Artificial Intelligence Large Language Models are Nearly Equivalent to Fourth Year Orthopaedic Surgery Residents on the Orthopaedic In-Training Examination: A Cause for Concern or Excitement?

Ashraf Nawari, Jamal Zahir, Sonal Kumar, Lovingly May Ocampo, Olivia A Opara, Hassan Rao Ahmad, Benjamin Crawford[1], Brian T Feeley
[1]St Marys Medical Center

INTRODUCTION:
The rapid improvement of artificial intelligence (AI) models with both supervised and unsupervised learning, along with previous success in answering board style questions, warrants further investigation regarding their utility and accuracy in answering orthopaedic surgery written board questions. Previous studies have analyzed the performance of ChatGPT alone on board exams but a head-to-head analysis of multiple current AI models has yet to be assessed. This study compared the utility and accuracy of various large language models (LLMs) in answering Orthopaedic Surgery In-Training Exam (OITE) written board questions against each other as well as orthopaedic surgery residents.
METHODS:
A complete set of questions from the OITE 2022 exam was inputted into various LLMs. The answers from each LLM were recorded, tabulated, and the percentage correct was calculated and compared against orthopaedic surgery residents nationally. Results were analyzed by overall performance and question type. Level A questions related to knowledge and recall of facts, Level B questions involved diagnosis and analysis of information, and Level C questions focused on the evaluation and management of diseases, requiring knowledge and reasoning to develop treatment plans. Significance was determined by p values less than or equal to 0.05.
RESULTS:
Google Gemini was the most accurate tool answering 69.9% of questions correctly, just shy of statistical significance ($p=0.052$). Google Gemini also performed superiorly to ChatGPT and Claude on Level A ($p=0.070$) and Level C ($p=0.116$) questions, with Claude performing superiorly on Level B questions (0.411). All LLMs performed above the average of a first-year orthopaedic surgery intern, with Google Gemini and Claude performance approaching that of fourth and fifth-year orthopaedic surgery residents.
DISCUSSION AND CONCLUSION:
The study assessed LLMs like Google Gemini, ChatGPT, and Claude against orthopaedic surgery residents on the OITE. Results showed that these LLMs perform as well as or better than orthopaedic surgery residents, with Google Gemini excelling in Type A and C questions and Claude in Type B questions.