Artificial intelligence in radiology examinations: a psychometric comparison of question generation methods

Article *in* Diagnostic and interventional radiology (Ankara, Turkey) · July 2025

DOI: 10.4274/dir.2025.253407

2 authors, including:

Emre Emekli
Eskişehir Osmangazi University
64 PUBLICATIONS **315** CITATIONS

SEE PROFILE

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.253407

ARTIFICIAL INTELLIGENCE AND INFORMATICS

ORIGINAL ARTICLE

Artificial intelligence in radiology examinations: a psychometric comparison of question generation methods

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

[1]Eskişehir Osmangazi University Faculty of Medicine, Department of Radiology, Eskişehir, Türkiye

[2]Eskişehir Osmangazi University, Translational Medicine Application and Research Center, Eskişehir, Türkiye

PURPOSE

This study aimed to evaluate the usability of artificial intelligence (AI)-based question generation methods–Chat Generative Pre-trained Transformer (ChatGPT)-4o (a non-template-based large language model) and a template-based automatic item generation (AIG) method–in the context of radiology education. The primary objective was to compare the psychometric properties, perceived quality, and educational applicability of generated multiple-choice questions (MCQs) with those written by a faculty member.

METHODS

Fifth-year medical students who participated in the radiology clerkship at Eskişehir Osmangazi University were invited to take a voluntary 15-question examination covering musculoskeletal and rheumatologic imaging. The examination included five MCQs from each of three sources: a radiologist educator, ChatGPT-4o, and the template-based AIG method. Student responses were evaluated in terms of difficulty and discrimination indices. Following the examination, students rated each question using a Likert scale based on clarity, difficulty, plausibility of distractors, and alignment with learning goals. Correlations between students' examination performance and their theoretical/practical radiology grades were analyzed using Pearson's correlation method.

RESULTS

A total of 115 students participated. Faculty-written questions had the highest mean correct response rate (2.91 ± 1.34), followed by template-based AIG (2.32 ± 1.66) and ChatGPT-4o (2.3 ± 1.14) questions ($P < 0.001$). The mean difficulty index was 0.58 for faculty, and 0.46 for both template-based AIG and ChatGPT-4o. Discrimination indices were acceptable (≥0.2) or very good (≥0.4) for template-based AIG questions. In contrast, four of the ChatGPT-generated questions were acceptable, and three were very good. Student evaluations of questions and the overall examination were favorable, particularly regarding question clarity and content alignment. Examination scores showed a weak correlation with practical examination performance ($P = 0.041$), but not with theoretical grades ($P = 0.652$).

CONCLUSION

Both the ChatGPT-4o and template-based AIG methods produced MCQs with acceptable psychometric properties. While faculty-written questions were most effective overall, AI-generated questions–especially those from the template-based AIG method–showed strong potential for use in radiology education. However, the small number of items per method and the single-institution context limit the robustness and generalizability of the findings. These results should be regarded as exploratory, and further validation in larger, multicenter studies is required.

CLINICAL SIGNIFICANCE

AI-based question generation may potentially support educators by enhancing efficiency and consistency in assessment item creation. These methods may complement traditional approaches to help scale up high-quality MCQ development in medical education, particularly in resource-limited settings; however, they should be applied with caution and expert oversight until further evidence is available, especially given the preliminary nature of the current findings.

KEYWORDS

Artificial intelligence, radiology education, multiple-choice questions, automatic item generation, clinical reasoning assessment

Corresponding author: Emre Emekli

E-mail: emreemekli90@gmail.com

Received 14 April 2025; revision requested 12 May 2025; last revision received 10 June 2025; accepted 14 June 2025.

Epub: 21.07.2025

Publication date:

DOI: 10.4274/dir.253407

You may cite this article as: Emekli E, Karahan BN. Artificial intelligence in radiology examinations: psychometric comparison of question generation methods. v. *Diagn Interv Radiol.* 21 July 2025 DOI: 10.4274/dir.2025.253407 [Epub Ahead of Print].

