European Journal of Clinical Pharmacology (2024) 80:729–735
https://doi.org/10.1007/s00228-024-03649-x

RESEARCH

ChatGPT for generating multiple-choice questions: Evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam

Yavuz Selim Kıyak[1,2] · Özlem Coşkun[1] · Işıl İrem Budakoğlu[1] · Canan Uluoğlu[3]

Received: 27 December 2023 / Accepted: 3 February 2024 / Published online: 14 February 2024

Abstract

Purpose Artificial intelligence, specifically large language models such as ChatGPT, offers valuable potential benefits in question (item) writing. This study aimed to determine the feasibility of generating case-based multiple-choice questions using ChatGPT in terms of item difficulty and discrimination levels.

Methods This study involved 99 fourth-year medical students who participated in a rational pharmacotherapy clerkship carried out based-on the WHO 6-Step Model. In response to a prompt that we provided, ChatGPT generated ten case-based multiple-choice questions on hypertension. Following an expert panel, two of these multiple-choice questions were incorporated into a medical school exam without making any changes in the questions. Based on the administration of the test, we evaluated their psychometric properties, including item difficulty, item discrimination (point-biserial correlation), and functionality of the options.

Results Both questions exhibited acceptable levels of point-biserial correlation, which is higher than the threshold of 0.30 (0.41 and 0.39). However, one question had three non-functional options (options chosen by fewer than 5% of the exam participants) while the other question had none.

Conclusions The findings showed that the questions can effectively differentiate between students who perform at high and low levels, which also point out the potential of ChatGPT as an artificial intelligence tool in test development. Future studies may use the prompt to generate items in order for enhancing the external validity of the results by gathering data from diverse institutions and settings.

Keywords ChatGPT · Artificial intelligence · Automatic item generation · Multiple-choice questions · Rational pharmacotherapy · Medical education

Introduction

The introduction of innovative artificial intelligence (AI) tools brings exciting possibilities. A noteworthy example of such tools is the Generative Pretrained Transformer (GPT), a large language (LLM) model developed by OpenAI. ChatGPT, a chatbot variant of GPT-3.5, was publicly introduced at the end of November 2022. It garnered a user base of one million within just five days [1]. Consequently, there have been suggestions to consider the release date of ChatGPT as a pivotal milestone, marking the division between the pre-ChatGPT era and the post-ChatGPT era [2].

GPT operates based on the principles of natural language processing (NLP), an area that has witnessed substantial advancements in recent years, particularly with the emergence of LLMs [3]. These models undergo extensive training with textual data, equipping them with the ability to produce text that closely resembles human writing, provide precise responses to queries, and perform other language-related tasks with a high level of accuracy [4].

AI has become integrated not only into medical education [5, 6] but also into higher education through diverse

✉ Yavuz Selim Kıyak
yskiyak@gazi.edu.tr

1 Department of Medical Education and Informatics, Faculty of Medicine, Gazi University, Ankara, Turkey

2 Gazi Üniversitesi Hastanesi E Blok 9, Kat 06500 Beşevler Ankara, Turkey

3 Department of Medical Pharmacology, Faculty of Medicine, Gazi University, Ankara, Turkey

applications [7, 8]. In the specific context of assessment in medical education, ChatGPT showed different levels of performance in various national medical exams [9–13]. A most recent study showed that GPT-4 version of ChatGPT answered more than 85% of the questions in the United States Medical Licensing Examination correctly [14]. Despite the presence of studies that focus on providing human-generated questions to ChatGPT, there has been a lack of research focused to asking ChatGPT-generated questions to humans.

Using ChatGPT to generate questions can be classified as a type of automatic item generation (AIG). There are two main groups of methods in AIG [15]: template-based and non-template-based. The template-based method, as demonstrated in the literature [16–18], has showed satisfying levels of validity evidence, even in national examinations [19]. Moreover, they have demonstrated promising outcomes not only in English but also in Chinese, French, Korean, Spanish, and Turkish [20, 21]. Despite this impressive success, the template-based AIG process continues to depend more on expert effort than on the use of NLP techniques (non-template-based) in AIG.

