An Empirical Study of Google Gemini's Potential in Generating NBME-Style Pharmacology Questions: Challenges and Improvements

Wesam Almasri[1], Marwa Saad[1], Changiz Mohiyeddini[1]

[1]Oakland University William Beaumont School of Medicine, United States

Introduction

Google Gemini is the latest multimodal AI tool by Google that launched in December 2023 [1]. Google Gemini was found to be especially useful for students, providing outputs that can be tailored and customized to an individual's specific requirements of a subject, as well as the user's level of understanding [2].

The National Board of Medical Examiners (NBME) is an organization which owns the United States Medical Licensing Examination (USMLE). The NBME is responsible for writing board exams for medical students throughout their medical education, including Step 1, an exam medical students take prior to beginning their clinical rotations [3].

As of Google Gemini's release in May 2023, there has been limited data on Google Gemini's capabilities to generate and answer NBME questions [4]. This poses the question if Google Gemini can be a free, cost-effective tool that medical students can use to advance their clinical knowledge using generated clinical vignettes [5].

If Google Gemini can sufficiently generate questions that are medically accurate and abide by the NBME Item-Writing Guide, this can pose as a free, valuable resource for medical students. This study aims to evaluate Google Gemini's capabilities to generate NBME-style pharmacology questions that sufficiently satisfy the NBME guidelines in question writing. This study's objective is to provide feedback to enhance Google Gemini's capabilities as a tool for medical education.

Methods

We selected 10 medications from the following organ systems, with their given drug in parentheses: hematology/lymphatics (warfarin), neurology (norepinephrine), cardiovascular (metoprolol), respiratory (albuterol), renal (lisinopril), gastrointestinal (bisacodyl), endocrine (metformin), reproductive (norethindrone), musculoskeletal (cyclobenzaprine), and behavioral medicine (trazodone).

We conducted 'prompt engineering' to generate questions that adhere closely to the NBME standards using 'GPT best practices'. In conjunction with IT experts and pharmacology experts who create NBME-standard questions for in-house exams, multiple stages were followed to find the most suitable prompt for the NBME questions. The standard prompt is as follows, with brackets indicating the variable of which different medication names were inputted to be evaluated:

''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?'

Using the above input, each AI engine was requested to generate a question for either the mechanism of action, indication, and side effect for each of the 10 above mentioned medications. Afterwards, the mechanism of action, indication, and side effect topics were randomly selected to be assessed and evaluated for each medication using the below 16 criteria extrapolated from the NBME writing guide:

Methods

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.

Lastly, the pharmacology experts were asked to estimate the complexity of each question. The panel was asked to rate the difficulty of each AI-generated question on a Lichert scale using the following question: "On a scale from very easy to very difficult, how would you rate the difficulty of this question for students?" The responses then ranged from 1 to 5, with 1 being very easy and 5 being very difficult.

Results

According to expert 1, the average score for Google Gemini's generated questions was 14.1 out of 16 (88.1%), while for expert 2 the average score was 14.5 out of 16 (90.6%). The average score among both experts was 14.3 out of 16 (89.3%). The scores for each criteria can be seen in Figure 1.

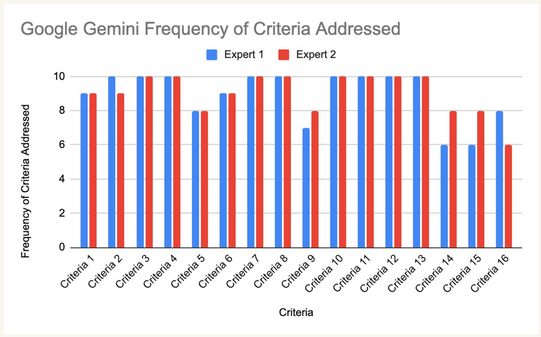
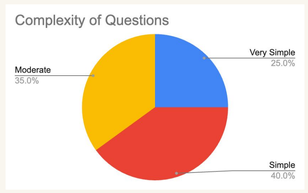


Figure 1. Frequency of Criteria Addressed by Google Gemini's Generated Questions

When it comes to the perceived difficulty of the NBME-standard questions, expert 1 found three questions very easy, five questions easy, two questions moderately difficult, and no questions hard or very hard. Expert 2 found two questions very easy, three questions easy, five questions moderately difficult, and no questions hard or very hard. The evaluation of the difficulty of these questions are depicted in Figure 2 and Figure 3.



