Revista Española de Educación Médica

Comparison of Automatic Item Generation Methods in the Assessment of Clinical Reasoning Skills

Emre Emekli [1]*, Betül Nalan Karahan [2]

[1] Eskişehir Osmangazi University, Faculty of Medicine, Eskişehir, Turkiye; emreemekli90@gmail.com , ORCID ID: 0000-0001-5989-1897

[2] Eskişehir Osmangazi University, Faculty of Medicine, Eskişehir, Turkiye; nalanksgl@gmail.com , ORCID ID: 0000-0003-4958-330X

* Correspondence: emreemekli90@gmail.com

Received: 11/11/24; Accepted: 25/11/24; Posted: 26/11/24

Summary: The use of automatic item generation (AIG) methods offers potential for assessing clinical reasoning (CR) skills in medical education, a critical skill combining intuitive and analytical thinking. In preclinical education, these skills are commonly evaluated through written exams and case-based multiple-choice questions (MCQs), which are widely used due to the high number of students, ease of standardization, and quick evaluation. This research generated CR-focused questions for medical exams using two primary AIG methods: template-based and non-template-based (using AI tools like ChatGPT for a flexible approach). A total of 18 questions were produced on ordering radiologic investigations for abdominal emergencies, alongside faculty-developed questions used in medical exams for comparison. Experienced radiologists evaluated the questions based on clarity, clinical relevance, and effectiveness in measuring CR skills. Results showed that ChatGPT-generated questions measured CR skills with an 84.52% success rate, faculty-developed questions with 82.14%, and template-based questions with 78.57%, indicating that both AIG methods are effective in CR assessment, with ChatGPT performing slightly better. Both AIG methods received high ratings for clarity and clinical suitability, showing promise in producing effective CR-assessing questions comparable to, and in some cases surpassing, faculty-developed questions. While template-based AIG is effective, it requires more time and effort, suggesting that both methods may offer time-saving potential in exam preparation for educators.

Keywords: clinical reasoning; automated item generation; template-based method; ChatGPT; multiple-choice questions

1. Introduction

Clinical reasoning involves both intuitive and analytical thinking processes. Physicians often synthesize these two methods to make clinical decisions (1,2). This process is iterative and complex, encompassing multiple cognitive steps such as acquiring information, generating hypotheses, and identifying problems (3,4). Teaching clinical reasoning processes is a critical component of training future physicians in medical education (5). In assessing clinical reasoning, the top three levels of Miller's pyramid (knows, knows how, shows how, does), which is widely accepted in medical education, are applicable (6). Assessing the "shows how" and "does" levels during the first three years of medical education is challenging, as clinical or simulated settings are typically required for these assessments. For preclinical students, written exams, case-based matching tests, and case-based multiple-choice questions (MCQs) are more feasible methods for evaluating clinical reasoning before they encounter real patients (7,8). MCQs are frequently preferred due to their advantages, including scalability for large numbers of students, ease of standardization, and rapid assessment.

Writing high-quality MCQs is a time-intensive process for medical schools. Preparing more complex questions to assess clinical reasoning skills, in particular, demands a significant amount of

educators' time (9). With extensive exam schedules and broad curricula in medical faculties, there is a constant need to add large numbers of new questions to question banks (10). Moreover, given the clinical responsibilities of educators, preparing exam questions adds a substantial workload, making it essential to develop efficient methods for producing exam-ready questions.

Two primary methods for automatically generating MCQs are described in the literature: template-based question generation and non-template-based automatic item generation (AIG) techniques (11). Each method has distinct advantages and disadvantages. Non-template-based AIG relies on widely used AI chatbots (12-14), where questions are generated without adhering to a predefined template. However, this approach has been criticized for similar reasons as chatbots in general; it functions as a "black box" and may occasionally produce incorrect outputs (15). Conversely, template-based question generation involves creating a cognitive model and question framework beforehand. Once variables are defined, questions are generated using non-AI-based software. When appropriate cognitive models are utilized, this method is reported to be more efficient and psychometrically robust than traditional methods (16). However, it requires more time and effort compared to chatbots, though the likelihood of errors is lower, and any errors can be easily identified and corrected by the developers (11).

Thus, the research question is how do template-based and non-template-based automatic item generation methods compare with faculty-prepared questions in assessing clinical reasoning skills in medical education, particularly in terms of clarity, clinical appropriateness, and effectiveness?

