Taylor & Francis
Taylor & Francis Group

Medical Teacher

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/imte20

ChatGPT to generate clinical vignettes for teaching and multiple-choice questions for assessment: A randomized controlled experiment

Özlem Coşkun, Yavuz Selim Kıyak & Işıl İrem Budakoğlu

MEDICAL TEACHER
https://doi.org/10.1080/0142159X.2024.2327477

MEDICAL TEACHER

Taylor & Francis
Taylor & Francis Group

Check for updates

RESEARCH ARTICLE

ChatGPT to generate clinical vignettes for teaching and multiple-choice questions for assessment: A randomized controlled experiment

Özlem Coşkun, Yavuz Selim Kıyak and Işıl İrem Budakoğlu

Department of Medical Education and Informatics, Gazi University, Ankara, Turkey

ABSTRACT

Aim: This study aimed to evaluate the real-life performance of clinical vignettes and multiple-choice questions generated by using ChatGPT.

Methods: This was a randomized controlled study in an evidence-based medicine training program. We randomly assigned seventy-four medical students to two groups. The ChatGPT group received ill-defined cases generated by ChatGPT, while the control group received human-written cases. At the end of the training, they evaluated the cases by rating 10 statements using a Likert scale. They also answered 15 multiple-choice questions (MCQs) generated by ChatGPT. The case evaluations of the two groups were compared. Some psychometric characteristics (item difficulty and point-biserial correlations) of the test were also reported.

Results: None of the scores in 10 statements regarding the cases showed a significant difference between the ChatGPT group and the control group ($p > .05$). In the test, only six MCQs had acceptable levels (higher than 0.30) of point-biserial correlation, and five items could be considered acceptable in classroom settings.

Conclusions: The results showed that the quality of the vignettes are comparable to those created by human authors, and some multiple-questions have acceptable psychometric characteristics. ChatGPT has potential in generating clinical vignettes for teaching and MCQs for assessment in medical education.

ARTICLE HISTORY
Received 6 December 2023
Accepted 4 March 2024

KEYWORDS
ChatGPT; artificial intelligence; automatic item generation; clinical vignette; medical education

Introduction

Medical education embarked on a breathtaking journey from the age of information to the age of artificial intelligence (AI) (Wartman and Combs 2018). One of the remarkable AI-powered tools is the language model called Generative Pretrained Transformer (GPT), which was created by OpenAI. As a variant of GPT-3.5, ChatGPT has been released for public use on 30 November 2022, and it reached one million users only in five days (Buchholz 2023). The rapid expansion affected not only medical education, but also other health professions education, such as nursing (Choi et al. 2023) and dental (Thurzo et al. 2023) education. Therefore, some have suggested considering the release date as a dividing line between the pre-ChatGPT world and post-ChatGPT world (Masters 2023).

ChatGPT works based on natural language processing (NLP) techniques. In the field of NLP, large language models like the GPT (Floridi and Chiriatti 2020) have made remarkable progress in recent years. As it is in ChatGPT (Cotton et al. 2023), these models are trained on extensive amounts of textual data and have the capability to produce text that closely resembles human writing, accurately answer questions, and accomplish other language-related tasks with an appropriate level of precision.

In the context of teaching in medical education, its potential use cases have been discussed in the literature (Seetharaman 2023; Tsang 2023). More specifically, a commentary stated that self-check quizzes with answer explanations can be generated by ChatGPT (Lee 2023). Members of a medical school's faculty have reported that they employ ChatGPT for diverse tasks, which encompass creating clinical vignettes and generating multiple-choice questions (Cross et al. 2023). Two studies showed that clinical vignette generation by using ChatGPT is possible (Benoit 2023; Bakkum et al. 2024), however, they did not evaluate the quality of vignettes in real educational settings. There is a scarcity of real-world implementations.

In the context of assessment in medical education, some studies demonstrated ChatGPT's ability to answer questions in national exams. It achieved a score higher than the cut-off mark on the national exam taken by medical students in Spain after completing medical school to pursue a career as

Practice points

- Medical students are unable to differentiate between the quality of clinical vignettes generated by ChatGPT and written by experts.
- Using ChatGPT can enhance the efficiency of the clinical vignette writing process.
- Psychometric evidence showed that it is possible to generate multiple-choice questions with acceptable item discrimination using ChatGPT.
- It is crucial to remember that ChatGPT may sometimes provide inaccurate content.