Although a study highlights the ongoing integration of ChatGPT into the practices of medical school members, including the use of ChatGPT for writing multiple-choice questions [22], none of the existing studies have made an effort to assess the quality of assessment content generated by ChatGPT. Only three studies proposed prompts for generating multiple-choice questions using ChatGPT [23–25]. Given that ChatGPT has exhibited academic hallucinations [2] and made inaccurate claims, such as asserting that "the human heart only has two chambers" [26], a thorough evaluation is necessary. Therefore, there is a need for studies that examine the quality of multiple-choice items generated by ChatGPT.

This study aimed to determine the feasibility of generating multiple-choice questions using ChatGPT in terms of item difficulty and discrimination levels.

Methods

Study setting and participants

This psychometric research was carried out at the Gazi University Faculty of Medicine, Ankara, Turkey. The psychometric analyses were conducted as a part of internal evaluation process to inform the related faculty board about the exam. This study was a part of a research project related to automatic item generation in different languages. Gazi University Institutional Review Board approved the project (code: 2023–1116). This study constitutes the part involved ChatGPT-generated questions in English.

During the fourth year of the six-year undergraduate medical program, a clerkship that consists of a series of small group activities were carried out to help students to learn the principles of rational prescribing using the WHO 6-Step Model [27]. These activities focused on cases primarily related to hypertension. Following the training, students participated in a written examination that consisted of multiple-choice questions. As the language of the program was English, both the training and the examination were carried out in English. As part of their curriculum, students were required to take the exam. A total of 99 fourth-year medical students enrolled in the undergraduate medical program were considered eligible for participation in the study. As our aim was to include all eligible students, we did not conduct a sample size calculation.

Question generation

The multiple-choice questions were created in August–September 2023 using the "Free Research Preview of ChatGPT" (August 3 Version). We opted not to utilize GPT-4, even though it offers enhanced capabilities compared to GPT-3.5 (offered as a free research preview), primarily due to GPT-4's monthly subscription cost, which could impede its accessibility in developing countries. Table 1 presents the prompt template that we utilized. The prompt asks users to fill these two parts: "[PLEASE INSERT A TOPIC]" and "[PLEASE INSERT A DIFFICULTY LEVEL (E.G. EASY, DIFFICULT)]."

The prompt's origins can be traced back to Esh Tatla, a medical student who initially developed it for medical students [28]. Subsequently, it was further refined and incorporated into the academic literature by a medical education researcher [24].

In the process of question generation, we took into account the specific requirements of the examination aligned with local needs. Given that the training primarily focused on essential hypertension cases, our goal was to generate questions by considering the subjects listed in Table 2. For each of these topics, we tasked ChatGPT with generating both an easy and a difficult multiple-choice question.

Expert panel and test administration

The questions generated by ChatGPT underwent a review process conducted by a panel of experts, comprising members of the rational pharmacotherapy board and also other subject matter experts. Each of these experts had over five years of experience in rational prescribing training and in the development of questions in medical school assessments. They evaluated each question based on two key criteria:

Table 1 The prompt template

You are developing a question bank for medical exams focusing on the topic of [PLEASE INSERT A TOPIC]. Please generate a high-quality single best answer multiple-choice question. Follow the principles of constructing multiple-choice items in medical education. Generate the questions using the following framework:
Case (write as a single narrative paragraph without providing each part separately):
Patient details (gender/age)
Presenting complaint
Relevant clinical history
Physical examination findings
Diagnostic test results (optional)
Question stem: [Insert relevant information from the above sections without compromising the answer] Acceptable question style: Ask for the BEST answer, NOT one that is TRUE/FALSE Answer options:
[Insert plausible answer option]
[Insert plausible answer option]
[Insert plausible answer option]
[Insert plausible answer option]
[Insert plausible answer option]
Explanation:
• Identify and explain the correct answer
• Explain why this is the most appropriate answer based on evidence-based guidelines or expert consensus
• Briefly explain why the other answer options are less correct or incorrect
Difficulty level: [PLEASE INSERT A DIFFICULTY LEVEL (E.G. EASY, DIFFICULT)]