2. Methods

The research is designed as a methodological study focusing on the automatic generation of questions, with expert physicians evaluating the generated questions anonymously. Since no human participants or personal data were involved, ethical approval was not sought. However, the use of AI-generated content raises potential ethical concerns, including the accuracy of the generated content, biases inherent in AI algorithms, and implications for fairness in educational assessments. To address these concerns, all AI-generated questions underwent expert review by experienced radiologists to ensure their clinical and educational appropriateness.

Study Design

In this study, questions were initially generated to assess clinical reasoning using two different automatic question generation techniques: template-based and non-template-based (using AI tools). Additionally, existing multiple-choice questions from medical school exams were included. The chosen topic was radiologic investigations in abdominal emergency pathologies, selected because abdominal pain is a common reason for emergency visits, requiring quick decision-making and clinical reasoning for ordering radiologic investigations. The conditions and appropriate radiologic tests were determined based on the National Core Education Program (17), the curriculum of Eskişehir Osmangazi University Faculty of Medicine, and the American College of Radiology's criteria for radiologic investigation appropriateness (18). Eighteen diagnoses or provisional diagnoses, deemed suitable for the target audience's knowledge level, were selected by radiologists with doctoral degrees in medical education and expertise in emergency radiology. For each of these topics, questions were generated using both template-based and non-template-based techniques. The topics included:Non-localized acute abdominal pain (with fever, no recent surgery, first imaging)

- Non-localized acute abdominal pain (post-surgical patient, first imaging).
- Suspected acute pancreatitis (epigastric pain, elevated amylase-lipase, less than 48-72 hours since symptom onset, first imaging).
- Epigastric pain (suspected acid reflux, esophagitis, gastritis, peptic ulcer, or duodenal ulcer, first imaging).
- Imaging for mesenteric ischemia (suspected acute mesenteric ischemia, first imaging).

- Left lower quadrant pain (left lower quadrant pain, first imaging).
- Non-variceal upper gastrointestinal bleeding (arterial bleeding source identified on endoscopy).
- Right lower quadrant pain (right lower quadrant pain, first imaging).
- Right lower quadrant pain (pregnant patient, right lower quadrant pain, fever, leukocytosis, suspected appendicitis, first imaging).
- Right upper quadrant pain (suspected biliary disease, first imaging).
- Suspected small bowel obstruction (acute presentation, first imaging).
- Abnormal uterine bleeding (abnormal uterine bleeding, first imaging).
- Acute pelvic pain (reproductive age, gynecological etiology suspected, positive β-hCG, first imaging).
- Acute pelvic pain (postmenopausal, acute pelvic pain, first imaging).
- Flank pain (acute onset, suspected stone disease, no history of stone disease or present history, first imaging).
- Flank pain (pregnant patient, acute onset, suspected stone disease, first or follow-up imaging).
- Hematuria (microscopic hematuria, no risk factors, no recent strenuous exercise, no infection, viral illness, recent or current menstruation, first imaging).
- Hematuria (macroscopic hematuria, first imaging).

Radiology training in our medical schools is typically offered during the fourth or fifth year; therefore, the questions were designed to match the fifth-year medical student level.

Template-Based Question Generation

Template-based question generation involves a three-step process (19). In the first step, content experts identify the necessary content for question generation and present it in a cognitive model that highlights the information, skills, and problem-solving processes required to reach a specific diagnosis. In the second step, a question model (template) is developed based on this cognitive model, structuring variables such as content and answer options for each generated question. Words or phrases for the variables in the template are then identified. In the third step, computer-based algorithms generate multiple questions from this model (20,21).

Step 1: Creating the Cognitive Model

As outlined in the study design, questions were generated on the topic of radiologic investigations in abdominal emergency pathologies for 18 diagnostic scenarios. For each scenario, variables such as patient information (age, gender) and disease information (symptoms, history, physical examination findings, and laboratory findings) were identified, and appropriate age ranges and genders were assigned. Potential symptoms, histories, physical findings, and laboratory results for each disease were documented and categorized into similar groups (e.g., Symptom A, B, C, D, E, F) (Table 1).