CONTACT Özlem Coşkun ocoskun@gazi.edu.tr. Department of Medical Education and Informatics, Gazi University, Hastanesi E Blok 9. Kat Beşevler, Ankara, Turkey

Supplemental data for this article can be accessed online at https://doi.org/10.1080/0142159X.2024.2327477.

a resident physician in a hospital (Carrasco et al. 2023). Considering its score, if ChatGPT were to be a student, it had the ability to choose from a wide range of specialties. It has also provided a high proportion of correct answers in the United States Medical Licensing Examination (Kung et al. 2023; Mihalache et al. 2023). However, its performance did not meet the passing requirements in Chinese National Medical Licensing Examination (Wang et al. 2023), as its performance varies across different national exams (Alfertshofer et al. 2023). While these studies focused on using ChatGPT to answer the questions generated by humans, there is a scarcity of research centered on presenting questions generated by ChatGPT to humans.

From a broader perspective, AI has been used for teaching and assessment in higher education in several ways such as automated essay scoring, automated formative assessment, evaluation of teaching, providing personalized content (Zawacki-Richter et al. 2019; Ouyang et al. 2022). However, none of the studies evaluated the psychometric characteristics of multiple-choice items generated by using ChatGPT in health professions education (Sallam 2023), and none has studied implementation of clinical vignettes in an authentic educational setting. Considering ChatGPT has provided some crucial errors such as claiming "the human heart only has two chambers" (Lee 2023) and it sometimes hallucinates (Masters 2023), the evaluation of the generated content in real-life settings becomes more important.

This study aimed to evaluate the real-life performance of clinical vignettes and multiple-choice questions generated by using ChatGPT.

The research questions are as follows:

1. From the perspective of medical students, how is the quality of clinical vignettes generated by ChatGPT compared to those written by humans?
2. How is the psychometric characteristics (difficulty and discrimination indices) of multiple-choice questions on evidence-based medicine generated by ChatGPT?

Methods

Study design

This was a randomized controlled study. Figure 1 presents the experiment process.

Study setting

We conducted this study in 2022–2023 education period in Gazi University Faculty of Medicine, Ankara, Turkey. The six-year undergraduate medical program teaches students evidence-based medicine through a specialized longitudinal curriculum. The first three years consist of theoretical and practical instruction, such as learning how to access and utilize PubMed. In the fourth-year practices, students actively participate in a structured training on evidence-based medicine that utilizes a modified version of the PEARLS (Presentations of Evidence Abstracted from Research Literature to Solve real people's problem) method (Stockler et al. 2009; Coşkun et al. 2022). Instead of working with actual patients, each fourth-year student receives a case, which is an ill-defined clinical vignette developed by the board of evidence-based medicine. Students use their assigned cases to present the evidence they have extracted from research literature.

During one and half month of training period, students are expected to follow a structured process that involves several steps. The first step is to identify the clinical question of the case. This involves formulating a clear and specific question that addresses the patient's problem or

