Evaluating Retrieval Augmented Generation and ChatGPT’s Accuracy on Orthopaedic Examination Assessment Questions

Jordan Eskenazi[1], Varun Krishnan[2], Maximilian Konarzewski, David S Constantinescu, Gilberto Lobaton, Seth D Dodds[3]
[1]Orthopedic Surgery, University of Miami Miller School of Medicine, [2]Orthopedic Surgery, [3]UNIVERSITY OF MIAMI, MILLER SCHOOL OF MEDICINE

INTRODUCTION: Since the introduction of Large Language Models (LLMs) such as ChatGPT, there has been a race to test its capability in medical problem solving across specialties to varying degrees of success. Retrieval Augmented Generation (RAG) allows LLMs to leverage subject specific knowledge to provide context, a greater number of sources, and the ability to cite medical literature to increase the accuracy and credibility of its answers. The use of LLM + RAG has not yet been used in the appraisal of Artificial Intelligence's capability of Orthopedic problem solving.

METHODS:

The AAOS OrthoWizard question bank was used as the source of questions. After 13 textbooks and 28 clinical guidelines were made available for RAG, text-only questions were presented in a zero-shot learning fashion to ChatGPT-4+RAG, ChatGPT-4, and ChatGPT-3.5.

RESULTS:

On 1023 questions tested, ChatGPT-3.5, ChatGPT-4, ChatGPT-4+RAG, and humans scored 52.98%, 64.91%, 73.80%, and 73.97%, respectively. There was no statistical difference between Orthopedic surgeons and ChatGPT-4+RAG on overall accuracy ($p=0.997$). Both Orthopedic surgeons and ChatGPT4+RAG scored better than ChatGPT-4 ($p<0.001$) and ChatGPT-3.5 ($p<0.001$). Of the 13 textbooks available to RAG, RAG used AAOS Comprehensive Review 3 Volume 3 for 39.6% of questions, more often than any other resource available to it.

DISCUSSION AND CONCLUSION:

ChatGPT-4+RAG was able to answer 1023 question from the OrthoWizard question bank at the same accuracy as Orthopedic surgeons. Both ChatGPT-4+RAG and Orthopedic surgeons had superior accuracy on these specialty exam questions compared to ChatGPT-4 and ChatGPT-3.5. Artificial intelligence is becoming increasingly accurate in its ability to answer orthopaedic surgery test questions with the guidance of orthopaedic surgery textbooks. RAG enables an LLM to effectively cite its sources after providing an answer to a question, which is an important tool for the integration of LLMs to orthopaedic surgery problem solving.