Original Paper

Integrating ChatGPT in Orthopedic Education for Medical Undergraduates: Randomized Controlled Trial

Wenyi Gan[1*], PhD; Jianfeng Ouyang[2*], PhD; Hua Li[3*], PhD; Zhaowen Xue[1*], PhD; Yiming Zhang[1*], PhD; Qiu Dong[1], PhD; Jiadong Huang[4], MD; Xiaofei Zheng[1], PhD; Yiyi Zhang[1], PhD

[1]The First Clinical Medical College of Jinan University, The First Affiliated Hospital of Jinan University, Guangzhou, China

[2]Department of Joint Surgery and Sports Medicine, Zhuhai People's Hospital (Zhuhai Hospital Affiliated With Jinan University), Zhuhai, Guangdong, China

[3]Department of Orthopaedics, Beijing Jishuitan Hospital, Beijing, China

[4]Jinan University-University of Birmingham Joint Institute, Jinan University, Guangzhou, China

[*]these authors contributed equally

Corresponding Author:
Yiyi Zhang, PhD
The First Clinical Medical College of Jinan University, The First Affiliated Hospital of Jinan University
No. 613, Huangpu Avenue West
Tianhe District
Guangzhou, 510630
China
Phone: 86 130 76855735
Fax: 86 020 38688563
Email: yiyizjun@126.com

Abstract

Background: ChatGPT is a natural language processing model developed by OpenAI, which can be iteratively updated and optimized to accommodate the changing and complex requirements of human verbal communication.

Objective: The study aimed to evaluate ChatGPT's accuracy in answering orthopedics-related multiple-choice questions (MCQs) and assess its short-term effects as a learning aid through a randomized controlled trial. In addition, long-term effects on student performance in other subjects were measured using final examination results.

Methods: We first evaluated ChatGPT's accuracy in answering MCQs pertaining to orthopedics across various question formats. Then, 129 undergraduate medical students participated in a randomized controlled study in which the ChatGPT group used ChatGPT as a learning tool, while the control group was prohibited from using artificial intelligence software to support learning. Following a 2-week intervention, the 2 groups' understanding of orthopedics was assessed by an orthopedics test, and variations in the 2 groups' performance in other disciplines were noted through a follow-up at the end of the semester.

Results: ChatGPT-4.0 answered 1051 orthopedics-related MCQs with a 70.60% (742/1051) accuracy rate, including 71.8% (237/330) accuracy for A1 MCQs, 73.7% (330/448) accuracy for A2 MCQs, 70.2% (92/131) accuracy for A3/4 MCQs, and 58.5% (83/142) accuracy for case analysis MCQs. As of April 7, 2023, a total of 129 individuals participated in the experiment. However, 19 individuals withdrew from the experiment at various phases; thus, as of July 1, 2023, a total of 110 individuals accomplished the trial and completed all follow-up work. After we intervened in the learning style of the students in the short term, the ChatGPT group answered more questions correctly than the control group (ChatGPT group: mean 141.20, SD 26.68; control group: mean 130.80, SD 25.56; *P*=.04) in the orthopedics test, particularly on A1 (ChatGPT group: mean 46.57, SD 8.52; control group: mean 42.18, SD 9.43; *P*=.01), A2 (ChatGPT group: mean 60.59, SD 10.58; control group: mean 56.66, SD 9.91; *P*=.047), and A3/4 MCQs (ChatGPT group: mean 19.57, SD 5.48; control group: mean 16.46, SD 4.58; *P*=.002). At the end of the semester, we found that the ChatGPT group performed better on final examinations in surgery (ChatGPT group: mean 76.54, SD 9.79; control group: mean 72.54, SD 8.11; *P*=.02) and obstetrics and gynecology (ChatGPT group: mean 75.98, SD 8.94; control group: mean 72.54, SD 8.66; *P*=.04) than the control group.

Conclusions: ChatGPT answers orthopedics-related MCQs accurately, and students using it excel in both short-term and long-term assessments. Our findings strongly support ChatGPT's integration into medical education, enhancing contemporary instructional methods.

Trial Registration: Chinese Clinical Trial Registry Chictr2300071774; https://www.chictr.org.cn/hvshowproject.html ?id=225740&v=1.0

(J Med Internet Res 2024;26:e57037) doi: 10.2196/57037

KEYWORDS

ChatGPT; medical education; orthopedics; artificial intelligence; large language model; natural language processing; randomized controlled trial; learning aid

Introduction

ChatGPT, a natural language processing model developed by OpenAI, is based on a sophisticated machine learning algorithm that can be iteratively updated and optimized to accommodate the changing and complex requirements of human verbal communication [1-3]. ChatGPT-4.0 is significantly superior to ChatGPT-3.5 in terms of language comprehension, context comprehension, generation speed, and interpretability [4]. It can be applied to a variety of natural language processing duties and provides individuals with more precise, efficient, and intelligent natural language processing services [5,6]. Numerous researchers have reported that ChatGPT can achieve satisfactory results on multidisciplinary medical practitioner examinations in a variety of countries as well as provide detailed evidence-based explanations when responding to input clinical scenarios [7-10]. These studies have investigated the function of ChatGPT in the vast field of medicine and demonstrated the feasibility of investigating the application of ChatGPT in medical education.

Due to the vast medical knowledge system, diverse content, and lengthy learning cycle, it is difficult to accomplish a quality breakthrough in medical education [11,12]. Both general practitioners and specialists must keep relearning medical knowledge to avoid forgetting some minor but essential knowledge points during clinical practice [13,14]. The internet has become a common learning resource for physicians and medical students due to its convenience. However, the internet's general search results are vast and complex, necessitating the use of very specific search terms so that users can find the answers they seek [15]. Designing a specific learning application based on the network can increase the output of knowledge points, but it may not be able to respond promptly to the personalized user query input [16]. By identifying keywords in queries and analyzing their relevance, search engines provide users with a variety of search results; however, users must frequently determine the authenticity and veracity of the answers [17,18]. It has been indicated that while ChatGPT and professional forum answers exhibit comparable levels of logic and accuracy in their specific responses, ChatGPT demonstrates significantly greater empathy than the answers provided by verified individuals on forums, despite the fact that ChatGPT relies on big data [8,19]. Therefore, ChatGPT's response is more akin to a web-based answer retrieval, with the veracity of the retrieved answer being filtered based on the user's personalized query and the expression being intensively processed prior to outputting the result.

Although many academics believe that it is a double-edged sword to use ChatGPT as an auxiliary instrument in medical education [20-22], the development and promotion of ChatGPT will undoubtedly fuel the innovation of medical education [23,24]. The major reason why academics are worried about ChatGPT's inadequacies in medical education is that its potent text creation feature brings the risk of making students too reliant on it as a writing tool [25,26]. However, through interactive learning and immediate feedback, ChatGPT as a virtual teaching assistant can increase learning efficiency [23]. Artificial intelligence (AI)–assisted learning breaks the mode of 1-way input of theoretical knowledge in the previous stage of basic medical knowledge education, and its ability to generate different clinical scenarios according to different diseases can assist medical students in constructing a bridge between professional medical theory and clinical practice [27-29].

