Dovepress

Taylor & Francis
Taylor & Francis Group

ADVANCES IN MEDICAL EDUCATION AND PRACTICE
OPEN ACCESS

Advances in Medical Education and Practice

ISSN: 1179-7258 (Online) Journal homepage: www.tandfonline.com/journals/dame20

# Quality of Large Language Model-Generated MCQs Across Three Medical Disciplines: An Expert Rater-Based Comparison of Gemini, GPT-4 and Perplexity Pro

**Nilesh Kumar Mitra, Thirupathirao Vishnumukkala, Thin Thin Win, Sasikala Devi Amirthalingam, Siew Kim Kwa, Gunasekar Thangarasu, Afshan Sumera & Skantha Kandiah**

**To cite this article:** Nilesh Kumar Mitra, Thirupathirao Vishnumukkala, Thin Thin Win, Sasikala Devi Amirthalingam, Siew Kim Kwa, Gunasekar Thangarasu, Afshan Sumera & Skantha Kandiah (2026) Quality of Large Language Model-Generated MCQs Across Three Medical Disciplines: An Expert Rater-Based Comparison of Gemini, GPT-4 and Perplexity Pro, Advances in Medical Education and Practice, , 618711, DOI: 10.2147/AMEP.S618711

**To link to this article:** https://doi.org/10.2147/AMEP.S618711



Published online: 19 Aug 2026.

Advances in Medical Education and Practice

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

# Quality of Large Language Model-Generated MCQs Across Three Medical Disciplines: An Expert Rater-Based Comparison of Gemini, GPT-4 and Perplexity Pro

Nilesh Kumar Mitra[1], Thirupathirao Vishnumukkala[1], Thin Thin Win[2], Sasikala Devi Amirthalingam[3], Siew Kim Kwa[3], Gunasekar Thangarasu[4], Afshan Sumera[5], Skantha Kandiah[6]

[1]Department of Human Biology, School of Medicine, IMU University, Bukit Jalil, Kuala Lumpur, 57000, Malaysia; [2]Department of Pathology and Pharmacology, School of Medicine, IMU University, Bukit Jalil, Kuala Lumpur, 57000, Malaysia; [3]Department of Family Medicine, School of Medicine, IMU University, IMU Clinical Campus, Seremban, Malaysia; [4]Department of Digital Health and Health Informatics, School of Business and Technology, IMU University, Bukit Jalil, Kuala Lumpur, 57000, Malaysia; [5]School of Medicine and Dentistry, University of Lancashire, Preston, PR1 2HE, UK; [6]Centre for Learning Anatomical Sciences, Faculty of Medicine, University of Southampton, Southampton, SO17 1BJ, UK

Correspondence: Nilesh Kumar Mitra, Department of Human Biology, School of Medicine, IMU University, Bukit Jalil, Kuala Lumpur, 57000, Malaysia, Tel +603-27317212, Fax +603-86567229, Email NileshKumar@imu.edu.my

**Background:** The increasing number of student cohorts has compelled academics to create a larger number of Multiple-Choice Question (MCQ) items. Large language models (LLMs) can help educators generate assessment items across multiple disciplines. The study compared the ability of LLMs (Gemini Advanced, Perplexity Pro, and ChatGPT 4.0) to generate high-quality, clinical-scenario-based MCQ items across three disciplines in a medical program, using an 8-point quality rubric.

**Materials and Methods:** Learning Outcomes (LOs) from Anatomy and Pathology disciplines of a pre-clinical semester 4 module and Family Medicine discipline of a clinical semester 6 module of a medical undergraduate program were selected. Using a pre-determined descriptive prompt, 63 MCQ items were generated from three AI tools. The quality of item construction was assessed by external content experts who were blinded to item generation using an 8-criterion rubric and a 4-point Likert scale. Mean scores and ranks for MCQs under each LLM were analysed, and a Friedman test was conducted to compare them. Kendall's W showed that all criteria except one demonstrated some effect and weak-to-fair inter-rater reliability.

