Artificial Intelligence in Orthopaedic Education: A Comparative Analysis of ChatGPT and Bing Artificial Intelligence’s Orthopaedic In-Training Examination Performance

Clark Jia-Long Chen[1], Duncan Van Nest, Vivek Koratkar Bilolikar, Solomon Samuel, James S Raphael[1], Gene W Shaffer
[1]Albert Einstein Med Ctr

INTRODUCTION: Large language models (LLMs) are artificial intelligence (AI) tools developed through machine learning algorithms that produce human-like responses. The purpose of this study is to explore how the potential application of AI tools can augment and enhance traditional educational methods. We evaluated the performance of these AI models on the Orthopaedic In-Training Examination (OITE), an annual exam administered to US orthopaedic residency programs by the American Academy of Orthopaedic Surgeons (AAOS).

METHODS:

The performance of two AI systems, ChatGPT and Bing AI were evaluated on a standardized set of multiple-choice questions drawn from the AAOS OITE online question bank spanning five years (2018 – 2022). A total of 1,165 questions were posed to each AI system. To prevent learning from question to question, a new chat was initiated for each response. Guided prompts were used in both models. When there was a radiographic image, the system was fed with a minimal description of the image. The performance of both systems was standardized using the latest version of ChatGPT 3.5 and Bing AI. As a comparison, historical data of resident scores were taken from the annual OITE technical reports. The results were analyzed using statistical methods to determine overall performance.

RESULTS: Across the five datasets, ChatGPT scored an average of 55.0% on the OITE questions. Bing AI scored higher, with an average of 80.0%. In comparison, the average performance of orthopaedic residents in national accredited programs under the Accreditation Council for Graduate Medical Education (ACGME) was 62.1% (Figure 1). Bing AI outperformed Chat GPT by an average of 25.0%, and outperformed ACGME examinees by 17.9%. ANOVA analysis demonstrated $p<0.001$ among groups. When comparing ChatGPT and ACGME averages results, there was a statistically significant difference ($p=0.004$). When comparing Bing AI and ACGME averages results, there was a statistically significant difference ($p<0.001$). When comparing ChatGPT and Bing AI, there was a statistically significant difference ($p<0.001$). The best performance was by Bing AI on OITE 2020 (Table 1). The model demonstrated a decline in ChatGPT performance over the later years of the OITE test. ChatGPT was limited by its inherent pretraining data up to September 2021. There was appropriate logical justification of responses in all answer from the two LLMs. An example output from each of the LLMs is described in Figure 2. Each response contained elements of internal information from the question as well as outside referenced information. In the case of ChatGPT, information is referenced from pretrained data, while Bing AI referenced relevant internet search results (e.g., PubMed, AAOS, etc.). Only Bing AI provided in-line accessible hyperlink citations for response justification.

DISCUSSION AND CONCLUSION: ChatGPT and Bing AI mark a significant keystone in natural language processing models. By performing at a greater than 63% threshold on the OITE question set, the Bing AI model can be anticipated to achieve the approximate of a passing score for the American Board of Orthopaedic Surgery Part 1 examination. In the rapidly evolving landscape of evidence-based medicine, the body of research grows day-by-day. We highlight the capacity of LLMs to provide logical and informational context across answer responses through its in-depth information search and citation of resources. This combination presents a convincing argument for the possible uses of LLMs in medical education as an interactive learning aid.

Figure 1: Average Performance of Large Language Models on OITE Exams

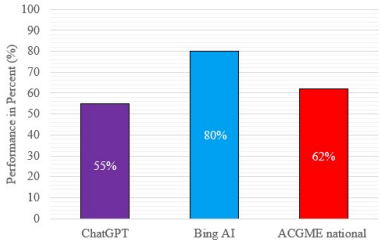

Figure 2

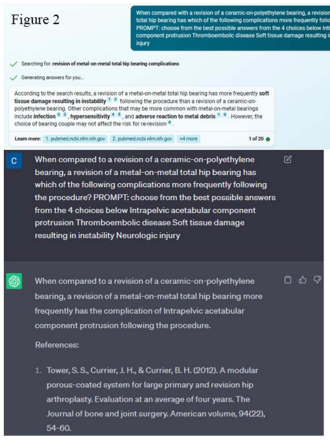

Table 1: ChatGPT, Bing AI, and ACGME national performance on OITE from 2018 to 2022

Year	ChatGPT (%)	Bing AI (%)	ACGME national (%)
2018	57.4	78.9	60.6
2019	54.2	78.9	60.8
2020	54.0	83.7	61
2021	50.7	77.9	62
2022	58.9	80.7	66
Mean	55.0	80.0	62.1