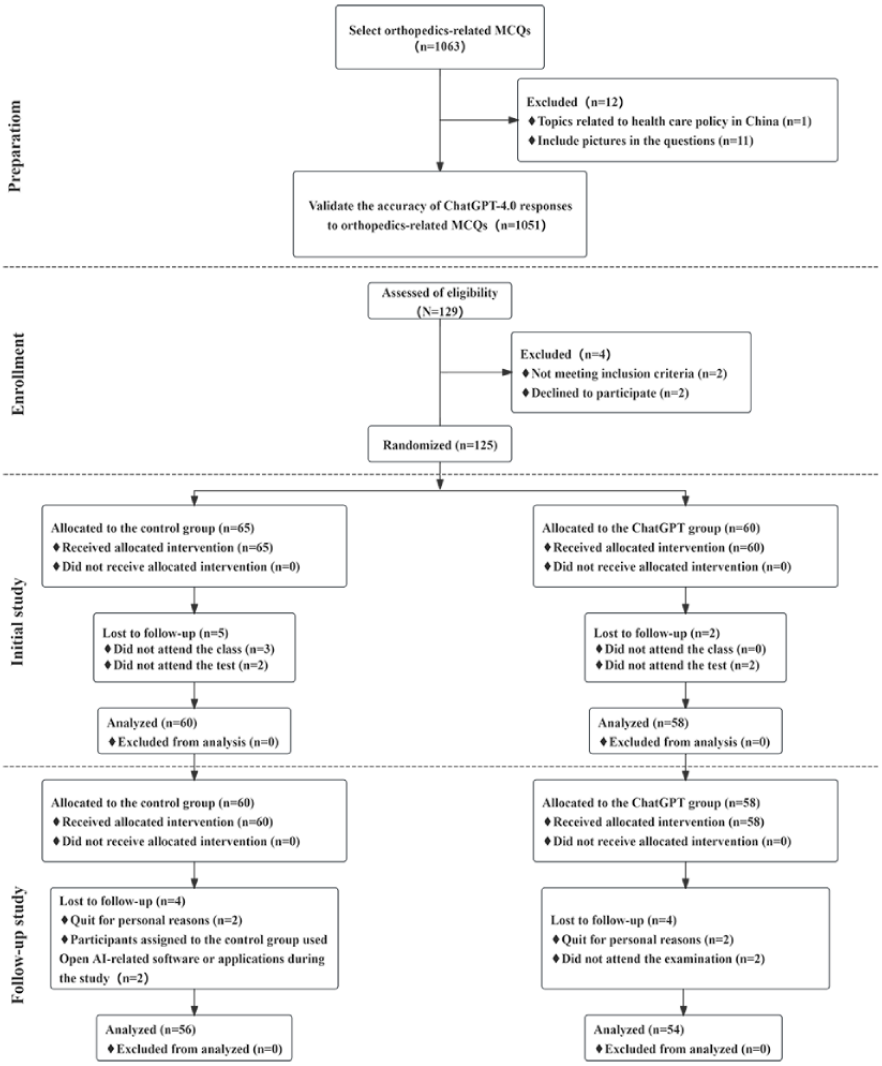
Therefore, we conducted a prospective randomized controlled trial concerning the application of ChatGPT in learning orthopedics. We first validated ChatGPT-4.0's accuracy in answering multiple-choice questions (MCQs) related to orthopedics and then used it as a learning aid intervention and conducted short- and long-term follow-up. Through the results of the randomized controlled trial and short-term follow-up, the ultimate objective was to determine whether ChatGPT can be used as an effective learning tool for undergraduate medical students.

Methods

Study Design

A parallel-design randomized controlled trial was used for this investigation. First, the accuracy of ChatGPT-4.0's responses to orthopedics-related MCQs was examined. In addition, for a group experiment, 129 third-year medical students from Jinan University's Medical College were recruited. They were divided into 2 groups at random, namely, the control group and the ChatGPT-4.0–assisted learning group (ChatGPT group). The internet-using students in the control group were not permitted to use any OpenAI-related software or programs, whereas those in the ChatGPT group used only ChatGPT-4.0 as the learning tool. Only after completing the orthopedics course, the orthopedics exercises, the review of the fundamental concepts in orthopedics, using the internet or ChatGPT, the orthopedic examination, and the final examination for the semester's teaching task did the participants finish the experiment. The detailed process of experimental arrangement is shown in Figure 1. Both the CONSORT (Consolidated Standards of Reporting Trials) checklist (Multimedia Appendix 1) and the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and online TeleHealth; version 1.6.1) checklist (Multimedia Appendix 2) were used for this trial [30,31].

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) 2010 flow diagram showing the randomized controlled trial process. MCQs: multiple-choice questions.



Participants

In China's undergraduate medical education curriculum, third-year students have already completed coursework in foundational subjects such as anatomy, histology, embryology, and physiology, and are presently immersed in acquiring theoretical knowledge pertaining to clinical disciplines. Consequently, to investigate the efficacy of ChatGPT-4.0 in aiding undergraduate medical students' learning of concepts in orthopedics, we recruited third-year undergraduate medical students from Jinan University.

We used a convenience sampling approach to recruit as many participants as possible from the target population. Recruitment was conducted through collective promotional efforts during class meetings in each individual class. This approach allowed us to efficiently disseminate information about the study and encourage widespread participation among the third-year undergraduate medical students.

Inclusion Criteria

Participants were included if they had completed courses in human anatomy, physiology, biochemistry, pathology, pathophysiology, and diagnostics.