**Results:** When measuring key problem-solving skills, Perplexity Pro-generated MCQ items received a higher percentage of "strongly agree" ratings in Anatomy (57.1%, n=63), Pathology (56.5%, n=63), and Family Medicine (71.4%, n=63). Perplexity Pro received a higher percentage of "strongly agree" ratings in Anatomy, at 47.6% (n = 63), when evaluating specific content. Gemini Advanced also scored highly, with 68.8% in Pathology and 81% in Family Medicine (n = 63). A comparative analysis of the higher mean scores and ranks across 24 quality criteria showed that Perplexity Pro, Gemini Advanced, and ChatGPT 4.0 achieved 13, 10, and 1 higher mean scores, respectively. There is no significant difference in mean scores and ranks among the three LLMs across different quality criteria.

**Conclusion:** MCQs created by Perplexity Pro and Gemini Advanced achieved comparatively higher percentages of "strongly agree" ratings across quality criteria for item construction. Assessing LLM-generated assessment items provides valuable insight into the quality of LLM-supported MCQ questions.

**Keywords:** large language models, MCQ, quality, rater-based comparison

## Introduction

Developing high-quality MCQ items for a medical undergraduate program requires time and attention to clarity, content accuracy, cognitive level, and alignment with learning outcomes.[1] To ensure the validity and reliability of the assessment tool, academics need to follow the procedures prescribed by educational institutions.[2] The assessment mapping ensures balanced

https://doi.org/10.2147/AMEP.S618711
Received: 28 April 2026
Accepted: 15 July 2026
Published: 19 August 2026

representation of the subject disciplines and student learning times (SLTs).[3] An assessment blueprint is a crucial step in improving assessment validity and ensuring constructive alignment between course learning outcomes (CLOs) and teaching methods.[4] The use of CLOs as the foundation for developing MCQ items ensures that the questions measure specific knowledge, skills, and professional competencies as per the blueprint.[5] Recent reforms in medical education, aimed at mitigating workforce shortages, have led to increased enrolment, with multiple cohorts of students admitted in a single academic year.[6] Academic staff are required to produce a higher volume of MCQ items over the course of an academic year. The increased workload can cause testing bias and lower assessment scores.[7] The demand for curriculum-aligned items has compelled the faculty to embrace the technological advancement in LLM-generated MCQ item creation and the consequent ethical and educational challenges.[8,9]

LLMs can rapidly produce large volumes of assessment items, which is appealing when faculty workload is high.[10] However, assessment quality supports fairness, validity, and accountability, and MCQs still need to map clearly to CLOs to achieve the outcome.[11–13] In parallel, LLMs such as GPT-4, Google Gemini, and other generative AI platforms have rapidly improved and are increasingly used to generate assessment items.[9,14] These systems are trained on large datasets and can generate questions that resemble expert-authored formats. Initial research indicates that LLM-assisted workflows have been reported to reduce question-generation time from 96 to 24.5 person-hours in high-stakes settings.[15] Other comparative pipeline work has shown that newer models (for example, GPT-4o) can generate quizzes quickly (about 16 seconds per question), with high topic diversity (89.8%), at a low approximate cost (~$0.51 per quiz).[14] LLMs can help educators generate assessment items on demand and at scale across multiple specialties. That said, whether LLM-generated MCQ items consistently meet quality and validity expectations remains an open question. When comparing LLM-generated items with human-authored questions, some psychometric studies report similar difficulty indices (0.67 vs 0.65) and discrimination parameters (0.29 vs 0.27), and students may not reliably identify whether an item was written by an LLM or a human.[14,15] However, other studies point to recurring weaknesses, including factual errors (6% vs 4% for human items), poor calibration of difficulty (14% vs 1% for human items), and misalignment with intended learning outcomes.[16,17] Approximately 31% of questions require considerable modifications due to issues related to content validity and insufficient clinical accuracy reasoning.[9] A recent study comparing LLM-generated and human-generated multiple-choice questions (MCQs) in medical education found the level of agreement among expert raters was moderate, with an intraclass correlation coefficient (ICC) of 0.62 (95% Confidence Interval: 0.12–0.84, p =0.01). LLM-generated questions often tested lower-order cognitive skills, whereas human-generated MCQs were better at assessing higher-order skills. This illustrates the potential of consistently applying the concept of "quality" throughout the review process for the items.[9] Educators should apply Kane's validity framework to LLM-generated MCQ items to assess whether the item scores translate into meaningful real-world decisions. Assessment of reliable scoring, generalisation to the overall cognitive domain, extrapolation to real-world competency, and implications of the test for the learning process should be evaluated by the application of human judgment.[18] One practical limitation in the current evidence is that there are still limited direct comparisons among modern LLMs across different medical disciplines. Many studies evaluate only a single model within a single specialty, and broader comparisons using the same methods and quality metrics are still limited.[14,19] This matters because commercial LLMs may not perform consistently across subject disciplines and clinical domains. At the same time, studies often use different prompts, review processes, and outcome measures, making findings hard to compare and limiting what educators can confidently take forward.[20] Because of this, a more rigorous expert review is needed to determine whether questions are appropriate for the intended learner level.[8,9,11]