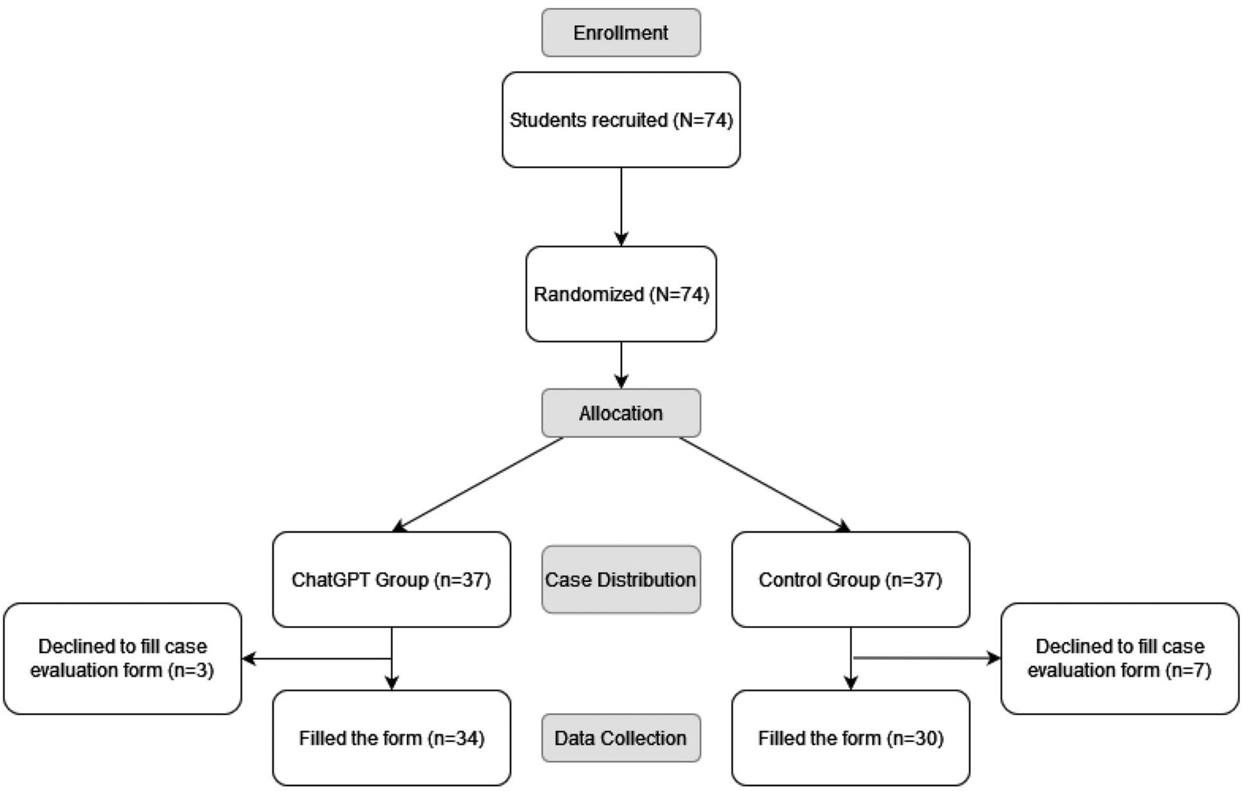

Figure 1. The experiment process.

concern. The next step is to develop a research strategy and use appropriate keywords to search for relevant studies. This involves searching electronic databases such as PubMed, and using appropriate search filters to refine the search results. Once students identify the studies, each student chooses the best ones based on pre-determined inclusion and exclusion criteria, which also considers evidence pyramid. The chosen studies are then evaluated for their quality, relevance, and applicability to the case at hand. After evaluating the studies, each student applies the information obtained to the case and makes a decision based on the available evidence. Finally, they convert the entire process into an oral presentation. The process also includes learning how to calculate and evaluate some measures, such as sensitivity, specificity, odds ratio, and number needed to treat.

Participants

Seventy-four fourth-year medical students ($N = 74$) enrolled in the undergraduate medical programme, where English is the language of instruction, were eligible to participate in the study. A group size of 35 students was chosen per group using power analysis with an effect size of 0.70, power of 0.80, and alpha of 0.05 (Creswell 2012). We randomly assigned the student population ($N = 74$) to either the ChatGPT group or the control group, and all underwent the same process during the training. The randomization was performed using SPSS 22.0 for Windows (Chicago, IL, USA). The only difference between the two groups was the origin of the cases they received, either ChatGPT or human-written. The students were blinded to the origin of the cases.

Cases

We generated the cases utilized in the ChatGPT group by using the Free Research Preview of ChatGPT (version 3.5) between December 2022 and January 2023. Since we conducted this study before ChatGPT-4 was introduced, we were unable to use the latest version. Furthermore, the subscription-based model of ChatGPT-4 poses inclusivity challenges for medical educators in developing countries due to the monthly payment requirement. Therefore, we believe that obtaining evidence on the free version would help to address the existing inequities in the world.

The prompt template we used was as follows:

> "We need an ill-defined medical case. Medical students will use this case to apply evidence-based medicine principles. For this reason, the case should include a dilemma. It should be on [THE DISEASE OR PROBLEM]. The case should consist of [THE NUMBER OF SENTENCES] sentences. Provide the age and gender of the patient."

The prompt template involved filling in the "[THE DISEASE OR PROBLEM]" section with the name of the disease or problem from the human-written cases. Additionally, we completed the "[THE NUMBER OF SENTENCES]" section by indicating the number of sentences in the human-written case. Table 1 provides an example of a human-written case, the prompt, and the case generated by ChatGPT.

We applied the template to each human-written case to generate a case on the respective disease/problem using ChatGPT, resulting in 37 cases produced by ChatGPT. The board of evidence-based medicine had developed and approved the human-written cases. The board members, who are subject-matter experts, also evaluated the cases produced by ChatGPT. Subject-matter experts carried out the evaluations to ensure the suitability of the cases for evidence-based medicine training. Out of 37 cases produced by ChatGPT, 22 cases were approved by the consensus of the board members without needing any revision. However, for the remaining 15 cases, they expressed some concerns such as being too vague for undergraduate medical students or mentioning the generic name of a medicine. Therefore, we entered new prompts as a response in the chat. These prompts are as follows:

- "Please use a specific treatment/medicine in the case."
- "Please remove the generic name of the medicine and use the active substance name instead."
- "Please mention a specific treatment option and write the case again."
- "Please mention the name of the test and write the case again."
- "Please mention for which genetic disease the test has been performed and write the case again."

