Analysis of ChatGPT-3.5's Potential in Generating NBME-Standard Pharmacology Questions: What Can Be Improved?

Marwa Saad, B.S.[1], Wesam Almasri, B.A.[1], Tanvirul Hye, Ph.D.[2], Monzurul Roni, Ph.D.[3], Changiz Mohiyeddini, Ph.D.[1]

[1]Oakland University William Beaumont School of Medicine, [2]Roseman University College of Medicine, [3]University of Illinois College of Medicine Peoria

Introduction

- Chat Generative Pre-trained Transformer (ChatGPT) is an artificial intelligence (AI) software developed by OpenAI, trained on information up to September 2021.[1,2]
- As of June 2023, ChatGPT has reached over 1.6 billion visitors and an average of 55 million visitors daily.[3]
- The National Board of Medical Examiners (NBME) writes Step exams for medical students to take during their medical education. Pharmacology makes up 15-22% of the questions on the USMLE Step 1 exam.[4]
- Previous studies found that ChatGPT answered NBME-Step 1 exam with 64.4% of questions correct.[5]
- The ability of ChatGPT to perform with a passing grade of over 60 percent on the NBME-Step 1 exam prompted our interest in assessing the ability of ChatGPT to generate NBME-style pharmacology multiple choice questions that meet the NBME guidelines.[6]

Aims and Objectives

Aims:

1. Identify the NBME standard guidelines for quality control of multiple choice questions.
2. Evaluate the ability of ChatGPT to generate NBME-style pharmacology multiple choice questions based on NBME standard guidelines.
3. Assess the ChatGPT-generated NBME-style pharmacology multiple choice questions and answer explanations for medical accuracy.

Objective:

Our objective is to evaluate ChatGPT-3.5's capability to generate NBME-style pharmacology multiple-choice questions, adhering to the NBME guidelines, and offer recommendations for enhancing and fine-tuning ChatGPT-generated NBME questions for medical education.

Methods

Stage 1: Defining a suitable prompt for ChatGPT to generate NBME-style pharmacology multiple-choice questions, adhering to the NBME guidelines.

- Selection of 10 medications from various organ systems and related drugs.
- We conducted prompt engineering to formulate questions that align with NBME standards using 'GPT best practices.'
- The standard prompt reads as follows: 'Can you provide an expert-level sample NBME question about the [mechanism of action, indication, OR side effect] of [drug] and furnish the correct answer along with a detailed explanation?'
- The questions and answers generated by ChatGPT were assessed using a grading system that was created following the NBME Item-Writing Guide.

Stage 2: Assessing the Quality of ChatGPT-Generated Questions

- All ChatGPT-generated questions and answers underwent evaluation by a panel consisting of two pharmacology education experts with a significant teaching background in medical schools in the USA.
- The experts independently utilized the aforementioned 16 different criteria to individually score each question and assigned a score of 0 or 1 for each criterion, resulting in a quality score ranging from 0 to 16 for each question.
- A higher score indicates better adherence to the NBME Item-Writing Guide. The experts were then asked to rate the difficulty of each questions from 1 (very easy) to 5 (very difficult).

NBME Criteria

1. Correct answer choice along with medically accurate explanations
2. Avoids describing the class of the drug in conjunction with the name of the drug.
3. Applies foundational medical knowledge, rather than basic recall.
4. Clinical vignette is in complete sentence format.
5. Question can be answered without looking at multiple-choice options
6. Avoids long or complex multiple choice options
7. Avoids vague frequency terms within the clinical vignette
8. Avoids "none of the above" in the answer choices
9. Avoids nonparallel, inconsistent answer choices.
10. Avoids negatively phrased lead-ins in the clinical vignette
11. Avoids general grammatical cues within the clinical vignette.
12. Avoids grouped or collectively exhaustive answer choices.
13. Avoids absolute terms, such as "always" or "never" in answer choices.
14. Avoids having the correct answer choice stand out
15. Avoids repeated words or phrases in the clinical vignette that clue the correct answer choice.
16. A balanced distribution of key terms in the answer choices.

Results

- According to expert 1, the average score for the ChatGPT-3.5 questions was 13.7 out of 16 (85.6%).
- According to expert 2, the average score for ChatGPT-3.5 questions was 14.5 out of 16 (90.6%).
- The average score among the two experts was 14.1 out of 16 (88.1%).