This study addresses these limitations by using a structured, rater-based approach to compare MCQ items generated by three leading LLM platforms (Gemini Advanced, Perplexity Pro, and ChatGPT 4.0) across three medical disciplines (Anatomy, Pathology, and Family Medicine). The objective of this study is to compare the ability of three LLM tools (Gemini Advanced, Perplexity Pro, and ChatGPT 4.0) to generate high-quality scenario-based MCQ items across three disciplines in a medical program.

# Materials and Methods

## Ethical Approval

The study was approved by the IMU Joint Committee on Research and Ethics at IMU University, Kuala Lumpur, Malaysia, under project number 631–2024. The study complies with the Declaration of Helsinki. Informed consent was obtained from the participants who joined the study as external assessors.

## Study Design

The study included a preclinical module (Neuroscience and Locomotion 1) (NL1) taught in semester 4 and a clinical module (Primary Care Medicine) (PCM) taught in semester 6 of the medical undergraduate degree program. With the help of the module coordinators, the lesson outcomes (LOs) mapped to the module examinations were identified, and assessment blueprints were developed for the Anatomy and Pathology components of the NL1 module and the Family Medicine components of the PCM module. Twenty-one higher-order LOs from the Anatomy and Pathology disciplines of the NL 1 module and from the Family Medicine discipline of the PCM module were selected. Two content experts from each discipline reviewed the LOs to assess the inclusion of the application of basic science and clinical reasoning concepts. Including disciplines such as Anatomy, Pathology, and Family Medicine facilitated integration across disease processes, encompassing pathophysiology, related anatomical structures, and primary care.

The project followed the institutional guidelines on the use of AI in teaching and learning, and the ethical guidelines prescribing the application of human judgment before using assessment items generated by LLM tools. Three LLM tools, Gemini Advanced, Perplexity Pro, and ChatGPT 4.0, were used to generate one higher-order single-best-answer MCQ with a clinical vignette for each of the 21 selected LOs in each of the three subjects. The three LLM tools were recognized by the university administration, and the university provided paid access to them. The structured prompt (Appendix 1) included the topic, student level (program and semester), learning outcome, difficulty level, Bloom's taxonomy, and four options. The prompt also requested that the clinical vignette be produced as a single narrative paragraph, including patient details (gender and age), presenting complaint, relevant clinical history, and physical examination findings. Researchers from the Anatomy, Pathology, and Family Medicine departments collaborated to standardize the prompt. The trial process focused on generating clinical scenario-based vignettes aligned with the LOs, using plausible descriptors that were structurally parallel and free of accidental clues. The LLMs were also instructed to select the correct answer and justify their choices, including explaining why the other options were incorrect. Once standardized, the prompt was not changed by the researchers from the Anatomy, Pathology, and Family Medicine departments. The sample size could not be determined because the project needed to follow the medical program module LOs to determine the number of MCQs.

