Potential Misinformation and Dangers Associated with Clinical Use of Large Language Model Chatbots

Branden Rafael Sosa[1], Michelle Cung, Vincentius J Suhardi, Kyle Morse[2], Andrew Thomson, Sravisht Iyer, Matthew B Greenblatt[1]
[1]Weill Cornell Medical College, [2]Hospital For Special Surgery

INTRODUCTION:

Over the last year, there has been great interest in the potential applications of artificial intelligence (AI) large language models (LLMs). [1] The performance and accuracy of these models will be highly dependent on the specific textural sources used during training, and it is unclear whether the relevant orthopaedic basic and clinical literature was adequately represented in the training dataset. Therefore, it is important that the performance of these LLMs be specifically assessed with respect to different clinical and biomedical research disciplines so that those in the field can understand both the relative utility of these tools and whether their use by patients or non-specialist colleagues could introduce bias or misconceptions. [2,3] In this study, we assess the accuracy of multiple large language models to explain basic orthopaedic concepts, synthesize clinical information, and address patient queries.

METHODS:

Publicly available LLM chatbots, Open AI ChatGPT 4.0, Google Bard, and BingAI chatbots were prompted to answer 45 orthopaedic-related questions spanning categories of “Bone Physiology,” “Referring Physician,” and “Patient Query” and assessed for accuracy. Two independent, blinded reviewers scored responses on a scale of 0-4 assessing for accuracy, completeness, and useability. Responses were analyzed for strengths and limitations within categories and across chatbots.

RESULTS:

ChatGPT was able to appropriately answer 83.3% of bone physiology while BingAI achieved 23.3%. (Fig.1, $p<0.01$). Google Bard refused to answer 73.3% of bone physiology questions stating the question is outside its capacity as a LLM. When providing clinical management suggestions, all chatbots displayed significant limitations deviating from the standard of care and omitting critical steps in work up. Nonetheless, when asked less complex patient queries, ChatGPT and Bard were able to provide mostly accurate responses but often failed to elicit critical medical history pertinent to fully addressing the question.

DISCUSSION AND CONCLUSION:

LLM chatbots possess remarkable capacity to provide concise summaries across a wide range of orthopaedic subject matter but have limited accuracy depending on the category of orthopaedic questions asked. ChatGPT outperformed Bard and Bing AI across all categories but exhibited significant limitations in addressing queries on contemporary research advancements and clinical decision making. When advising on clinical management, all LLM chatbots made similar errors such as ordering antibiotics before cultures or neglecting to include key studies in diagnostic work up. In some instances, executing the suggested patient management plans could have led to adverse outcomes for patients as they significantly diverge from the standard of care.

A careful analysis of the citations provided by the chatbots revealed major flaws. ChatGPT provided 10 faulty links across 15 questions that were non-functional or led to incorrect articles. The chatbot references varied broadly from peer-reviewed clinical trials to Wikipedia, with citations often repeated between questions despite a small question set, suggesting the presence of bias or “oversampling” of a small number of references. This suggests that creation of dedicated LLMs that utilize more tightly curated, level-1 peer-reviewed training datasets with a focus on clinical applications may be needed to improve their performance when addressing biomedical topics, particularly if these LLMs are going to be used formally or informally to guide clinical decision making.

Despite having the potential to serve as a powerful tool, these chatbots do not currently possess the capacity to replace the expertise of an orthopaedic surgeon. Consulting a LLM chatbot when initially exploring a subject allows users to obtain quick summaries curated from multiple resources in ways not allowed by conventional search engine queries. Nonetheless, this study highlights the limitations embedded into chatbots and their responses should be interpreted with caution.

References:

1. Intelligence, S.U.H.-C.A., *Artificial Intelligence Index Report*, in *Measuring Trends in Artificial Intelligence*. 2023, Stanford University.
2. Gilson, A., et al., *How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment.* JMIR Med Educ, 2023. **9**: p. e45312.
3. Patel, S.B. and K. Lam, *ChatGPT: the future of discharge summaries?* The Lancet Digital Health, 2023. **5**(3): p. e107-e108.

Fig. 1: Performance of LLM chatbots ChatGPT 4.0, Google Bard, and Bing AI when prompted with queries related to three categories: A) Bone Physiology, B) Patient Queries, or C) Physician Management. Two independent, blinded reviewers rated responses on a scale 0-4. Scores from both reviewers were plotted and median values are indicated by horizontal bar. (** p<0.05, **p<0.01, ***p<0.001*)

Figure 1