- Criterion 1: "Is there any problem in terms of scientific/clinical knowledge? Is the question clear? Is there only one correct answer? Is the information provided in the question sufficient to find the correct answer? Is the question high-quality?"
- Criterion 2: "Is this question suitable for the unique context of rational drug prescribing training carried out in the school?"

Reviewers were tasked with evaluating the scientific acceptability of the questions through their expertise (Criterion 1) and verifying their suitability for integration into the official clerkship exam (Criterion 2). Importantly, it was explicitly emphasized that they were not authorized to make any changes to the questions. All ten questions were considered scientifically sound and clear. Each question had only one correct answer. The information provided in the questions was sufficient to find the correct answer. However, eight of them were excluded due to their unsuitability for our medical school context (Criterion 2). This decision was based on various factors, one of which was the inclusion of a correct option related to "The Dietary Approaches to Stop Hypertension" (DASH diet). DASH diet is a terminology based on the USA and was not covered in our training. Two questions (#3 and #10) left for inclusion in the exam.

We integrated the questions (Table 3) generated by ChatGPT into the test. To address cultural considerations,

Table 2 The inserted elements to the prompt template

[PLEASE INSERT A TOPIC] was filled with:	[PLEASE INSERT A DIFFICULTY LEVEL (E.G. EASY, DIFFICULT)] was filled with:	ID of the question
"rational drug prescribing in managing the initial diagnosis of essential hypertension"	Easy	1
	Difficult	2
"rational drug prescribing in essential hypertension in a primary care setting"	Easy	3
	Difficult	4
"rational drug prescribing in gestational hypertension"	Easy	5
	Difficult	6
"non-pharmacological treatment in managing essential hypertension"	Easy	7
	Difficult	8
"rational drug prescribing in essential hypertension in a patient with gout"	Easy	9
	Difficult	10

Table 3 The ChatGPT-generated questions that were included in the exam

Question #3	A 58-year-old male patient visits your primary care clinic with complaints of persistent high blood pressure. He reports occasional headaches and mild dizziness. His medical history includes type 2 diabetes, for which he takes metformin, and dyslipidemia, managed with atorvastatin. He denies any significant renal or cardiac issues. On physical examination, his blood pressure is consistently elevated at 160/95 mmHg. Cardiac and respiratory exams are unremarkable. Routine blood work and urinalysis show no signs of end organ damage. **In the management of the patient's essential hypertension, which antihypertensive class would be the most appropriate initial choice?** A. Beta-blockers B. Angiotensin-converting enzyme inhibitors (ACE inhibitors) C. Calcium channel blockers D. Alpha-blockers E. Thiazide diuretics Correct answer: B
Question #10	A 58-year-old male patient presents to your clinic with a complaint of recurrent episodes of joint pain in his left big toe, which he describes as sudden, excruciating, and accompanied by redness and swelling. His medical history includes a diagnosis of essential hypertension for the past 10 years, which has been well-controlled with lifestyle modifications. He has recently been experiencing worsening of his joint symptoms, and his physical examination reveals an erythematous and swollen left big toe with limited range of motion. Serum uric acid levels are elevated. He has no history of kidney disease, diabetes, or any other significant medical conditions. **In managing the patient's essential hypertension alongside his gout, which of the following antihypertensive agents should be chosen with caution due to the potential to exacerbate his gout symptoms?** A. Amlodipine B. Hydrochlorothiazide C. Lisinopril D. Metoprolol E. Losartan Correct Answer: B

we replaced patient names with generic terms such as "a patient" or "the patient" by eliminating specific names like "Mr. Johnson." The questions themselves remained unchanged without any further modifications. In total, the test comprised 25 single best answer multiple-choice questions, combined with questions written by human authors. This test was conducted in physical classroom settings, supervised by proctors.