In medical education, structured examination techniques are of great importance in objectively assessing students' knowledge, skills, and clinical decision-making competencies. In this regard, multiple-choice questions (MCQs) are widely used to measure theoretical knowledge. These questions also contribute to examination security and the standardization of the assessment process.[1,2] However, generating high-quality MCQs requires attention to criteria such as appropriate difficulty level, discrimination index, distractor quality, and measurement validity, which demand significant time and expertise. Considering the intensity of educational programs, this creates a need for more efficient and sustainable methods in MCQ generation.[3]

In recent years, the rapid development of artificial intelligence (AI) has introduced potential alternatives to meet this need. With systems developed using large language models (LLMs), it is possible to generate a large number of questions in a short period. LLMs stand out as non-template-based methods for direct question generation and can create a high volume of case-based scenarios with minimal effort.[4] In addition, template-based automatic item generation (AIG) techniques provide more structured and instructor-controllable outputs.[5] Although both methods offer their advantages, varying levels of expert oversight are required in terms of quality, accuracy, and contextual appropriateness.[6,7] The literature suggests that questions created using these methods are psychometrically valid; however, direct comparative studies remain limited. This study distinguishes itself by directly comparing Chat Generative Pre-trained Transformer (ChatGPT)-4o and template-based AIG with faculty-written questions in the same real-life examination, while incorporating both objective psychometric analyses and subjective student evaluations in the context of radiology education.

Main points

- Artificial intelligence (AI)-generated questions using both Chat Generative Pre-trained Transformer (ChatGPT)-4o and template-based automatic item generation (AIG) methods had acceptable or very good discrimination indices, highlighting their consistency in performance difficulty and discrimination indices, making them suitable for use in radiology education.
- Faculty-authored questions outperformed generated ones in terms of student success, although template-based AIG questions showed the highest consistency in psychometric performance.
- Students rated the generated questions favorably, indicating that these methods can produce clear, plausible, and educationally aligned content.
- AI tools, such as ChatGPT-4o and template-based AIG systems, can alleviate faculty workload and support scalable question generation with proper validation and oversight.
- Combining AI-generated questions with expert oversight may improve the quality and efficiency of assessment development.

In this study, the usability of questions generated by ChatGPT-4o (a non-template-based method) and questions developed through template-based AIG was examined within the context of assessment and evaluation. Specifically within radiology education, both AIG methods were compared with questions written by a faculty member. The evaluation was conducted based on students' examination performances, the psychometric properties of the questions, and perceived quality levels, aiming to provide empirical data on the applicability of AI-based tools in medical education.

Methods

Prior to the study, approval was obtained from the Non-Interventional Clinical Research Ethics Committee of Eskişehir Osmangazi University (decision number: 34, date: October 22, 2024). Participation in the study was voluntary, and consent was obtained from students who were to take the examination.

Participants

The study was conducted with 5th-year medical students enrolled in the Radiology internship at Eskişehir Osmangazi University Faculty of Medicine during the 2024–2025 academic year. A total of 163 students were invited to participate in the study.

Examination topics

Within the scope of the radiology internship, students were taught musculoskeletal radiology topics prepared in accordance with the European Society of Radiology[8] and the National Core Education Program[9], including "Radiology of the Skeletal System and Soft Tissue Tumors," "Radiology in Metabolic and Endocrine System Diseases," and "Radiology in Rheumatologic Diseases." In addition to the routine radiology internship covering these subjects, an examination consisting of 15 MCQs was administered. Participation in this examination was voluntary and had no impact on students' internship grades. The examination was scheduled 7–10 days before the end of each internship rotation to ensure that students had received sufficient radiology training.

Question generation

Of the 15 questions, 5 were created by a faculty member, 5 by ChatGPT-4o, and 5 by the template-based AIG technique. The sources of the questions were not disclosed to the participants prior to the examination. The questions prepared by the faculty member were written by a physician with nine years of radiology experience and a PhD in medical education.

