How does a Large Language Model Artificial Intelligence Fare with Spine Knowledge?

Zachary Lum[1], Dylon Patrick Collins, Lohitha Guntupalli, Stanley Robert Dennison, Soham Choudhary, Augustine M Saiz, R Lor Randall[2]
[1]South Florida Institution Sports Med, [2]UCDavis Health

INTRODUCTION: As the capabilities of artificial intelligence (AI) continue to advance, it is important to regularly evaluate competency to maintain high standards and prevent potential errors or biases which could deliver misinformation that could harm patients or spread inaccurate information. A new AI model using large language models (LLM) and non-specific domain areas has gained recent attention in its novel way to process information. We sought to test its performance to correctly answer sports medicine questions compared to other subject types and taxonomy question type (recall, interpretation, knowledge application).
METHODS: We asked ChatGPT, 3,173 questions based on the Orthopaedic In-Training Exam (OITE) and 757 questions from the real OITE. Questions were categorized by subject type, and by taxonomy type. These questions were then entered into the AI chatbot and score was recorded. Multivariate logistic regression analysis was performed comparing spine questions with other question types, and taxonomy type.
RESULTS: After exclusions, ChatGPT answered 960/1,871 (51%) of total questions correctly and 67/141 (48%) of spine correctly, which was a lower-range performing subspecialty type. Spine exhibited worse performance than basic science ($p<0.001$, OR 0.42) and pathology/oncology ($p=0.007$, OR 0.48). When evaluating subgroup taxonomy analysis, univariate logistic regression demonstrated the AI's lower performance in taxonomy type 3 compared to type 1 (50% vs. 41%, $p=0.049$).
DISCUSSION AND CONCLUSION: This AI LLM may be less effective in answering orthopaedic questions related to spine knowledge. Furthermore, the study's taxonomy analysis highlights the importance of considering the question structure when evaluating AI performance. Ultimately, as AI continues to evolve and advance, it will be important to consider its limitations and potential biases to ensure its responsible and ethical use.