Statistical analysis

We conducted a psychometric analysis based on Classical Test Theory. We performed item-level analysis to determine two parameters: item difficulty and item discrimination indices. Item difficulty was calculated by dividing the cumulative score of examinees by the maximum attainable score. Item discrimination was calculated by using point-biserial correlation (using the Spearman correlation in SPSS 22.0 for Windows, Chicago, IL, USA). This allowed us to determine an individual item's capacity to effectively differentiate between high-performing students and their lower-performing students.

Although large-scale standardized tests require a point-biserial correlation of no less than 0.30 for an item, values in the mid to high 0.20 s can be considered acceptable for locally written classroom-type assessments [29]. Furthermore, we assessed the response distribution for each answer option to identify non-functioning distractors. We adhered to the established criterion on functional distractors as those chosen by examinees at a rate exceeding 5% [29].

Table 4 Item difficulty and discrimination values and response percentages

Question number	Indices		Response percentages in answer options				
	Difficulty	Discrimination (point-biserial correlation)	A	B	C	D	E
3	0.78	0.41*	4.1	**78.7**	15.1	0	2.1
10	0.58	0.39*	8.1	**58.6**	14.1	5.1	14.1

Bold options are correct

*$p < 0.001$

Results

Both of the items demonstrated point-biserial correlations that exceeded the acceptable threshold of 0.30. Although Question #3 had three options (A, D, and E) that did not exceed 5% level of response, Question #10 did not have any non-functional distractors. The specific values for these indices and response percentages can be found in Table 4. The mean difficulty and discrimination levels of remained 23 items were 0.68 and 0.21, respectively.

Discussion

In automatic item generation (AIG) for medical education, template-based AIG methods have been favored over non-template-based ones by researchers because non-template-based methods could not provide feasible multiple-choice questions [21]. In this study, we found that ChatGPT is able to generate multiple-choice questions with acceptable levels of item psychometrics. Our findings showed that the questions can effectively differentiate between students who perform at high and low levels. To our best knowledge, this is the first study that reveals psychometric properties of ChatGPT-generated questions in the English language administered within an authentic medical education context.

The findings point out the beginning of an AI-driven era for AIG instead of using template-based methods. This transformation is readily observable in our new ability, as humans, to produce appropriate case-based multiple-choice questions with minimal human efforts, accomplished by the simple process of inputting a prompt and hitting the enter key. The efficiency achieved through AI would appear remarkable to test developers from a decade ago, who were engaged in the effortful task of manually writing multiple-choice questions.

The increased potential observed in GPT-3 is likely due to its ten times larger dataset compared to previous models [30]. The data may not perfectly align with the test's purpose, but it suggests that specialized language models could help generate better multiple-choice questions. However, it is essential to recognize that the quality of multiple-choice questions is closely tied to prompt quality. This emphasizes the importance of prompt engineering skills for medical teachers and test developers in the future. They can take the prompt we used as a starting point because it is customizable to generate various types of single best answer multiple-choice questions, in contrast to the prompt developed only for generating NBME-style (National Board of Examiners) questions [23].

AI-based AIG has some drawbacks as well. While our findings have shown ChatGPT's ability to generate multiple-choice questions with acceptable psychometrics, it is not infallible. It is crucial for test developers to remember that ChatGPT, like any AI model, relies on the data it has been trained on, and it may sometimes provide inaccurate or outdated content. For instance, certain explanations provided by ChatGPT contained contradicting content regarding the effect of beta-blockers in gout patients, despite the absence of issues with the questions themselves. We did not encounter any problem because we did not use the explanations but if they are used for, for example, formative purposes, using it with caution and expert oversight remain essential [25] to ensure the correctness and relevance. Hence, while it is efficient to generate questions with ChatGPT, constructing an entire exam without any revisions can be challenging. The questions still necessitate subject matter experts to review and revise [31]. This difficulty arises because the generated questions may, for example, lack scientific validity or include elements that do not align with the specific context of a medical school. For instance, the term "DASH diet" is unfamiliar within our training, which led subject matter experts to opt against including that question in the exam.

