Surgical and Radiologic Anatomy (2025) 47:158
https://doi.org/10.1007/s00276-025-03667-z

ORIGINAL ARTICLE

# Anatomy exam model for the circulatory and respiratory systems using GPT-4: a medical school study

**Ayla Tekin**[1] · **Nizameddin Fatih Karamus**[1,2] · **Tuncay Çolak**[1]

Received: 24 March 2025 / Accepted: 22 May 2025
© The Author(s), under exclusive licence to Springer-Verlag France SAS, part of Springer Nature 2025

## Abstract

**Purpose** The study aimed to evaluate the effectiveness of anatomy multiple-choice questions (MCQs) generated by GPT-4, focused on their methodological appropriateness and alignment with the cognitive levels defined by Bloom's revised taxonomy to enhance assessment.

**Methods** The assessment questions developed for medical students were created utilizing GPT-4, comprising 240 MCQs organized into subcategories consistent with Bloom's revised taxonomy. When designing prompts to create MCQs, details about the lesson's purpose, learning objectives, and students' prior experiences were included to ensure the questions were contextually appropriate. A set of 30 MCQs was randomly selected from the generated questions for testing. A total of 280 students participated in the examination, which assessed the difficulty index of the MCQs, the item discrimination index, and the overall test difficulty level. Expert anatomists examined the taxonomy accuracy of GPT-4's questions.

**Results** Students achieved a median score of 50 (range, 36.67-60) points on the test. The test's internal consistency, assessed by KR-20, was 0.737. The average difficulty of the test was 0.5012. Results show difficulty and discrimination indices for each AI-generated question. Expert anatomists' taxonomy-based classifications matched GPT-4's 26.6%. Meanwhile, 80.9% of students found the questions were clear, and 85.8% showed interest in retaking the assessment exam.

**Conclusion** This study demonstrates GPT-4's significant potential for generating medical education exam questions. While it effectively assesses basic knowledge recall, it fails to sufficiently evaluate higher-order cognitive processes outlined in Bloom's revised taxonomy. Future research should consider alternative methods that combine AI with expert evaluation and specialized multimodal models.

**Keywords** Anatomy assessment · Bloom's revised taxonomy · GPT-4 · Multiple-choice questions

**Abbreviations**

| | |
|---|---|
| AI | Artificial intelligence |
| LLM | Large language model |
| MCQs | Multiple-choice questions |
| NCC | National Core Curriculum |
| KR-20 | Kuder-Richardson 20 |

## Introduction

Advancements in medicine have led to a rising role for artificial intelligence (AI). AI-based LLMs, such as Google Gemini, Llama 3.3, and ChatGPT, are now being used to facilitate medical education and evaluation. Many researchers have stated that it can have different benefits, including cost, efficiency, and time savings. However, it is also emphasized that particular concerns must be considered, such as compliance with ethical standards, potential copyright infringement, plagiarism in writing, data security, and the risk of providing inaccurate information [1–4].

Anatomy constitutes a primary priority within the realm of medical education. Nonetheless, the development of

✉ Ayla Tekin
ayla.tekinorha@kocaeli.edu.tr

✉ Nizameddin Fatih Karamus
fatih.karamus@altinbas.edu.tr

✉ Tuncay Çolak
tuncayc@kocaeli.edu.tr

[1] Faculty of Medicine, Anatomy Department, Kocaeli University, Umuttepe Campus, Kocaeli 41001, Turkey

[2] Faculty of Medicine, Anatomy Department, Altinbas University, İstanbul, Turkey

Published online: 10 June 2025

innovative assessment and evaluation methodologies for this discipline, as well as the adaptation to evolving educational requirements, presents various challenges. There is an increasing need for innovative tools that introduce new approaches to evaluation processes while enhancing the accuracy and efficiency of examinations.

ChatGPT exhibits considerable potential in supporting medical research, education, and clinical practice, especially in anatomy education. Recent studies have increasingly emphasized its ability to deliver precise anatomical descriptions and clarify the intricate relationships among anatomical structures. Furthermore, the incorporation of ChatGPT into anatomy education has garnered attention for its potential to elevate student engagement and stimulate research initiatives. However, the positioning of ChatGPT as a co-author in academic publications has given rise to significant ethical dilemmas, particularly concerning academic integrity and the appropriateness of generating scholarly work without direct human intellectual contribution [5–7]. Although the large language models (LLMs) that have emerged recently were not explicitly designed for medicine, they can be quite effective in many medical fields due to their capabilities [8, 9]. In a study that reviewed 50 different works conducted with ChatGPT, five main functions for educators were categorized under two primary themes: instructional preparation (material development, suggestions, translation, etc.) and evaluation processes (task creation and performance assessment). However, it has been observed that there are a limited number of studies addressing its role in supporting both instructional and evaluation processes [3]. It has been noted that educators can conduct evaluation processes more efficiently by using ChatGPT, saving both time and effort [10]. Also, many educators spend a significant amount of time preparing quizzes, monthly tests, and exams, and thus, ChatGPT offers an opportunity to alleviate the assessment burden. ChatGPT could contribute to the evaluation process by providing a semi-automated grading system and reducing instructors' workload [11]. Assessing student feedback is crucial since their views greatly affect engagement, motivation, and overall learning results in relation to AI-generated assessments. For instance, ChatGPT's ability to generate questions about anatomy was evaluated, and it was noted that it was quite effective [12]. However, the methodologies used and the commands given may prevent questions from being created at a level appropriate for the purpose. Such situations raise some questions about integrating artificial intelligence into the assessment processes in the field of anatomy. Following the rise of artificial intelligence language models, a recent study has started to explore their ability to generate exam questions for relevant medical fields [13]. These studies have examined the accuracy of AI models' responses to the same questions in different languages and evaluated the comparative results of multiple language models. However, research regarding the effective and efficient use of LLMs in the medical field remains limited, particularly concerning the reliability, validity, and difficulty level of AI-generated questions. Automating the creation of purposeful questions can both reduce the workload of experts in the field of anatomy and increase the scope and depth of the assessment process. Bloom's revised taxonomy is a framework used to classify the cognitive levels an educator expects a student to utilize when learning a subject or responding to a question [14]. However, the use of Bloom's revised taxonomy in anatomy, particularly in studies focused on assessment, has remained limited [15].

This study makes a unique contribution to the literature by systematically analyzing anatomy MCQs generated by GPT-4 using Bloom's revised taxonomy. This approach effectively addresses significant gaps related to the validity and reliability of questions and their cognitive appropriateness. In this respect, our study evaluates the effectiveness of GPT-4 in creating anatomy MCQs, using Bloom's revised taxonomy and concentrating specifically on the circulatory and respiratory systems because of their essential roles in medical education.

## Materials and methods

### Ethics committee approval

The necessary permissions for the study's conduct were obtained from the Kocaeli University Non-Interventional Clinical Research Ethics Committee, with approval number GOKAEK-2023/17.13.

### Learning environment

The study was conducted within the Department of Anatomy at Kocaeli University Faculty of Medicine. An anatomy education program comprising 21 theoretical and 28 practical lecture hours was implemented as part of the 7-week circulatory and respiratory systems course. These lectures were distributed by the National Core Curriculum (NCC). Faculty members from the Department of Anatomy delivered the education in person. There are 313 students in the second-year medical faculty, 165 female and 148 male students. The average age of the students was $20.57 \pm 0.93$ years.

## Question preparation process

### AI training and commands

The questions to be used in evaluating medical students were prepared using GPT-4, an AI-based LLM, following the subcategories of Bloom's revised taxonomy. Before generation, we provided GPT-4 with detailed prompts that included:

- the course aims and learning objectives in circulatory and respiratory systems anatomy;
- the second-year students' level of professional competence;
- relevant background on previously covered anatomy topics;
- explicit mention of the need to align questions with the National Core Curriculum.

240 multiple-choice questions were created, with 40 questions for each subcategory [14]. Commands (prompts) were provided to GPT-4 to generate questions at the desired level. These commands included the following key instructions:

- The AI must identify itself as an expert in anatomy, referred to as AnatomyGPT.
- It should behave as if it is a specialist in the field of anatomy.
- The purpose of the committee; cardiovascular, blood, bone marrow, lymphoid organs, and respiratory systems is the examination of developmental, structural, macroscopic, cellular, and functional aspects.
- The learning objectives of this committee include the following information:

1. Learns basic information about the anatomical structure and functions of the heart, vessels, and lungs.
2. Can explain the organization of the circulatory system, systemic and pulmonary circulations.
3. Defines the structures in the respiratory system and knows the respiratory system mechanisms.
4. With its applications, (i) can take venous blood, (ii) can apply intravenous injection, (iii) and perform urinary catheterization on men, (iv) suturing, and (v) breast examination on women.
5. Relates basic medical courses with clinical correlation courses.

- The difficulty level of the questions should align with the professional competency of second-year medical students. These students received 48 lesson hours of anatomy training from the Skeletal and Tissue Systems Committee in the first year, 80 lesson hours from the Muscle and Nerve Tissues Committee, and 49 lesson hours of anatomy education from the Circulatory and Respiratory Systems Committee in the second year.
- The topics must cover the anatomy of the circulatory and respiratory systems.
- The questions should consist of five options.
- The questions must be original and free of repetition.
- The AI must demonstrate a full understanding of the importance of clinical anatomy.
- The question content should not contain synonymous terms that would directly reveal the correct answer.
- The correct answer must be highlighted underneath each question.
- The question prompts explicitly mentioned "avoid negative phrasing or double negatives" and "provide only one correct answer to minimize ambiguity."
- 240 questions, 40 questions suitable for each subheading of the taxonomy, should be prepared. There should be an equal number of questions for each subheading.
- Our purpose in creating these questions is to verify the hypothesis that "In the field of medicine, especially in the field of anatomy, the use of large language models can enable more effective use of artificial intelligence by improving training and evaluation processes."
- The options in the questions must comply with FIPAT Terminologia Anatomica v.2.07 and Latin terminology should be used for anatomical terms [16] (Fig. 1).

### Randomization of AI-generated questions

The questions in the pool were reviewed by two expert anatomists to ensure their content complied with FIPAT Terminologia Anatomica v2.07 and to verify the accuracy of the answers. Five questions from each subcategory of the taxonomy were randomly selected to form six sets of 30-question tests using the online randomization service Research Randomizer v4.0 [17].

## Manual verification of questions and Bloom's classification

Two expert anatomists, each with over five years of teaching experience, reviewed the 30 generated questions from the randomizer. They ensured that each question met the following criteria:

1. ***Content validity and accuracy***: The questions were thoroughly examined to verify their alignment with recognized anatomical standards (FIPAT Terminologia Anatomica v2.07).

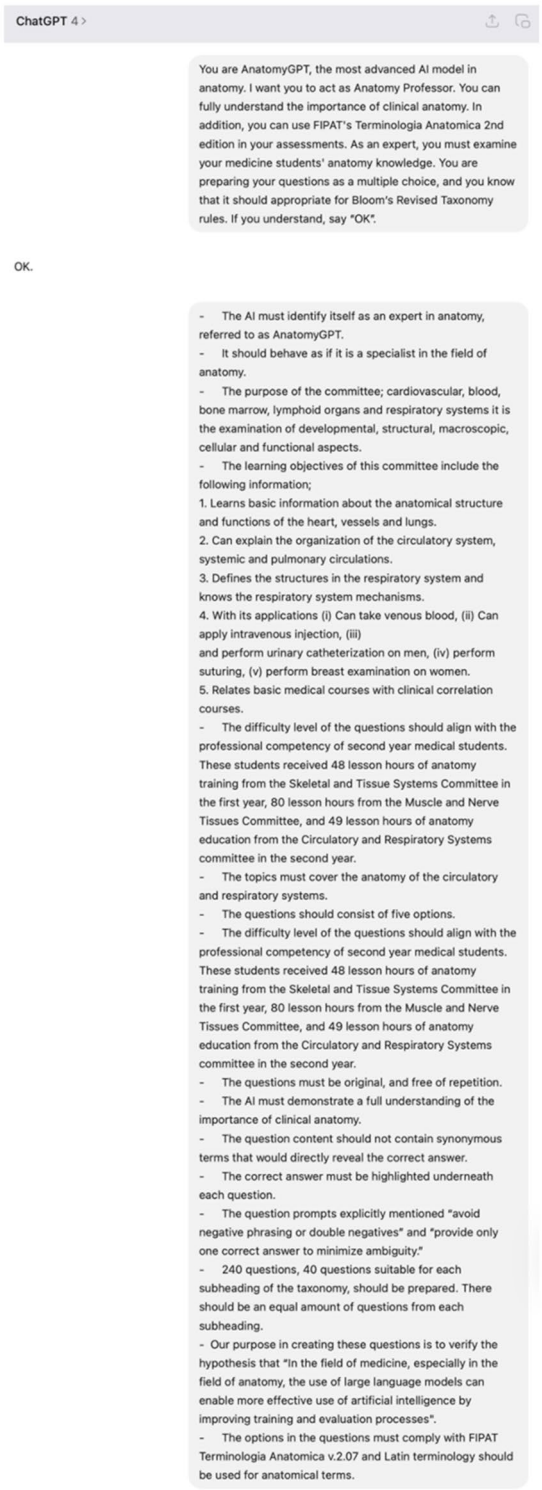



**Fig. 1** Commands given to GPT-4 and questions created according to Bloom's revised taxonomy

2. ***Bloom's revised taxonomy classification***: As suggested by GPT-4, the questions were classified using Bloom's revised taxonomy. The experts will evaluate the accuracy of GPT-4's taxonomy classification. Considering this assessment, they have opted not to modify the subgroups in which the questions have been categorized. However, their actual feedback regarding the classification of questions based on Bloom's revised taxonomy was documented in tables.
3. ***Clarity and phrasing***: The questions were carefully reviewed to ensure they were unambiguous, avoiding potential confusion or multiple correct answers.
4. ***Alignment with learning objectives***: The questions were thoroughly assessed to ensure they aligned with the learning objectives and the second-year curriculum level.

### Feedback on the AI-based questions by medicine faculty students

After the test was administered to the students, feedback was received on the following questions and topics from 42 students who voluntarily participated in the study, with a survey prepared using a 5-point Likert scale.

- How did you find the difficulty level of the prepared questions?
- I would like this kind of evaluation exams to be repeated.
- The questions were clear and understandable enough.

### Participant information and test administration

Out of 313 s-year medical students, 280 voluntarily participated in the study. The questions were distributed to the students in printed form in the lecture hall two days before the committee exam, and the answers were marked on optical sheets. The confidentiality of the students was ensured, and the answer sheets were stored in sealed envelopes without collecting any identifying information. Of these students, 225 completed the answer sheets accurately by marking their responses on the optical forms.

**Table 1** Students' scores in Bloom's revised taxonomy subgroups were based on questions classified by GPT-4

| | Median (IQR) |
|---|---|
| Analyzing questions (1) | 10 (6.67–13.33) |
| Applying questions (2) | 3.33 (3.33–6.67) |
| Creating questions (3) | 6.67 (3.33-10) |
| Evaluating questions (4) | 10 (6.67–13.33) |
| Remembering questions (5) | 10 (6.67–13.33) |
| Understanding questions (6) | 6.67 (6.67-10) |
| Total point | 50 (36.67-60) |
| p | $<0.001$ |
| Post-Hoc test | 1,3,4,5,6 > 2 for all<br>1,4,5,6 > 3<br>1,4,5 > 6 |

### Analysis and statistical methods

Statistical analysis was performed using IBM SPSS 29.0 (IBM Corp., Armonk, NY, USA). The normality of the data distribution was assessed using the Kolmogorov-Smirnov test. Since the assumption of normal distribution was unmet, numerical variables were presented as medians (25th-75th percentiles), and categorical variables were presented as frequencies (percentages). The Friedman two-way ANOVA was used for dependent group comparisons, and multiple comparisons were conducted using the Dunn test. The percentage of correct and incorrect responses for each question was calculated for item analysis. Success rates for the subcategories of questions aligned with Bloom's revised taxonomy were also evaluated. The scores obtained by the students on the AI-generated test were recorded, and overall averages were calculated. Using the top-bottom group method, the top 27% of students who scored the highest were classified as the "high group," while the bottom 27% were classified as the "low group" [18]. The item difficulty index was used to assess whether each question demonstrated high performance in evaluating students, while the item discrimination index was used to determine the quality of each question [19]. Reliability analysis was examined using the Kuder-Richardson 20 (KR-20) [20]. A $p$-value of $<0.05$ was considered sufficient for statistical significance in hypothesis testing.

## Results

### Scores on test questions and subcategories

The 30 questions generated by GPT-4 are grouped into five questions and show the categories of analyzing, applying, creating, evaluating, remembering, and understanding, respectively. The table below displays the scores achieved by students on questions classified into six subcategories based on Bloom's revised taxonomy, as organized by GPT-4 (Table 1). The results are expressed as medians and percentiles because the data did not follow a normal distribution. It was found that students scored 50 (36.67-60) points on the administered test. According to the Post-Hoc test, which analyzed significant differences between groups, the median score for application-type questions in other subcategories was noted. Furthermore, the scores from creativity-based questions were lower than those from understanding, evaluating, remembering, and analyzing questions in line with GPT-4. The scores from understanding-type questions were

also lower than those from evaluating, remembering, and analyzing questions ($p<0.001$).

The KR-20 test checks the internal consistency of measurements with dichotomous choices. A correct answer is scored 1 point, and an incorrect answer is scored 0 points. The most common use of KR-20 is in analyzing students' abilities or learning tests. These tests contain one correct answer for each item, which means that even though multiple options are presented to the respondent, only one is considered correct, while the others are deemed incorrect. The ability to categorize responses into correct and incorrect categories makes such questions inherently dichotomous, even if the item has more than two potential answers [20]. The analysis conducted using this method found that the test's internal consistency, according to the KR-20 formula, was 0.737. Since the ideal range for KR-20 is between 0.70 and 0.80, the test's internal consistency was considered to be within an acceptable range. It was determined that the highest possible value if the 10th question were removed to improve the internal consistency, would be 0.749.

In item analysis, no universally agreed-upon range values exist for item difficulty and discrimination indices in the literature. However, it is generally desired that the item difficulty index for a test's questions falls between 0.20 and 0.80, with an average difficulty of 0.50, and that the item discrimination index exceeds 0.30. According to Wood, the item difficulty index can range from 0 to 1 [21] (Table 2). When the *p*-value is multiplied by 100, it converts into a percentage representing the proportion of students who answered the item correctly. These ratios are generally expressed with the letter "p" [22]. The higher the *p*-value, the easier the items are. A higher difficulty index indicates that the item is easier to understand. Items with a *p*-value between 20% and 90% are considered good and acceptable. Within this range, items with a *p*-value between 40% and 60% are deemed excellent, as the difficulty index is at its maximum within this range. Items with a *p*-value less than 20% (very difficult) or greater than 90% (very easy) are considered unacceptable and need revision. It should be conceptualized that the *p*-value is essentially a behavioral measure. Rather than explaining difficulty as an intrinsic characteristic of an item, difficulty is defined based on how frequently test-takers select the correct answer [23].

The item difficulty and discrimination indices for each question are provided in the tables below. Items with a discrimination index between 0.30 and 0.40 are considered good, while items with a discrimination index above 0.40 are regarded as very good [18] (Table 3).

The percentage of correct and incorrect answers for each question is presented in the Table 4. It was observed that the question with the most incorrect responses, indicating outlier behavior among students, was question 28, while the question with the highest percentage of correct responses was question 3. Although question 3 had a very easy item difficulty index, it was found to have a high discrimination index. On the other hand, question 28 was calculated to be very difficult regarding the item difficulty index, and its discrimination index was found to be weak.

**Table 2** Item difficulty index

| Item difficulty index | Interpretation |
|---|---|
| 0.00–0.20 | Extremely difficult |
| 0.20–0.40 | Difficult |
| 0.40–0.60 | Moderately difficult |
| 0.60–0.80 | Relatively easy |
| 0.80–1.00 | Extremely easy |

**Table 3** Item discrimination index

| Item discrimination index | Interpretation |
|---|---|
| -1.00–0.00 | Invalid (Measures the opposite variable) |
| 0.00–0.19 | Weak (Does not measure the intended construct) |
| 0.20–0.29 | Acceptable (Can be revised and used) |
| 0.30–0.39 | Moderately Valid |
| 0.40–1.00 | Highly Valid |

Table 5 presents the item difficulty indices for the AI-generated questions. The average difficulty of the test is calculated by dividing the sum of these values by the total number of questions, which is 30. The average difficulty value of the test was found to be 0.50. In terms of the item difficulty index, six questions (20%) were classified as extremely difficult, 4 (13.3%) as difficult, 9 (30%) as moderately difficult, 8 (26.6%) as relatively easy, and 3 (10%) as extremely easy.

In terms of the item discrimination index, twenty questions (66.6%) were found to have a high validity rate. Three questions (10%) were deemed acceptable, while six (20%) required revision due to weak discrimination. One question (3.3%) did not adequately assess the intended objective and, therefore, needed to be removed from the test.

We received feedback about AI-generated questions from students. 16.7% of volunteer students found the anatomy questions to be very difficult, 47.6% found them hard questions, and 35.7% found them moderately difficult (Fig. 2).

On the other side, 85.8% of volunteer students expressed interest in retaking such assessment exams. While 7.1% were undecided, 7.2% of the students did not agree with this opinion (Table 3).

Of the students surveyed, 80.9% provided feedback indicating that the questions were clear and understandable. Conversely, 14.3% expressed undecision on this matter, while 4.8% expressed dissatisfaction with the clarity and comprehensibility of the questions (Table 4).

**Table 4** Correct and incorrect answer rates of medical faculty students for each question

| | | No. | Percent | | | No. | Percent |
|---|---|---|---|---|---|---|---|
| 1 | False | 35 | 15.6 | 16 | False | 94 | 41.8 |
| | True | 190 | 84.4 | | True | 131 | 58.2 |
| 2 | False | 96 | 42.7 | 17 | False | 55 | 24.4 |
| | True | 129 | 57.3 | | True | 170 | 75.6 |
| 3 | False | 34 | 15.1 | 18 | False | 91 | 40.4 |
| | True | 191 | 84.9 | | True | 134 | 59.6 |
| 4 | False | 193 | 85.8 | 19 | False | 86 | 38.2 |
| | True | 32 | 14.2 | | True | 139 | 61.8 |
| 5 | False | 103 | 45.8 | 20 | False | 182 | 80.9 |
| | True | 122 | 54.2 | | True | 43 | 19.1 |
| 6 | False | 163 | 72.4 | 21 | False | 36 | 16.0 |
| | True | 62 | 27.6 | | True | 189 | 84.0 |
| 7 | False | 136 | 60.4 | 22 | False | 128 | 56.9 |
| | True | 89 | 39.6 | | True | 97 | 43.1 |
| 8 | False | 158 | 70.2 | 23 | False | 88 | 39.1 |
| | True | 67 | 29.8 | | True | 137 | 60.9 |
| 9 | False | 179 | 79.6 | 24 | False | 106 | 47.1 |
| | True | 46 | 20.4 | | True | 119 | 52.9 |
| 10 | False | 193 | 85.8 | 25 | False | 110 | 48.9 |
| | True | 32 | 14.2 | | True | 115 | 51.1 |
| 11 | False | 58 | 25.8 | 26 | False | 66 | 29.3 |
| | True | 167 | 74.2 | | True | 159 | 70.7 |
| 12 | False | 188 | 83.6 | 27 | False | 44 | 19.6 |
| | True | 37 | 16.4 | | True | 181 | 80.4 |
| 13 | False | 125 | 55.6 | 28 | False | 209 | 92.9 |
| | True | 100 | 44.4 | | True | 16 | 7.1 |
| 14 | False | 104 | 46.2 | 29 | False | 176 | 78.2 |
| | True | 121 | 53.8 | | True | 49 | 21.8 |
| 15 | False | 186 | 82.7 | 30 | False | 98 | 43.6 |
| | True | 39 | 17.3 | | True | 127 | 56.4 |

## Discussion

As medical school curricula grow increasingly integrated and regulated, careful planning (blueprinting) of course and exam content is essential. Educators must thoughtfully create exam tasks and criteria that evaluate the knowledge and cognitive skills students must develop, ensuring that assessment methods align with the current curriculum [15, 24]. Our study evaluated the effectiveness of the prompts for GPT-4 in generating MCQs for the circulatory and respiratory systems in anatomy education. The findings revealed significant differences in student performance across questions. Although the questions are uniquely created, they are not sufficient in terms of creating questions that are fully compatible with the taxonomy subgroups. The MCQ format may struggle to capture complex cognitive processes, which appears to hinder the AI's ability to generate taxonomy-compliant questions [25, 26]. Indeed, previous studies have highlighted similar constraints, demonstrating that MCQs often fail to properly capture students' abilities in analyzing, evaluating, and creating because they rely primarily on recognition rather than production of knowledge. This reliance limits the assessment of critical thinking and complex cognitive reasoning [26, 27, 28].

Our study also examined each question based on item difficulty and discrimination indices. The analyses indicated that five questions were inadequate for use as multiple-choice questions. Specifically, 80% of these extremely difficult questions also exhibited poor discrimination indices. Furthermore, the negation phrase ('which one is not') used in questions 7, 14, 17, 18, and 29 likely introduced unnecessary complexity. In the 'understanding' category, the terms 'widely' and 'primary' in questions 26 and 27 undermined the distractors, reducing overall question quality and limiting effective assessment. Although GPT-4 offers instructors a quick way to generate numerous questions, manual revisions and additional instructions based on student performance are advised to enhance question suitability and ensure alignment with Bloom's revised taxonomy [24, 29].

It was observed that the questions generated by GPT-4 and classified in the "analyzing" subcategory showed a balanced distribution in terms of difficulty and item

**Table 5** Item difficulty index and item discrimination index values and classifications of each question generated by GPT-4

| Question No. | Cronbach's Alpha if Item Deleted | Item difficulty index | Item difficulty classification | Item discrimination index | Item discrimination index classification |
|---|---|---|---|---|---|
| 1 | 0.731 | 0.84 | Extremely easy | 0.52 | High validity |
| 2 | 0.726 | 0.56 | Moderately difficult | 0.58 | High validity |
| 3 | 0.728 | 0.84 | Extremely easy | 0.56 | High validity |
| 4 | 0.732 | 0.15 | Extremely difficult | 0.23 | Acceptable |
| 5 | 0.724 | 0.54 | Moderately difficult | 0.64 | High validity |
| 6 | 0.749 | 0.28 | Difficult | 0.14 | Weak |
| 7 | 0.733 | 0.40 | Moderately difficult | 0.41 | High validity |
| 8 | 0.737 | 0.31 | Difficult | 0.24 | Acceptable |
| 9 | 0.741 | 0.21 | Difficult | 0.19 | Weak |
| 10 | 0.749 | 0.14 | Extremely difficult | -0.03 | Invalid |
| 11 | 0.719 | 0.72 | Relatively easy | 0.65 | High validity |
| 12 | 0.745 | 0.17 | Extremely difficult | 0.09 | Weak |
| 13 | 0.728 | 0.45 | Moderately difficult | 0.52 | High validity |
| 14 | 0.732 | 0.53 | Moderately difficult | 0.52 | High validity |
| 15 | 0.749 | 0.16 | Extremely difficult | 0.04 | Weak |
| 16 | 0.732 | 0.60 | Relatively easy | 0.53 | High validity |
| 17 | 0.720 | 0.74 | Relatively easy | 0.63 | High validity |
| 18 | 0.724 | 0.60 | Relatively easy | 0.65 | High validity |
| 19 | 0.716 | 0.63 | Relatively easy | 0.77 | High validity |
| 20 | 0.740 | 0.18 | Extremely difficult | 0.14 | Weak |
| 21 | 0.723 | 0.82 | Extremely easy | 0.63 | High validity |
| 22 | 0.717 | 0.46 | Moderately difficult | 0.64 | High validity |
| 23 | 0.723 | 0.60 | Relatively easy | 0.68 | High validity |
| 24 | 0.724 | 0.54 | Moderately difficult | 0.63 | High validity |
| 25 | 0.722 | 0.51 | Moderately difficult | 0.67 | High validity |
| 26 | 0.719 | 0.69 | Relatively easy | 0.76 | High validity |
| 27 | 0.719 | 0.79 | Relatively easy | 0.72 | High validity |
| 28 | 0.739 | 0.06 | Extremely difficult | 0.05 | Weak |
| 29 | 0.738 | 0.23 | Difficult | 0.25 | Acceptable |
| 30 | 0.733 | 0.56 | Moderately difficult | 0.49 | High validity |

discrimination. In contrast, the questions categorized by GPT-4 under the "applying" subcategory revealed inconsistencies, with some questions more fitting for the "remembering" subgroup. For instance, question 7 displayed moderate difficulty and high validity, but the use of negation in the question stem added unnecessary complexity, reducing its alignment with Bloom's revised taxonomy. Manual revisions in such instances could improve the accuracy and suitability of these questions. Furthermore, the questions accurately categorized by GPT-4 as "remembering" exhibited high validity and an appropriate difficulty distribution. This observation may indicate the effectiveness of students' study strategies or the inherently memorable quality of anatomical knowledge. These findings align with previous research suggesting that multiple-choice items, particularly those assessing lower-order cognitive skills such as remembering and understanding, can perform reliably in evaluating basic knowledge when well-constructed [30].

It appears that two questions in the GPT-4's creating subcategory demonstrate a weak level of discrimination. When we examine the structure of these questions, it becomes clear that they connect anatomical knowledge with surgical methods, requiring a general understanding of surgical branches within the context of clinical anatomy. A more suitable solution might involve reviewing the courses taken by second-year students in clinical or surgical fields, re-evaluating the questions in relation to NCC, or adding different questions related to the creating subcategory instead of these.

Approximately 80% of the evaluating questions based on GPT-4 are perceived as easy, yet they demonstrate high validity regarding item discrimination. Although the questions were deemed appropriate in terms of discrimination index, analyses by anatomy experts indicated that four of these questions belonged to the remembering category and one to the understanding category. This shows that the classification consistency is low in categories that require interpretation and analysis, such as "evaluating".

However, a study indicates the productivity of GPT-4 in anatomy questions; it has been noted that the questions were not prepared within a proper methodology or systematic

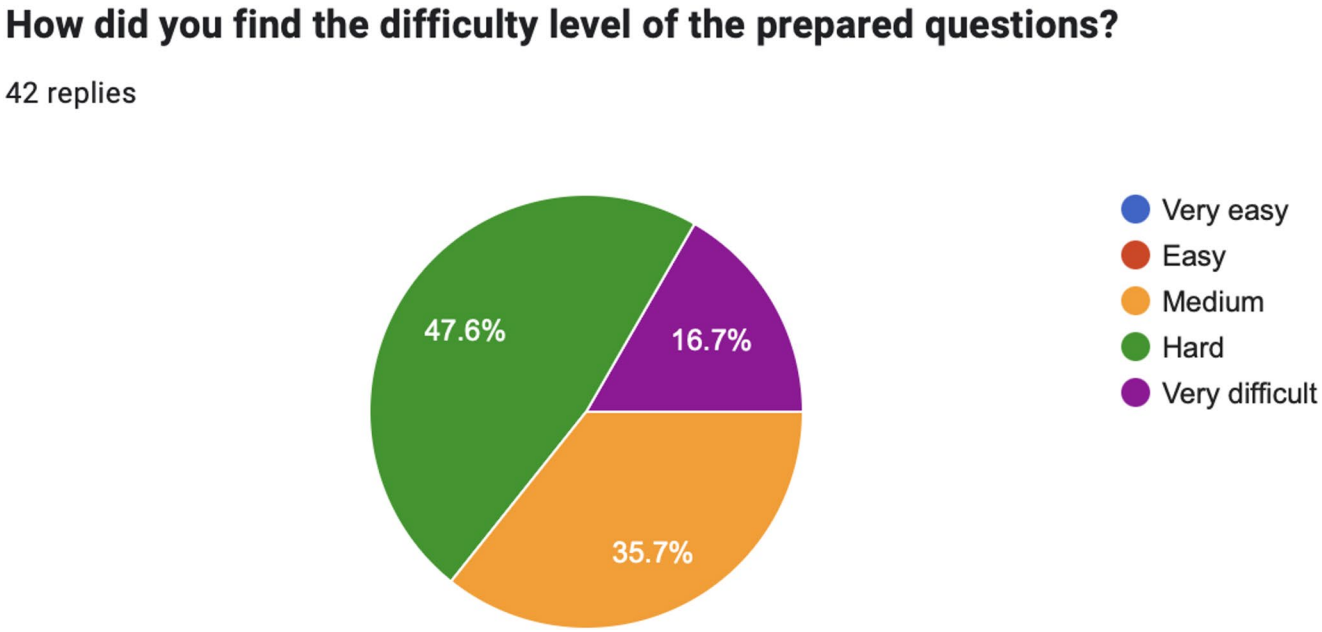



**Fig. 2** Students' opinions about the difficulty of exam questions

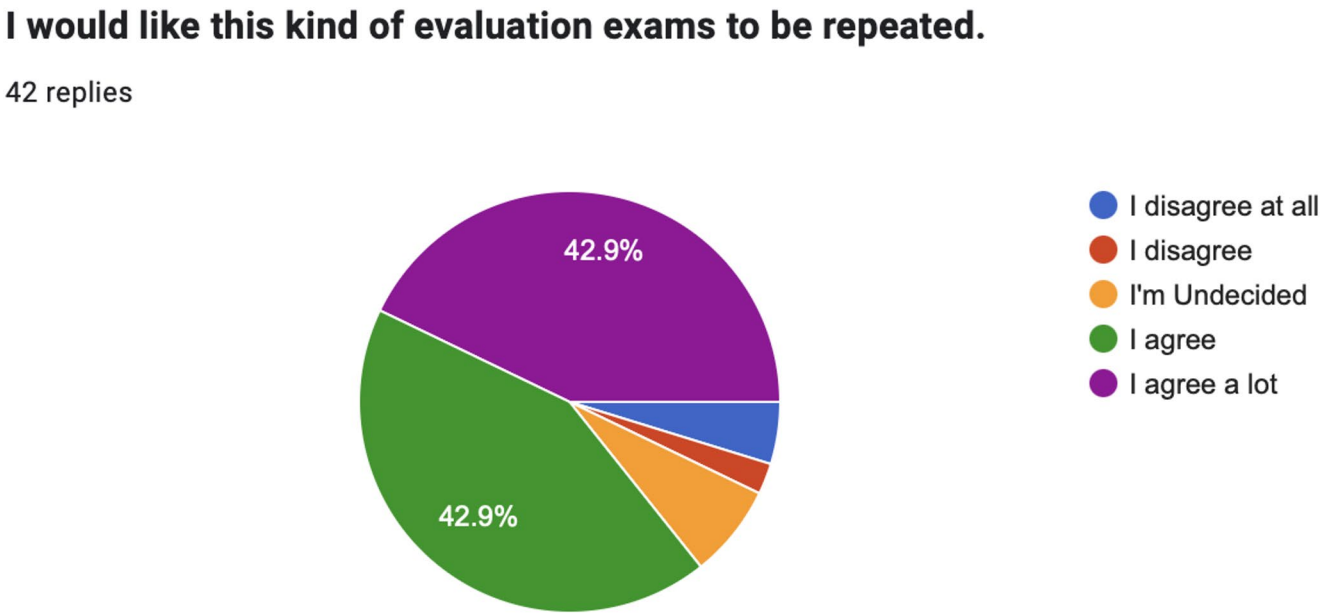



**Fig. 3** Students' opinions about the repeatability of such exams

framework [8]. Additionally, there is a lack of evaluation regarding the quality of the generated questions. This study analyzed the questions produced by GPT-4 at both the question and test levels, showing objective results, that anatomy questions can be created based on artificial intelligence in the future with appropriate commands and grounds.

Our findings show how complex the AI-generated questions were. On average, the test was moderately difficult, with an average difficulty value of 0.5012. Out of 30 questions, six were very difficult, four were difficult, nine were moderately difficult, eight were easy, and three were very easy. In terms of the item discrimination index, twenty questions had a high validity rate, three were acceptable, six required revisions due to weak discrimination, and one needed to be removed from the test as it didn't correctly assess the intended objective. The results obtained reveal the usability of GPT-4 in creating a valid test of the question distribution ratio in a created test. Additionally, the internal consistency of the test was found to be generally acceptable; however, some questions were identified as having a

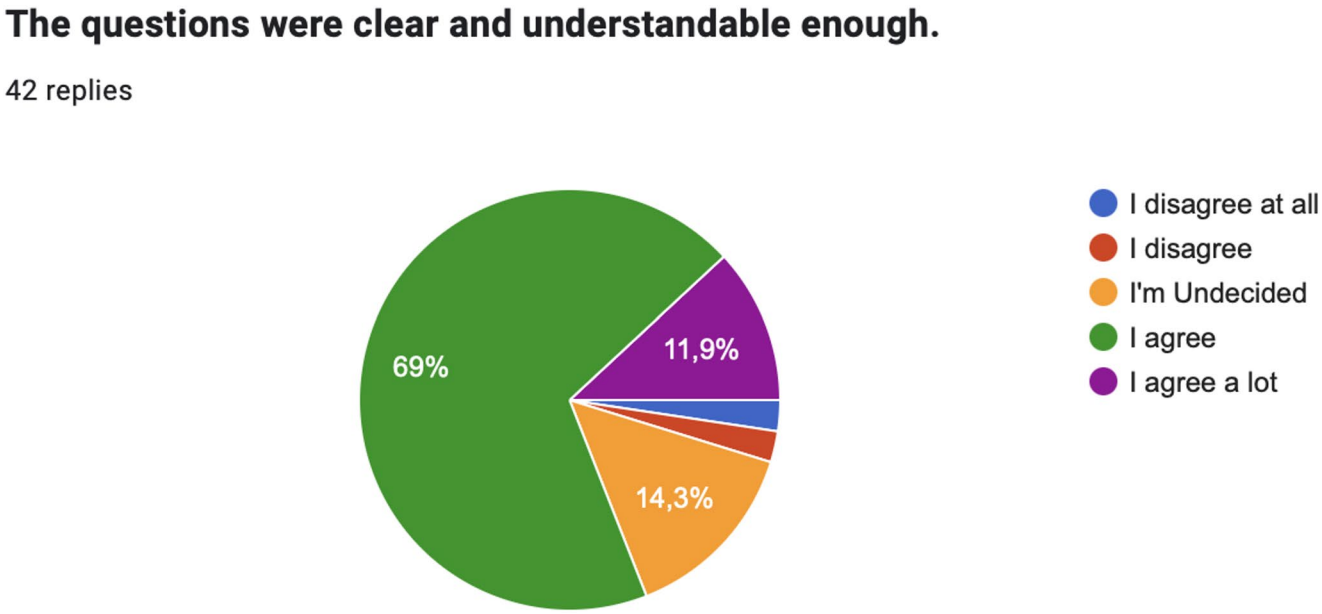



**Fig. 4** Students' opinions about the understandability of these questions According to the opinion of two expert anatomists in the department of anatomy, the categories in which the questions created by GPT-4 should be classified are detailed in Table 6. It has been reported that 24 of these questions should be included in the remembering category, 1 in the understanding category, and 5 in the applying category. When the classification made by GPT-4 was compared with the taxonomy classification of anatomy experts, a 26.6% agreement was observed. It was determined that 2 of these questions were in the applying subgroup and 1 in the understanding group, while all of the questions in the remembering subcategory were found to be compatible with the classification.

negative impact on the test's overall reliability. In order to increase the usability and validity of the questions in the test, instead of selecting questions from the question pool with the randomization method, it may be a more effective method for anatomy experts to choose the questions that comply with the taxonomy rules. Thus, the rates in tests based on questions with high-quality levels may be higher.

In our study, student feedback showed that the AI-generated questions were considered challenging, closely aligning with the objective analysis indicating that 63.33% of the items were classified as not easy based on their difficulty indices. This consistency highlights GPT-4's ability to generate contextually rich and complex anatomy questions. Furthermore, over 85% of students expressed a willingness to retake ChatGPT-based assessments, highlighting their positive experiences and the perceived value of diverse questions. Such exams seem to foster greater student interest and active participation compared to traditional assessments. Most students also reported that the AI-generated questions were clear, understandable, and suitably challenging, showcasing GPT-4's skill in accurately using complex medical terminology. These findings suggest that AI-generated questions can effectively help students focus on content rather than question structure. Combining objective data with student feedback further strengthens the validity of our results and underscores the potential of GPT-4 to significantly enhance student engagement and motivation in educational assessments [31, 32].

## Study limitations

### Limited student sample

This study's conclusions are restricted to a particular subset of Kocaeli University Faculty of Medicine students. The number of students taking part in the research and their demographics place restrictions on how far the findings may be applied. Using a bigger sample size could help the results become more broadly applicable. Other limitations of this study may include not grouping students by gender and age. Evaluating the relationship between genders may be necessary in revealing differences between student groups. In addition, although noting the opinions of these groups on the questions created will constitute a subjective analysis, it may be essential to obtain the opinions of the student groups regarding the questions.

### Limitations of the LLMs

LLMs have significant limits regarding their application in medical education despite their promise to be an extremely effective teaching tool. These restrictions include giving false information, depending too much on technology,

**Table 6** Classification of questions developed by GPT-4 following Bloom's revised taxonomy by expert anatomists of the anatomy department

| No. | Question | Taxonomy Classification |
|---|---|---|
| 1. | During a thoracic surgery, a surgeon identifies a vessel that originates from the aortic arch and supplies the head and neck. Which of the following vessels is being described? | Remembering |
| 2. | Which pharyngeal structure serves as the opening for both the respiratory and digestive tracts? | Remembering |
| 3. | The cricothyroid muscle plays a crucial role in adjusting the pitch of the voice. It functions by: | Remembering |
| 4. | Which of the following structures is found in the posterior part of the heart, making it most likely to be injured during procedures on the esophagus? | Remembering |
| 5. | Which nasal concha has a direct relation to the opening of the maxillary sinus? | Remembering |
| 6. | Patient complains of difficulty swallowing solid foods but can swallow liquids with ease. An obstruction in which part of the esophagus might explain this? | Applying |
| 7. | A thoracic surgeon needs to access the posterior mediastinum. Through which of the following would the surgeon NOT need to maneuver? | Applying |
| 8. | A patient who has had damage to their posterior cricoarytenoid muscle would most likely have difficulty in: | Remembering |
| 9. | Which muscle assists in tensing the vocal cords? | Remembering |
| 10. | During a surgical procedure on the larynx, a surgeon notes damage to a nerve that solely innervates the cricothyroid muscle. | Remembering |
| 11. | Given the anatomical positioning and function of the heart's chambers, which chamber would primarily be responsible for pumping oxygenated blood to the body tissues? | Remembering |
| 12. | In developing a new surgical technique, a surgeon seeks to minimize the risk of myocardial infarction. Which vessel, if occluded, would pose the highest risk to the left ventricular myocardium? | Applying |
| 13. | A new surgical technique aims to alleviate pressure on the recurrent laryngeal nerve by the aortic arch. Which lung's related structures would most likely be in close proximity to this surgery? | Applying |
| 14. | A physician is attempting to locate the thymus in an adult patient. In which region of the mediastinum would the physician expect to find a relatively small and regressed thymus? | Remembering |
| 15. | Imagine that a new thoracic surgical procedure requires an entry point through the intercostal space. Which of the following landmarks would best guide the surgeon to avoid damage to the intercostal nerve and vessels? | Applying |
| 16. | The absence of which structure would significantly impede the function of the vocal cords? | Remembering |
| 17. | Which structure is NOT found in the intercostal space? | Remembering |
| 18. | Which of the following muscles would NOT play a role in elevating the ribs during forced inhalation? | Remembering |
| 19. | A tumor has been found pressing against the only muscle responsible for abducting the vocal cords. Which muscle is compromised? | Remembering |
| 20. | A singer presents with a strained voice and undergoes laryngoscopy. Damage is noted on the arytenoid cartilage. Which muscle is primarily responsible for adducting the vocal cords, which may be affected? | Remembering |
| 21. | Which structure lies immediately posterior to the tracheal bifurcation? | Remembering |
| 22. | The torus tubarius is related to which structure? | Remembering |
| 23. | Which layer of the heart contains Purkinje fibers? | Remembering |
| 24. | Which rib articulates with the sternum at the sternal angle? | Remembering |
| 25. | Which structure separates the two atria of the heart? | Remembering |
| 26. | Which bronchi is wider, shorter, and more vertical, making it more susceptible to foreign object entry? | Remembering |
| 27. | A foreign object lodged in the trachea would likely enter which primary bronchus due to its more vertical orientation? | Remembering |
| 28. | The oropharynx is bordered superiorly by the: | Remembering |
| 29. | Which of the following structures is NOT located in the posterior mediastinum? | Remembering |
| 30. | Why is the location of the azygos vein clinically relevant in relation to the mediastinum? | Understanding |

hurting critical thinking, and raising issues with academic integrity. Therefore, it's critical that the data supplied by LLMs be verified and that the focus on practical skills never changes. Furthermore, it is believed that regular LLM updates are essential to guaranteeing timeliness and correctness. The GPT-4's shortcomings were particularly noticeable regarding questions that assessed higher-order cognitive functions like application and creativity. This implies that additional model development is required to offer a more thorough assessment of exam questions based on artificial intelligence. Although GPT-4 seems sufficient in preparing sentence-based anatomy questions, there is a serious inadequacy in preparing visual questions and describing anatomical structures. The analysis of visual structures holds significant importance in the evaluation of anatomy education. This clearly demonstrates that this area is poised for future development and enhancement, which can lead to improved learning outcomes. Considering the publications that emphasize the importance of visuality in anatomy, the importance of this subject can be clearly understood [33]. Moreover, although we made randomization among the questions we created, repeatability was observed regarding question style. This situation may pose a problem in creating question pools of very large scope and creating various types of questions. Furthermore, comparing the validity and dependability of AI-produced questions with those generated by humans is possible, and more research is necessary.

### Feedback on questions from anatomy department experts

As a result of the examinations, it was emphasized that MCQs generated by GPT-4 were generally focused on the categories of remembering, understanding, and applying, but did not address the levels of analyzing, creating, and evaluating. The key factors contributing to this outcome seem to be the exam format and assessment objectives [26, 28]. The MCQ format has limitations in evaluating higher-order cognitive processes like analyzing, evaluating, and creating. Because students can only choose from set options, they cannot adequately showcase their critical thinking, creative problem-solving, or evaluative reasoning skills. Therefore, alternative assessment methods such as open-ended questions, clinical case discussions, or oral exams may better assess these intricate cognitive abilities. Our experts observed that multiple-choice questions are better suited for assessing students' abilities in remembering and comprehending information. Moreover, since anatomy education emphasizes learning fundamental structures at the preclinical level, ChatGPT-4 based MCQs could not fully address questions in the analyzing, creating and evaluating category. Educational literature further supports these expert findings by demonstrating that multiple-choice formats effectively assess knowledge retention and comprehension, yet they remain insufficient for assessing analytical and evaluative reasoning [34, 35].

### Future research goals

Future studies can conduct similar experiments at multiple medical schools and with larger student populations to explore the effectiveness of AI-based exam questions in greater depth. This research can provide valuable insights into how artificial intelligence (AI) can be applied more broadly and successfully in educational settings. Furthermore, by contrasting the artificial intelligence language model's skills and flexibility with the questions developed using the "Blooming Anatomical Tool" created for the anatomy branch, it will be easier to see how strong it is [15].

Additionally, further research is needed to develop artificial intelligence models that can more effectively generate creative and practice-intensive questions. This further maximizes AI's educational potential by assisting students in acquiring a greater variety of cognitive capabilities and streamlining the evaluation procedure. Due to the increasing LLM competition between technology and companies today, it will be valuable to see the results of more comprehensive studies that include other language models to compare the capabilities of LLMs. In addition, AI models trained with unique data sets can be crucial in creating more qualified anatomy questions and eliminating terminology errors. Although it can be interpreted as a limitation in this study, it is a good idea for future studies.

## Conclusion

This study demonstrates that advanced artificial intelligence models, like GPT-4, have significant potential in creating exam questions for medical education. GPT-4 is highly effective at generating anatomy MCQs that test factual recall consistently but struggles with prompts that require higher-order thinking skills. To leverage its strengths and mitigate its weaknesses, we strongly recommend exploring alternative assessment methods and hybrid approaches that combine AI generation with expert human evaluation. Specialized LLMs trained specifically with medical and anatomical data could further improve the quality and applicability of AI-generated assessment tools. Future research should prioritize comparative studies between AI-generated and conventionally developed assessment materials, as well as investigate multimodal AI models that can address the essential visual components of anatomy education.

**Author contributions** A.T.: project development, first draft of the manuscript, writing - review and editing, data analysis; N.F.K.: project de-

velopment, manuscript writing/editing, design of study methodology, data collection or management; T.C.: manuscript editing, approval of the final version.

**Funding** This research did not receive any specific financial support from funding agencies within the public, commercial, or not-for-profit sectors.

**Data availability** No datasets were generated or analysed during the current study.

## Declarations

**Ethics approval** The study's necessary permissions were obtained from the Kocaeli University Non-Interventional Clinical Research Ethics Committee, with approval number GOKAEK-2023/17.13.

**Consent to participate** Written informed consent was obtained from the participants.

**Consent for publication** Not applicable.

**Competing interests** The authors declare no competing interests.

## References

1. Arun G, Perumal V, Urias FPJB et al (2024) ChatGPT versus a customized AI chatbot (Anatbuddy) for anatomy education: a comparative pilot study. Anat Sci Educ 17:1396–1405. https://doi.org/10.1002/ase.2502
2. Gordon M, Daniel M, Ajiboye A et al (2024) A scoping review of artificial intelligence in medical education: BEME guide 84. Med Teach 46:446–470. https://doi.org/10.1080/0142159X.2024.2314198
3. Lo CK (2023) What is the impact of ChatGPT on education? A rapid review of the literature. Educ Sci 13:410. https://doi.org/10.3390/educsci13040410
4. Sallam M (2023) ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthc 11:887. https://doi.org/10.3390/healthcare11060887
5. Lee H (2024) The rise of ChatGPT: exploring its potential in medical education. Anat Sci Educ 17:926–931. https://doi.org/10.1002/ase.2270
6. Totlis T, Natsis K, Filos D et al (2023) The potential role of ChatGPT and artificial intelligence in anatomy education: a conversation with ChatGPT. Surg Radiol Anat 45:1321–1329. https://doi.org/10.1007/s00276-023-03229-1
7. Uzun GB (2024) In the anatomic studies, is it correct to add an artificial intelligence such as ChatGPT as a co-author? Eur J Ther 30:89–90. https://doi.org/10.58600/eurjther1826
8. Collins BR, Black EW, Rarey KE (2024) Introducing AnatomyGPT: a customized artificial intelligence application for anatomical sciences education. Clin Anat 37:661–669. https://doi.org/10.1002/ca.24178
9. Mantzou N, Ediaroglou V, Drakonaki E et al (2024) ChatGPT efficacy for answering musculoskeletal anatomy questions: a study evaluating quality and consistency between raters and timepoints. Surg Radiol Anat 46:1885–1890. https://doi.org/10.1007/s00276-024-03477-9
10. Zhai X (2023) ChatGPT for next generation science learning. XRDS Crossroads ACM Mag Stud 29:42–46. https://doi.org/10.1145/3589649
11. Sok S, Heng K (2023) ChatGPT for education and research: a review of benefits and risks. Cambodian J Educ Res 3:110–121. https://doi.org/10.2139/ssrn.4378735
12. Leng L (2024) Challenge, integration, and change: ChatGPT and future anatomical education. Med Educ Online 29:2304973. https://doi.org/10.1080/10872981.2024.2304973
13. Coraci D, Maccarone MC, Regazzo G et al (2023) ChatGPT in the development of medical questionnaires. The example of the low back pain. Eur J Transl Myol 33(4):12114. https://doi.org/10.4081/ejtm.2023.12114
14. Anderson LW, Krathwohl DR (2001) A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives: complete edition. Longman, New York
15. Thompson AR, O'Loughlin VD (2015) The blooming anatomy tool (BAT): a discipline-specific rubric for utilizing Bloom's taxonomy in the design and evaluation of assessments in the anatomical sciences. Anat Sci Educ 8:493–501. https://doi.org/10.1002/ase.1507
16. Federative International Programme for Anatomical Terminology (FIPAT) (2019) FIPAT.library.dal.ca, 2nd edn. Federative International Programme for Anatomical Terminology
17. Urbaniak GC, Plous S (2013) Research Randomizer (version 4.0) [computer program]. Available at: http://www.randomizer.org/
18. Ebel RL, Frisbie DA (1991) Essentials of educational measurement, 5th edn. Prentice-Hall of India, New Delhi
19. Kelley TL (1939) The selection of upper and lower groups for the validation of test items. J Educ Psychol 30:17. https://doi.org/10.1037/h0057123
20. Kuder GF, Richardson MW (1937) The theory of the Estimation of test reliability. Psychometrika 2:151–160. https://doi.org/10.1007/BF02288391
21. Wood DA (1960) Test construction: development and interpretation of achievement tests. Charles E. Merrill, Columbus, Ohio
22. Crocker L, Algina J (2008) Introduction to classical and modern test theory. Cengage Learning, Mason, Ohio
23. Thorndike RM, Cunningham GK, Thorndike RL, Hagen EP (1991) Measurement and evaluation in psychology and education, 5th edn. Macmillan, New York
24. Cheung BHH, Lau GKK, Wong GTC et al (2023) ChatGPT versus human in generating medical graduate exam multiple choice questions—A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the united Kingdom). PLoS ONE 18:e0290691. https://doi.org/10.1371/journal.pone.0290691
25. Haladyna TM, Downing SM (1989) A taxonomy of multiple-choice item-writing rules. Appl Meas Educ 2:37–50. https://doi.org/10.1207/s15324818ame0201_3
26. Stanger-Hall KF (2012) Multiple-choice exams: an obstacle for higher-level thinking in introductory science classes. CBE—Life Sci Educ 11:294–306. https://doi.org/10.1187/cbe.11-11-0100
27. Haataja ESH, Tolvanen A, Vilppu H et al (2023) Measuring higher-order cognitive skills with multiple choice questions–potentials and pitfalls of Finnish teacher education entrance. Teach Teach Educ 122:103943. https://doi.org/10.1016/j.tate.2022.103943
28. Race P (2019) The lecturer's toolkit: a practical guide to assessment, learning and teaching, 5th edn. Routledge, London
29. Ahmed A, Kerr E, O'Malley A (2025) Quality assurance and validity of AI-generated single best answer questions. BMC Med Educ 25:300. https://doi.org/10.1186/s12909-025-06881-w
30. Simkin MG, Kuechler WL (2005) Multiple-choice tests and student understanding: what is the connection? Decis Sci J Innov Educ 3:73–98. https://doi.org/10.1111/j.1540-4609.2005.00053.x

31. Birks S, Gray J, Darling-Pomranz C (2024) Using artificial intelligence to provide a 'flipped assessment' approach to medical education learning opportunities. Med Teach. https://doi.org/10.1080/0142159X.2024.2434101
32. Monzon N, Hays FA (2025) Leveraging generative artificial intelligence to improve motivation and retrieval in higher education learners. JMIR Med Educ 11:e59210. https://doi.org/10.2196/59210
33. Sagoo MG, Vorstenbosch MA, Bazira PJ et al (2021) Online assessment of applied anatomy knowledge: the effect of images on medical students' performance. Anat Sci Educ 14:342–351. https://doi.org/10.1002/ase.1965
34. Ozuru Y, Briner S, Kurby CA, McNamara DS (2013) Comparing comprehension measured by multiple-choice and open-ended questions. Can J Exp Psychol 67:215. https://doi.org/10.1037/a0032918
35. Palmer EJ, Devitt PG (2007) Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper. BMC Med Educ 7:49. https://doi.org/10.1186/1472-6920-7-49

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.