Exclusion Criteria

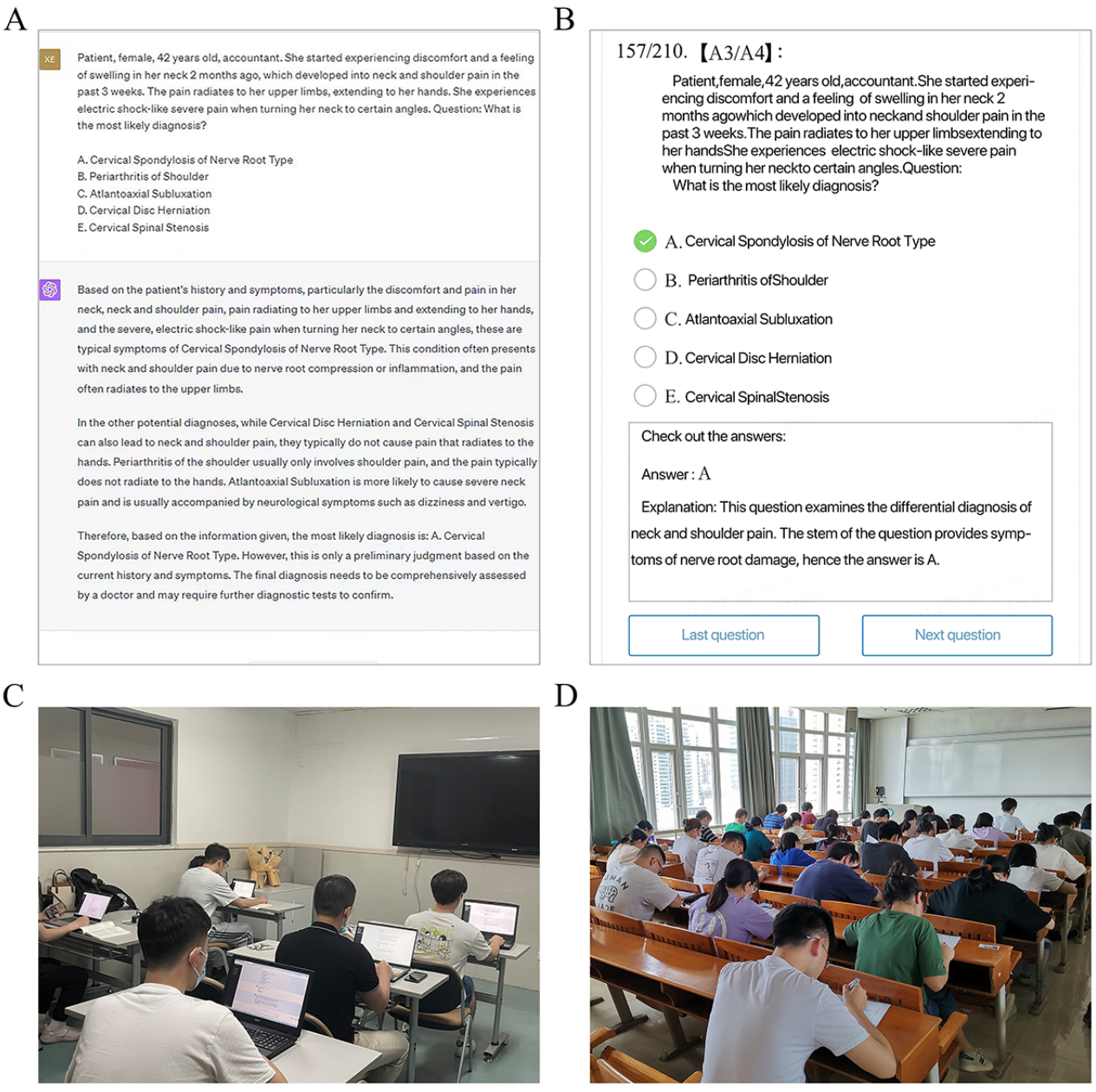
The exclusion criteria were as follows: failing final examinations for 6 courses in human anatomy, physiology, biochemistry, pathology, pathophysiology, and diagnostics that were completed before; switching majors, stopping classes, or dropping out during the current academic year; refusing to join or leaving in the middle for private reasons; failing to finish the multiple-choice exercises or the orthopedics course; using OpenAI-related software or apps in the control group; not completing the orthopedics multiple-choice examination; and missing subjects on the semester's final test.

Preparation

The "National Medical Electronic Schoolbag" (People's Military Medical Press) is the multiple-choice practice software for the standardized training theory examination of China's medical industry (Figure 2B). It is the first digital teaching reform project

in medical higher education funded by China. After screening by 2 orthopedics specialists (WYG and ZWX), we removed 12 questions from the 1075 orthopedics-related MCQs (Multimedia Appendix 3) that were downloaded from the "National Medical Electronic Schoolbag" software (1 question was related to medical policies with Chinese characteristics, and 11 inquiries were disqualified because they contained images). A1, A2, A3/A4, and case analysis questions are the different categories of inquiries. Type A1 questions are primarily about the fundamental knowledge of orthopedics. A2 questions have a brief medical history as the topic stem. Type A3/A4 questions describe a simple patient-centered clinical situation (eg, questions 169 and 170 from the first set of questions in Multimedia Appendix 3), and case analysis questions describe a clinical situation focused on a single patient or family (eg, questions 208-217 from the first set of questions in Multimedia Appendix 3). Chinese was used as the text input language for ChatGPT in this investigation. Finally, 1051 orthopedics-related MCQs were sequentially entered into ChatGPT-4.0 (Figure 2A), and ChatGPT-4.0's responses were recorded.

Figure 2. Scenarios for performing the educational clinical trial. (A) ChatGPT interface for answering orthopedics-related multiple-choice questions; (B) the "National Medical Electronic Schoolbag" is the multiple-choice practice software (English schematic diagram); (C) the participants in the ChatGPT group using ChatGPT to assist in learning orthopedics knowledge; and (D) the participants in 2 groups doing the orthopedics examination.



Intervention

Initial Study

Participants were enrolled in a 1-week orthopedics course. Upon completion of the course, they took a multiple-choice test on orthopedics-related topics, and each participant's accuracy rate was recorded. The control group and the ChatGPT group were subsequently formed out of the participants. In the final step, participants in various groups used the internet and ChatGPT (Figure 2C) to evaluate the omissions both in the course and the exercises during the week following the completion of the orthopedic course.

Follow-Up Study

For orthopedics, there was a 7-day review period. The ChatGPT group's students were required to use ChatGPT-4.0 to search up knowledge points to help their learning. Other web-based search engines and message boards were not allowed to be used by the students in this group to search for information. The students in the control group were required to use the internet to support their study by scouring forums and search engines for relevant information. However, they were not allowed to use any software connected to OpenAI. After a week of study using the internet or the ChatGPT, the subjects completed an orthopedics multiple-choice examination. There were 66 questions of type A1, 88 questions of type A2, 29 questions of type A3/A4, and 31 questions requiring case analysis. Participants in diverse groups provided the correct response and total score for each kind of question, which were recorded.

Following the orthopedics examination, the final examination review period was scheduled in accordance with the semester learning plan. We did not obstruct the participants' learning to complete their reviews at this time. The results of this semester's internal medicine, surgery, pediatrics, obstetrics and gynecology, and infectious diseases final examinations (Figure 2D) were systematically gathered via the academic affairs office with the participants' informed consent. Multimedia Appendix 4 contrasts all interventions between the experimental group and the control group.

Outcomes

In the preparation phase, 2 orthopedics specialists (WYG and ZWX) curated and inputted 1075 orthopedics-related MCQs into ChatGPT-4.0, with the aim of evaluating the accuracy of ChatGPT-4.0 in answering these orthopedics MCQs, while also recording its performance across different question types. We had logged answers from "National Medical Electronic Schoolbag" MCQs and ChatGPT-4.0 responses, uploading the data as Multimedia Appendix 5.

Next, after students completed a 1-week prestudy course in orthopedics and used different methods to review the knowledge for 1 week, we assigned ChatGPT-4.0 as an auxiliary learning review tool to students in the ChatGPT-4.0 group, while students in the control group were not allowed to use large language models (LLMs) for auxiliary review. After the review, we tested all students with the same test questions, enabling us to gauge the quality of short-term revision using different learning auxiliary tools. We recorded the performance of the different student groups in the orthopedics MCQ test, which served as the basis for the short-term impact analysis of the learning intervention.

Subsequently, at the end of the current semester, we accessed and recorded the clinical subject examination scores of the different student groups through the academic system. During the period from the end of our intervention to the end of the final examination, our follow-up content consisted of recording the use of LLMs by the ChatGPT group and the control group. We did not carry out any additional interventions during this time. The scores in the subject examinations at the end of the semester served as the basis for long-term impact analysis of the changes in students' learning methods after using ChatGPT-4.0. In this study, the accuracy of ChatGPT-4.0, short-term impact analysis, and long-term impact analysis are all primary outcomes.

Blinding

To eradicate subjective bias in the grading process, the collector was unaware of the classification of the participants when collecting the results from the orthopedics-related multiple-choice exercises and examinations. Other information relied on data from the school's pedagogical administration system, and the personnel who collected the data did not know the group information of the experiment participants.