Table 1. Cognitive Model Variables for Automatic Item Generation

Variable	Descriptions
Symptom A	abdominal pain, fever, nausea, vomiting, diarrhea
Symptom B	sudden onset flank pain, fever, nausea, vomiting, blood in urine
Symptom C	vomiting blood, stomach cramps, fatigue, dark stools
Symptom D	right upper quadrant pain, nausea, vomiting, fever, chills
Symptom E	sudden onset pelvic pain, foul-smelling vaginal discharge, fever, chills

Symptom F	prolonged vaginal bleeding, clotted vaginal bleeding, excessive vaginal bleeding filling a pad, fatigue, weakness, palpitations
History A	no known illness, diabetes, hypertension, no known disease, smoking, social alcohol consumption
History B	arrhythmia, previous heart attack, high cholesterol, known heart disease, heart failure
History C	abdominal surgery last year, recent abdominal surgery
History D	frequent acid reflux after meals, increased symptoms after fatty and spicy foods
History E	endoscopy was performed, but no bleeding source was identified
History F	pain started suddenly yesterday, no previous similar complaints
Physical Exam A	generalized abdominal pain, pain in all quadrants, generalized tenderness in the abdomen
Physical Exam B	costovertebral angle tenderness
Physical Exam C	epigastric pain
Physical Exam D	left lower quadrant pain, tenderness in the left lower quadrant
Physical Exam E	guarding in the right lower quadrant, guarding and rebound in the right lower quadrant, tenderness in the right lower quadrant
Physical Exam F	positive Murphy sign, guarding in the right upper quadrant
Laboratory A	leukocytosis, elevated CRP, elevated ESR
Laboratory B	leukocytosis, low hemoglobin
Laboratory C	elevated red blood cells in urinalysis, positive bacteria in urinalysis
Laboratory D	elevated amylase-lipase
Laboratory E	leukocytosis, elevated CRP, elevated total-direct bilirubin, elevated ALP and GGT
Laboratory F	positive βhCG and leukocytosis, positive βhCG and elevated CRP

At this stage, variables that could not be grouped were excluded from the study. The cognitive model developed by the authors was then reviewed by a physician with academic expertise in question generation and a radiology faculty member. Following minor adjustments, the final version of the cognitive model was established (Table 2).

Table 2. Comprehensive Cognitive Model for Abdominal Emergency Scenarios

Diagnosis	Age	Gender	Symptom A	Symptom B	Symptom C	Symptom D	Symptom E	Symptom F	History A	History B	History C	History D	History E	History F	PE A	PE B	PE C	PE D	PE E	PE F	Laboratory A	Laboratory B	Laboratory C	Laboratory D	Laboratory E	Laboratory F
Acute abdominal pain (No surgery)	18-80	F/M	Y	N	N	N	N	N	Y	N	N	N	N	N	Y	N	N	N	N	N	Y	N	N	N	N	N
Acute abdominal pain (Surgery)	18-80	F/M	Y	N	N	N	N	N	N	N	Y	N	N	N	Y	N	N	N	N	N	Y	N	N	N	N	N
Acute pancreatitis	18-80	F/M	Y	N	N	N	N	N	N	N	N	Y	N	N	N	N	Y	N	N	N	N	N	N	Y	N	N
Epigastric pain	18-80	F/M	Y	N	N	N	N	N	N	N	N	N	N	N	N	N	Y	N	N	N	Y	N	N	N	N	N
Mesenteric ischemia imaging	65-90	F/M	Y	N	N	N	N	N	N	Y	N	N	N	N	Y	N	N	N	N	N	Y	N	N	N	N	N
Left lower quadrant pain	50-90	F/M	N	N	Y	N	N	N	Y	N	N	N	N	N	N	N	N	Y	N	N	Y	N	N	N	N	N
Upper gastrointestinal bleeding	18-80	F/M	N	N	Y	N	N	N	N	N	N	N	Y	N	N	N	Y	N	N	N	N	Y	N	N	N	N
Right lower quadrant pain	18-80	F/M	Y	N	N	N	N	N	N	N	N	N	N	N	N	N	Y	N	N	N	N	Y	N	N	N	N
Right lower quadrant pain (Pregnant)	18-80	F/M	Y	N	N	N	N	N	N	N	N	N	N	Y	N	N	N	N	Y	N	Y	N	N	N	N	N
Right upper quadrant pain	18-35	F/M	N	N	N	Y	N	N	N	N	N	N	N	Y	N	N	N	N	Y	N	N	N	N	N	N	Y
Small bowel obstruction	18-80	F/M	Y	N	N	N	N	N	Y	N	N	N	N	N	N	N	N	N	N	Y	N	N	N	N	Y	N
Uterine bleeding	18-80	F	N	N	N	N	N	Y	N	N	Y	Y	Y	Y	Y	N	N	N	N	N	Y	N	N	N	N	N