These prompts led ChatGPT to revise the cases generated by itself in order to enhance the cases' suitability. All revisions have been made by ChatGPT considering the prompts. Subject-matter experts only evaluated and approved the revised cases, no single change was made directly by humans in the ChatGPT-generated cases.

The students evaluated the cases at the end of the training. They accessed an online version of the survey form through Google Forms. The survey form included 10 statements with a 5-point Likert scale (1: Definitely not agree, 5: Definitely agree). Table 2 presents the statements. There was no missing response in any of the statements.

Multiple-choice questions

ChatGPT generated the questions to assess students' knowledge on evidence-based medicine. Supplementary material presents the prompts developed based on the training

Table 1. An example of the cases and the prompt.

The human-written case	The prompt used in ChatGPT	The case generated by ChatGPT
A two-year-old girl applies to your outpatient clinic because of a barking cough. In the anamnesis obtained from her mother, she appears to have stridor the night before. The medical resident who works with you asks for your opinion on whether nebulized budesonide reduces hospitalization compared to placebo in a mild to moderate episode of croup.	We need an ill-defined medical case. Medical students will use this case to apply evidence-based medicine principles. For this reason, the case should include a dilemma. It should be on croup. The case should consist of three sentences. Provide the age and gender of the patient.	A three-year-old male presents with a barking cough and hoarseness. He has a history of croup and his parents are unsure whether to bring him to the emergency room or wait and see if symptoms improve. The medical team must decide whether to administer a corticosteroid to reduce inflammation.

Table 2. Evaluation of the clinical vignettes by 64 medical students.

Statements[a]	Vignette Group	Number of participants	Mean (SD)	p-value[b]	Effect Size (Cohen's d)
1. The case was written in an understandable way.	ChatGPT	34	4.11 (1.12)	>.05	0.13
	Human	30	3.96 (1.06)		
2. There were no spelling errors in the case.	ChatGPT	34	4.73 (.51)	>.05	0.30
	Human	30	4.53 (.77)		
3. There were no contradictions in the case.	ChatGPT	34	4.29 (.79)	>.05	0.26
	Human	30	4.50 (.82)		
4. The case was fun to deal with.	ChatGPT	34	4.02 (.86)	>.05	0.13
	Human	30	3.90 (.88)		
5. The case required me to use my clinical reasoning skills.	ChatGPT	34	4.44 (.61)	>.05	0.44
	Human	30	4.10 (.88)		
6. The case was suitable for generating keywords to search the literature.	ChatGPT	34	3.97 (1.14)	>.05	0.32
	Human	30	4.30 (.87)		
7. A research question could be developed using the case.	ChatGPT	34	4.35 (.88)	>.05	0.21
	Human	30	4.16 (.87)		
8. I liked the case.	ChatGPT	34	3.76 (1.20)	>.05	0.28
	Human	30	4.06 (.90)		
9. The case was not conducive to learn evidence-based medicine processes.	ChatGPT	34	2.29 (1.26)	>.05	0.13
	Human	30	2.13 (1.19)		
10. The case needs to be corrected.	ChatGPT	34	2.44 (1.41)	>.05	0.17
	Human	30	2.20 (1.29)		

[a]Likert scale; 1: Definitely not agree, 5: Definitely agree.
[b]Independent-Samples *T*-Test.

programme and the questions, with the numbers mentioned in the results section. Subject-matter experts did not make any change in the questions. In total, there were 15 multiple-choice questions. The test was administered face-to-face in classroom settings with proctors. Out of 74 students, 63 of them accepted to participate in.

Statistical analysis

We assessed the difference between the mean scores of two groups on the cases using Independent Samples T-Test on SPSS 22.0 for Windows (Chicago, IL, USA), given that it is acceptable to perform parametric tests when analyzing Likert scale responses (Norman 2010). We considered a p-value less than 0.05 statistically significant. We also calculated the effect sizes using Cohen's d. Effect sizes of 0.2 are small, while effect sizes of 0.5 are medium (Sullivan and Feinn 2012).