Randomization

After completing the orthopedics-related multiple-choice exercises, the participants were randomly assigned to different groups using the sealed envelope method to minimize systemic bias. To ensure balanced group sizes, a clinician not involved in the program prepared sealed envelopes containing group assignment information. The participants then selected envelopes to determine their group allocation without knowing the group information beforehand. This approach served as a blocking method to achieve balanced group sizes while reducing artificial bias.

Statistics

For statistical analysis, SPSS software (version 26.0; IBM Corp) was used. The chi-square test was used to analyze gender differences between different groups. The Kolmogorov-Smirnov test was used to determine whether the data exhibited a normal distribution. If the data did not conform to a normal distribution, the Mann-Whitney *U* test was used for analysis. In addition, the data were processed according to the Levene test, and it was determined that the variance was homogeneous; accordingly, an independent-samples 2-tailed *t* test was conducted. When the *P* value was <.05, the difference was considered statistically significant. GraphPad Prism 8 was used to create bar charts. The results of continuous variables were displayed as follows: mean difference between the experimental group and the control group (mean, SD of the difference, *P* value), whereas the accuracy rate of ChatGPT was displayed as: correct number/total number.

Ethical Considerations

This study was approved by the First Affiliated Hospital of Jinan University's Ethics Committee (KY-2023-171), and it

was also registered in the Chinese Clinical Trials Registry (Chictr2300071774). To ensure confidentiality of the participants, access to the original experimental data necessitates a valid request that is sent to the email address of the corresponding author. Prior to manuscript submission, the research team was required to submit the content of the uploaded materials to the Science and Technology Department of the First Affiliated Hospital of Jinan University for review. This step ensured that no personal information of the participants was disclosed inadvertently. Concurrently, the department was also able to verify the implementation of the participants' rewards. In accordance with the Declaration of Helsinki, the participants' written informed consent was obtained before any information about them was acquired.

Results

Overview

We began collectively recruiting the Jinan University third-year undergraduate class of medical students on April 1, 2023, and finished the recruitment process on April 7, 2023. We finished the "initial study" and short-term follow-up through the course of the next 2 weeks. The long-term follow-up work was then conducted by gathering the final examination results of the participating students after the semester ended in June 2023. From April 1, 2023, to April 7, 2023, a total of 129 eligible candidates from the Medical College of Jinan University were assessed for participation. During the recruitment phase, 4 individuals dropped out, 7 more during the initial study, and 8 during the follow-up, resulting in a total of 110 participants who completed the study, with 56 in the control group and 54 in the ChatGPT group (Figure 1). As part of our follow-up study, we carried out telephone interviews to assess the extent of ChatGPT usage among the participants in both the control and experimental groups. The interviews revealed that, prior to the final examination, only 2 individuals from the control group had engaged with LLMs. In contrast, every participant in the experimental group had used different types of LLMs to varying degrees.

All of the participants mentioned their ages and grade point averages as of the school year to the researchers after completing an informed consent form for this study. We analyzed age (mean –0.02, SD 0.14 years; *P*=.89), sex (*P*=.44), grade point average (mean 0.10, SD 0.11; *P*=.38), and orthopedic practice accuracy rate (mean 0.12, SD 1.34; *P*=.93) across the 2 groups and found no significant differences after omitting those who were lost to follow-up (Table 1).

Table 1. Baseline characteristics of the final participants.

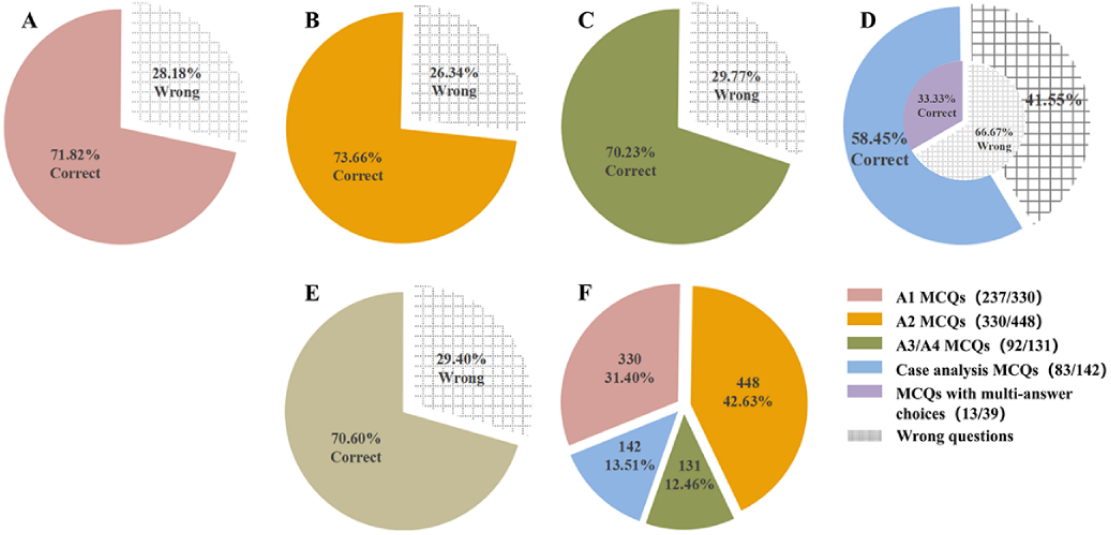
Characteristics	Control group (n=56)	ChatGPT group (n=54)	Total (n=110)	*P* value
Age (years), mean (SD), range	22.46 (0.79), 21-24	22.44 (0.69), 21-24	22.45 (0.74), 21-24	.89
Male sex, n (%)	28 (50)	23 (43)	51:59	.44
Grade point average, mean (SD), range	3.23 (0.51), 2.5-4.3	3.23 (0.68), 2.4-4.4	3.28 (0.60), 2.4-4.4	.38
Orthopedic MCQs[a] practice accuracy rate, mean (SD), range	56.08 (7.34), 42.94-70.91	56.19 (6.74), 41.44-73.92	56.13 (7.02), 41.44-73.92	.93

[a]MCQ: multiple-choice question.

The Accuracy of ChatGPT-4.0 Responses to Orthopedics-Related MCQs

A total of 330 A1 MCQs, 448 A2 MCQs, 131 A3/A4 MCQs, and 142 case analysis MCQs made up to a total of 1051 questions (Figure 3). ChatGPT-4.0 answered the 1051 orthopedics-related MCQs with a satisfactory accuracy rate of 70.60% (742/1051; Figure 3E and 3F). Specifically, the accuracy was 71.8% (237/330) for all A1 MCQs (Figure 3A), 73.7% (330/448) for all A2 MCQs (Figure 3B), and 70.2% (92/131) for all A3/A4 MCQs (Figure 3C), whereas it was only 58.5% (83/142) for all case analysis MCQs (Figure 3D). Among the case analysis MCQs, the accuracy of questions with multiple correct answers was as low as 33% (13/39) (Figure 3D). We found that ChatGPT explained each answer option and enumerated the associated knowledge points (Figure 2A).