Acute pelvic pain	18-80	F	N	N	N	N	Y	N	Y	N	N	N	N	N	Y	N	N	N	N	N	N	Y	N	N	N	N
Acute pelvic pain (Postmenopausal)	15-55	F	N	N	N	N	N	N	Y	N	N	N	N	N	Y	N	N	N	N	N	N	N	N	N	N	Y
Flank pain	55-90	F/M	N	N	N	N	N	N	Y	N	N	N	N	N	Y	N	N	N	N	N	Y	N	N	N	N	N
Flank pain (Pregnant patient)	18-80	F	N	Y	N	N	N	N	Y	N	N	N	N	N	N	Y	N	N	N	N	N	N	Y	N	N	N
Microscopic hematuria	18-80	F/M	Y	N	N	N	N	N	N	N	N	N	N	Y	N	Y	N	N	N	N	N	N	Y	N	N	N
Macroscopic hematuria	18-80	F/M	N	Y	N	N	N	N	Y	N	N	N	N	N	Y	N	N	N	N	N	N	N	Y	N	N	N

PE, physical exam

Step 2: Creating the Question Template (Model)

At this stage, the goal was to develop a question template in a format suitable for question generation. Initially, a sample question was created, which served as the foundation for designing the question template. The template was structured into four parts: presenting symptom, patient history, physical examination findings, and laboratory results. To enhance question variety, each section of the template was written in two different ways, allowing for the random selection of sentences for each part, thereby diversifying the question stems (Table 3).

Table 3. Template-Based Question Model

CONTENT OF THE QUESTIONS TO BE GENERATED: [Presenting Symptom] [History] [Physical Examination Findings] [Laboratory Findings]		
	Template 1	**Template 2**
Symptom	A <AGE> -year-old <GENDER> patient presents to the emergency department with <SYMPTOM1> and <SYMPTOM2>.	A <AGE>-year-old <GENDER> patient arrives at the emergency department with complaints of <SYMPTOM1> and <SYMPTOM2>.
History	In the patient's history, it is stated that <HISTORY>.	The patient's history reveals <HISTORY>.
Physical Exam	Physical examination reveals <PE>.	During the physical examination, <PE> is identified.
Laboratory	In laboratory findings, <LAB> is observed. Which of the following imaging modality is most appropriate to request as a priority?	Based on the results, <LAB> is observed. Among the options provided, which imaging should be prioritized over the others?

During the study design phase, questions were planned to be generated for 18 topics, with answer options established for each diagnosis. The American College of Radiology (ACR) guidelines on appropriateness criteria for radiologic investigations were used to determine these options. These guidelines specify the recommended and non-recommended tests for each diagnosis. One recommended test was randomly selected as the correct answer, while four non-recommended tests were randomly chosen as distractors.

Step 3: Question Generation Using Software

At this stage, a software application was created to generate questions based on the template developed in the first two steps. The authors developed a Python script specifically for this study to automate the process of assigning variables from the cognitive model to the question template, generating randomized question stems and distractors. The script operates in three main steps:

- Inputs: Cognitive model variables (e.g., symptoms, history, lab findings) are imported into the script.
- Processing: Variables are randomly selected and inserted into the predefined question templates.
- Outputs: Complete questions are generated with one correct answer and four distractors based on predefined appropriateness criteria. The generated questions are output in a structured format for evaluation.

Generation of Non-Template-Based Questions

In non-template-based question generation, no fixed format or structure is used. Instead, questions are generated using AI tools without adhering to specific rules or templates. In this study, ChatGPT-4o, a widely

used chatbot, was employed for question generation (the questions were generated in August 2024). Research has shown that the quality of generated questions can be influenced by the prompts used (14). The detailed prompt created by Yavuz (22), which incorporated the variables "topic" and "difficulty level," was utilized. By setting the difficulty level to the fifth-year medical undergraduate level, 18 questions were generated for each of the 18 topics.