The external content experts, recruited from universities or institutes with no direct ties to the university involved in the study, were blinded to the process used to generate the MCQs from the three Gen AI tools and tasked with assessing the quality of all generated MCQ items. The quality of each of the 63 questions in Anatomy, Pathology, and Family Medicine generated by Gemini Advanced, Perplexity Pro, and ChatGPT 4.0 was assessed using a rubric adapted from Haladyna (2002) (Table 1).[21] The rubric comprised eight criteria designed to minimize item writing flaws (IWFs) that can undermine the validity and reliability of assessments. These criteria emphasize outcome-focused, single-content-specific

**Table 1** Kendall's Coefficient (W) Indicates the Effect Size and Inter-Rate Agreement for the Quality Criteria Used to Score the MCQ Items

| No | Quality Criteria | Kendall's W | Chi-Square ($\chi 2$) | df | *p*-value | Interpretation |
|---|---|---|---|---|---|---|
| 1 | The question measures specific content, as outlined in the syllabus. | 0.22 | 27.98 | 2 | 0.001 | Small effect, Fair agreement |
| 2 | The question measures key thinking and problem-solving skills. | 0.43 | 38.77 | 2 | 0.001 | Medium effect, Moderate agreement |

(*Continued*)

**Table 1** (Continued).

| No | Quality Criteria | Kendall's W | Chi-Square ($\chi 2$) | df | *p*-value | Interpretation |
|---|---|---|---|---|---|---|
| 3 | The question is carefully edited and formatted using correct grammar and spelling. | 0.14 | 17.6 | 2 | 0.065 | Small effect, Weak agreement |
| 4 | The central idea is included in the stem, not the options. | 0.44 | 45.54 | 2 | 0.001 | Medium effect, Moderate agreement |
| 5 | The stem of the question is worded positively and avoids negatives such as NOT and EXCEPT. | 0.37 | 26.88 | 2 | 0.001 | Medium effect, Fair agreement |
| 6 | Only 1 of the options is clearly correct. | 0.29 | 36.44 | 2 | 0.001 | Small effect, Fair agreement |
| 7 | The correct option is not cued by item writing errors | 0.067 | 8.43 | 2 | 0.015 | No effect, No Agreement |
| 8 | All of the distractors are plausible | 0.19 | 23.24 | 2 | 0.001 | Small effect, Weak agreement |

**Notes**: Test statistic (Chi-Square) (X2); Degrees of freedom (df); Significance level (p-value<0.05).

STEM construction, the development of plausible homogeneous options without cues, and correct formatting free of grammatical errors to ensure MCQs assess content knowledge rather than test-taking skills. The subject-specific content experts in Anatomy, Pathology, and Family Medicine were blinded to the process used to generate the MCQ items. There was no mixed scoring. Initially, each of them received 30 MCQ items (10 from each of the three LLM tools) and scored them using the 8 criteria shown in Figure 1. The scoring was conducted using a 4-point Likert scale (1 = strongly disagree, 4 = strongly agree). To ensure the rigor of the quality check process, the internal consistency of the 8-point MCQ evaluation criteria was statistically validated. Because the three content experts were assigned to distinct, subject-specific domains (Anatomy, Pathology, and Family Medicine) with no overlap in evaluations, the 4-point Likert scale

| | Criteria | Strongly Disagree (1) | Disagree (2) | Agree (3) | Strongly Agree (4) |
|---|---|---|---|---|---|
| 1. | The question measures specific content, as outlined in the syllabus. | | | | |
| 2. | The question is designed to measure key thinking and problem-solving skills. | | | | |
| 3. | The question is carefully edited and formatted using correct grammar and spelling. | | | | |
| 4. | The central idea is included in the stem, not the options. | | | | |
| 5. | The stem of the question is worded positively and avoids negatives. | | | | |
| 6. | Only 1 of the options is clearly correct. | | | | |
| 7. | The correct option is not cued by item writing errors. | | | | |
| 8. | All of the distractors are plausible. | | | | |

**Figure 1** Rubric to evaluate the quality of MCQ items generated by AI tools (adapted from Haladyna 2002 et al).[21]

responses to 30 MCQ items across all eight quality criteria for three LLM tools were evaluated using Cronbach's alpha (α). The analysis yielded a standardized Cronbach's alpha of 0.795 (unstandardised α = 0.766). These demonstrated that the 8-point quality criteria remained a reliable and coherent measurement tool.

