A Comparative Study of Artificial Intelligence Models (ChatGPT, Bard, Bing), Medical Students, and Orthopaedic Surgery Residents on the Orthopaedic In-Training Examination

Gage A Guerra BA[1], Hayden L Hofmann BS[1], Alexander M Wong BS[1], Jonathan L Le BS, MS[1], Amir K Fathi BS[1], Jacob L Kotlier, BA[1], Cory K Mayfield MD[1], Frank A Petrigliano MD[1], Joseph N Liu MD[1]

[1]USC Epstein Family Center for Sports Medicine at Keck Medicine of USC, Los Angeles, CA USA; gageguer@usc.edu

Disclosures: There are no financial disclosures or conflicts of interest to report from all authors listed.

INTRODUCTION: Advances in artificial intelligence (AI) like OpenAI's Chat Generative Pre-Trained Transformer (ChatGPT), Google's Bard, and Microsoft's Bing Chat have the potential to revolutionize medicine. These models have demonstrated the ability to generate unique and nuances responses to a wide range of topics, notably passing the United States Medical Licensing Exam (USMLE). The present study aims to assess the three AI models' ability to synthesize clinical information in the field of orthopaedic surgery by evaluating their performance on the Orthopaedic In-Training Examination (OITE).

METHODS: The OITE question sets from 2021 and 2022 were compiled to form a set of 420 questions. Questions requiring the interpretation of an image to be answered were eliminated; the remaining questions were answered by ChatGPT (model GPT-3.5), Bard, and Bing Chat. Overall accuracy was determined by calculating the weighted average between the two question banks. A chi-squared test was used for a comparative analysis between ChatGPT, Bard, Bing Chat, 4000 orthopaedic residents, and a small cohort of medical students using Microsoft Excel. Significance was determined to be $p < 0.05$.

RESULTS SECTION: **Figure 1** shows the comparative accuracy of the AI models and human cohorts on text-only questions. ChatGPT correctly answered 49.1% of questions (115/234), Bard correctly answered 52.4% of questions (118/225), and Bing Chat correctly answered 53.5% of questions (123/230). By PGY1-5, orthopaedic residents correctly answered 53.1%, 60.4%, 66.6%, 70.0%, and 71.9%, respectively. The medical student cohort correctly answered 30.8% of the composite OITE questions, an accuracy significantly lower than ChatGPT ($p=0.0012$), Bard ($p<0.001$), and Bing ($p<0.001$). There was no significant difference in composite accuracy between PGY-1 and ChatGPT ($p=0.12$), Bard ($p=0.13$), and Bing ($p=0.077$). All three AI models were less accurate on image-associated questions than on text-only questions. ChatGPT correctly answered 42.6% of image-associated questions (75/176), whereas Bard and Bing Chat correctly answered 41.8% (74/177) and 50.9% (82/161), respectively. ChatGPT, Bard, and Bing Chat rejected a response to an image-associated question 10 (2.3%), 18 (4.3%), and 29 (6.9%) times, respectively **(Table 1)**.

DISCUSSION: ChatGPT, Bard, and Bing Chat completed OITE questions with an accuracy similar to first-year orthopaedic residents. All three AI models demonstrated a capacity to synthesize clinical orthopaedic information with similar accuracy. Our results demonstrate the clinical potential of these technologies while also indicating a lack of orthopaedic expertise seen in senior residents. Physicians should be attuned to the improving clinical accuracy of future AI models.

SIGNIFICANCE/CLINICAL RELEVANCE: The AI models' accuracy on orthopaedic board-style questions suggests potential applications in medical educational settings and clinical decision-making.

IMAGES AND TABLES:

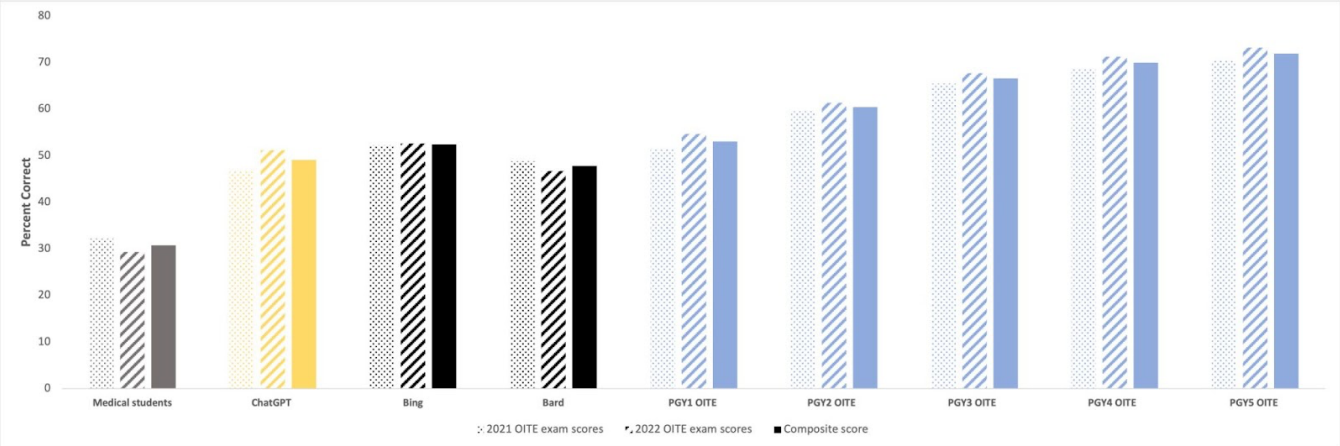

Figure 1: Comparative performance of ChatGPT, Bing, Bard, and the human cohorts on 2021 OITE, 2022 OITE, and composite examinations.

	Number of 2021 OITE available questions attempted (%)	Number of 2022 OITE available questions attempted (%)	Total number of available questions attempted (%)
ChatGPT	205 (96.2)	205 (99.0)	410 (97.6)
Bard	207 (97.2)	195 (94.2)	402 (95.7)
Bing	197 (92.5)	194 (93.7)	391 (93.1)
Medical students	213 (100)	207 (100)	420 (100)
PGY1-5 OITE test-takers	263 (100)	264 (100)	527 (100)

Table 1: Number of questions that were attempted and scored by AI models and human cohorts.