Table 4. Sample question generated by ChatGPT

A 60-year-old male patient presents to the emergency department with complaints of hematemesis. He has a history of peptic ulcer disease. Vital signs reveal a blood pressure of 90/60 mmHg and a pulse rate of 110 beats per minute. Emergency endoscopy identifies a non-variceal upper gastrointestinal arterial bleeding source. His hemoglobin level is measured at 8 g/dL, and the patient receives a blood transfusion.
Which of the following is the most appropriate next imaging method for this patient?
A. Abdominal ultrasound B. Non-contrast helical CT C. Angiography D. Magnetic resonance imaging (MRI) E. Scintigraphy

Selection of Questions for Evaluation

In addition to the questions generated by both methods, questions were selected from fifth-year radiology exams previously created by faculty members. These questions were chosen from those used in internship exams for fifth-year students. Working retrospectively, the first questions encountered on the topic, with psychometric data deemed "usable" by our institution, were included in the study. In this study, a total of 7,260 questions were generated using the template-based method, of which six were selected for evaluation. Similarly, six questions were generated by ChatGPT using a non-template-based approach, and six faculty-prepared questions were included. In total, 18 questions were evaluated.

Evaluation Form

An evaluation form was created using Google Forms, comprising the 18 selected questions. The questions were randomly ordered, and the correct answers were highlighted in bold. The evaluation criteria included seven parameters: clarity of the question text, clinical appropriateness, the presence of a single correct answer, sufficiency of information to determine the correct answer, quality of distractors, whether the question was challenging even for experts, and whether the difficulty level was appropriate for medical students. Each parameter was rated on a three-point Likert scale ('agree,' 'neutral,' 'disagree'). These criteria were adapted from existing literature and finalized through consensus among the authors (12, 23). The questions, formatted as an evaluation form, were distributed to resident physicians, specialist physicians, and faculty members in the radiology department who had been practicing for at least two years. In total, 14 radiologists evaluated the questions. The evaluators assessed the questions without knowledge of their origin.

3. Results

A total of 7,260 questions were generated using the template-based method, from which six questions were selected, one for each of the six predefined topics. For the non-template-based method, six questions on the selected topics were generated by ChatGPT, and an additional six questions were chosen from past exam questions prepared by faculty members on the same topics. The responses from the evaluators are presented in Table 5.

When evaluators were asked whether the questions measured clinical reasoning or rote knowledge, they indicated that template-based questions measured clinical reasoning at a rate of 66/84 (78.57%), faculty-generated questions at 69/84 (82.14%), and ChatGPT-generated questions at 71/84 (84.52%). The average total and percentage of "Yes" responses for each evaluation question by question source were as follows: for "The question text is clear," 11.83 (84.52%) / 11.67 (83.33%) / 13 (92.86%) for template-based, faculty-written, and ChatGPT questions, respectively; for "The question is clinically appropriate," 8.33 (59.52%) / 9.17 (65.48%) / 11.5 (82.14%); for "The question has only one correct answer," 7.17 (51.19%) / 7.67 (54.76%) / 9 (64.29%); for "The information provided is sufficient to find the correct answer," 8 (57.14%) / 9.33 (66.67%) / 10.17 (72.62%); for "The distractors are reasonable," 6.67 (47.62%) / 8.67 (61.90%) / 10.67 (76.19%); for "The question was challenging even for experts," 2.83 (7.14%) / 1 (7.14%) / 0.5 (3.57%); and for "The question difficulty was appropriate for medical students," 6.17 (44.05%) / 6.5 (46.43%) / 8.17 (58.33%).

Table 5. Assesment of questions

		Template-Based Item Generation		**Faculty-Prepared Questions**		**ChatGPT**	
		n (84)	%	n (84)	%	n (84)	%
The question text is clear.	Yes	71	84.52 %	70	83.33 %	78	92.86 %
	No	1	1.19 %	4	4.76 %	0	0
	Neutral	12	14.29 %	10	11.90 %	6	7.14 %
The question is clinically appropriate.	Yes	50	59.52 %	55	65.48 %	69	82.14 %
	No	6	7.14 %	11	13.10 %	0	0
	Neutral	27	32.14 %	18	21.43 %	15	17.86 %
The question has only one correct answer.	Yes	43	51.19 %	46	54.76 %	54	64.29 %
	No	10	11.90 %	16	19.05 %	4	4. 76%
	Neutral	31	36.90 %	22	26.19 %	26	30.95%
The information provided is sufficient to find the correct answer.	Yes	48	57.14 %	56	66.67 %	61	72.62%
	No	7	8.33 %	9	10.71 %	2	2.38 %
	Neutral	29	34.52 %	19	22.62 %	21	25%
The distractors are reasonable.	Yes	40	47.62 %	52	61.90 %	64	76.19 %
	No	13	15.48 %	15	17.86 %	4	4.76 %
	Neutral	31	36.90 %	17	20.24 %	16	19.05 %
The question was challenging even for experts.	Yes	6	7.14 %	6	7.14 %	3	3.57 %
	No	54	64.29 %	52	61.90 %	63	75%
	Neutral	23	27.38 %	26	30.95 %	18	21.43 %
The question difficulty was appropriate for medical students.	Yes	37	44.05 %	39	46.43 %	49	58.33 %
	No	16	19.05 %	16	19.05 %	6	7.14 %
	Neutral	31	36.90 %	29	34.52 %	29	34.52 %

