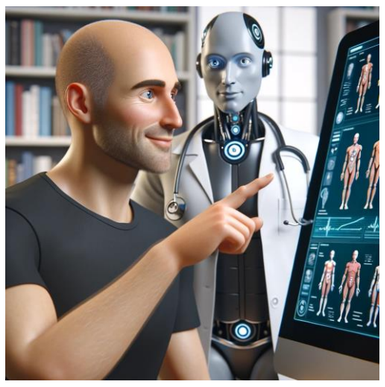


Comparative Analysis of LLMs in Developing Medical Board Style Questions and a GPT-4 Based Interactive Clinical Vignette Bot

Sage Arbor, PhD[1], Raquel Ritchie PhD[2], Verena Van Fleet PhD[3], Mariluz Henshaw PhD[4], Tafline Arbor PhD[5].
[1]Duke Clinical Research Inst. (DCRI), [2]Michigan State U., [3]Northwestern Health Sciences U., [4]Touro U., [5]Wake Forrest U.

Introduction/Background

Exploring the capabilities of different Large Language Models (LLMs) in medical education, this project compares GPT-3.5, GPT-4, and Llamma-v1, and Llamma-v2 models in generating medical board style exam questions. Each question, structured as a clinical vignette, includes multiple answer options, explanations for the correct & incorrect answers, as well as relevant medical objectives and disciplines. Additionally, a GPT-4 based bot was developed to create clinical vignettes interactively based on user input, offering a dynamic educational tool. Exam questions for each LLM were rated my medical school faculty and the GPTs slightly outperformed the Claude models. Faculty with less Ai experience expressed greater interest in using LLMs after their evaluation process.

Methods

The study involved a comparative analysis of GPT-3.5, GPT-4, Claude-V1, and Claude-V2 models in generating high-quality medical questions (GPTs madeby openAI, Claudes by Anthropic[2]). Each model's output to the prompt shown in Figure-1 was evaluated for accuracy and educational value. Three biochemistry faculty evaluated 20 questions, five from each model with the results in Figure-2 and Figure-3.

The ability to create clinical questions without domain specific prompts was tested by providing one of the biochemistry question as a template with the same prompt but asking for anatomy questions which was then reviewed by an anatomist, results in Figure-4.

Faculty reviewers were asked questions about their view of AI before and after evaluating the questions, results in Figure-5. 120 questions were made, the bulk using Claude-v1 the last week of October, the rest were made the last week of november (time given since models can change temporally).

A GPT-4 based interactive bot was also developed, capable of generating clinical vignettes in real-time based on user queries (Figure-6), with additional prompts such as:

- Not given example questions which could shift it to that discipline.
- Start by asking simple "Would you like the question or summary first?" and order the content in line with request
- includes a learning objective database previously created to cover all medical school disciplines and content.[3]
- If it is useful for learning to make an image use DALLE to make a medically accurate diagram.
- they guess the answers BEFORE you output the answer explanation and objective.

Figure-4: Score by Discipline.

Note 3 biochem and only 1 anatomy faculty reviewed 15 and 5 GPT4 questions respectively.

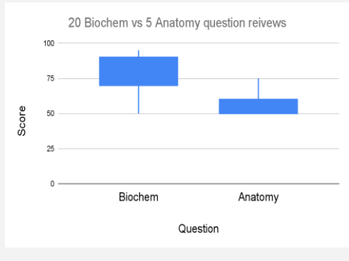
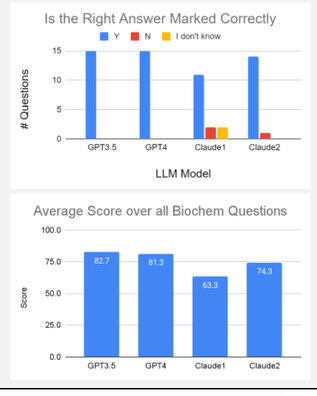


Figure-3: Avg Score

The average biochemistry scores across all 60 reviews (5 questions reviewed in triplicate for each LLM).