The ChatGPT-4o questions were prepared in August 2024. For each lecture topic, course materials (slide presentations used in lectures and lecture notes provided to students as resources) were shared with ChatGPT-4o. Subsequently, using prompts described in the literature, 5 questions were generated for each lecture topic.[4,10] Each of the 5 questions used in the study was generated in a separate ChatGPT conversation to avoid memory retention effects and ensure independence between items. For each session, only one prompt was submitted, and the first complete set of 5 questions generated by ChatGPT-4o was directly included in the study without further selection or filtering. No modifications, edits, or refinements were made to the wording, content, or structure of these questions after generation.

In the template-based AIG method, three stages for question generation are defined in the literature. In the first stage, topic headings were identified.[11] For this study, it was planned to generate questions using the template-based AIG method within the lecture content "Radiology in Rheumatologic Diseases." Accordingly, it was decided to generate questions related to rheumatoid arthritis (RA), gout, psoriatic arthritis (PA), ankylosing spondylitis (AS), and osteoarthritis (OA). For these diseases, information variables were defined as patient age, sex, symptoms, medical history, and radiographic images. For each diagnosis, suitable age ranges (OA: 55–85 years; AS: 22–45; PA: 30–58; RA: 40–80; gout: 35–65),

appropriate gender (OA, PA: female/male; AS, gout: male; RA: female), and potentially relevant radiographs were specified. Subsequently, appropriate symptoms and histories were written for each diagnosis. In the second stage, a question template was created. Two types of question templates were developed to ensure variety (Table 1). In this phase, the correct answer option was not directly tied to each diagnosis to assess clinical reasoning skills. Radiographic findings for each diagnosis were defined (e.g., OA: asymmetric joint space narrowing, subchondral sclerosis, osteophyte formation; AS: bamboo spine, shiny corner sign, vertebral squaring). One diagnosis was randomly assigned as the correct answer, and distractors were randomly selected from among the other diagnoses. Finally, questions were generated using Python-based software previously validated for use in item generation.[12] A total of 322 questions were generated using the template-based AIG method (5 for each diagnosis). One question was randomly selected from each diagnostic category to ensure diversity across clinical conditions, resulting in 5 questions used in the examination. These 5 questions were reviewed by two faculty members for accuracy, clarity, and appropriateness before inclusion.

The questions written by the faculty member were reviewed by a radiology research assistant for factual and grammatical errors before being included in the examination. Of the 5 questions generated by ChatGPT-4o and the template-based AIG method, 1 randomly selected question for each diagnosis (total of 5) was evaluated by two faculty members in terms of clarity, clinical appropriateness, presence of a single correct answer, factual accuracy, and distractor quality[13]; all were deemed appropriate and included in the study. The number of questions per method was limited to 5 to avoid placing an excessive testing burden on students during their radiology internship. Given that participation in the study was voluntary and that the examination was administered in addition to the standard curriculum assessments, it was deemed ethically and practically necessary to keep the examination length manageable. The primary aim was to explore the feasibility of applying AI-based question generation in a real-world educational setting. The study methodology is illustrated in the flowchart presented in Figure 1.

Question evaluation

After the examination, anonymous examination and question evaluation forms were distributed to students. Each question was presented individually, and students were asked to evaluate them using a previously developed form with a 5-point Likert scale ("the question text is clear"; "the question is of appropriate difficulty"; "the question has only one correct answer"; "the information provided is sufficient to find the correct answer"; "the distractors are logical"). Although students may not possess expert-level judgment on technical item quality, their feedback is valuable in assessing the clarity, plausibility, and perceived appropriateness of MCQs. Since the examination was designed for undergraduate education, student perceptions reflect real-world usability. Moreover, the Likert items used had been adapted from previous studies and paired with objective psychometric analyses to provide a more comprehensive evaluation.