4. Discussion

Case-based learning plays a significant role in both practical and theoretical learning in medical education. Through case scenarios, students are expected to learn diagnoses, differential diagnoses, and adopt clinical reasoning processes as they work toward diagnoses. In assessing clinical reasoning, these case scenarios are used as tools in both theoretical and practical exams. Considering the high course load and frequency of exams in medical education, multiple-choice questions (MCQs) are the most commonly used assessment method (24). However, generating MCQs, especially those that contain case scenarios and

measure clinical reasoning skills, is challenging. This study evaluates automatic question generation techniques aimed at assessing clinical reasoning.

The study's 18 sample questions were evaluated by radiologists with academic experience. When asked whether the generated questions assessed clinical reasoning skills, it was found that ChatGPT, faculty members, and template-based techniques measured clinical reasoning at rates of 84.52%, 82.14%, and 78.57%, respectively. If faculty-prepared questions are taken as a reference, it can be inferred that ChatGPT generates questions more effectively for clinical reasoning, while template-based questions measure clinical reasoning at a comparable rate. This suggests that both methods can produce questions that are practically useful for assessing clinical reasoning. Studies in the literature also indicate that both methods can generate questions capable of assessing clinical reasoning in different languages (11,13).

Among the evaluation parameters, question clarity, clinical appropriateness, and sufficiency of information for finding the correct answer were rated highly for both question generation techniques. Regarding question difficulty, few "Yes" responses were recorded for the statement that the questions were challenging even for experts across all three methods. However, noteworthy findings include relatively lower "Yes" responses for the statements "The distractors are reasonable" and "The question has only one correct answer" for template-based, faculty-prepared, and ChatGPT-generated questions (6.67 [47.62%] / 8.67 [61.90%] / 10.67 [76.19%]; 7.17 [51.19%] / 7.67 [54.76%] / 9 [64.29%]). This may be because, although the question stem inquired about the most likely investigation, evaluators sought a definitive answer, potentially influenced by the inherent uncertainty and interpretive nature of some scenarios in medicine. If faculty-prepared questions are used as a benchmark, similar results were observed with the other two methods, and ChatGPT questions performed even better.

A key contribution of this study to the literature is its comparison of all three methods by generating questions on the topic of radiologic investigations for abdominal emergencies and evaluating them with the same group of evaluators. While studies on template-based question generation and ChatGPT have been conducted (25, 26), these studies typically assessed only these techniques in isolation or compared them to human-made questions (12, 19, 26, 27). Notably, in this study, ChatGPT and template-based questions were found to be non-inferior to faculty-prepared questions in each evaluation parameter. Although prior studies indicate that chatbot-generated questions may have inaccuracies or limitations, it can be assumed that question quality has improved as this field advances (28).

This study has some limitations. First, only 18 questions were evaluated, with six questions from each method. Second, while the evaluator group comprised experienced radiologists in academia, evaluations conducted by a group with more expertise in question writing could provide further insights. This study specifically targeted the assessment of clinical reasoning skills and selected a single topic for this purpose. Different topics may yield varied results. Additionally, evaluations assumed the faculty-prepared questions were appropriately written, although they may also have deficiencies. ChatGPT-4o (questions were generated in August 2024), one of the most advanced chatbots currently available, was used in the non-template-based method; however, other chatbots might produce different or improved results. Although a detailed prompt defined in the literature was employed for ChatGPT-generated questions, it should be noted that prompt quality significantly impacts the output of chatbots (29). A significant limitation of this study is the restricted scope of the seven quality criteria used to evaluate the questions. While these criteria provide a structured framework for assessment, relying solely on them can be reductionistic when evaluating the overall quality of MCQs. The complexity of question design, especially for assessing clinical reasoning skills, extends beyond these parameters. Readers are therefore advised to interpret the results with caution, acknowledging that the evaluation does not capture the full spectrum of factors influencing MCQ quality. Future studies should incorporate a broader range of evaluation metrics to provide a more comprehensive understanding of question quality. Additionally, future studies should focus on evaluating

automatically generated MCQs using detailed psychometric parameters such as difficulty index, discrimination index, and item-total correlation.