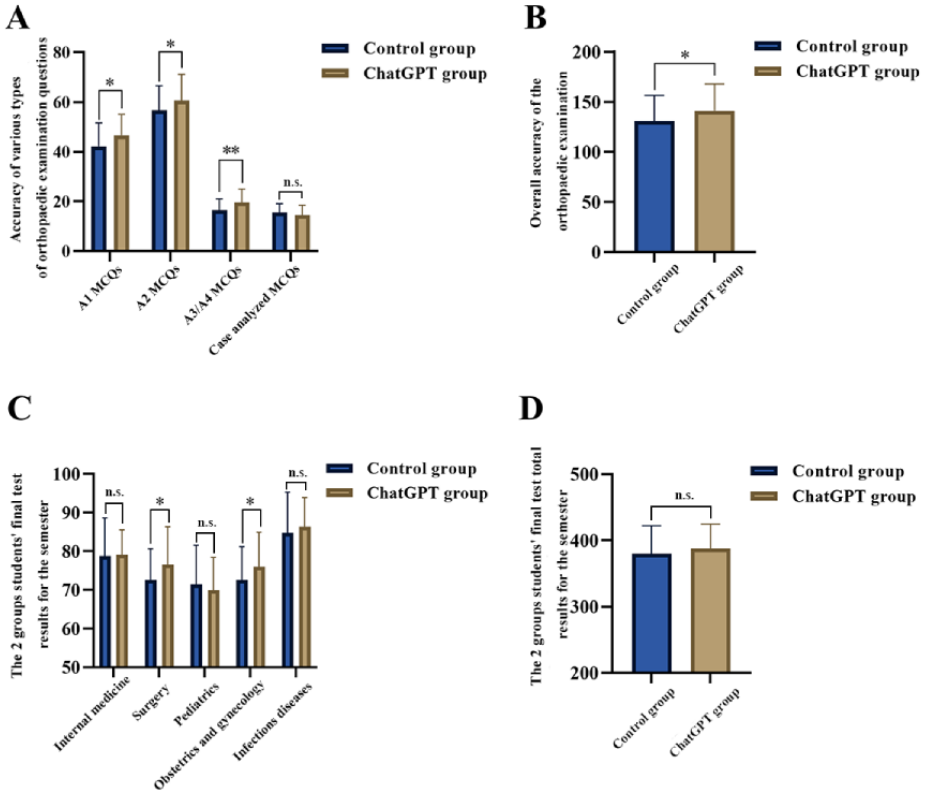
Figure 3. ChatGPT's performance in answering multiple-choice questions related to orthopedics. (A-D) Accuracy of various types of questions answered by ChatGPT; (E) overall accuracy of questions answered by ChatGPT; and (F) the percentage of each question type in the overall questions. MCQs: multiple-choice questions.



The Accuracy of the 2 Groups of Participants in the Orthopedics Examination

In the orthopedics examination, we were surprised to find out that the participants in the ChatGPT group were typically able to respond to 138.46 (SD 26.97) questions correctly, whereas those in the control group were able to respond to only 130.80 (SD 25.56) questions correctly on average, a significant decline in accuracy (mean 10.40, SD 4.98; *P*=.04; Figure 4B). As shown in Figure 4A, correct answers to A1 (mean 4.40, SD 1.75; *P*=.01), A2 (mean 3.93, SD 1.95; *P*=.047), and A3/4 (mean 3.11, SD 0.96; *P*=.002) among them demonstrated that the ChatGPT group performed significantly better than the control group. Despite the fact that the ChatGPT group showed a lower accuracy in case analysis questions than the control group (mean –1.04, SD 0.72; *P*=.16), the difference was not statistically significant (Figure 4A).

Figure 4. Short-term orthopedics test results and long-term final subjects examination results. (A) Accuracy of various types of orthopedics examination questions; (B) overall accuracy of the orthopedics examination; (C) the 2 groups' final test results for the semester; and (D) the 2 groups' final test total results for the semester. ChatGPT group: ChatGPT-4.0–assisted learning group; MCQs: multiple-choice questions; ns: not significant. **P* value less than .05; ***P* value less than .01.



The Score of the Final Examination for the Semester's Teaching Task

The statistical findings and the participants' scores of the final examination for the semester are shown in Figures 4C and 4D. We were pleasantly surprised to find that the ChatGPT-assisted participants scored higher on the final examinations in surgery (mean 4.00, SD 1.72; *P*=.02), internal medicine (mean 0.23, SD 1.58; *P*=.88), obstetrics and gynecology (mean 3.45, SD 1.68; *P*=.04), and infectious diseases (mean 1.60, SD 1.75; *P*=.36) than the control group, although the difference was statistically significant only for surgery and obstetrics and gynecology (Figure 4C). In the pediatric examination, the ChatGPT group scored worse than the control group (mean –1.58, SD 1.79; *P*=.38), but the difference was not statistically significant (Figure 4C). The mean total final examination scores of the ChatGPT group were also higher than those of the control group (mean 7.71, SD 7.53; *P*=.31), but the difference was not statistically significant (Figure 4D).

Discussion

Principal Findings

This is the first prospective clinical trial using ChatGPT as an intervention in the instruction of medical undergraduates. We discovered that ChatGPT has a high rate of accuracy when responding to orthopedics-pertinent MCQs, and that it explains each answer option and lists the corresponding knowledge points. After successfully enrolling the final 110 participants in a clinical study, the students using ChatGPT-4.0 were able to have a better comprehension of orthopedics-related knowledge points and performed better on end-of-semester examinations in other disciplines. These findings provide a firm foundation for medical students to use ChatGPT as an auxiliary learning aid in order to improve learning efficacy and promote the use of ChatGPT in medical education.

The purpose of undergraduate medical education is to assist students in developing a strong theoretical foundation in medicine and fundamental clinical abilities. The MCQs test has grown to be one of the most popular tools for objectively evaluating medical theoretical knowledge [32,33]. The aim of our research was to determine whether undergraduate medical students who use ChatGPT as a learning tool in orthopedics can learn more effectively. It is the first prospective randomized controlled trial that used ChatGPT as the main intervention. Through this research, we discovered that using ChatGPT as a learning tool might increase the effectiveness of acquiring information relevant to orthopedics and lead to improved performance on the MCQs test. The short-term follow-up revealed that the students who used ChatGPT as an additional tool for learning also performed better on assessments for the majority of the subjects at the end of the semester, indicating that ChatGPT users might show willing to alter their learning strategies and include ChatGPT as one of their daily learning tools.

We fed ChatGPT-4.0 with the orthopedics-pertinent practice questions and found that it had a respectable accurate answer rate for all of them (A1 [237/330, 71.8%], A2 [330/448, 73.7%], A3/A4 [92/131, 70.2%], and case analysis [83/142, 58.5%]). While doing so, we saw that ChatGPT provided not just knowledge points but also explanations for each answer choice. ChatGPT was also performed at a level of 54.96% accuracy on the Plastic Surgery Inservice Training Examination, which consisted of 242 questions [7]. Because correct rates of 58.5% and 54.96% are not acceptable in clinical work, we believe that ChatGPT is better suited as an auxiliary learning tool for the basic teaching of medical knowledge than for the auxiliary clinical diagnosis and treatment, relying on the powerful and rapid ability to collect and collate data and the ability to answer questions instantly. In addition to textual information such as the patient's complaint, current history, and prior history, the clinician's physical examination and imaging must be incorporated into the clinical diagnosis and treatment [34,35]. Although some academics enter the results of physical examinations and imaging examinations into ChatGPT as text for big data analysis [7], the quality of such input content is inconsistent, and systematic errors cannot be ruled out.