We evaluated the items (multiple-choice questions) based on Classical Test Theory. We carried out item analysis using Microsoft Excel to determine item difficulty (calculated by dividing the total score of test-takers by the maximum possible score). We calculated point-biserial correlation values using Spearman correlation on SPSS 22.0 for Windows (Chicago, IL, USA). It helps to determine whether an individual test item effectively discriminates between students who perform well on the overall test and those who do not. Large-scale standardized test developers usually require an item's point-biserial correlation to be at least 0.30 or higher to be considered effective (Downing and Yudkowsky 2009). However, for classroom-type tests developed locally, the values in the mid to high 0.20s may be satisfactory (Downing and Yudkowsky 2009). Therefore, we considered the values 0.30 or higher acceptable.

Results

Evaluation of the cases

Out of 74 students, 10 students declined to fill the evaluation form. In total, 64 students participated in the study. While there were 34 students in the ChatGPT group, there were 30 students in the control group.

None of the scores showed a significant difference between the ChatGPT group and the control group ($p >$ .05). Nine out of the 10 statements had a small effect size. However, one statement, "The case required me to use my clinical reasoning skills", had a medium effect size (Cohen's d: 0.44). The mean score for this statement was 4.44 in the ChatGPT group and 4.10 in the control group (Table 2).

Evaluation of the multiple-choice questions

Out of 64 students who accepted to participate in the study, one student could not participate in the exam.

Table 3 presents item difficulty and point-biserial correlation values.

Only six items (#3, #8, #11, #12, #13, #15) had acceptable levels (higher than 0.30) of point-biserial correlation. Addition to those items, five items (#2, #4, #6 #7, #14) can be considered acceptable in classroom settings due to having the values in the mid to high 0.20s. However, the remained items (#1, #5, #9, #10) had unacceptable values.

Discussion

This study aimed to investigate the feasibility of using a publicly available language model, ChatGPT-3.5, to generate clinical vignettes and multiple-choice questions for medical education. To our knowledge, this study is the first of its kind to evaluate the performance of ChatGPT-generated vignettes and multiple-choice questions in real-world educational settings.

The results regarding vignettes showed that there was no significant difference in the quality rated by medical students between the ChatGPT vignettes and the control vignettes. However, the clinical vignettes related to the statement "The case required me to use my clinical reasoning skills" had a medium effect size and showed that the vignettes generated by ChatGPT scored higher than the human-written vignettes. In addition, nine out of the ten evaluation criteria had small effect sizes, indicating that the

Table 3. Item difficulty and point-biserial correlation values of 15 multiple-choice questions answered by 63 medical students.

Item (Multiple-Choice Question)	Difficulty	Point Biserial Correlation
1	0.87	0.16
2	0.96	0.25*
3	0.47	0.42*
4	0.96	0.26*
5	0.98	−0.02
6	0.33	0.24
7	0.74	0.27*
8	0.61	0.35*
9	1	N/A
10	0.82	0.17
11	0.42	0.42*
12	0.68	0.40*
13	0.38	0.55*
14	0.31	0.27*
15	0.50	0.42*

*$p < 0.05$.

clinical vignettes generated by ChatGPT were comparable to those created by human authors. This finding extends the findings of a previous research to medical education vignettes, which showed there have been reports by users stating that the text generated by the GPT-3 model is difficult to distinguish from text written by humans (Elkins and Chun 2020). Similarly, in a recent study, academic physicians struggled to consistently distinguish between letters of recommendation authored by humans and those generated by ChatGPT (Preiksaitis et al. 2023). In clinical vignette context, two studies showed that ChatGPT is capable of generating clinical vignettes (Benoit 2023; Bakkum et al. 2024). Our study complemented the findings of these studies by demonstrating the performance in authentic medical education settings.

While the vignettes rated by medical students were not significantly different, the questions generated by ChatGPT were able to achieve acceptable levels of point-biserial correlation for some of the multiple-choice questions. Six items had acceptable levels of point-biserial correlation, while five items could be considered acceptable only in classroom settings. These results suggest that ChatGPT has potential in generating multiple-choice questions that can be used in medical education.

Automatic item generation (AIG) has been performed using mainly one of these three different methods (Kurdi et al. 2020): Syntax-based, semantic-based, and template-based. Template-based methods have shown a good success (Falcão et al. 2022) even in national exams, for example, in Canadian exams (Pugh et al. 2020), and have also provided promising results in various languages such as Chinese (Gierl et al. 2016) and Turkish (Kiyak et al. 2023). However, the process relies more on expert effort and time than generating items based on NLP techniques. According to Gierl et al. (2021), non-template-based approaches such as NLP techniques were not preferred over template-based methods in AIG. However, a recent study found promising results in an expert evaluation of questions generated by ChatGPT using simple prompts (Cheung et al. 2023). Our study took a step further by presenting evidence from an exam setting. These recent findings indicate the potential shift away from template-based methods. This potential may have arisen from the fact that GPT-3's corpus is ten times larger than any previous models (Brown et al. 2020), even if the data is not directly related to the test's purpose. Therefore, it is reasonable to assume that larger and specialized models could result in higher quality items. Additionally, utilizing more complex prompts instead of our simple ones also may lead to better items. Incorporating factors such as learner type, competency level, and preferred difficulty level would probably have enhanced the suitability of the generated multiple choice questions. As a consequence of that, prompt engineering could potentially become a crucial skill for medical teachers and test developers. Furthermore, there already are well-designed prompts for generating multiple-choice questions in the published literature (Kıyak 2023; Zuckerman et al. 2023).