This study specifically focused on radiologic investigations for abdominal emergencies to assess clinical reasoning. While the findings provide valuable insights, they may not be directly generalizable to other medical topics or specialties. Future studies should examine diverse topics to validate the effectiveness of these automatic item generation methods in broader contexts.

5. Conclusions

- ChatGPT and template-based question generation techniques are effective tools for creating multiple-choice questions that assess clinical reasoning skills in medical education, with ChatGPT showing slightly higher performance in evaluation metrics.
- Case-based learning remains fundamental in medical education, and automated item generation can help meet the high demand for quality questions that support diagnostic and clinical reasoning skills.
- Faculty-prepared questions serve as a valuable benchmark, but findings suggest that ChatGPT-generated questions may sometimes exceed faculty-generated questions in assessing clinical reasoning effectively.
- Automated item generation techniques, especially ChatGPT, hold potential to save time for educators in question development, though careful prompt design and evaluation are essential for maximizing question quality.

Funding: There has been no funding

Declaration of conflict of interest: The authors declare that they have no conflict of interest

Author Contributions : Emre Emekli conceptualized the idea, methodology,data validation, data curation, writing and preparation of the research design stages, review and final editing of the document. Betül Nalan Karahan participated data validation, review and final editing of the document.

References

1. Bonilauri Ferreira AP, Ferreira RF, Rajgor D, Shah J, Menezes A, Pietrobon R. Clinical reasoning in the real world is mediated by bounded rationality: implications for diagnostic clinical practice guidelines. *PLOS ONE.* 2010; 5(4): e10265. https://doi.org/10.1371/journal.pone.0010265
2. Pelaccia T, Plotnick LH, Audétat MC, Nendaz M, Lubarsky S, Thomas A, Young M, Dory VA. A Scoping Review of Physicians' Clinical Reasoning in Emergency Departments. *Annals of Emergency Medicine.* 2020;75(2): 206-217. https://doi.org/10.1016/j.annemergmed.2019.06.023
3. Gruppen LD. Clinical Reasoning: Defining It, Teaching It, Assessing It, Studying It. *West Journal of Emergency Medicine.* 2017; 18(1): 4-7. https://doi.org/10.5811/westjem.2016.11.33191
4. Simmons B. Clinical reasoning: concept analysis. *Journal of Advanced Nursing*. 2010;66(5): 1151-1158. https://doi.org/10.1111/j.1365-2648.2010.05262.x
5. Schmidt HG, Mamede S. How to improve the teaching of clinical reasoning: a narrative review and a proposal. *Medical Education.* 2015;49(10): 961-973. https://doi.org/10.1111/medu.12775
6. Miller GE. The assessment of clinical skills/competence/performance. *Academic Medicine.* 1990; 65(9 Suppl): 63-67. https://doi.org/10.1097/00001888-199009000-00045
7. Brentnall J, Thackray D, Judd B. Evaluating the Clinical Reasoning of Student Health Professionals in Placement and Simulation Settings: A Systematic Review. *Int J Environ Res Public Health.* 2022;19(2). https://doi.org/10.3390/ijerph19020936
8. Modi JN, Anshu Gupta P, Singh T. Teaching and Assessing Clinical Reasoning Skills. *Indian Pediatrics.* 2015;52(9): 787-794. https://doi.org/10.1007/s13312-015-0718-7
9. Schuwirth LW, van der Vleuten CP. Different written assessment methods: what can be said about their strengths and weaknesses? *Medical Education.* 2004;38(9): 974-979. https://doi.org/10.1111/j.1365-2929.2004.01916.x

