Introducing Graduate and Medical Students to an Ethical Use of AI, Making Item Writing and Self-Assessment More Effective

B. Kaw (DO27)[1] , N. Yusufzai (MHS24)[1] , D. Shelton (DO27)[1] , E. Aigbe (MHS24)[1] , A. Sadik[1] , S. Koshy Chenthittayil[2]

[1]TUN, College of Osteopathic Medicine [2]TUN, Office of Institutional Effectiveness

TOURO UNIVERSITY NEVADA Research Day 2024

TOURO UNIVERSITY NEVADA Research Day 2024

Background

Artificial intelligence (AI) tools, such as ChatGPT-4, are novel in their use in medical education. This study pioneers ChatGPT-4 usage for medical science students' assessment. Its goal is to adopt a more efficient method of item writing while ensuring adequate coverage of basic science concepts. To increase engagement with the content, graduate students in the medical biochemistry course must write multiple choice questions (MCQs) based on learning objectives to create a question bank for self-assessment.

Traditionally, MCQ development is often time-consuming and labor-intensive. With faculty oversight, introducing students to AI for MCQ creation may streamline this process, making it more efficient and effective, hence this study.

Methods

I. Using funds from the Masters of Health Science program, involved students and faculty were given access to GPT-4.
   A. Four research assistants were selected based on their performance on the first exam. They were tasked to create vignette style MCQs using ChatGPT-4, while the rest of the class (n = 41) continued to generate questions manually.
   B. All students were asked to create questions that are congruent with the specific learning objectives from the lectures.
   C. Students using ChatGPT-4 were required to upload the lecture material in PowerPoint format to exclusively generate questions from the material provided.

II. Data Collection and Analysis:
   A. The process included recording prompt iterations, time for generated MCQs, and tracking hallucinations for more than 100 questions.
   B. Faculty reviewed the AI-generated MCQs for coverage of basic science concepts before their inclusion in a formative and summative course assessment.

III. Evaluation Survey:
   A. A 51 item survey was prepared, reviewed, and improved by two specialists in the Office of Institutional Effectiveness.
   B. The survey was deployed to 41 graduate students at the end of the fall semester. In order to evaluate the perceptions of ChatGPT4 generated MCQs in terms of usability, accuracy, and effectiveness.
   C. Part of the survey addressed the students' ability to differentiate between manually created and AI-generated MCQs. .
   D. A statistician in the Office of Institutional Effectiveness conducted the preliminary analysis.

Results

Research Assistant	Average Iterations per Question	Average Time per Question (min)	Rate of Hallucinations (%)
1	2.1	2.3	3.5%
2	1.7	1.8	33.3%
3	1.1	7.5	0.0%
4	1.6	3.2	0.0%

Table 1: Comparing ChatGPT-4 Users Rate of Hallucinations. The rate of hallucinations varies based on the researcher's understanding of prompt engineering as well as dept of knowledge of the content.

Learning Objective	Formative Assessment Point Biserial	Summative Assessment Point Biserial
Determine what metabolites/enzymes are used as markers to assess liver function impairments.	0.36	0.36
Specify the usefulness of determining the levels of acute-phase response proteins i.e (CRP)	0.47	
Identify the effects of alcohol abuse on the major liver and brain functions	0.22	0.44
Diagnose the type of impairment based on a decrease or increase of proteins/compounds unique to the liver	0.48	

Table 2: Performance of ChatGPT 4.0- generated vignette style MCQs in formative and summative assessments in ExamSoft. More challenging and discriminating questions have a higher point biserial. The point biserial average is based on the differentiation between the upper quartile and lower quartile of the class.

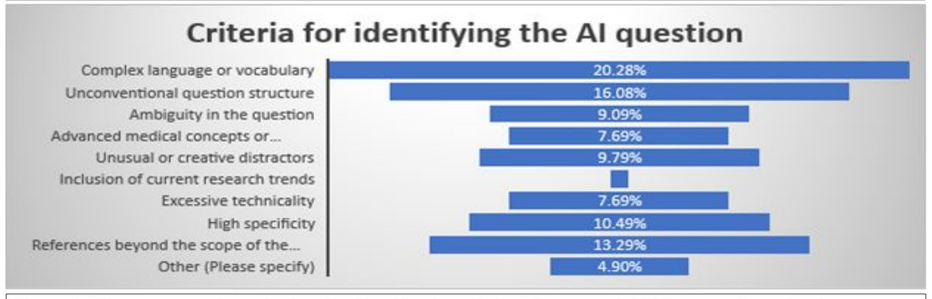

Figure 1. The most popular criterion for identification of the AI-generated MCQs was "complex language or vocabulary" (20%) and the least was "inclusion of current research trends" (0.7%). This will help faculty identify the provenance of item writing assignments when students utilize AI unethically.