Statistical analysis

The theoretical examination consisted of 32 MCQs, with 1 question corresponding to each hour of radiology instruction. The practical examination included 10 radiological cases, each accompanied by three components: identification of the imaging modality, description of radiological findings, and determination of the most likely diagnosis. Each component was scored separately to calculate the total practical score. Each student's scores from the radiology theoretical and practical examinations were recorded. The correlation between these scores and the number of correct answers in the study examination was assessed using Pearson's correlation method. Item discrimination was calculated using both the traditional 27% upper–lower group method and the Pearson item–rest correlation method. The latter was computed as the correlation between individual item scores and total test scores with the target item removed. This dual approach aligns with current psychometric recommendations and allows for more robust interpretation of item quality. Discrimination index thresholds were interpreted based on current guidelines in medical education literature, with values ≥0.30 considered very good, 0.20–0.29 acceptable, and <0.20 poor[14] students' perceptions of the examination were collected using a 5-point Likert scale. Due to the limited number of items (n = 5 per group), internal consistency measures, such as Cronbach's alpha, were not computed, as such estimates are considered statistically unreliable with small item sets.

Results

A total of 115 students (70.6%) agreed to participate in the study. The mean number of correct answers out of 15 questions in the examination was 7.53 ± 3.21. When evaluated according to the source of the questions, the mean number of correct answers was 2.91 ± 1.34 for the faculty-written questions, 2.3 ± 1.14 for the ChatGPT questions, and 2.32 ± 1.66 for the template-based AIG questions, with a statistically significant difference ($P < 0.001$). The faculty-written questions were answered correctly more frequently than those generated by ChatGPT-4o and the template-based AIG method ($P = 0.010$; $P = 0.001$).

The average difficulty indices were calculated as 0.58 for the faculty-written questions, 0.46 for the template-based AIG questions, and 0.46 for the ChatGPT-4o questions. Regarding the discrimination index (27% upper–lower group), 5 of the faculty-written questions were classified as acceptable and 3 as very good; 4 of the ChatGPT questions were acceptable and 3 were very good; all of the template-based AIG questions were found to be appropriate (5 acceptable, 5 very good). In addition, item–rest correlation (Pearson) analyses were performed to further assess discrimination power. According to this method, 4 of the faculty-written questions, 3 of the ChatGPT-generated questions, and 5 of the template-based AIG questions showed acceptable or higher item–total correlations ($r \geq 0.20$), supporting the results obtained via the traditional method, providing preliminary support for the psychometric performance of the items. The difficulty and discrimination indices of the questions by source are presented in Table 2.

Students' evaluations of the questions and the examination are presented in Tables 3 and 4. Although the scores obtained from the prepared examination did not correlate with the theoretical examination scores ($P = 0.652$), a weak correlation was found with the practical examination scores ($P = 0.041$). Given the limited number of items per method, all psychometric results should be interpreted with caution and considered exploratory.

Table 1. Template-based question model

CONTENT TO BE GENERATED: [Reason for Admission] [History] [Radiography] [Question Stem]				
	Template 1		Template 2	
Reason for admission	<SYMPTOM> complaint brought the <AGE>-year-old <GENDER> patient to the outpatient clinic		The <AGE>-year-old <GENDER> patient presents to the outpatient clinic with a <SYMPTOM> complaint	
History	In the patient's history, <HISTORY> is noted		The history reveals <HISTORY>	
Radiography	A <RADIOGRAPHY> is requested		A <RADIOGRAPHY> is performed for this patient	
Question stem	Which of the following findings is more likely to be seen in this patient compared with the others?		Which of the following radiographic findings would you most expect to see in this patient?	
Diagnosis	Symptoms	History	Radiography	Options
Osteoarthritis	• Pain in the right knee and restricted movement at both knees • Pain and inability to bend the left knee	• Pain worsens in the late afternoon • Pain worsens with physical activity	Knee /shoulder	Asymmetric joint space narrowing, subchondral sclerosis, osteophyte formation, Heberden's nodes, Bouchard's nodes
Ankylosing spondylitis	• Lower back pain • Morning stiffness	• Morning stiffness lasting for about an hour after waking up • Lower back stiffness decreases with physical activity	Vertebral/sacroiliac joints	Bamboo spine, vertebral squaring, shiny corner sign, syndesmophytes, ossification of the anterior longitudinal ligament
Psoriatic arthritis	• Red eyes and eye itching • Rash on extensor surfaces of extremities	• Morning stiffness lasting for about an hour after waking up • Pitting in the nails with joint pain	Hand	Arthritis mutilans, dactylitis, feather-like periostitis, pencil-in-cup deformity
Rheumatoid arthritis	• Swelling in fingers of both hands • Swelling and pain in both wrists	• Morning stiffness lasting for about an hour after waking up • Pain decreases with exercise	Hand/vertebral	Marginal erosions, rheumatoid nodules, juxta-articular osteoporosis, swan-neck deformity, Boutonnière deformity
Gout	• Severe swelling in the toe • Redness and increased warmth in the toe	• Symptoms started after alcohol consumption last night • Symptoms started after eating a rich meal	Elbow/foot	Rat-bite erosions, eccentric erosions, soft tissue tophi, Martel's sign