The vast and intricate medical knowledge system is the primary source of medical education's complexity. In addition to cramming for examinations and studying for examinations, medical students must devote significant effort to accumulating and organizing information and making causal links between different domains of knowledge [36,37]. Medical students often reread the medical material many times to ensure a thorough grasp of the material. Moreover, it is also very challenging to realize the transformation of medical knowledge from theory to practice [38,39] because the process of clinical practice requires the interaction between doctors and patients, that is, the immediate output and immediate feedback between them. With the evolution of teaching concepts and learning aids, medical education model has undergone significant transformations over time [40,41]. The traditional education model consists of a 1-way transmission from professors to students [42]. Mass Open Online Course, a nonfixed multidirection input model, has emerged progressively with the advent of the internet [43]. To enhance students' capacity for self-exploration, education sector constructs problem-based learning and promotes interaction during the learning process [44,45]. Simultaneously, virtual reality–assisted instruction is being developed to enhance medical students' perceptual comprehension of medical knowledge through the interaction and immediate feedback of programmed audiovisual operations [46]. Through big data analysis and data aggregation, ChatGPT interacts with users via text and provides immediate feedback [5,6]. As LLMs technology continues to evolve and improve, its application prospects in medical education will become even more extensive. Educators should actively embrace this technological innovation [29]. These advanced AI tools can provide personalized learning support, helping students better understand complex medical concepts such as anatomy [47], physiology [48], and biochemistry [48]. Through interactive dialogues with LLMs, students can ask questions, obtain detailed explanations and examples, and deepen their understanding of abstract concepts [49,50]. Moreover, LLMs can generate engaging learning materials, such as clinical case studies, question-and-answer exercises, and knowledge summaries, to enhance student participation and learning motivation [49,51].

This immersive learning experience can stimulate students' curiosity and encourage them to actively explore knowledge in the biomedical field [51]. Hence, as a supplementary learning aid for medical students, ChatGPT might be able to realize nonprocedural multidirectional interaction and continuously enhance students' knowledge systems [29] via immediate feedback.

As an auxiliary instrument for medical education, ChatGPT is currently viewed by many academics as a 2-way street [52]. Notably, ChatGPT as a supplementary aid can assist users in collecting the correct target information more efficiently, thereby enhancing work and learning efficiency [5]. However, many academics believe that ChatGPT is merely a plagiarist that lacks initiative and can only collect information, which has become a slang term for slothful people [53-55]. GPTZero (accessed on June 6, 2023), software created by Edward Tian, a student at Princeton University, for statistical analysis of whether text has been generated by AI, has gone a long way toward alleviating the public's anti-AI sentiment. Instead of emphasizing excessively the "writing shortcuts" provided by ChatGPT's text-generation function, users should take advantage of its big data rapid retrieval summary and immediate feedback to improve their study and work efficiency. As society, science, and technology advance, it is inevitable that students will take the initiative to adapt and combine their own learning mode with the updates and iteration of learning media and supplementary learning tools [56]. This general trend cannot be reversed. Therefore, society should hold higher standards for teaching methods, teaching content, and student assessment in this irreversible trend.

Limitations

First, although MCQs are often used to gauge medical students' fundamental theoretical knowledge in many nations, China may have a distinct custom for framing questions and placing a different priority on knowledge points than other nations. Second, this study focuses more on orthopedic intervention than on multidisciplinary intervention, which might result in some limitations in the research findings. Third, during the preparation stage of this study, while assessing ChatGPT-4.0's proficiency in answering MCQs pertaining to orthopedics, we made the decision to omit MCQs that included visual elements. This choice might have, to some degree, constrained the breadth of our investigation and consequently fell short of delivering an all-encompassing appraisal of ChatGPT-4.0's aptitude in tackling a wide spectrum of medical inquiries. This study was a single-center randomized controlled trial; thus, more prospective multicenter and interdisciplinary investigations are required to fully examine ChatGPT's potential as a teaching tool for medical education.

Conclusions

ChatGPT has a high rate of accuracy in answering orthopedics-related MCQs, and it explains each answer choice and lists the corresponding knowledge points. Compared with the students assigned to the control group, those in the ChatGPT group had a better understanding of the knowledge points related to orthopedics after a short intervention period and performed better on the end-of-semester examinations in some other disciplines. These findings provided a solid foundation for medical students to use ChatGPT as an auxiliary learning aid to enhance learning efficacy and help promote the use of ChatGPT in medical education.

Acknowledgments

We would like to thank all the Jinan University medical students who took part in this research and short-term follow-up. The Academic Affairs Office of Jinan University, the Basic Medical College of Jinan University, and the First Clinical Medical College of Jinan University are thanked for their assistance with the planning and compilation of data for this study. We are grateful to Dr Guorong She, who assisted us in completing the randomization process for the experiment. We thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Disclaimer

All authors affirm that neither ChatGPT nor any AI-assisted software was used in the text creation of this article. The screenshot example used to illustrate learning with ChatGPT unequivocally demonstrates that the interface originates from ChatGPT. ChatGPT was used exclusively as an experimental intervention and not as a paper-writing instrument in this study.

Authors' Contributions

WG led the conceptualization, investigation, visualization, and writing (original draft, review, and editing). ZX initiated the data curation, formal analysis, and writing (original draft). Yiming Z and HL initiated the data curation and formal analysis. QD, and JH led the assistance with finishing the study and providing orthopedics instructions. JO, XZ, and Yiyi Z led the conceptualization, writing (review and editing), and study supervision. All authors concur that neither ChatGPT nor any other OpenAI-related software or application was used during the writing of the paper's text. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

CONSORT (Consolidated Standards of Reporting Trials) 2010 Checklist.
[DOC File , 197 KB-Multimedia Appendix 1]

Multimedia Appendix 2

CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and online TeleHealth; version 1.6.1) checklist.
[PDF File (Adobe PDF File), 2914 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Multiple-choice questions.
[DOCX File , 176 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Illustration showing different parts of the training were undertaken by each group.
[DOCX File , 13 KB-Multimedia Appendix 4]

Multimedia Appendix 5

The reference answers and the responses from ChatGPT-4.0.
[XLSX File (Microsoft Excel File), 37 KB-Multimedia Appendix 5]