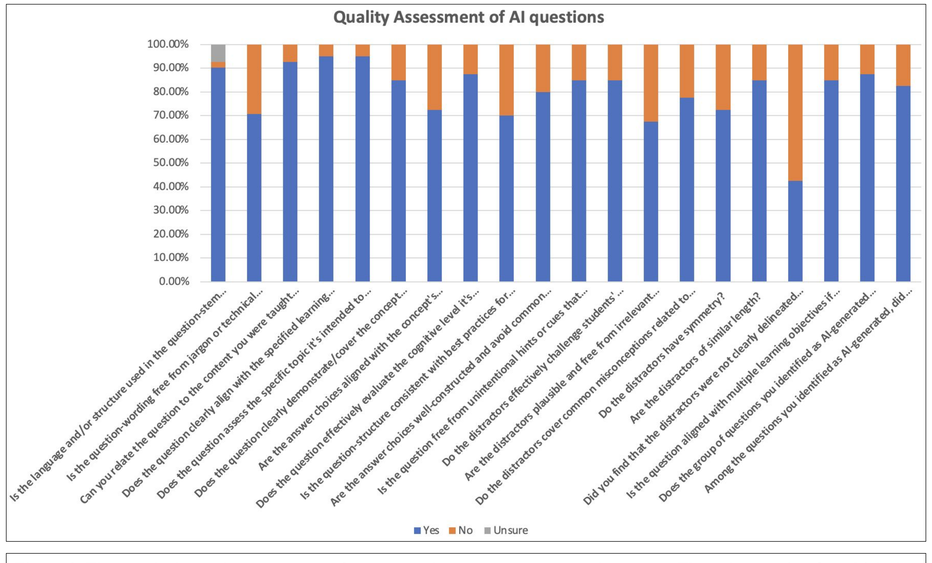

Figure 2. The graph shows the percentage of yes, no, and unsure responses to the survey. The only question for which the percentage of "no" responses were higher than "yes" was "Did you find that the distractors were not clearly delineated or thought there was more than one answer?" Consequently, the ChatGPT- generated MCQs were perceived as clear.

Discussion and Conclusion

- Impact of Prompt Engineering:
  - There was significant variance in MCQ quality between research assistants, emphasizing the familiarity of prompt engineering between participants and their depth of content knowledge (Table 1).
  - Rate of hallucination depends on the user, highlighting the need for oversight and validation of AI-generated content (Table 1).
- Students' Ability to Identify AI-Generated Content:
  - When given a series of vignette style of questions, half of which were AI-generated, students showed varying abilities to differentiate between them. This underlines the high quality of ChatGPT MCQs in relation to those generated by the faculty (Fig 1).
- Perception of AI-Generated Questions:
  - Almost 95% of students reported that ChatGPT-4 questions were "highly satisfactory" in assisting them understand complex concepts, while assessing their critical thinking skills (Fig 2).
- Efficiency of ChatGPT-4 in creating MCQs
  - Our preliminary data show that the use of ChatGPT-4 to create MCQs is an effective method as long as students know their content in depth to identify hallucinations.
  - It is noteworthy to indicate that initiating new conversations with ChatGPT-4 will lead to reducing the rate of hallucinations.
  - The performance (point biserial) of these type of questions indicate that faculty can use ChatGPT-4 to create MCQs not only for formative but also for summative assessments.

Future Considerations

- Develop comprehensive guidelines and training for prompt engineering to optimize AI use in graduate and medical education.
- Increase faculty involvement in not only the evaluation process for accuracy and educational relevance, but also the ethical use of AI.

Acknowledgments

Our gratitude go to
- Masters in Medical Health Science (MHS) program for the financial support of this project.
- MHS '24 for completing the survey
- Dr. Murukutla and Dr. Chenthittayil for their feedback on the survey.

References

Available Upon Request.