Assessing ChatGPT's Viability as a Cost-Effective Educational Resource in Orthopedic Surgery

Neil Jain[1], Caleb Gottlich[1], Ruthvik Allala[1], Travis Winston[1]
[1]Texas Tech University Health Sciences Center, Lubbock, TX
Njain5466@gmail.com

Disclosures: Neil Jain (None), Caleb Gottlich (None), Ruthvik Allala (None), Travis Winston (None)

INTRODUCTION: Costs associated with current study materials and question banks available to Orthopedic resident physicians to prepare for the Orthopaedic In-Training Exam (OITE) can range upwards of $500 per resource. ChatGPT has gained widespread media attention as a free to use artificial intelligence system with conversational chatbot capabilities able to provide human like responses to inputs. This study assessed ChatGPT's question writing ability and its potential feasibility as a supplemental cost-effective resource in the education of orthopedic surgeons.

METHODS: We directed ChatGPT to produce high-yield multiple-choice questions with accompanying rationales and references across 11 tested domains during the OITE. Questions produced by ChatGPT were analyzed by content experts to determine the accuracy of the question, its answer, and explanation. A 22-question exam featuring one ChatGPT-authored question and one previous OITE question classified by domain was created and administered to the Orthopaedic residents at our institution. Performance on questions authored by ChatGPT was compared to that of the OITE questions. Questions were subjectively assessed by testtakers on a 1-5 scale on the basis of relevance, clarity, specificity, and originality. Paired t-test comparisons were used to understand resident perception of questions and analyze percentage answered correctly.

RESULTS SECTION: ChatGPT repeated itself most frequently in the domains of Sports and Oncology with 9 repeats each. A total of 96 references were used among the 55 generated questions. 41 (42.7%) were found to be fabricated.

In the distributed survey, the average rating given to a ChatGPT written question was 3.12. The average rating given to an OITE question was 3.23. The highest rated questions were found in the Basic Science (n=1), and Oncology (n=2) domains. The lowest rated questions were found in the Shoulder and Elbow (n=1), Hand (n=1), and Rehabilitation (n=1) domains. Questions which had the highest percentage of correct answers were found in the Joints and Rehabilitation domains respectively. Residents overall answered 60.43% of ChatGPT authored questions correctly as opposed to 50.27% of OITE questions correctly.

Paired t-test comparison between ChatGPT and OITE groups on question perception by residents and percentage correct were not significant ($p = 0.2296$, $p = 0.3635$ respectively)

DISCUSSION: The study recognizes the promising nature of utilizing artificial intelligence, represented by ChatGPT, as a supplementary education resource in a highly specialized field like Orthopedic Surgery. The lack of difference in question perception and scoring suggests that ChatGPT has the potential to be a valuable cost-effective resource, offering questions of comparable difficulty to traditional standardized exams. However, it also underscores the importance of addressing limitations, refining the model, and validating its performance before widespread integration. Future research should focus on enhancing ChatGPT's knowledge base, reducing fabrication rates, and establishing ethical content generation practices.

SIGNIFICANCE/CLINICAL RELEVANCE: The rising prominence of artificial intelligence technology, including ChatGPT, has created new avenues for education delivery. Within the context of graduate medical education, ChatGPT provides an exciting opportunity to both address inequities in access to resources and provide regularly up-to-date educational content. The findings contribute to the ongoing discourse on the evolving role of AI in medical education and highlight the need for a balanced approach to harnessing the benefits while mitigating potential challenges.

IMAGES AND TABLES:

Domain	Repeat Count
Basic Science	1
Sports	9
Joints	3
Oncology	9
Rehab	1

Table 1: ChatGPT question repeat counts classified by domain

	PGY-1	PGY-2	PGY-3	PGY-4	PGY-5
Number/Count	4	4	4	4	1
Previous year's OITE score	N/A (n=4)	N/A (n=1) 0-25% (n=1) 26-50% (n=1) 51-75% (n=1)	0-25% (n=1) 26-50% (n=2) 51-75% (n=1)	51-75% (n=3) 76-100% (n=1)	26-50% (n=1)
Survey Score (Average)	11.75/22	13.25/22	13/22	12.5	15/22

Table 2: Survey Data

Domain	Writer	Average Rating (1-5)	Correct Responses (%)
Basic Science	ChatGPT	3.24	76.47
Basic Science	OITE	**3.59**	58.82
Foot & Ankle	ChatGPT	3.18	82.35
Foot & Ankle	OITE	3.24	52.94
Trauma	ChatGPT	3.05	11.76
Trauma	OITE	3.29	17.65
Sports	ChatGPT	3.41	82.35
Sports	OITE	3.29	52.94
Shoulder & Elbow	ChatGPT	2.94	52.94
Shoulder & Elbow	OITE	3.18	76.47
Joints	ChatGPT	3.24	82.35
Joints	OITE	3.05	**94.12**
Oncology	ChatGPT	**3.53**	11.76
Oncology	OITE	**3.53**	29.41
Hand	ChatGPT	3.18	11.76
Hand	OITE	2.71	70.59
Spine	ChatGPT	2.65	52.94
Spine	OITE	3.18	58.82
Pediatrics	ChatGPT	3.18	76.47
Pediatrics	OITE	3.47	58.82
Rehabilitation	ChatGPT	2.76	11.76
Rehabilitation	OITE	3.05	**94.12**

Table 3: Detailed breakdown of question rating by author and domain