Another significant drawback is the "black box" nature of AI models. Although this enables us to generate unique questions each time, we cannot know how our input will precisely affect the output. In contrast, template-based AIG offers an appropriate level of control and customization that can be valuable in revising and correcting hundreds of questions at once [21]. While generating questions using AI can be efficient, test developers must consider a balance between the efficiency offered by AI and the need for the level of control.

There are some limitations in our study. The first limitation is that it is based on a limited set of questions and low number of participants from a single university. Although the inclusion of more questions would have been preferable, the need for

compliance with official regulations constrained our ability to expand the number of questions included in the exam. Future studies with more questions are necessary to determine the applicability of the findings across a wider range fields. However, it is important to recognize that extending these results to different subjects or medical institutions may present challenges due to the constantly evolving nature of LLMs. Another limitation is that relying solely on the point-biserial correlation as the primary measure of quality may not have encompassed all relevant quality measures. A detailed qualitative analysis of item content would provide valuable information.

Conclusion

This study investigated the feasibility of a ChatGPT prompt for generating clinical multiple-choice questions because a major challenge faced by medical schools is the labor-intensive task of writing case-based multiple-choice questions for assessing the higher-order skills. The findings showed that ChatGPT-generated questions exhibit psychometric properties that meet acceptable standards. It presents a significant opportunity to make test development more efficient. However, further research is essential to either corroborate or question these findings.

Acknowledgements We express our gratitude to the medical students who participated in this study.

Author contribution Conceptualization: Yavuz Selim Kıyak, Özlem Coşkun, and Canan Uluoğlu. Methodology: Yavuz Selim Kıyak, Özlem Coşkun, and Işıl İrem Budakoğlu. Data collection: Özlem Coşkun and Canan Uluoğlu. Statistical analysis: Yavuz Selim Kıyak. Writing—original draft preparation: Yavuz Selim Kıyak. Writing—review and editing: Işıl İrem Budakoğlu, Özlem Coşkun, and Canan Uluoğlu.

Data availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval The study was performed in accordance with the ethical standards as laid down in the 1964 Declaration of Helsinki. This study has been approved by Gazi University Institutional Review Board (code: 2023-1116).

Consent to participate Informed consent was obtained from the data owners.

Competing interests The authors declare no competing interests.

References

1. 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
2. Masters K (2023) Ethical use of artificial intelligence in health professions education: AMEE Guide No.158. Med Teach 45:574–584. https://doi.org/10.1080/0142159X.2023.2186203
3. Floridi L, Chiriatti M (2020) GPT-3: its nature, scope, limits, and consequences. Mind Mach 30:681–694. https://doi.org/10.1007/s11023-020-09548-1
4. Cotton DRE, Cotton PA, Shipway JR (2023) Chatting and cheating: ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International 1–12. https://doi.org/10.1080/14703297.2023.2190148
5. Masters K (2019) Artificial intelligence in medical education. Med Teach 41:976–980. https://doi.org/10.1080/0142159X.2019.1595557
6. Zhang W, Cai M, Lee HJ et al (2023) AI in medical education: global situation, effects and challenges. Educ Inf Technol. https://doi.org/10.1007/s10639-023-12009-8
7. 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 27:7893–7925. https://doi.org/10.1007/s10639-022-10925-9
8. 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 16:39. https://doi.org/10.1186/s41239-019-0171-0
9. Gilson A, Safranek CW, Huang T et al (2023) How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 9:e45312. https://doi.org/10.2196/45312
10. Kung TH, Cheatham M, Medenilla A et al (2023) Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2:e0000198. https://doi.org/10.1371/journal.pdig.0000198
11. Carrasco JP, García E, Sánchez DA et al (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 4:55–69. https://doi.org/10.6018/edumed.556511
12. Wang X, Gong Z, Wang G et al (2023) ChatGPT performs on the chinese national medical licensing examination. J Med Syst 47:86. https://doi.org/10.1007/s10916-023-01961-0
13. Alfertshofer M, Hoch CC, Funk PF et al (2023) Sailing the Seven Seas: a multinational comparison of ChatGPT's performance on medical licensing examinations. Ann Biomed Eng. https://doi.org/10.1007/s10439-023-03338-3
14. 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. Medical Teacher 1–7. https://doi.org/10.1080/0142159X.2023.2249588
15. Kurdi G, Leo J, Parsia B et al (2020) A systematic review of automatic question generation for educational purposes. Int J Artif Intell Educ 30:121–204. https://doi.org/10.1007/s40593-019-00186-y
16. Falcão F, Costa P, Pêgo JM (2022) Feasibility assurance: a review of automatic item generation in medical assessment. Adv in Health Sci Educ 27:405–425. https://doi.org/10.1007/s10459-022-10092-z
17. Shappell E, Podolej G, Ahn J et al (2021) Notes from the field: automatic item generation, standard setting, and learner performance in mastery multiple-choice tests. Eval Health Prof 44:315–318. https://doi.org/10.1177/0163278720908914
18. Westacott R, Badger K, Kluth D et al (2023) Automated item generation: impact of item variants on performance and standard setting. BMC Med Educ 23:659. https://doi.org/10.1186/s12909-023-04457-0

19. Pugh D, De Champlain A, Gierl M et al (2020) Can automated item generation be used to develop high quality MCQs that assess application of knowledge? RPTEL 15:12. https://doi.org/10.1186/s41039-020-00134-8
20. Kıyak 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 22:72–90. https://doi.org/10.25282/ted.1225814
21. Gierl MJ, Lai H, Tanygin V (2021) Advanced methods in automatic item generation, 1st edn. Routledge
22. Cross J, Robinson R, Devaraju S et al (2023) Transforming medical education: assessing the integration of ChatGPT into faculty workflows at a caribbean medical school. Cureus. https://doi.org/10.7759/cureus.41399
23. Zuckerman M, Flood R, Tan RJB et al (2023) ChatGPT for assessment writing. Med Teach 45:1224–1227. https://doi.org/10.1080/0142159X.2023.2249239
24. Kıyak YS (2023) A ChatGPT prompt for writing case-based multiple-choice questions. Rev Esp Educ Méd 4:98–103. https://doi.org/10.6018/edumed.587451
25. Han Z, Battaglia F, Udaiyar A et al (2023) An explorative assessment of ChatGPT as an aid in medical education: Use it with caution. Medical Teacher 1–8. https://doi.org/10.1080/0142159X.2023.2271159
26. Lee H (2023) The rise of ChatGPT : exploring its potential in medical education. Anatomical Sciences Ed ase.2270. https://doi.org/10.1002/ase.2270
27. Tichelaar J, Richir MC, Garner S et al (2020) WHO guide to good prescribing is 25 years old: quo vadis? Eur J Clin Pharmacol 76:507–513. https://doi.org/10.1007/s00228-019-02823-w
28. Tatla E (2023) 5 Essential AI (ChatGPT) Prompts every medical student and doctor should be using to 10x their…. In: Medium. https://medium.com/@eshtatla/5-essential-ai-chatgpt-prompts-every-medical-student-and-doctor-should-be-using-to-10x-their-de3f97d3802a. Accessed 18 Sep 2023
29. Downing SM, Yudkowsky R (2009) Assessment in health professions education. Routledge
30. Brown T, Mann B, Ryder N et al (2020) Language models are few-shot learners. In: Larochelle H, Ranzato M, Hadsell R et al (eds) Advances in neural information processing systems. Curran Associates, Inc., pp 1877–1901
31. Indran IR, Paramanathan 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. Medical Teacher 1–6. https://doi.org/10.1080/0142159X.2023.2294703