References

1. Baumgarten KM, Gerlach D, Galatz LM, Teefey SA, Middleton WD, Ditsios K, et al. Cigarette smoking increases the risk for rotator cuff tears. Clin Orthop Relat Res. 2010;468(6):1534-1541. [FREE Full text] [doi: 10.1007/s11999-009-0781-2] [Medline: 19283436]
2. Agathokleous E, Saitanis CJ, Fang C, Yu Z. Use of ChatGPT: what does it mean for biology and environmental science? Sci Total Environ. 2023;888:164154. [doi: 10.1016/j.scitotenv.2023.164154] [Medline: 37201835]
3. Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science. Nature. 2023;614(7947):214-216. [doi: 10.1038/d41586-023-00340-6] [Medline: 36747115]
4. Ghosh A, Bir A. Evaluating ChatGPT's ability to solve higher-order questions on the competency-based medical education curriculum in medical biochemistry. Cureus. 2023;15(4):e37023. [FREE Full text] [doi: 10.7759/cureus.37023] [Medline: 37143631]
5. Ahn C. Exploring ChatGPT for information of cardiopulmonary resuscitation. Resuscitation. 2023;185:109729. [doi: 10.1016/j.resuscitation.2023.109729] [Medline: 36773836]
6. Cahan P, Treutlein B. A conversation with ChatGPT on the role of computational systems biology in stem cell research. Stem Cell Reports. 2023;18(1):1-2. [FREE Full text] [doi: 10.1016/j.stemcr.2022.12.009] [Medline: 36630899]
7. Gupta R, Herzog I, Park JB, Weisberger J, Firouzbakht P, Ocon V, et al. Performance of ChatGPT on the plastic surgery inservice training examination. Aesthet Surg J. 2023;43(12):NP1078-NP1082. [doi: 10.1093/asj/sjad128] [Medline: 37128784]
8. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183(6):589-596. [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
9. Alfertshofer M, Hoch CC, Funk PF, Hollmann K, Wollenberg B, Knoedler S, et al. Sailing the seven seas: a multinational comparison of ChatGPT's performance on medical licensing examinations. Ann Biomed Eng. 2024;52(6):1542-1545. [FREE Full text] [doi: 10.1007/s10439-023-03338-3] [Medline: 37553555]
10. Borchert RJ, Hickman CR, Pepys J, Sadler TJ. Performance of ChatGPT on the situational judgement Test-A professional dilemmas-based examination for doctors in the United Kingdom. JMIR Med Educ. 2023;9:e48978. [FREE Full text] [doi: 10.2196/48978] [Medline: 37548997]
11. Malau-Aduli BS, Lee AY, Cooling N, Catchpole M, Jose M, Turner R. Retention of knowledge and perceived relevance of basic sciences in an integrated case-based learning (CBL) curriculum. BMC Med Educ. 2013;13:139. [FREE Full text] [doi: 10.1186/1472-6920-13-139] [Medline: 24099045]

12. Atchley TJ, Vukic B, Vukic M, Walters BC. Review of cerebrospinal fluid physiology and dynamics: a call for medical education reform. Neurosurgery. 2022;91(1):1-7. [doi: 10.1227/neu.0000000000002000] [Medline: 35522666]
13. Hilty DM, Turvey C, Hwang T. Lifelong learning for clinical practice: how to leverage technology for telebehavioral health care and digital continuing medical education. Curr Psychiatry Rep. 2018;20(3):15. [doi: 10.1007/s11920-018-0878-y] [Medline: 29527637]
14. Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. Acad Med. 2004;79(10 Suppl):S70-S81. [doi: 10.1097/00001888-200410001-00022] [Medline: 15383395]
15. Ullrich PF, Vaccaro AR. Patient education on the internet: opportunities and pitfalls. Spine (Phila Pa 1976). 2002;27(7):E185-E188. [doi: 10.1097/00007632-200204010-00019] [Medline: 11923675]
16. Skryd A, Lawrence K. ChatGPT as a tool for medical education and clinical decision-making on the wards: case study. JMIR Form Res. 2024;8:e51346. [FREE Full text] [doi: 10.2196/51346] [Medline: 38717811]
17. Vallée A, Blacher J, Cariou A, Sorbets E. Blended learning compared to traditional learning in medical education: systematic review and meta-analysis. J Med Internet Res. 2020;22(8):e16504. [FREE Full text] [doi: 10.2196/16504] [Medline: 32773378]
18. Rouleau G, Gagnon M, Côté J, Payne-Gagnon J, Hudson E, Dubois C, et al. Effects of e-learning in a continuing education context on nursing care: systematic review of systematic qualitative, quantitative, and mixed-studies reviews. J Med Internet Res. 2019;21(10):e15118. [FREE Full text] [doi: 10.2196/15118] [Medline: 31579016]
19. Xue Z, Zhang Y, Gan W, Wang H, She G, Zheng X. Quality and dependability of ChatGPT and DingXiangYuan forums for remote orthopedic consultations: comparative analysis. J Med Internet Res. 2024;26:e50882. [FREE Full text] [doi: 10.2196/50882] [Medline: 38483451]
20. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. ChatGPT and other large language models are double-edged swords. Radiology. 2023;307(2):e230163. [doi: 10.1148/radiol.230163] [Medline: 36700838]
21. Gordijn B, Have HT. ChatGPT: evolution or revolution? Med Health Care Philos. 2023;26(1):1-2. [doi: 10.1007/s11019-023-10136-0] [Medline: 36656495]
22. NA. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. Nature. 2023;613(7945):612. [doi: 10.1038/d41586-023-00191-1] [Medline: 36694020]
23. Lee H. The rise of ChatGPT: exploring its potential in medical education. Anat Sci Educ. 2024;17(5):926-931. [doi: 10.1002/ase.2270] [Medline: 36916887]
24. Seetharaman R. Revolutionizing medical education: can ChatGPT boost subjective learning and expression? J Med Syst. 2023;47(1):61. [doi: 10.1007/s10916-023-01957-w] [Medline: 37160568]
25. The Lancet Digital health. ChatGPT: friend or foe? Lancet Digit Health. 2023;5(3):e102. [FREE Full text] [doi: 10.1016/S2589-7500(23)00023-7] [Medline: 36754723]
26. Krügel S, Ostermaier A, Uhl M. ChatGPT's inconsistent moral advice influences users' judgment. Sci Rep. 2023;13(1):4569. [FREE Full text] [doi: 10.1038/s41598-023-31341-0] [Medline: 37024502]
27. Arif TB, Munaf U, Ul-Haque I. The future of medical education and research: is ChatGPT a blessing or blight in disguise? Med Educ Online. 2023;28(1):2181052. [FREE Full text] [doi: 10.1080/10872981.2023.2181052] [Medline: 36809073]
28. Feng S, Shen Y. ChatGPT and the future of medical education. Acad Med. 2023;98(8):867-868. [doi: 10.1097/ACM.0000000000005242] [Medline: 37162219]
29. Karabacak M, Ozkara BB, Margetis K, Wintermark M, Bisdas S. The advent of generative language models in medical education. JMIR Med Educ. 2023;9:e48163. [FREE Full text] [doi: 10.2196/48163] [Medline: 37279048]
30. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. CONSORT. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. Int J Surg. 2012;10(1):28-55. [FREE Full text] [doi: 10.1016/j.ijsu.2011.10.001] [Medline: 22036893]
31. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res. 2011;13(4):e126. [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
32. DeSantis M, McKean TA. Efficient validation of teaching and learning using multiple-choice exams. Adv Physiol Educ. 2003;27(1-4):3-14. [FREE Full text] [doi: 10.1152/advan.00016.2001] [Medline: 12594068]
33. Gupta V, Williams ER, Wadhwa R. Multiple-choice tests: A-Z in best writing practices. Psychiatr Clin North Am. 2021;44(2):249-261. [doi: 10.1016/j.psc.2021.03.008] [Medline: 34049647]
34. Thorborg K, Reiman MP, Weir A, Kemp JL, Serner A, Mosler AB, et al. Clinical examination, diagnostic imaging, and testing of athletes with groin pain: an evidence-based approach to effective management. J Orthop Sports Phys Ther. 2018;48(4):239-249. [doi: 10.2519/jospt.2018.7850] [Medline: 29510653]
35. Chou R, Qaseem A, Snow V, Casey D, Cross JT, Shekelle P, Clinical Efficacy Assessment Subcommittee of the American College of Physicians, American College of Physicians, et al. American Pain Society Low Back Pain Guidelines Panel. Diagnosis and treatment of low back pain: a joint clinical practice guideline from the American College of Physicians and the American Pain Society. Ann Intern Med. 2007;147(7):478-491. [FREE Full text] [doi: 10.7326/0003-4819-147-7-200710020-00006] [Medline: 17909209]