10. Wrigley W, van der Vleuten CP, Freeman A, Muijtjens A. A systemic framework for the progress test: strengths, constraints and issues: AMEE Guide No. 71. *Medical Teacher. 2012; 34*(9): 683-697. https://doi.org/10.3109/0142159x.2012.704437
11. Gierl MJ, Lai H, Tanygin V. *Advanced methods in automatic item generation*: Routledge, 2021.
12. Cheung BHH, Lau GKK, Wong GTC, Lee EYP, Kulkarni D, Seow CS, Wong R, Co MT. 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(8): e0290691. https://doi.org/10.1371/journal.pone.0290691
13. Kıyak YS, Coşkun Ö, Budakoğlu I, Uluoğlu C. ChatGPT for generating multiple-choice questions: Evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam. *Eurapean Journal of Clinical Pharmacoogyl.* 2024;80(5): 729-735. https://doi.org/10.1007/s00228-024-03649-x
14. Kıyak YS, Emekli E. ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: a literature review. *Postgraduate Medical Journal*. 2024;6: qgae065 https://doi.org/10.1093/postmj/qgae065
15. Williamson SM, Prybutok V. The Era of Artificial Intelligence Deception: Unraveling the Complexities of False Realities and Emerging Threats of Misinformation. *Information.* 2024;15(6): 299 https://doi.org/10.3390/info15060299
16. Pugh D, De Champlain A, Gierl M, Lai H, Touchie C. Using cognitive models to develop quality multiple-choice questions. *Medical Teacher*. 2016;38(8): 838-843. https://doi.org/10.3109/0142159x.2016.1150989
17. Ulusal Cep-2020 UCG, Ulusal Cep-2020 UYVYCG, Ulusal Cep-2020 DSBBCG. Medical Faculty - National Core Curriculum 2020. *Tıp Eğitimi Dünyası* 2020; 19: 141-146. https://doi.org/10.25282/ted.716873
18. American College of Radiology. *Appropriateness Criteria* Available online: https://www.acr.org/Clinical-Resources/ACR-Appropriateness-Criteria (15.08.2024)
19. Gierl MJ, Lai H, Turner SR. Using automatic item generation to create multiple-choice test items. *Medical Education.* 2012;46(8): 757-765. https://doi.org/10.1111/j.1365-2923.2012.04289.x
20. Gierl MJ, Lai H. Evaluating the quality of medical multiple-choice items created with automated processes. *Medical Education.* 2013;47(7): 726-733. https://doi.org/10.1111/medu.12202
21. Gierl MJ, Lai H. Using Automated Processes to Generate Test Items And Their Associated Solutions and Rationales to Support Formative Feedback. *IxD&A.* 2015;25: 9-20. https://doi.org/10.1177/0146621617726788
22. Kıyak YS. A ChatGPT Prompt for Writing Case-Based Multiple-ChoiceQuestions. *Revista Española de Educación Médica.* 2023; 4(3). https://doi.org/10.6018/edumed.587451
23. Pugh D, De Champlain A, Gierl M, Lai H, Touchie C. Can automated item generation be used to develop high quality MCQs that assess application of knowledge? *Research and Practice in Technology Enhanced Learning.* 2020;15: 11-13. https://doi.org/10.1186/s41039-020-00134-8
24. Cansever Z, Acemoğlu H, Avşar Ü, Hoşoğlu S. Tıp fakültesindeki çoktan seçmeli sınav sorularının değerlendirilmesi. *Tıp Eğitimi Dünyası.* 2016;14(44): 44-55. https://doi.org/10.25282/ted.228764
25. Kıyak YS, Emekli E. A Prompt for Generating Script Concordance Test Using ChatGPT, Claude, and Llama Large Language Model Chatbots. *Revista Española de Educación Médica.* 2024;5(3). https://doi.org/10.6018/edumed.612381
26. Kurdi G, Leo J, Parsia B, Sattler U, Al-Emari S. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education.* 2020; 30: 121-204. https://doi.org/10.1007/s40593-019-00186-y
27. Kıyak YS, Budakoğlu Iİ, Coşkun Ö, Koyun E. The first automatic item generation in Turkish for assessment of clinical reasoning in medical education. *Tıp Eğitimi Dünyası.* 2023; 22(66): 72-90. https://doi.org/10.25282/ted.1225814
28. Ngo A, Gupta S, Perrine O, Reddy R, Ershadi S, Remick D. ChatGPT 3.5 fails to write appropriate multiple choice practice exam questions. *Academic Pathology*. 2024;11(1): 100099. https://doi.org/10.1016/j.acpath.2023.100099
29. Kıyak YS. Beginner-Level Tips for Medical Educators: Guidance on Selection, Prompt Engineering, and the Use of Artificial Intelligence Chatbots. *Medical Science Educator*. 2024; 1-6. https://doi.org/10.1007/s40670-024-02146-1