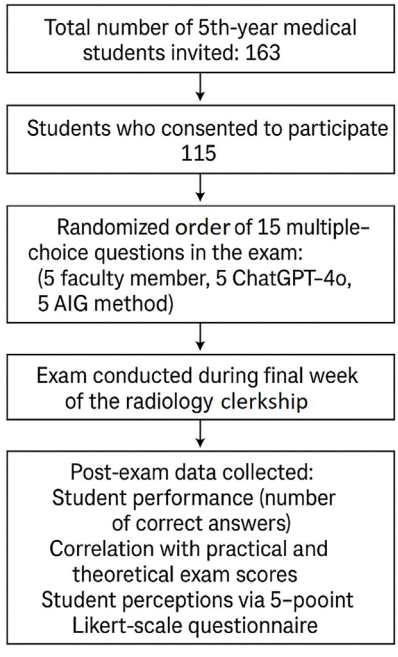

Figure 1. Flowchart of the study design. ChatGPT, Chat Generative Pre-trained Transformer; AIG, automatic item generation.

Table 2. Difficulty and discrimination indices of the generated questions

		Difficulty	Discrimination (27% upper–lower group method)	Item–rest correlation (Pearson)
Faculty member	1	0.60	0.53	0.29
	2	0.76	0.57	0.45
	3	0.61	0.36	0.27
	4	0.67	0.71	0.41
	5	0.28	0.28	0.12
Automatic item generation	1	0.57	0.76	0.48
	2	0.42	0.64	0.37
	3	0.59	0.74	0.54
	4	0.31	0.80	0.54
	5	0.43	0.74	0.46
ChatGPT-4o	1	0.28	0.65	0.41
	2	0.49	0.52	0.29
	3	0.14	0.00	–0.80
	4	0.77	0.43	0.33
	5	0.62	0.25	0.06

ChatGPT, Chat Generative Pre-trained Transformer.

Table 3. Students' evaluation of the questions [rated on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree)]

	Faculty member					Automatic item generation					ChatGPT-4o				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
The question text is clear	4.73	4.76	4.88	4.68	4.54	4.65	4.49	4.73	4.58	4.63	4.53	4.63	4.28	4.73	4.69
The question is of appropriate difficulty	4.35	4.55	4.52	4.15	3.88	4.2	3.88	4.24	3.96	4.18	3.75	4.13	3.79	4.44	4.54
The question has only one correct answer	4.69	4.79	4.68	4.64	4.24	4.62	4.37	4.55	4.5	4.46	4.4	4.7	4.32	4.71	4.67
The information provided in the question is sufficient to identify the correct answer	4.52	4.65	4.81	4.48	4.3	4.51	4.24	4.5	4.2	4.41	4.28	4.54	4.19	4.73	4.57
The distractors are reasonable	4.5	4.52	4.69	4.3	4.23	4.42	4.29	4.49	4.39	4.43	4.27	4.46	4.23	4.64	4.59

ChatGPT, Chat Generative Pre-trained Transformer.