36. Maniar KP, Arva N, Blanco LZ, Mao Q, Morency EG, Rodriguez R, et al. Accreditation council for graduate medical education self-study for pathology: one institution's experience and lessons learned. Arch Pathol Lab Med. 2019;143(10):1271-1277. [FREE Full text] [doi: 10.5858/arpa.2018-0467-RA] [Medline: 31017451]
37. McAdams CD, McNally MM. Continuing medical education and lifelong learning. Surg Clin North Am. 2021;101(4):703-715. [doi: 10.1016/j.suc.2021.05.015] [Medline: 34242611]
38. Mann K, Gordon J, MacLeod A. Reflection and reflective practice in health professions education: a systematic review. Adv Health Sci Educ Theory Pract. 2009;14(4):595-621. [doi: 10.1007/s10459-007-9090-2] [Medline: 18034364]
39. Slotnick HB. How doctors learn: physicians' self-directed learning episodes. Acad Med. 1999;74(10):1106-1117. [doi: 10.1097/00001888-199910000-00014] [Medline: 10536633]
40. Wong G, Greenhalgh T, Pawson R. Internet-based medical education: a realist review of what works, for whom and in what circumstances. BMC Med Educ. 2010;10:12. [FREE Full text] [doi: 10.1186/1472-6920-10-12] [Medline: 20122253]
41. Venkatesan M, Mohan H, Ryan JR, Schürch CM, Nolan GP, Frakes DH, et al. Virtual and augmented reality for biomedical applications. Cell Rep Med. 2021;2(7):100348. [FREE Full text] [doi: 10.1016/j.xcrm.2021.100348] [Medline: 34337564]
42. Sattar MU, Palaniappan S, Lokman A, Hassan A, Shah N, Riaz Z. Effects of Virtual Reality training on medical students' learning motivation and competency. Pak J Med Sci. 2019;35(3):852-857. [FREE Full text] [doi: 10.12669/pjms.35.3.44] [Medline: 31258607]
43. Milbourn B, Black MH, Afsharnejad B, Snyman Z, Baker-Young E, Thompson C, et al. The "Talk-to-Me" MOOC intervention for suicide prevention and mental health education among tertiary students: protocol of a multi-site cross-over randomised controlled trial. Contemp Clin Trials. 2022;112:106645. [doi: 10.1016/j.cct.2021.106645] [Medline: 34861409]
44. Al-Azri H, Ratnapalan S. Problem-based learning in continuing medical education: review of randomized controlled trials. Can Fam Physician. 2014;60(2):157-165. [FREE Full text] [Medline: 24522680]
45. Jin J, Bridges SM. Educational technologies in problem-based learning in health sciences education: a systematic review. J Med Internet Res. 2014;16(12):e251. [FREE Full text] [doi: 10.2196/jmir.3240] [Medline: 25498126]
46. Gan W, Mok T, Chen J, She G, Zha Z, Wang H, et al. Researching the application of virtual reality in medical education: one-year follow-up of a randomized trial. BMC Med Educ. 2023;23(1):3. [FREE Full text] [doi: 10.1186/s12909-022-03992-6] [Medline: 36597093]
47. Temsah M, Alhuzaimi AN, Almansour M, Aljamaan F, Alhasan K, Batarfi MA, et al. Art or artifact: evaluating the accuracy, appeal, and educational value of AI-generated imagery in DALL·E 3 for illustrating congenital heart diseases. J Med Syst. 2024;48(1):54. [doi: 10.1007/s10916-024-02072-0] [Medline: 38780839]
48. Luke WANV, Seow Chong L, Ban KH, Wong AH, Zhi Xiong C, Shuh Shing L, et al. Is ChatGPT 'ready' to be a learning tool for medical undergraduates and will it perform equally in different subjects? Comparative study of ChatGPT performance in tutorial and case-based learning questions in physiology and biochemistry. Med Teach. 2024:1-7. [doi: 10.1080/0142159X.2024.2308779] [Medline: 38295769]
49. Sauder M, Tritsch T, Rajput V, Schwartz G, Shoja MM. Exploring generative artificial intelligence–assisted medical education: assessing case-based learning for medical students. Cureus. 2024;16(1):e51961. [FREE Full text] [doi: 10.7759/cureus.51961] [Medline: 38333501]
50. Meng X, Yan X, Zhang K, Liu D, Cui X, Yang Y, et al. The application of large language models in medicine: a scoping review. iScience. 2024;27(5):109713. [FREE Full text] [doi: 10.1016/j.isci.2024.109713] [Medline: 38746668]
51. Safranek CW, Sidamon-Eristoff AE, Gilson A, Chartash D. The role of large language models in medical education: applications and implications. JMIR Med Educ. 2023;9:e50945. [FREE Full text] [doi: 10.2196/50945] [Medline: 37578830]
52. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel). 2023;11(6):887. [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
53. Dergaa I, Chamari K, Zmijewski P, Ben Saad H. From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing. Biol Sport. 2023;40(2):615-622. [FREE Full text] [doi: 10.5114/biolsport.2023.125623] [Medline: 37077800]
54. Graham F. Daily briefing: will ChatGPT kill the essay assignment? Nature. 2022. [doi: 10.1038/d41586-022-04437-2] [Medline: 36517680]
55. Stokel-Walker C. AI bot ChatGPT writes smart essays—should professors worry? Nature. 2022. [doi: 10.1038/d41586-022-04397-7] [Medline: 36494443]
56. Xu J, Lio A, Dhaliwal H, Andrei S, Balakrishnan S, Nagani U, et al. Psychological interventions of virtual gamification within academic intrinsic motivation: a systematic review. J Affect Disord. 2021;293:444-465. [doi: 10.1016/j.jad.2021.06.070] [Medline: 34252688]

Abbreviations

AI: artificial intelligence
CONSORT: Consolidated Standards of Reporting Trials

CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and online TeleHealth
LLM: large language model
MCQ: multiple-choice question

Edited by T de Azevedo Cardoso, G Eysenbach; submitted 02.02.24; peer-reviewed by Z Liao, F Tang; comments to author 22.05.24; revised version received 10.06.24; accepted 27.06.24; published 20.08.24

Please cite as:
Gan W, Ouyang J, Li H, Xue Z, Zhang Y, Dong Q, Huang J, Zheng X, Zhang Y
Integrating ChatGPT in Orthopedic Education for Medical Undergraduates: Randomized Controlled Trial
J Med Internet Res 2024;26:e57037
URL: https://www.jmir.org/2024/1/e57037
doi: 10.2196/57037
PMID: