OXFORD

Postgraduate Medical Journal, 2024, 1–8
https://doi.org/10.1093/postmj/qgae065
Education and Learning

ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: a literature review

Yavuz Selim Kıyak [1,*], Emre Emekli[2]

[1]Department of Medical Education and Informatics, Faculty of Medicine, Gazi University, Ankara 06500, Turkey
[2]Department of Radiology, Faculty of Medicine, Eskişehir Osmangazi University, Eskişehir 26040, Turkey
*Corresponding author. Department of Medical Education and Informatics, Gazi Üniversitesi Hastanesi, E Blok 9. Kat, 06500 Beşevler, Ankara, Turkey.
E-mail: yskiyak@gazi.edu.tr

Abstract

ChatGPT's role in creating multiple-choice questions (MCQs) is growing but the validity of these artificial-intelligence-generated questions is unclear. This literature review was conducted to address the urgent need for understanding the application of ChatGPT in generating MCQs for medical education. Following the database search and screening of 1920 studies, we found 23 relevant studies. We extracted the prompts for MCQ generation and assessed the validity evidence of MCQs. The findings showed that prompts varied, including referencing specific exam styles and adopting specific personas, which align with recommended prompt engineering tactics. The validity evidence covered various domains, showing mixed accuracy rates, with some studies indicating comparable quality to human-written questions, and others highlighting differences in difficulty and discrimination levels, alongside a significant reduction in question creation time. Despite its efficiency, we highlight the necessity of careful review and suggest a need for further research to optimize the use of ChatGPT in question generation.

Main messages

- Ensure high-quality outputs by utilizing well-designed prompts; medical educators should prioritize the use of detailed, clear ChatGPT prompts when generating MCQs.
- Avoid using ChatGPT-generated MCQs directly in examinations without thorough review to prevent inaccuracies and ensure relevance.
- Leverage ChatGPT's potential to streamline the test development process, enhancing efficiency without compromising quality.

Keywords: ChatGPT; artificial intelligence; automatic item generation; ChatGPT prompts; multiple-choice questions

Introduction

ChatGPT has emerged as a useful companion to humans with potential benefits in medical education [1]. Its adoption in medical education can be particularly significant, where medical educators are continually seeking innovative methods to enhance learning [2] and assessment. Among the potential innovations, the use of ChatGPT for generating multiple-choice questions (MCQs) has captured interest due to this format's ubiquity in written assessments. The format is ubiquitous because MCQs serve as a useful, which means efficient, scalable, and cost effective, materials in assessing medical students' and residents' knowledge, clinical reasoning, and problem-solving skills [3].

However, while ChatGPT's utility in creating educational content is promising, the validity of the MCQs generated through this artificial intelligence (AI)-driven approach remains unexplored. Validity, in the context of educational assessments, is crucial, ensuring that the questions accurately measure what they are intended to. The quality of the prompts given to ChatGPT directly influences the quality and relevance of generated MCQs, thereby impacting the validity of MCQs. Despite its critical importance, there exists a noticeable gap in the literature regarding the validity of MCQs generated by ChatGPT in medical education.

This gap is significant as the integration of AI tools in educational settings accelerates. Cross-sectional studies showed that medical educators and students, both in undergraduate and postgraduate levels, use ChatGPT for various purposes [4–6] including for question generation [7]. Therefore, it necessitates an understanding of assessment validity of ChatGPT-generated MCQs. The previous reviews were too broad and have not fully explored and documented the prompts whether they generate high-quality MCQs [8].

Addressing this gap is crucial. It will provide medical educators with insights into the effective use of AI for MCQ generation. It also will contribute to the broader discourse on the integration of artificial intelligence in educational practices, particularly the field of medicine requiring high levels of accuracy. By identifying the strengths and limitations of ChatGPT-generated MCQs, this

Received: February 28, 2024. **Revised:** April 29, 2024. **Accepted:** May 23, 2024

research can guide the development of optimal strategies for AI-assisted test development in medical education.

Therefore, this rapid review aims to answer the following research questions:

1. What prompts have been used for MCQ generation using ChatGPT in medical education?
2. What is the validity evidence of the MCQs generated using ChatGPT in medical education?

Materials and methods

We opted for conducting a rapid review due to the urgent need for information about the use of AI, specifically ChatGPT, in generating MCQs for medical education. Given the fast-paced developments in AI and its growing application in medical education settings, there was a pressing demand for immediate guidance and understanding that could inform educators and test developers. This approach allowed for a quicker synthesis of available evidence, providing early, actionable findings to stakeholders in a rapidly evolving field. We followed the practical guide for rapid reviews [9] to ensure a methodologically sound and efficient approach.

Search strategy

The literature search was designed to identify studies involving the use of ChatGPT for the generation of MCQs in medical education. Therefore, the search terms used were ("ChatGPT" OR "Chat GPT") AND ("medicine" OR "medical"). The search was conducted without any time restrictions and up until February 15, 2024. The databases searched included PubMed and Web of Science, adhering to the recommendation of searching at least two electronic databases [9]. The search included all fields in the studies. The initial search was conducted by one researcher, yielding 1690 studies from PubMed and 1063 studies from Web of Science. After removing duplicates with the help of a reference management tool, 1920 articles remained for title and abstract screening. After the screening process, we also searched the reference lists of the eligible studies [9].

Inclusion and exclusion criteria

The selection criteria included any study that reported the use of a prompt for generating MCQs within a medical education context. We did not limit the type of publications (e.g. research articles, reviews, commentaries). Exclusion criteria were studies not published in English and those outside of medical education (other health professions education, such as, nursing, dentistry, and pharmacy).

Screening of the studies

To enhance the reliability, two reviewers involved in the screening process. The screening began with titles and abstracts together, resulting in 158 studies selected for full-text screening in terms of the criteria. Following the full-text screening, only 21 articles met the criteria. The search in the reference lists added two more studies suitable for full review, bringing the total to 23 studies. Since we invested considerable effort to achieve a shared understanding before screening, we encountered minimal number of discrepancies between the two reviewers. These were resolved through discussion.

Extraction of the data

Data extraction was carried out by the experienced researcher on ChatGPT for question generation (YSK), focusing on key information such as author(s), date of publication, publication type, ChatGPT version used, the prompts employed, the number of MCQs generated, any further modifications made to the questions, the basis of evaluation presented by the studies, and the outcomes. Although the guide recommended as a reasonable approach that a second reviewer should check a 10% random sample for accuracy [9], in this study, the second reviewer (EE) verified the extracted data from all 23 studies. This process identified and corrected five instances of inconsistency or error overlooked by the first reviewer in the initial review process.

Assessment of the results

Prompts were evaluated considering the prompt engineering strategies and tactics described in a guide provided by OpenAI [10]. Some of them are: Write clear instructions, provide/mention examples, instructing to answer using a reference text, and ask the model to adopt a persona. The complete list of these strategies and tactics is presented in Table 1.

Validity was assessed by considering five major sources of validity evidence in testing [11]: First, "content" examines the relationship between test content and the intended construct. Second, "response process" focuses on individual responses and their alignment with intended interpretations. Third, "internal structure" focuses on the reliability and statistical characteristics of items within the assessment, such as, item difficulty and discrimination, and functionality of distractor options. Fourth, "relations to other variables" involves external data analysis, such as correlations with independent measures. Fifth, "evidence based on consequences of testing" focuses on the impact of assessments on individuals, institutions, and society, considering both intended and unintended effects.

Results

What prompts have been used for MCQ generation using ChatGPT in medical education?

We presented the prompts in Table 2. Four studies did not provide the prompt they used [12–15] and five studies did not report the version of ChatGPT [12, 16–19].

Six studies asked ChatGPT to generate MCQs based on a known style, such as, American Board of Dermatology Applied Exam [20], The United States Medical Licensing Examination (USMLE) [12, 17, 21, 22], and The National Board of Medical Examiners (NBME) [19]. This strategy aligns with providing/mentioning examples as a prompt engineering tactic.

Three studies used another prompt engineering tactic by asking the model to adopt a persona. These prompts were "You are a developer of teaching materials …" [23] and "You are developing a question bank for medical exams …" [24, 25]. Three studies submitted text for ChatGPT to generate responses using the provided text as a reference [18, 20, 26], a tactic that is also recommended. Five studies were so kind in their prompts as they used "please" to ask the model to generate questions [22, 27–30], even if it is not a prompt engineering tactic.

Table 1. OpenAI guidance on prompt engineering.

Strategy	Tactics	When/Why to use
Write clear instructions	Include details in your prompt	When specificity is needed for relevance.
	Ask the model to adopt a persona	To guide the model's tone or perspective.
	Use delimiters to clearly indicate parts	For complex inputs requiring clear separation.
	Specify steps to complete a task	For tasks that can be broken down into sequential actions.
	Provide examples	To model a desired format or approach.
	Specify desired output length	When output needs to be within certain length constraints.
Provide reference Text	Instruct to answer using a reference text	To base responses on provided material for accuracy.
	Instruct to answer with citations from reference text	When referencing specific information directly from provided texts.
Split complex tasks into simpler subtasks	Use intent classification	To identify and address specific parts of a complex query.
	Summarize/filter previous dialog	In long dialogs, to maintain relevance and context.
	Summarize long documents piecewise	When dealing with content that exceeds model's context window (allowed input load).
Give time to "think"	Instruct model to work out its own solution before concluding	For reasoning tasks requiring step-by-step processing.
	Use inner monolog or sequence of queries	To internally process reasoning before providing an answer.
	Ask if anything was missed previously	To ensure comprehensiveness in tasks requiring thorough exploration.
Use external tools	Implement embeddings-based search	For enhancing model's knowledge with external information.
	Use code execution for calculations or API calls	For tasks requiring precise calculations or external data.
	Give access to specific functions	To benefit from specialized functionalities in specific contexts.
Test changes systematically	Evaluate outputs with reference to gold-standard answers	Testing is always helpful.

What is the validity evidence for the MCQs generated using ChatGPT in medical education?

Content

The generated questions were on various domains and diseases: Physiology [16, 31], dermatology [20], anatomy [13, 14, 30], immunology [32], internal medicine [26], surgery [26] and neurosurgery [14], anesthesia [15], clinical pharmacology [25], carpal tunnel syndrome [12], syndrome of inappropriate antidiuretic hormone secretion [21], diabetes [28], hyperlipidemia [29], anterior cutaneous nerve entrapment syndrome [23], urinary tract infection [33], hypertension [24, 25], chondroid tumors [18], and reproductive system [19].

The rates of MCQs that have inaccuracies based on expert reviews were 60% [20], 21% and 37% [34], 1% and %15 [27], and 16% [31]. One study reported that all MCQs were considered acceptable in an expert panel [25]. However, each study used different criteria with small number of experts. One study found that a large proportion of the ChatGPT-generated questions were nearly identical [34].

Response process

All studies generated single best answer MCQs in English, some with five options, and others with four options.

Internal structure

Three studies conducted item analyses [19, 25, 31]. In one study, two case-based MCQs produced ideal discrimination levels (above 0.30), which are 0.41 and 0.39 [25], in an exam consists of 25 MCQs (23 were human-written). Difficulty levels were 0.78 and 0.58 [25], which are easy and moderate, respectively. In another study, the average of discrimination scores of 21 MCQs was 0.24, which is acceptable (above 0.20) but some particular MCQs had unacceptable discrimination levels [31]. Average of difficulty levels of 21 MCQs was 0.62 [31]. In a study that carried out an editing process after generating MCQs, 29 MCQs' discrimination level was 0.23, which is not ideal but acceptable, and average difficulty level was 0.71 [19].

Relations to other variables

Some studies compared different aspects of ChatGPT-generated and human-written MCQs. In one study based only on expert reviews, "appropriateness, clarity and specificity, discriminative power, and suitability for medical school exams" of MCQs generated by using ChatGPT were comparable to human-written MCQs but inferior in the relevance criterion [26]. In one study based on administering MCQs in exams, human-written MCQs had significantly higher levels of discrimination but similar level of difficulty [31]. In another study, ChatGPT-based MCQs had similar levels of discrimination and difficulty with human-written MCQs but this study carried out further modifications on the MCQs before administering them [19].

