BELMONT UNIVERSITY
THOMAS F. FRIST, JR. COLLEGE OF MEDICINE

Evaluating the utility of ChatGPT as a study aid for biochemistry in medical school

Aneesa Jones[1], Isaiah Ware[1], Kieffer Hellmeister[1], **Peter J. Huwe[2]**

[1]Mercer University School of Medicine
[2]**Frist College of Medicine at Belmont University**

Abstract

Some medical students have used ChatGTP as a study aid in medical school without understanding the limitations of this new technology. Here, we provide ChatGPT with prompts from actual medical school syllabi and analyze the accuracy and adequacy of its responses using a Likert scale. Our results demonstrate that ChatGPT is not a sufficiently reliable source of information for biomedical studies.

Methods

All biochemistry learning objective prompts from the MUSM Block 1, Module 1 syllabus were submitted to ChatGTP. AI-generated responses were independently scored by one biochemistry professor and three medical students (two MS1 and one MS2) for perceived accuracy and adequacy. A total of 53 prompts were evaluated using the following Likert scale:

1. The answer provided by ChatGPT does not cover the necessary information or is inaccurate
2. The answer provided by ChatGPT only partially covers the necessary information or may contain some inaccuracies
3. The answer provided by ChatGPT is generally accurate but may not provide all the necessary information
4. The answer provided by ChatGPT covers most of the necessary information, but there may be some gaps or areas where more detail is needed.
5. The answer provided by ChatGPT covers all the necessary information and is accurate.

Results

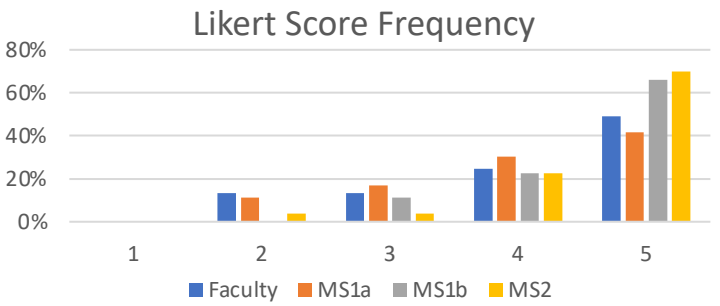

Figure 1. Frequency of Likert evaluations for AI responses to medical school biochemistry learning objective prompts (n = 53) by each of four evaluators.

	Avg score	MS Avg – Fac Avg	Frequency [MS – Fac] > 1
Faculty	4.09		
MS1a	4.01	-0.08	25%
MS1b	4.55	+0.46	21%
MS2	4.58	+0.49	23%

Table 1. Comparison of of evaluations by a biochemistry faculty member (n = 1) versus students (n = 3) who completed their evaluations at the end of their first (MS1a and MS1b) and second (MS2) years of medical school.

Discussion

None of the AI responses received a score of "1", and 41-70% of the prompts received a perfect score of 5. The faculty and MS1a evaluators determined that more than half of the AI responses either lacked important details or contained inaccuracies. Lower scores were predominantly assign based on perceived lack of important information, however several inaccuracies were noted.

In 21-25% of the responses, that faculty and student evaluators differed by a score of 2 or more. This may be due to differences in level of subject area mastery or differences in perceptions of of the depth of knowledge appropriate for first year medical students.

Conclusions

While the AI generated responses to medical school learning objective prompts were usually accurate, they often lacked important details and occasionally made mistakes. We conclude that ChatGPT is not currently a sufficiently reliable primary study aid for MS1 biochemistry curriculum.

We also noted significant (>1) scoring differences between faculty and students for >20% of responses. This suggests that students' appraisal of the adequacy of learning resources differs from that of the subject area faculty.

This study is being expanded to other subject areas.