Performance of ChatGPT vs. HuggingChat on OB-GYN Topics

Gabrielle Kirshteyn BS[1] and Mark Chaet MD[1]

Introduction

- Large language models (LLMs), a subset of artificial intelligence (AI), are constructed of vast computational algorithms that analyze data and train the machine to make autonomous conclusions [1].
- ChatGPT is an LLM that operates on a proprietary framework [2].
- Hugging Face's chatbot, HuggingChat, is developed on an open-source platform, allowing for community-based contributions and enhancements [3].
- The consensus to use LLMs as a tool in medicine has not yet been achieved, as the reliability of the information they provide is not yet fully understood [4].
- Studies have examined ChatGPT's capabilities in undertaking medical exams [5], however, there is a notable lack of comparisons between its scores and those of other AI entities, like HuggingChat [6].
- Analyzing the knowledge of AI through its responses to medical questioning could elucidate the strengths of using its resources in the medical field.

Purpose

- To quantify and compare the performance of two AI modalities, ChatGPT and HuggingChat, on medical examination questions testing OB-GYN proficiency.

Methods

- ChatGPT and HuggingChat were each subjected to two standardized question banks: Test 1 and Test 2.
- Test 1 was composed of 20 questions and was compiled from the free online Obstetrics & Gynecology Sample Items developed by the National Board of Medical Examiners (NBME).
- Test 2 was composed of 50 questions gathered from the Comprehensive 1: 50 question exam (2022) created by the Association of Professors of Gynecology & Obstetrics (APGO) Web-Based Interactive Self-Evaluation (uWISE).
- The passing score for each exam is 70%, which was determined by each of the test makers (NBME and APGO).
- The 70 questions included in the examination data set only included textual prompts and did not include images.
- All questions were multiple-choice formatted in that the question prompt was followed by its associated set of multiple-choice answers.
- The ChatGPT (GPT-3.5) August 3, 2023 version was used.
- The HuggingChat model meta-llama/Llama-2-70b-chat-hf was used.
- We manually entered questions into the ChatGPT and HuggingChat chat prompts. We then manually recorded the AI's multiple-choice answer and directly copied its explanation into a spreadsheet.
- We employed a two-proportion z-test for each examination.
- The significance level was set at $\alpha = 0.05$ for determining statistical significance.

Results

- The two-proportion z-test revealed no statistically significant difference in performance between ChatGPT and HuggingChat on both medical examinations.
- For Test 1, ChatGPT correctly answered 18 out of 20 questions (90%), while HuggingChat correctly answered 17 out of 20 questions (85%) ($p = 0.6$).
- For Test 2, ChatGPT correctly answered 35 out of 50 questions (70%), in contrast to HuggingChat's 31 correct answers out of 50 questions (62%) ($p = 0.4$)
- On Test 1, a wrong answer was commonly generated by both AI modalities on one question.
- On Test 2, there were seven questions that both HuggingChat and ChatGPT got incorrect.

Conclusions

- Determining the reliability of AI's information database can help justify its use as a resource in the medical field.
- We found that the two LLMs can compute medical questioning and formulate responses.
- ChatGPT outperformed HuggingChat on both examinations, however, the differences in their performances were not statistically significant.
- There were a handful of questions where both AI programs produced an incorrect answer. We believe that the programs struggle to make assumptions and read between the lines, as they did not analyze aspects that were not explicitly stated.
- There are numerous advantages to the use of LLMs' in the medical field. It provides accessible, detailed information tailored to individualized topics and allows for conversation (with the AI) about the subject, which could enhance comprehension and fill knowledge gaps
- The programs serve to provide relevant information on a particular subject, but the results should be verified before their acceptance, as some may be incorrect.

AI system	Test	Questions answered correctly	Performance (%)	Outcome	p-value
ChatGPT	Test 1	18/20	90%	Pass	0.6
HuggingChat	Test 1	17/20	85%	Pass	0.6
ChatGPT	Test 2	35/50	70%	Pass	0.4
HuggingChat	Test 2	31/50	62%	Fail	0.4

Table 1. Performance of ChatGPT and HuggingChat on Test 1 and Test 2

References

1. Machine learning and artificial intelligence: definitions, applications, and future directions. Helm JM, Swiergosz AM, Haeberle HS, et al. Curr Rev Musculoskelet Med. 2020;13:69–76.
2. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Tian S, Jin Q, Yeganova L, et al. Brief Bioinform.
3. HuggingChat. [Sep; 2023]. https://huggingface.co/chat
4. Artificial intelligence in academic writing: a paradigm-shifting technological advance. Golan R, Reddy R, Muthigi A, Ramasamy R. Nat Rev Urol. 2023;20:327–328.
5. ChatGPT performance on the American Urological Association Self-assessment Study Program and the potential influence of artificial intelligence in urologic training. Deebel NA, Terlecki R. Urology. 2023;177:29–33.
6. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. JMIR Med Educ. 2023;9:0.