Table 4. Students' evaluation of the examination [rated on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree)]

	Number of participants	1	2	3	4	5	Mean
The examination is clear	111	0	1	6	18	86	4.73
The examination is of appropriate difficulty	111	7	3	25	35	41	4.35
The examination is appropriate for the radiology internship	111	1	6	7	33	64	4.69
The examination aligns with the learning objectives	95	0	4	10	23	58	4.52

Discussion

There are publications in the literature indicating that high-quality questions can be generated using LLMs.[15,16] In many of these studies, the questions were subject to expert review, and it was concluded that the use of these questions in examinations is appropriate.[17,18] In studies involving student participation, the questions were found to be psychometrically acceptable when compared with human-written questions, and LLMs were shown to have the potential to serve as test developers for student assessment.[19] Some studies have also calculated difficulty and discrimination indices for the questions.[5,6,20,21] In a study by Emekli and Karahan[20] conducted with radiology technician students, the difficulty index for questions generated by ChatGPT was found to be 0.50, and 73.33% of the questions were deemed acceptable in terms of discrimination index. In another study in the field of anatomy, the discrimination index of the generated questions ranged between 0.29 and 0.54, and 11 out of 12 questions were reported to have a high discrimination index. In the same study, difficulty indices were calculated that ranged between 0.41 and 0.89.[6] Similarly, in this study, 4 out of 5 ChatGPT-generated questions were found acceptable in terms of discrimination index, with an average difficulty index of 0.46. These findings appear consistent with the existing literature on questions generated with LLMs, though further validation is required.

The template-based AIG method is an automatic question generation technique that was studied and developed earlier than LLMs. The literature shows that questions of sufficient quality for assessment and evaluation can be generated using this method.[22] Questions have been produced using this technique in different languages and medical specialties, and its effectiveness has been demonstrated.[12,23,24] Recent studies have also implemented template-based AIG in multiple languages, including polish.[25] In this study, the difficulty index was 0.46, and the discrimination index ranged between 0.64 and 0.8, which is classified as very good. Additionally, with the recent emergence of LLMs, hybrid methods that combine both question generation techniques have begun to be developed. This approach, combining the strengths of AI-based and template-based methods, presents a promising alternative in the field. In expert evaluations of questions generated using the hybrid method, it was reported that the correct answer could be identified in 96.2% of cases, and consensus was reached among experts on all questions.[26]

In addition to the findings in the literature, this study applied questions generated by both methods to the same student population and compared them with faculty-written questions. However, no prior study in radiology education has evaluated both question generation techniques in an examination format alongside faculty-authored questions. In one study, clinical reasoning skills in radiology were evaluated by experts, and questions generated by ChatGPT were rated 84.52% successful, faculty-written questions 82.14%, and template-based AIG-generated questions 78.57%. Based on these findings, both AI-based question generation methods were considered effective in assessing clinical reasoning skills.[27] Recent work by Mistry et al.[28] also supports the feasibility of LLMs, particularly GPT-4, in generating radiology board-style MCQs with high expert ratings in clarity, difficulty, and rationale quality. Although their study focused on expert review rather than student performance data, our findings complement theirs by providing psychometric and student-based validation of AI-generated items in undergraduate radiology education. Together, these studies contribute to a growing body of evidence supporting the role of LLMs in medical assessment design. In this study, consistent with the aforementioned research, both methods were found to produce questions with appropriate difficulty and discrimination indices. Additionally, the questions generated using the template-based AIG method demonstrated particularly high discrimination indices.

ChatGPT and template-based AIG offer significant advantages in accelerating and diversifying the process of preparing MCQs, while also reducing the workload on faculty members. The fact that the questions produced using these methods have acceptable levels of difficulty and discrimination suggests potential utility in assessment processes, although further research is needed to confirm this. However, due to the limited sample size of items in this study, these findings should be considered preliminary and exploratory, in line with similar small-scale investigations in the literature. Further large-scale studies are required to confirm these initial observations. These methods also have certain limitations. Questions generated by LLMs may sometimes contain content errors or semantic ambiguities, whereas the template-based AIG method requires extensive preparation and expertise, making it time-consuming. To minimize these limitations, hybrid methods that combine the speed and flexibility of AI with the controllability of template-based approaches may offer a more balanced and reliable question generation model.