While our study demonstrated that ChatGPT can be effective in generating clinical vignettes and multiple-choice questions in the context of medical education, it is important to consider the potential disadvantages of using this technology. One major disadvantage is the lack of control over the generated content. There is always a risk that some generated vignettes or questions may contain inaccuracies or biases that could lead to incorrect or misleading information being presented to students, as once ChatGPT claimed that "the human heart only has two chambers" (Lee 2023). Similarly, in one of our questions, an option included "none of the above", an option discouraged by test development guidelines for multiple-choice questions. While developing materials, it is still essential for medical educators and test developers to bear in mind that any AI model depends on the data it was trained on. Therefore, using well-designed prompts (Kıyak 2023; Zuckerman et al. 2023) rather than simple ones, exercising caution, and incorporating expert oversight remain imperative (Han et al. 2023; Indran et al. 2023). To mitigate the risk in teaching, we initially provided the vignettes for approval by subject-matter experts.

This study has some limitations. The first limitation is that the sample size was small. Moreover, the participants' specific characteristics, such as gender and age, might affect generalizability. Unfortunately, the lack of demographic data collection prevented us from assessing their impact. Another limitation is that the study was conducted in a single institution, and the results may not be generalizable to other settings. Therefore, future studies can be conducted in different institutions and settings to increase the external validity of the findings. However, it is important to note that extrapolating the results to different subjects or medical schools could be challenging due to the ongoing rapid evolution of language models. For example, a newer version, which is GPT-4, was released on March 2023. The current findings are indicative of its capabilities as of January 2023. Another limitation arises from using only the difficulty and discrimination indices as the primary indicator of question quality. An in-depth analysis of the question content could identify any problems in the questions, such as "none of the above" option.

Conclusion

The present study provides evidence that ChatGPT has potential in generating clinical vignettes for teaching and multiple-choice questions for assessment in medical education. The results showed that the quality of the vignettes are comparable to those created by human authors, and some multiple-questions have acceptable psychometric characteristics. However, further research is needed to support or refute these findings and to address the limitations of this study. Despite the limitations, this study sheds light on the role of artificial intelligence in medical education and highlights the potential of large language models in generating educational and assessment materials.

Acknowledgements

The authors are grateful to the students who participated in this study.

Ethical approval

The students were required to participate in evidence-based medicine practices as part of their curriculum, but participation in the randomized controlled study was voluntary. The institution gave permission to present the results. Gazi University Institutional Review Board approved the study on 10 January 2023 (code: 2023 – 49).

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

No funding was received for conducting this study.

Notes on contributors

Özlem Coşkun is a medical doctor and an associate professor of medical education in the Department of Medical Education and Informatics, Gazi University Faculty of Medicine, Ankara, Turkey.

Yavuz Selim Kıyak is a medical doctor and has a PhD degree in medical education. His research interests include clinical reasoning, formative assessment, and distributed ledger technologies such as Holochain. He dedicated himself spreading the knowledge regarding Medical Education using social media. YouTube: https://www.youtube.com/c/MedicalEducationFlamingo X (Twitter): @MedEdFlamingo

Işıl İrem Budakoğlu is a medical doctor, specialist of Public Health, professor of medical education, and chair of the Department of Medical Education and Informatics, Gazi University Faculty of Medicine, Ankara, Turkey.

ORCID

Özlem Coşkun http://orcid.org/0000-0001-8716-1584
Yavuz Selim Kıyak http://orcid.org/0000-0002-5026-3234
Işıl İrem Budakoğlu http://orcid.org/0000-0003-1517-3169

Data availability statement

The datasets generated and analyzed during the current study are available in the Zenodo repository at https://doi.org/10.5281/zenodo.7920148

References

Alfertshofer M, Hoch CC, Funk PF, Hollmann K, Wollenberg B, Knoedler S, Knoedler L., 2023. Sailing the Seven Seas: a multinational comparison of ChatGPT's performance on medical licensing examinations. Ann Biomed Eng. doi: 10.1007/s10439-023-03338-3.