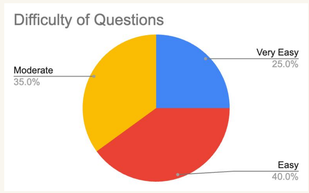


Figure 2, Figure 3. Evaluation of Question Complexity & Difficulty

Discussion

Based on the average score of 14.3 out of 16 amongst the two experts, Google Gemini's seems to be proficient in generating pharmacology questions that align with the NBME guidelines. However, beyond the face-value of the high average score, the questions seem to not be up to the same level of quality as other board exam material resources medical students use today.

Google Gemini received 100% fulfillment in criteria 3, 4, 7, 8, and 10-13, showcasing its ability to provide structurally and grammatically sound questions according to the NBME guidelines. However, by delving deeper into the answer choices, it struggled significantly.

On multiple occasions, Google Gemini failed to utilize the clinical vignette it provided within the question stem to encourage the medical student to deduce the answer choice with application of their medical knowledge. These were termed "pseudo vignettes", in which the student only needed to rely on basic recall to answer the question, and did not need to utilize any clinical reasoning [6]. In order to become a strong tool for medical education, it is imperative that Google Gemini has the capacity to integrate the clinical vignette properly into the question stem to allow the medical student to apply their clinical knowledge.

In addition to the pseudo vignettes, Google Gemini's quality in generating NBME-standard pharmacology questions fell short in the formatting of its generated answer choices. In some cases, answer choices generated by Google Gemini were excessively long, which could give away the correct option. Additionally, Google Gemini occasionally failed to maintain parallelism in answer choices or maintain balancing of key terms between answer choices, unintentionally highlighting the correct answer choice. These consistent cues led to easier questions in which application of medical knowledge was not needed, and was merely pattern recognition to identify the correct answer choice.

Overall, Google Gemini does not provide pharmacology questions that adhere to the NBME-standard compared to other leading resources that medical students use to prepare for board exams. Their general ease can be primarily attributed to their lack of integrating clinical vignettes and cues that lead to the correct answer choice, which does not allow the question to test a medical student's clinical and medical knowledge. . By providing questions that integrate clinical vignettes with the medical student's knowledge base, Google Gemini can provide excellent pharmacology questions that can be a great tool for medical students' preparation for board exams.

Acknowledgments: The authors would like to extend their gratitude to Dr. Steve Wilson who made substantial contribution into making this project a success.

References

1. Coles G. Google DeepMind Gemini AI release date: The world's new most powerful language model. PCguide.com. 2023 Dec 25. Available from: https://www.pcguide.com/apps/google-deepmind-gemini-release-date/
2. Aktay S, Seçkin G, Uzunoğlu D. ChatGPT in education. Türk Akademik Yayınlar Dergisi (TAY Journal). 2023;7(2):378–406.
3. United States Medical Licensing Examination. Step 1 Overview. Available from: https://www.usmle.org/step-exams/step-1. Accessed 2023 Jul 29.
4. Mihalache A, Grad J, Patil NS, et al. Google Gemini and Bard artificial intelligence chatbot performance in ophthalmology knowledge assessment. Eye. 2024;38:2530–2535. doi:10.1038/s41433-024-03067-4.
5. Benoit J. ChatGPT for Clinical Vignette Generation, Revision, and Evaluation. medRxiv. 2023. Preprint.
6. Vanderbilt AA, Feldman M, Wood IK. Assessment in undergraduate medical education: a review of course exams. Med Educ Online. 2013;18:1-5.