Duke Clinical Research Institute

Results

The comparison revealed varying levels of efficacy among the LLMs, with GPT3.5 and GPT-4 showing superior performance in creating comprehensive and accurate medical questions, while Claude-v2 ranked 3rd and Claude-V1 was worst (Figure-2 & 3).

Initially GPT4 created biochemistry questions when prompted to make anatomy questions, likely due to the significant number of biochem words in the example question. However with one additional single sentence prompt it generated high quality anatomy questions, the GPT4bot prompt always returned questions correct to the discipline and disease.

The faculty that had not used AI to develop questions in the past were more likely to use AI in the future after the evaluation process, suggesting efforts to expose faculty to AI capabilities will increase utilization rates.

The GPT-4 based interactive bot successfully generated clinical vignettes on demand, demonstrating its potential as an innovative educational tool for medical students and professionals (but was not rigorously evaluated).

Figure-1: Prompt to LLMs

I would like you to make board style examination questions for medical students that are becoming doctors. Questions should be a clinical vignette, with a patient presentation, usually have 5 answers options (allowing 4-6 answers if that makes more sense), include an explanation of why the correct answer is correct compared to other answers, a learning objective related to the question (try to reuse the same learning objectives across some questions if it makes sense to), and also tag each question with the biological "System" from the system list below. There are two example questions to give you an idea of the format and caliber.

System start | [list of biomed systems: e.g. Biochemistry, Nervous System] | System end

An example is below:
Q1 start | [high quality biochem Q, with explanation, system & objective tag] | Q1 end
Q2 start | [high quality biochem Q, with explanation, system & objective tag] | Q2 end

Figure-2: LLM Model quality.

Biochemistry faculty evaluated questions overall (A), if the answers are correctly labeled, are explained well (B), with a good objective (C), and have a good discipline tag. The

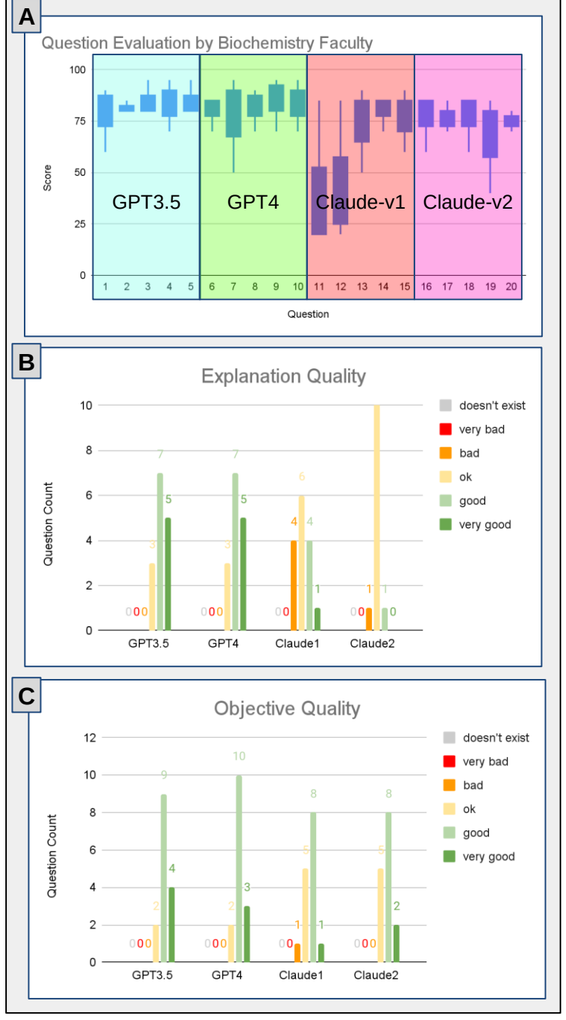
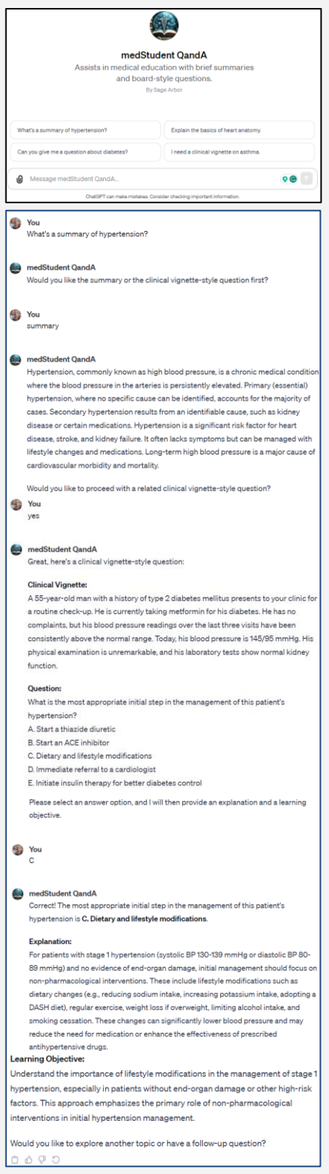


Figure-6: MedQ bot



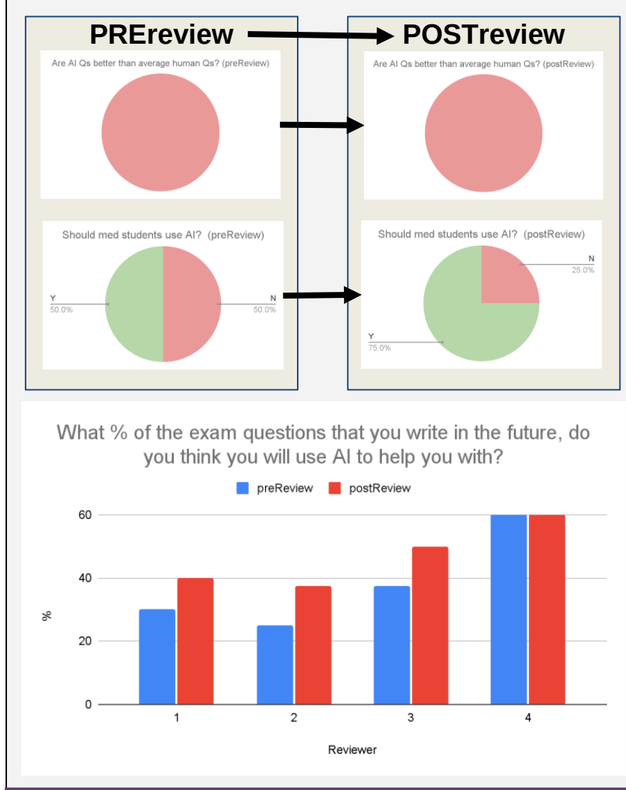
Conclusion

This study highlights the effectiveness of advanced LLMs, particularly the GPTs, in enhancing medical education resources. The GPT-4 based interactive bot represents a significant step forward in creating personalized and dynamic learning experiences. Future research could focus on refining the bot's capabilities and exploring its integration into medical education curricula.

The clinical vignette bot is open for anyone with a chaGPT+ account at https://chat.openai.com/g/g-BChrQDLES-medstudent-qanda , but not freely available as each time it is used there are marginal API costs.

Figure-5: Reviewer LLM Model quality

Biochemistry faculty answered a survey before and after reviewing AI generated questions.



References

[1] OpenAI. GPT-4 Technical Report. Preprint at https://doi.org/10.48550/arXiv.2303.08774 (2023).

[2] Anthropic [Internet] Claude 2. Available from: https://www.anthropic.com/index/claude-2

[3] Arbor, S. et al. Development and Validation of a Shared Secure Biochemistry Test Bank for Medical, Dental, and Pharmacy Schools. Med.Sci.Educ. (2020) doi:10.1007/s40670-020-00919-y.

DukeHealth

Health Data Science Poster Showcase

Duke University, December 8, 2023