Bakkum MJ, Hartjes MG, Piët JD, Donker EM, Likic R, Sanz E, de Ponti F, Verdonk P, Richir MC, van Agtmael MA, et al. 2024. Using artificial intelligence to create diverse and inclusive medical case vignettes for education. Br J Clin Pharmacol. 90 (3):640–648. doi: 10.1111/bcp.15977.

Benoit JRA. 2023. ChatGPT for clinical vignette generation. Re Eval. medRxiv

Brown T, Mann B, Ryder N, et al. 2020. Language models are few-shot learners. In: Larochelle H, Ranzato M, Hadsell R,editors. Advances in Neural Information Processing Systems. Vancouver (Canada): Curran Associates, Inc.; p. 1877–1901.

Buchholz K. 2023. Infographic: chatGPT Sprints to One Million Users. In: Statista Infographics. https://www.statista.com/chart/29174/time-to-one-million-users. Accessed 28 Apr 2023.

Carrasco JP, García E, Sánchez DA, Porter E, De La Puente L, Navarro J, Cerame A. 2023. ¿Es capaz "ChatGPT" de aprobar el examen MIR de 2022? Implicaciones de la inteligencia artificial en la educación médica en España. Rev Esp Edu Med. 4(1): 55–69. doi: 10.6018/edumed.556511.

Cheung BHH, Lau GKK, Wong GTC, Lee EYP, Kulkarni D, Seow CS, Wong R, Co MT-H. 2023. ChatGPT versus human in generating medical graduate exam multiple choice questions—A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). PLOS One. 18(8):e0290691. doi: 10.1371/journal.pone.0290691.

Choi EPH, Lee JJ, Ho M-H, Kwok JYY, Lok KYW., 2023. Chatting or cheating? The impacts of ChatGPT and other artificial intelligence language models on nurse education. Nurse Educ Today. 125: 105796. doi: 10.1016/j.nedt.2023.105796.

Coşkun Ö, Kıyak YS, Budakoğlu Iİ, et al. 2022. A novel approach to teach evidence-based medicine: modified PEARLS. GMJ. 33:329–336. doi: 10.12996/gmj.2022.75.

Cotton DRE, Cotton PA, Shipway JR. 2023. Chatting and cheating: ensuring academic integrity in the era of ChatGPT. Innov Educ Teach Int. 61(2):228–239. doi: 10.1080/14703297.2023.2190148.

Creswell JW. 2012. Educational research: planning, conducting, and evaluating quantitative and qualitative research. Boston:Pearson

Cross J, Robinson R, Devaraju S, Vaughans A, Hood R, Kayalackakom T, Honnavar P, Naik S, Sebastian R. 2023. Transforming medical education: assessing the ıntegration of chatgpt ınto faculty workflows at a Caribbean Medical School. Cureus. 15(7):e41399., doi: 10.7759/cureus.41399.

Downing SM, Yudkowsky R. 2009. Assessment in Health Professions Education. New York (NY): Routledge

Elkins K, Chun J. 2020. Can GPT-3 pass a writer's turing test? J Cult Anal. 5(2):1–16. doi: 10.22148/001c.17212.

Falcão F, Costa P, Pêgo JM. 2022. Feasibility assurance: a review of automatic item generation in medical assessment. Adv Health Sci Educ Theory Pract. 27(2):405–425. doi: 10.1007/s10459-022-10092-z.

Floridi L, Chiriatti M. 2020. GPT-3: ıts nature, scope, limits, and consequences. Minds Machines. 30(4):681–694. doi: 10.1007/s11023-020-09548-1.

Gierl MJ, Lai H, Fung K, Zheng B. 2016. Using technology-enhanced processes to generate test items in multiple languages. In: F. Drasgow, editor. Technology and testing: improving educational and psychological measurement. New York: Routledge; p. 109–127.

Gierl MJ, Lai H, Tanygin V. 2021. Advanced methods in automatic item generation. 1st ed. New York (NY): Routledge.

Han Z, Battaglia F, Udaiyar A, Fooks A, Terlecky SR., 2023. An explorative assessment of ChatGPT as an aid in medical education: use it with caution. Med Teach. doi: 10.1080/0142159X.2023.2271159.

Indran IR, Paranthaman P, Gupta N, Mustafa N. 2023. Twelve tips to leverage AI for efficient and effective medical question generation: a guide for educators using Chat GPT. Med Teach. doi: 10.1080/0142159X.2023.2294703.