This study has some limitations. First, it was conducted at a single medical school and only with 5th-year radiology internship students, which may limit the generalizability of the findings. Additionally, the examination administered during the study was voluntary and not included in official grading, which may have resulted in variable student motivation levels. The limited number of questions in the examination also restricted the evaluation of a broader range of item characteristics for each method. In particular, key psychometric metrics, such as reliability coefficients, could not be calculated due to the small item set, limiting the interpretability of overall examination consistency. Therefore, the findings should be viewed as preliminary and not generalizable without replication in larger datasets. The inclusion of only 5 questions per method constrains the reliability and generalizability of the psychometric analyses. Future research should consider increasing the number of items–ideally approaching the scale of studies such as Law et al.[29], which suggests 100 questions per method–to strengthen statistical power and validity. Although the prompt structures used in ChatGPT-4o were systematically controlled, the model's inherent variability may have led to some content differences. In the template-based AIG technique, questions were generated based on only one specific topic, limiting the evaluation of the method's effectiveness in other radiologic subfields. Some items in the examination showed discrimination index values >0.70, which may indicate high heterogeneity among participants or inconsistent motivation, as the test was voluntary and ungraded. Additionally, we used both the 27% upper–lower method and item–rest (Pearson) correlation for discrimination analysis. The latter provided more moderate values, suggesting that the items performed reasonably well across ability levels. Finally, psychometric analyses were conducted solely within the framework of classical test theory, limiting comparisons with alternative analytical methods. Therefore, future research is recommended to include larger participant groups, be conducted in different medical schools, and involve multicenter applications across various clinical domains. Another limitation is relat-

ed to the nature of the student evaluations of the MCQs. Although students provided valuable insights regarding question clarity, plausibility of distractors, and perceived alignment with learning goals, their ability to assess critically the psychometric and clinical quality of the questions may be limited. As undergraduate learners, they may not have the expertise required to fully evaluate item quality or the nuances of the clinical content presented.

In conclusion, this study demonstrates that AI-based question generation techniques in radiology education perform comparably with faculty-authored questions in a limited pilot setting. Both ChatGPT-4o and template-based AIG-generated questions possess acceptable levels of discrimination. However, due to the small number of items and single-institution design, these results should be interpreted with caution, given the small sample of items used and the single-institution design. The findings are preliminary and exploratory and should not be generalized without replication in larger item sets and multicenter studies. AI-based question generation methods may serve as supplementary tools in student assessment by reducing the time and expertise requirements traditionally needed in question development. Given the small item sample used in this study, the results should be interpreted as exploratory and hypothesis-generating rather than conclusive. Future research with larger item sets and multicenter designs is warranted to further validate the effectiveness of these approaches.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. 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(5):569-577. [Crossref]
2. 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. *Med Teach*. 2012;34(9):683-697. [Crossref]
3. Schuwirth LW, van der Vleuten CP. Different written assessment methods: what can be said about their strengths and weaknesses? *Med Educ*. 2004;38(9):974-679. [Crossref]
4. Kıyak YS, Emekli E. ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: a literature review. *Postgrad Med J*. 2024;100(1189):858-865. [Crossref]
5. Abujabal A, Yahya M, Riedewald M, Weikum G. Automated template generation for question answering over knowledge graphs. In: Proceedings of the 26th International Conference on World Wide Web (WWW '17). Geneva, Switzerland; 2017:1191-1200. [Crossref]
6. Kıyak YS, Soylu A, Coşkun Ö, Budakoğlu İİ, Peker T. Can ChatGPT generate acceptable case-based multiple-choice questions for medical school anatomy exams? A pilot study on item difficulty and discrimination. *Clin Anat*. 2025;38(4):505-510. [Crossref]
7. Leslie T, Gierl MJ. Using automatic item generation to create multiple-choice questions for pharmacy assessment. *Am J Pharm Educ*. 2023;87(10):100081. [Crossref]
8. ESR Curriculum for Undergraduate Radiological Education. (Accessed April 4, 2025.) [Crossref]
9. Medical Faculty - National Core Curriculum 2020. *Tıp Eğitimi Dünyası*. 2020;19(57-1):1-146. [Crossref]
10. Kıyak YS. A ChatGPT prompt for writing case-based multiple-choice questions. *Span J Med Educ*. 2023;4(3). [Crossref]
11. Gierl MJ, Lai H, Turner SR. Using automatic item generation to create multiple-choice test items. Med Educ. 2012;46(8):757-765. [Crossref]
12. Emekli E, Emekli E, Kıyak YS, Hoşgören Alıcı Y, Coşkun Ö, Budakoğlu İİ. Clinical reasoning in psychiatric education: development of multiple-choice questions with automatic item generation in Turkish. *Turk Psikiyatri Derg*. 2025;36:336-343. [Crossref]
13. 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? *Res Pract Technol Enhanc Learn*. 2020;15:12. [Crossref]
14. Downing SM, Yudkowsky R Assessment in health professions education (1st ed.). Routledge. 2009. [Crossref]
15. Elkins S, Kochmar E, Serban I, Cheung JCK. How useful are educational questions generated by large language models? In: Wang N, Rebolledo-Mendez G, Dimitrova V, et al., eds. Artificial Intelligence in Education. Communications in Computer and Information Science. Vol 1831. Springer; 2023:603-607. [Crossref]
16. 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(8):e0290691. [Crossref]
17. 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(6):e40977. [Crossref]
18. Ayub I, Hamann D, Hamann CR, Davis MJ. Exploring the potential and limitations of chat generative pre-trained transformer (ChatGPT) in generating board-style dermatology questions: a qualitative analysis. *Cureus*. 2023;15(8):e43717. [Crossref]
19. Lin Z, Chen H. Investigating the capability of ChatGPT for generating multiple-choice reading comprehension items. *System*. 2024;123:103344. [Crossref]
20. Emekli E, Karahan BN. AI in radiography education: evaluating multiple-choice questions difficulty and discrimination. *J Med Imaging Radiat Sci*. 2025;56(4):101896. [Crossref]
21. Zuckerman M, Flood R, Tan RJB, Kelp N, Ecker DJ, Menke J, Lockspeiser T. ChatGPT for assessment writing. *Med Teach*. 2023;45(11):1224-1227. [Crossref]
22. Gierl MJ, Lai H. Evaluating the quality of medical multiple-choice items created with automated processes. *Med Educ*. 2013;47(7):726-733. [Crossref]
23. Kurdi G, Leo J, Parsia B, Sattler U, Al-Emari S. A systematic review of automatic question generation for educational purposes. *Int J Artif Intell Educ*. 2020;30(1):121-204. [Crossref]
24. Gierl MJ, Lai H, Tanygin V. Advanced Methods in Automatic Item Generation. 1st ed. New York, NY: Routledge; 2021. [Crossref]
25. Kıyak YS, Kononowicz A, Górski S. Multilingual template-based automatic item generation for medical education supported by generative artificial intelligence models ChatGPT and claude. *Bio-Algorithms and Med-Systems*. 2024;20(1):81-96. [Crossref]
26. Kıyak YS, Emekli E, Coşkun Ö, Budakoğlu İİ. Keeping humans in the loop efficiently by generating question templates instead of questions using AI: Validity evidence on Hybrid AIG. *Med Teach*. 2025;47(4):744-747. [Crossref]
27. Emekli E, Karahan BN. Comparison of automatic item generation methods in the assessment of clinical reasoning skills. *Span J Med Educ*. 2024;6(1). [Crossref]
28. Mistry NP, Saeed H, Rafique S, Le T, Obaid H, Adams SJ. Large language models as tools to generate radiology board-style multiple-choice questions. *Acad Radiol*. 2024;31(9):3872-3878. [Crossref]
29. Law AK, So J, Lui CT, et al. AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination. *BMC Med Educ*. 2025;25(1):208. [Crossref]