The framework for developing test items with LLM tools is shown in Figure 2.

## Data Analysis

The quality of the 63 multiple-choice questions (MCQs) in Anatomy, Pathology, and Family Medicine, generated by the three LLM tools, was evaluated by external content experts using a 4-point Likert scale (1 = strongly disagree, 4 = strongly agree). Categorical scores for the MCQ items were compiled across Gemini Advanced, Perplexity Pro, and ChatGPT 4.0. Consequently, within each subject area, scores for a total of 63 MCQ items (21 MCQ items for each of the 21 LOs generated by the three LLMs) were recorded. Each MCQ received a single score for each of the eight criteria outlined in the quality rubric (Figure 1). Data for each criterion were analysed using IBM SPSS Statistics for Windows, version 30, for descriptive statistics, and a Friedman test was conducted to compare the mean ranks of 63 questions across three LLM tools within each discipline. Differences were considered statistically significant at $p < 0.05$. Pair-wise post-hoc comparisons (Gemini Advanced vs Perplexity Pro, ChatGPT4.0 vs Perplexity Pro, ChatGPT4.0 vs Gemini Advanced) using Bonferroni correction were conducted in case of a significant p-value. The mean score and 95% confidence interval (CI) for each of the three LLM tools against each quality criterion were tabulated. The percentage scores on the Likert scale for MCQs generated by Gemini Advanced, Perplexity Pro, and ChatGPT 4.0 were also tabulated. These 63 MCQs assessed specific knowledge and essential problem-solving skills within a similar discipline.