Kiyak YS, Budakoğlu Iİ, Coşkun Ö, Koyun E. 2023. The first automatic item generation in Turkish for assessment of clinical reasoning in medical education. Tıp Eğitimi Dünyası. 22(66):72–90. doi: 10.25282/ted.1225814.

Kıyak YS. 2023. A ChatGPT prompt for writing case-based multiple-choice questions. Rev Esp Edu Med. 4(3):98–103. doi: 10.6018/edumed.587451.

Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, et al. 2023. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2(2): e0000198. doi: 10.1371/journal.pdig.0000198.

Kurdi G, Leo J, Parsia B, Sattler U, Al-Emari S. 2020. A systematic review of automatic question generation for educational purposes. Int J Artif Intell Educ. 30(1):121–204., doi: 10.1007/s40593-019-00186-y.

Lee H. 2023. The rise of ChatGPT : exploring its potential in medical education. Anat Sci Ed. :2270. doi: 10.1002/ase.2270.

Masters K. 2023. Ethical use of artificial intelligence in health professions education: AMEE Guide No.158. Med Teach. 45(6):574–584. doi: 10.1080/0142159X.2023.2186203.

Mihalache A, Huang RS, Popovic MM, Muni RH. 2023. ChatGPT-4: an assessment of an upgraded artificial intelligence chatbot in the United States Medical Licensing Examination. Med Teach. 46(3): 366–372. doi: 10.1080/0142159X.2023.2249588.

Norman G. 2010. Likert scales, levels of measurement and the “laws” of statistics. Adv Health Sci Educ Theory Pract. 15(5):625–632. doi: 10.1007/s10459-010-9222-y.

Ouyang F, Zheng L, Jiao P. 2022. Artificial intelligence in online higher education: a systematic review of empirical research from 2011 to 2020. Educ Inf Technol. 27(6):7893–7925. doi: 10.1007/s10639-022-10925-9.

Preiksaitis C, Nash C, Gottlieb M, Chan TM, Alvarez A, Landry A. 2023. Brain versus bot: distinguishing letters of recommendation authored by humans compared with artificial intelligence. AEM Educ Train. 7(6):10924. doi: 10.1002/aet2.10924.

Pugh D, De Champlain A, Gierl M, Lai H, Touchie C. 2020. Can automated item generation be used to develop high quality MCQs that assess application of knowledge? RPTEL. 15(1):12. doi: 10.1186/s41039-020-00134-8.

Sallam M. 2023. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare. 11(6):887. doi: 10.3390/healthcare11060887.

Seetharaman R. 2023. Revolutionizing medical education: can ChatGPT boost subjective learning and expression? J Med Syst. 47(1):61. doi: 10.1007/s10916-023-01957-w.

Stockler MR, March L, Lindley RI, Mellis C. 2009. Students’ PEARLS: successfully incorporating evidence-based medicine in medical students’ clinical attachments. Evid Based Med. 14(4):98–99. doi: 10.1136/ebm.14.4.98-a.

Sullivan GM, Feinn R. 2012. Using effect size—or why the *P* value ıs not enough. J Grad Med Educ. 4(3):279–282. doi: 10.4300/JGME-D-12-00156.1.

Thurzo A, Strunga M, Urban R, Surovková J, Afrashtehfar KI. 2023. Impact of artificial ıntelligence on dental education: a review and guide for curriculum update. Educ Sci. 13(2):150., doi: 10.3390/educsci13020150.

Tsang R. 2023. Practical applications of ChatGPT in undergraduate medical education. J Med Educ Curric Dev. 10(238212052311784): 23821205231178449. doi: 10.1177/23821205231178449.

Wang X, Gong Z, Wang G, Jia J, Xu Y, Zhao J, Fan Q, Wu S, Hu W, Li X, et al. 2023. ChatGPT performs on the Chinese National Medical Licensing Examination. J Med Syst. 47(1):86. doi: 10.1007/s10916-023-01961-0.

Wartman SA, Combs CD. 2018. Medical education must move from the ınformation age to the age of artificial ıntelligence. Acad Med. 93(8):1107–1109. doi: 10.1097/ACM.0000000000002044.

Zawacki-Richter O, Marín VI, Bond M, Gouverneur F. 2019. Systematic review of research on artificial intelligence applications in higher education – where are the educators? Int J Educ Technol High Educ. 16(1):1–27. doi: 10.1186/s41239-019-0171-0.

Zuckerman M, Flood R, Tan RJB, Kelp N, Ecker DJ, Menke J, Lockspeiser T. 2023. ChatGPT for assessment writing. Med Teach. 45(11):1224–1227. doi: 10.1080/0142159X.2023.2249239.