Evidence based on consequences of testing

Although several studies mentioned that ChatGPT can expedite item writing process, one study reported that ChatGPT reduced MCQ creation time from an estimated 30–60 minutes per MCQ to 5–15 minutes but these were case-based MCQs [19]. Another study reported that ChatGPT generated 50 MCQs (not case-based) within 20 minutes and 25 seconds, whereas humans spent 211 minutes and 33 seconds for writing 50 MCQs [26].

Table 2. ChatGPT prompts for generating multiple-choice questions and their characteristics published in the literature.

Author(s) and date	Publication type	ChatGPT version	Prompt	Number of MCQs	Further modification	Evaluation based on	Outcome
Agarwal et al., January 2023	Research article	Unspecified	"Generate 5 difficult reasoning-based MCQs for MBBS undergraduates on [Topic]"	110	No	Expert review by two physiologists	The questions were "highly valid", "somewhat difficult or easy", and "somewhat required reasoning".
Ayub et al., August 2023	Research article	An app based on ChatGPT-3.5	"Create five American Board of Dermatology Applied Exam (ABD-AE)-style questions"	40	No	Expert review by two dermatologists	"… only 16 (40%) questions created using ChatGPT 3.5 were accurate and at an appropriate level of complexity for a trainee studying for ABD-AE."
Benitez et al., January 2024	Special article (Perspective)	Unspecified	Unspecified but it might include "USMLE-style"	4	No	Informal review	"the complexity of the questions appeared to be of a lower order"
Biswas, May 2023	Special article (Editorial)	Unspecified	"Provide some MCQs for USMLE step 1 exams."	5	No	None	None
Cheung, August 2023	Research article	ChatGPT-3.5	"Can you write a multiple-choice question based on the following criteria, with the reference I am providing for you and your medical knowledge?"	50	"Interaction is only allowed for clarification but not any modification."	Expert review by five experts (for a comparison to 50 human-written MCQs)	Appropriateness, clarity and specificity, discriminative power, and suitability for medical school exams of AI MCQs are comparable to human-written MCQs, AI was inferior only in the relevance criterion.
Divito et al., November 2023	Special article (New Wave)	ChatGPT-3.5	"For each learning issue, can you provide a USMLE style multiple choice (A-E) question that tests student knowledge on the topic? Provide the correct answer and a justification of the correct choice for each question."	10	No	None	None
Doggett et al., February 2024	Special article (Commentary)	ChatGPT-3.5	"Create a 5 answer Single Best Answer question at the level of a [final year/preclinical] medical student on the topic of [topic] where only one answer is correct. Give the answer to this question"	100 pre-clinical, 100 final-year	No	Expert review by six experts (two for preclinical, four for final-year)	"Of 100 preclinical questions, 21 had factual inaccuracies. Of 100 final-year questions, 37 had factual inaccuracies. … a large proportion of the regenerated questions were near identical. … Questions were also generally first-order (recall)…"
E et al., October 2023	Research article	ChatGPT-4	"Could you please generate Multiple Choice Questions (MCQs) on Internal Medicine for medical students? They should be of the same difficulty level as the examples I've provided. Please start numbering from 1, and label the choices from a-d, marking the correct answer with an asterisk. Please write 5 as knowledge questions and 5 with 'clinical history' questions"	210	Two different prompts were used, the description of the process is unclear.	Expert review by five specialists (the number of questions reviewed by each expert is unclear)	"Only 1 question (0.5%) was defined as false; … approximately 15% of the questions generated from the detailed prompt required some correction, primarily due to inaccuracies in content or faulty methodology."
Eysenbach, March 2023	Special article (Editorial)	ChatGPT-3.5	"Please generate a quiz that asks students to identify the symptoms of diabetes."	5	No	None	None

(Continued)

Table 2. Continued.

Author(s) and date	Publication type	ChatGPT version	Prompt	Number of MCQs	Further modification	Evaluation based on	Outcome
Han et al., October 2023	Research article	ChatGPT-3.5	"Please write multiple-choice questions with a vignette containing lab values"	3	Additional prompts "improve this question by adding more lab values in the stem" and "Can you change this question into one for which familial combined hyperlipidemia is the best answer?" were used.	None	None
Hirosawa and Shimizu, October 2023	Special article (Short communication)	ChatGPT-4	"You are a developer of teaching materials. Create a multiple-choice question with one correct answer regarding the symptoms and physical findings in a patient's medical history that would lead you to suspect anterior cutaneous nerve entrapment syndrome (ACNES). Do not generate the answer yet."	1	No	None	None
Ilgaz and Çelik, September 2023	Research article	ChatGPT-3.5	Unspecified	3	Yes but unspecified	Informal review	"… ChatGPT generated one incorrect question"
Indran et al., December 2023	Special article (Twelve Tips)	ChatGPT-3.5 and ChatGPT-4	"1. Generate a multiple-choice question requiring the learner to choose the single best answer out of 5 options. 2. Provide an answer and rationale. 3. The question should test multiple pharmacology-related concepts using a clinical scenario on nitrofurantoin.	2	No	None	None
Kıyak, October 2023	Special article	ChatGPT-3.5	Since the prompt is too detailed, we could not fit in this cell. Please refer to the study [24].	1	No	Informal review	"… it seems the MCQ is both plausible and well-constructed."
Kıyak et al., February 2024	Research article	ChatGPT-3.5	Since the prompt is too detailed, we could not fit in this cell. Please refer to the study [25].	10	No but the change in the patient names for cultural considerations.	Psychometric analysis (item difficulty and discrimination levels, and distractor functionality) in an administration together with 23 human-written questions	All MCQs were considered acceptable in the expert panel. Eight MCQs were not suitable to the context of the medical school. The MCQs had ideal levels of discrimination (0.41 and 0.39), which were above 0.30. Difficulty indices were 0.78 and 0.58. One MCQ had non-functional distractors.
Koga, May 2023	Special article (Letter to Editor)	ChatGPT-4	"Please generate multiple-choice questions formatted in the style of USMLE Step 1, each containing a clinical vignette of about 5 to 6 sentences"	1	No	None	None

(Continued)

Table 2. Continued.

Author(s) and date	Publication type	ChatGPT version	Prompt	Number of MCQs	Further modification	Evaluation based on	Outcome
Laupichler et al., December 2023	Special article (Innovation Report)	ChatGPT-3.5	"Create an MCQ with 5 answer options and a single correct answer. The question is to be answered by medical students. The topic of the question is [specific topic]. The question should be about [learning objective]. However, the terms do not have to appear in each answer option. Make sure that the answer options do not exclude each other—that is, different answer options must not contain directly opposite statements. Provide the correct answer at the end."	25	No	Psychometric analysis (item difficulty and discrimination levels) in an administration together with 25 human-written questions.	Four MCQs (16%) included completely or partially incorrect answer options. In difficulty averages, "… the difference between human questions and LLM questions was not statistically significant." Average discrimination level was acceptable (0.24) but not ideal (not above 0.30). Some MCQs had poor discrimination levels (below zero).
Meşe et al., December 2023	Special article (Review)	Unspecified	"Create a multiple-choice quiz based on the given text about chondroid tumors. Ensure that the correct answer positions vary and aren't always at the same index. Try to keep answer options under 35 characters. Generate a 10-question quiz, ensuring a mix of difficulty levels, with some questions being easy and others hard."	10	No	None	None
Ngo et al., December 2023	Research article	ChatGPT-3.5	"Write 4 multiple choice questions with 4 answers and explanations for the incorrect and correct answers."	60	No	Expert review by four reviewers	32% required minor revision, 43% required major revision, 25% unacceptable
Sevgi et al., April 2023	Research article	ChatGPT-3.5	Unspecified but it might include "at the level of a neurosurgery board exam."	3	No	Informal review	The answers to two questions were correct, one was incorrect.
Subbaramaiah, November 2023	Special article (Letter to Editor)	ChatGPT-3.5	Unspecified	200	No but the best MCQs were selected by experts among 1200 MCQs.	Student perception	Perceived exam difficulty is 7.3/10, half of the students perceived the exam is superior, 44% perceived the exam is comparable to the previous ones.
Totlis et al., August 2023	Research article	ChatGPT-4	"Please generate a multiple-choice quiz for medical students to test their knowledge of musculoskeletal anatomy"	16	An additional prompt "Could you please create another one with increased difficulty?" was used.	None	None
Zuckerman et al., October 2023	Special article (New Wave)	Unspecified	"Write a multiple-choice question in NBME format with 4 sentence clinical vignette, vital signs, and exam findings. Avoid pseudovignettes. [Item-Specific Test Point as a learning objective or disease state or physical exam finding]"	29	"Content expert editing for focus, appropriateness, language, style."	Psychometric analysis (item difficulty and discrimination levels) in an administration together with 250 human-written questions	The average difficulty (0.71) and discrimination (0.23) levels are similar to human-written MCQs. The average discrimination level is not ideal (not above 0.30) but acceptable (above 0.20).

Discussion

This rapid review seeks to explore the use of ChatGPT in generating MCQs for medical education, focusing on the prompts used for generation and the validity of the generated questions. This is in response to the increasing integration of AI tools like ChatGPT into material generation in medical schools [7], especially for creating MCQs that are very common and useful in assessing medical students' knowledge and clinical reasoning [3].

The findings highlight the diversity in the application of ChatGPT for MCQ generation across different medical fields such as physiology, dermatology, anatomy, and internal medicine among others. The findings show that the prompts align with recommended prompt engineering tactics [10], such as, referencing to specific exam styles [19], providing a reference text [26], directing the AI to adopt a test developer role [23], or providing clear and detailed explanations on what is needed in an MCQ [25], could yield better outputs. In parallel with this, a prompt such as "Write 4 multiple choice questions with 4 answers and explanations for the incorrect and correct answers" resulted in ChatGPT's failure to generate suitable MCQs [32], due to the absence of a reference to an existing format or detailed explanation in the prompt. Therefore, medical educators looking to create MCQs should either leverage prompt engineering strategies or employ existing, well-designed prompts. A useful example is a customized ChatGPT for generating MCQs, called Case-based MCQ Generator, which works based on published prompts in the literature [35].

In terms of validity evidence, the studies found varying levels of content accuracy among the MCQs generated by ChatGPT, with some studies noting significant inaccuracies [20, 32]. Although some of these problems can be attributed to poorly designed prompts, they also underscore the importance of expert review and the need for paying attention to the possible inconsistencies in the AI-generated MCQs. Moreover, the internal structure validity of the MCQs, including their difficulty and discrimination levels, was assessed in some studies, revealing mixed outcomes with some questions performing well [19, 25, 31] while others did not meet acceptable standards [31]. However, the results are promising because it significantly reduces the amount of time required for developing an MCQ [19, 26]. Although there are limitations, the efficiency achieved through AI would seem remarkable to test developers from a decade ago who were accustomed to the effortful task of manually creating MCQs. Despite the efficiency of AI in creating MCQs, it is crucial to note that the necessity to review each MCQ remains. This is due to AI's tendency to produce misleading or incorrect information [36], thus ensuring their accuracy becomes a cumbersome process. Consequently, unlike template-based MCQ generation, which permits simultaneous review and revision of hundreds of questions [37], this method does not support such efficient bulk editing yet.

While we interpret the variability in question performance mainly through the lens of prompt engineering, the early stage of research limits our capacity to establish connections with other influencing factors. Some may suggest that factors such as learner level, the category of medicine, and question length could also play an important role. Their potential impact implies the need for further accumulation of research. Such studies should aim to assess how varying these parameters might influence the performance of ChatGPT-generated MCQs. The results of these studies can shed light on developing optimal strategies for using AI for test development.

There are several limitations in this rapid review. The first limitation stems from the inherent characteristics of rapid reviews, which prioritize speed over comprehensive analysis. Second, its focus is on ChatGPT, while justified given its prominence among AI tools, may not provide insights applicable to other AI models, thus limiting the generalizability of the findings. However, it remains a valuable resource for those specifically interested in the capabilities of ChatGPT. Third, the limited number of published studies might not provide a solid foundation for drawing definitive conclusions yet. Despite this, the gathered evidence could still offer fruitful insights for medical educators seeking to understand the potential and limitations of using ChatGPT to generate MCQs. Additionally, we revealed the need for conducting more studies on different aspects of MCQs generation using ChatGPT. Instead of asking human-generated questions to ChatGPT over and over, researchers should focus also on asking ChatGPT-generated questions to humans in order for assessing validity evidence and finding ways for an optimal use of ChatGPT in question generation. Additionally, there is a need for research focused on generating questions in languages other than English.

Conclusions

The findings from this review are crucial for medical educators considering the use of ChatGPT for MCQ generation. We point out the need for careful design of prompts and thorough review of AI-generated questions to ensure they meet educational standards.

Conflict of interest statement: None declared.

References

1. Heng JJY, Teo DB, Tan LF. The impact of chat generative pre-trained transformer (ChatGPT) on medical education. *Postgrad Med J* 2023;**99**:1125–7.
2. Sardesai N, Russo P, Martin J. *et al.* Utilizing generative conversational artificial intelligence to create simulated patient encounters: a pilot study for anaesthesia training. *Postgrad Med J* 2024;**100**:237–41.
3. Pugh D, De Champlain A, Touchie C. Plus ça change, plus c'est pareil: making a continued case for the use of MCQs in medical education. *Med Teach* 2019;**41**:569–77.
4. Eppler M, Ganjavi C, Ramacciotti LS. *et al.* Awareness and use of ChatGPT and large language models: a prospective cross-sectional global survey in urology. *Eur Urol* 2024;**85**:146–53.
5. Sallam M, Salim NA, Barakat M. *et al.* Assessing health students' attitudes and usage of ChatGPT in Jordan: validation study. *JMIR Med Educ* 2023;**9**:e48254.
6. Weidener L, Fischer M. Artificial intelligence in medicine: cross-sectional study among medical students on application, education, and ethical aspects. *JMIR Med Educ.* 2024;**10**:e51247.
7. Cross J, Robinson R, Devaraju S. *et al.* Transforming medical education: assessing the integration of ChatGPT into faculty workflows at a Caribbean medical school. *Cureus* 2023;**15**:e41399.
8. Sahu PK, Benjamin LA, Singh Aswal G. *et al.* ChatGPT in research and health professions education: challenges, opportunities, and future directions. *Postgrad Med J* 2023;**100**:50–5.
9. Tricco AC, Langlois EV, Straus SE (eds). *Rapid Reviews to Strengthen Health Policy and Systems: A Practical Guide*. Genève:World Health Organization, 2017
10. OpenAI. Guide: Prompt engineering. 2024. https://platform.openai.com/docs/guides/prompt-engineering (22 February 2024, date last accessed)

11. Downing SM, Haladyna TM. Validity and its threats. In: Downing SM, Haladyna TM (eds.), *Assessment in Health Professions Education*, New York: Routledge, 2009, 21–56.
12. Benítez TM, Xu Y, Boudreau JD. *et al.* Harnessing the potential of large language models in medical education: promise and pitfalls. *J Am Med Inform Assoc* 2024;**31**:776–83.
13. Ilgaz HB, Çelik Z. The significance of artificial intelligence platforms in anatomy education: an experience with ChatGPT and Google bard. *Cureus*. 2023;**15**:e45301.
14. Sevgi UT, Erol G, Doğruel Y. *et al.* The role of an open artificial intelligence platform in modern neurosurgical education: a preliminary study. *Neurosurg Rev* 2023;**46**:86.
15. Subbaramaiah MT, Dixit A, Sivashanmugam T. Impact of ChatGPT support on conducting regional anaesthesia examination. *Indian J Anaesth* 2023;**67**:1025–6.
16. Agarwal M, Sharma P, Goswami A. Analysing the applicability of ChatGPT, Bard, and Bing to generate reasoning-based multiple-choice questions in medical physiology. *Cureus*. 2023;**15**: e40977.
17. Biswas S. Passing is great: can ChatGPT conduct USMLE exams? *Ann Biomed Eng* 2023;**51**:1885–6.
18. Meşe İ, Altıntaş Taşlıçay C, Kuzan BN. *et al.* Educating the next generation of radiologists: a comparative report of ChatGPT and e-learning resources. *Diagn Interv Radiol* 2024;**30**:163–74. https://doi.org/10.4274/dir.2023.232496.
19. Zuckerman M, Flood R, Tan RJB. *et al.* ChatGPT for assessment writing. *Med Teach*. 2023;**45**:1224–7.
20. Ayub I, Hamann D, Hamann CR. *et al.* Exploring the potential and limitations of chat generative pre-trained transformer (ChatGPT) in generating board-style dermatology questions: a qualitative analysis. *Cureus* 2023;**15**:e43717.
21. Divito CB, Katchikian BM, Gruenwald JE. *et al.* The tools of the future are the challenges of today: the use of ChatGPT in problem-based learning medical education. *Med Teach*. 2023; **46**:320–2.
22. Koga S. The potential of ChatGPT in medical education: focusing on USMLE preparation. *Ann Biomed Eng* 2023;**51**:2123–4.
23. Hirosawa T, Shimizu T. Enhancing clinical reasoning with chat generative pre-trained transformer: a practical guide. *Diagnosis (Berl)* 2024;**11**:102–5.
24. Kıyak YS. A ChatGPT prompt for writing case-based multiple-choice questions. *Rev Esp Edu Med* 2023;**4**:98–103.
25. Kıyak YS, Coşkun Ö, Budakoğlu Iİ. *et al.* ChatGPT for generating multiple-choice questions: evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam. *Eur J Clin Pharmacol*. 2024;**80**:729-35. https://doi.org/10.1007/s00228-024-03649-x.
26. Cheung BHH, Lau GKK, Wong GTC. *et al.* 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* 2023;**18**:e0290691.
27. Klang E, Portugez S, Gross R. *et al.* Advantages and pitfalls in utilizing artificial intelligence for crafting medical examinations: a medical education pilot study with GPT-4. *BMC Med Educ* 2023;**23**:772.
28. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. *JMIR Med Educ* 2023;**9**:e46885.
29. Han Z, Battaglia F, Udaiyar A. *et al.* An explorative assessment of ChatGPT as an aid in medical education: use it with caution. *Med Teach* 2024;**46**:657–64.
30. Totlis T, Natsis K, Filos D. *et al.* The potential role of ChatGPT and artificial intelligence in anatomy education: a conversation with ChatGPT. *Surg Radiol Anat* 2023;**45**:1321–9.
31. Laupichler MC, Rother JF, Grunwald Kadow IC. *et al.* Large language models in medical education: comparing ChatGPT- to human-generated exam questions. *Acad Med* 2024;**99**:508–12. https://doi.org/10.1097/ACM.0000000000005626.
32. Ngo A, Gupta S, Perrine O. *et al.* ChatGPT 3.5 fails to write appropriate multiple choice practice exam questions. *Acad Pathol* 2024;**11**:100099.
33. Indran IR, Paramanathan P, Gupta N. *et al.* Twelve tips to leverage AI for efficient and effective medical question generation: a guide for educators using chat GPT. *Med Teach*. 2023;**1–6**. https://doi.org/10.1080/0142159X.2023.2294703.
34. Doggett T, Warr H, Johnson J-A. *et al.* Response to: "ChatGPT for assessment writing". *Med Teach* 2024;**1**. https://doi.org/10.1080/0142159X.2024.2311269.
35. Kıyak YS, Kononowicz AA. Case-based MCQ generator: a custom ChatGPT based on published prompts in the literature for automatic item generation. *Med Teach*. 2024;1–3. https://doi.org/10.1080/0142159X.2024.2314723.
36. Masters K. Medical Teacher's first ChatGPT's referencing hallucinations: lessons for editors, reviewers, and teachers. *Med Teach*. 2023;**45**:673–5.
37. Gierl MJ, Lai H, Tanygin V. *Advanced Methods in Automatic Item Generation*. 1st edn. New York: Routledge, 2021. https://doi.org/10.4324/9781003025634.