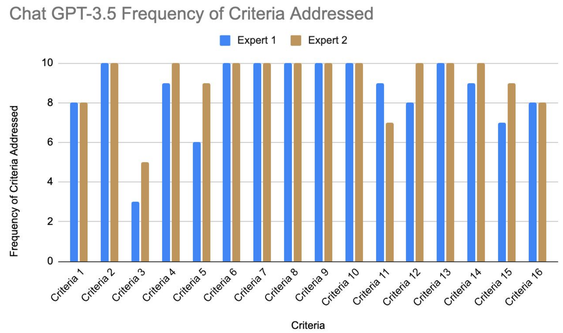

Figure 1. Frequency of Criteria Addressed by Google Gemini's Generated Questions

- As depicted in Figure 2, expert 1 found four questions very easy, two questions easy, two questions moderately hard, two questions hard, and no questions very hard.
- In contrast, expert 2 found three questions very easy, four questions easy, three questions moderately hard, and no questions hard or very hard.

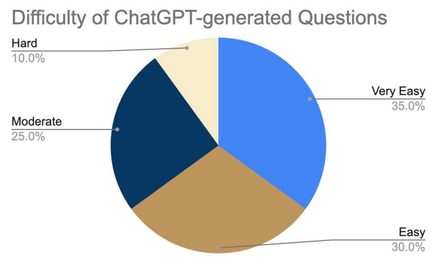

Figure 2. Evaluation of Question Difficulty

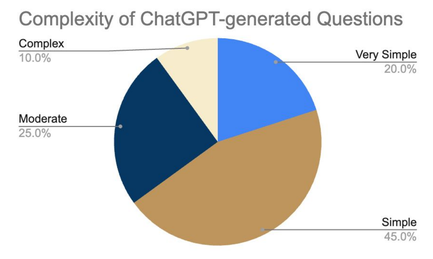

Figure 3. Evaluation of Question Complexity

Conclusions

- ChatGPT-3.5 scored highly on NBME Item-Writing Guide criteria, averaging 14.1/16 (88.1%) between two experts, but notable weaknesses emerged on closer review.
- Both experts found that 2/10 (20%) questions contained medically inaccurate information. For instance, ChatGPT-3.5 incorrectly stated that warfarin causes thrombocytopenia.[7] This can lead to the student unknowingly make incorrect associations with drugs and adverse effects, such as abciximab actually causing thrombocytopenia, rather than warfarin.[7] Such inaccuracies may stem from AI "hallucinations", which are plausible but incorrect information is generated by mixing up of facts.[8,9]
- Criteria 3, testing application of foundational medical knowledge rather than recall, was most frequently missed with 60-70% questions failing to meet this standard. Furthermore, this constitutes most of the questions as "pseudo vignettes" which are vignette that do not require the need for the clinical information provided to answer the question and rather tests basic recall of material.[10]
- Although the scores suggest that the ChatGPT-3.5 questions meet NBME guidelines, closer review shows they fall short of Step exam standards. The questions are oversimplified and rely on repetitive wording that allows test-takers to answer them without true medical knowledge. These findings highlight the need for AI models specifically trained on medical and scientific information.

Acknowledgements

The authors would like to extend their gratitude to Dr. Steve Wilson who made substantial contribution in the development of the design guidelines and provision of software expertise.

References

1. Natalie. What is ChatGPT? OpenAI Help Center. 2023. Available from: https://help.openai.com/en/articles/6783457-what-is-chatgpt
2. Niles R. GPT-3.5 Turbo Updates. OpenAI Help Center. 2023. Available from: https://help.openai.com/en/articles/8555514-gpt-3-5-turbo-updates
3. Brandl R. ChatGPT Statistics and User Numbers 2023 - OpenAI Chatbot. Tooltester. 2023 Feb 15. Available from: https://www.tooltester.com/en/blog/chatgpt-statistics/#:~:text=diagrams%2C%20and%20illustrations.-
4. Step 1 Content Outline and Specifications | USMLE. Available from: https://www.usmle.org/prepare-your-exam/step-1-materials/step-1-content-outline-and-specifications
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.
6. Scoring & Score Reporting | USMLE. Available from: https://www.usmle.org/bulletin-information/scoring-and-score-reporting
7. Warfarin: Drug Information. UpToDate. Available from: https://www-uptodate-com.huaryu.kl.oakland.edu/contents/warfarin-drug-information?search=warfarin&source=panel_search_result&selectedTitle=1~148&usage_type=panel&kp_tab=drug_general&display_rank=1
8. Shuster K, Poff S, Chen M, Kiela D, Weston J. Retrieval augmentation reduces hallucination in conversation. arXiv.org. 2021 Apr 15. Available from: https://arxiv.org/abs/2104.07567
9. Tian K, Mitchell E, Yao H, Manning CD, Finn C. Fine-tuning language models for factuality. arXiv.org. 2023 Nov 14. Available from: https://arxiv.org/abs/2311.08401
10. Vanderbilt AA, Feldman M, Wood IK. Assessment in undergraduate medical education: a review of course exams. Med Educ Online. 2013;18:1-5.