertensive drug group. The difference in incidence is statistically significant. Which of the following is the most accurate interpretation of these results? (A) Absolute risk reduction is 1% (B) Absolute risk reduction is 3% (C) Number needed to treat is 10 (D) Number needed to treat is 33 (E) Relative risk reduction is 33.3% (F) Relative risk reduction is 50%

Caption: (Left) A sample USMLE Step 2 CK question generated by GPT-4 where there was unanimous consensus amongst physicians that the question was generated by a human (i.e., consensus was incorrect). The answer selected by GPT-4 is (B) Thoracentesis. (Right) A sample USMLE Step 2 CK question generated by GPT-4 where there was unanimous consensus amongst physicians that the question was generated by GPT-4 (i.e., consensus was correct). All physicians noted the presence of multiple correct answer choices as a reason for their determination of the question's source. Notably, in its explanation GPT-4 also identified answers B, D, and F as all being correct answers but included them all in the answer choices anyway. This illustrates a common failure mode for GPT-4 with our prompting strategy (it can generate questions with multiple correct answers).

References

1. United States Medical Licensing Examination. Performance Data. *usmle.org* https://www.usmle.org/performance-data.

2. Bhatnagar, V., Diaz, S. R., Bucur, P. A. & Diaz, S. The Cost of Board Examination and Preparation: An Overlooked Factor in Medical Student Debt. *Cureus* **11**, (2019).

3. Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of GPT-4 on Medical Challenge Problems. *arXiv:2303.13375 [cs.CL]* (2023).

4. United States Medical Licensing Examination. Step 2 CK Sample Test Questions. *usmle.org* https://www.usmle.org/prepare-your-exam/step-2-ck-materials/step-2-ck-sample-test-questions (2023).

5. Wu, T., Terry, M. & Cai, C. J. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* 1–22 (Association for Computing Machinery, 2022).

6. Melissa S Billings, Kristine DeRuchie, Kieran Hussie, Allison Kulesher, Jacquelyn Merrell, Amy Morales, Miguel A Paniagua, Jennifer Sherlock, Kimberly A Swygert, Julie Tyson. NBME Item-Writing Guide. Preprint at https://www.nbme.org/item-writing-guide (2021).

Note: ChatGPT was used to generate the title, with the following prompt:
"Suggest a title for: [content of the introduction]"

Data Availability: All data produced in the present study will be made available upon reasonable request to the authors

Conflicts of Interest: Dr. Patel reports receiving equity compensation from Google LLC

Funding: This work is supported by a National Science Foundation grant DMS-1916163; the Mark and Debra Leslie endowment for AI in Healthcare; the Stanford University Department of Medicine; Stanford Healthcare; the Stanford Medicine Program for AI in Healthcare; a NSF CAREER Grant (Emma Brunskill); and a Stanford Graduate Fellowship (Scott Fleming). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding bodies.