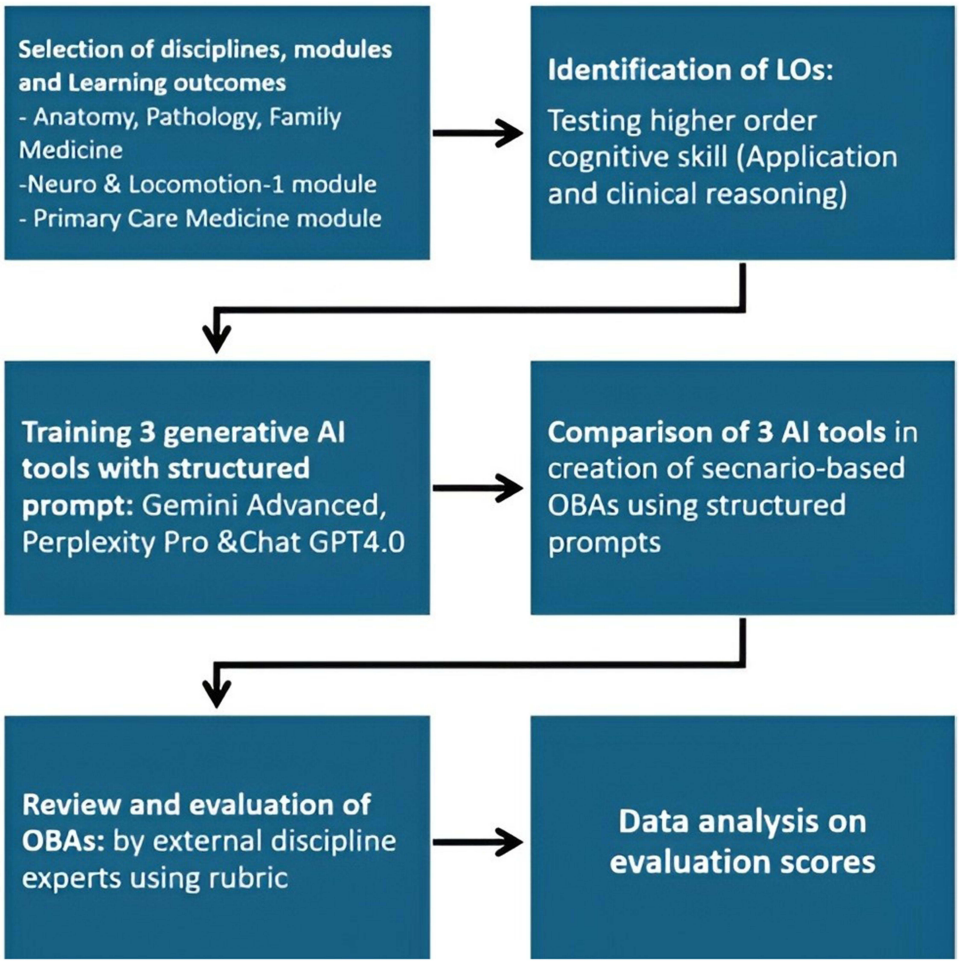


**Figure 2** The framework for developing test items with LLM tools.


https://doi.org/10.2147/AMEP.S618711

To assess inter-rater reliability among the three scorers across the 63 LLM-generated questions, Kendall's coefficient of concordance (W) was computed for each of the eight evaluation criteria. This non-parametric test was selected because of the ordinal nature of the 4-point Likert scale. Kendall's W serves directly as an effect size, with 0 indicating no consensus and 1 indicating perfect consensus. Following Cohen's (1988) guidelines, Kendall's W was used to assess the effect size and the degree of agreement among the raters.[22]

## Results

External content experts rated each MCQ item generated by LLMs on a 4-point Likert scale (strongly agree, agree, disagree, or strongly disagree) for each of the eight criteria. Kendall's Coefficient of Concordance (Kendall's W) demonstrated effect size and inter-rater reliability across all eight quality criteria (Table 1). All criteria except one showed some effect and weak-to-fair agreement. The criterion of the correct option not cued by item-writing errors showed no effect or agreement. Under the first criterion, which assesses whether the item measures specific content, Perplexity Pro-created Anatomy MCQ items received 47.6% of "strongly agree" ratings (n=63), compared with items created by Gemini Advanced (33.3%, n=63) and ChatGPT 4.0 (28.6%, n=63. Under the first criterion, 80.9% of items generated by Perplexity Pro, 76.2% of items created by Gemini, and 71.5% of items created by ChatGPT 4.0 showed agreement in assessing specific content. Under the first criterion of the Pathology discipline, Gemini Advanced items received 68.8% of strongly agree ratings (n=63). In comparison, 56.5% of Perplexity Pro-created items and 44% of ChatGPT 4.0-created items received a "strongly agree" rating. In the Family Medicine discipline, 81% of Gemini Advanced-generated items received a "strongly agree" rating, compared with 66.7% of Perplexity Pro-generated items and 47.6% of ChatGPT 4.0-generated items.

Figure 3 shows the percentage distribution of ratings for Anatomy MCQ items according to the criterion for assessing key thinking and problem-solving skills. The Perplexity Pro-created items received 57.1% of strongly agree ratings (n=63). Figure 4 shows the percentage distribution of ratings for Pathology MCQ items on the criterion for assessing key thinking and problem-solving skills. The Perplexity Pro-generated items received 56.5% of "strongly agree" ratings (56.5%, n=63). Figure 5 shows the percentage distribution of ratings for Family Medicine MCQ items in the criterion measuring key thinking and problem-solving skills. The Perplexity Pro-generated items received 71.4% of "strongly agree" ratings (71.4%, n=63). Gemini Advanced-generated MCQ items received 38.1% of "strongly agree" ratings in Anatomy (38.1%, n=63), Pathology (39.3%, n=63), and Family Medicine (61.9%, n=63) MCQ items in the criterion for measuring problem-solving skills (Figures 3–5).

A Friedman test was conducted to assess differences in mean ranks across 63 MCQ items (n=63) rated on 4-point Likert scales, each generated by Gemini Advanced, Perplexity Pro, and ChatGPT 4.0. Table 2 presents the mean score,

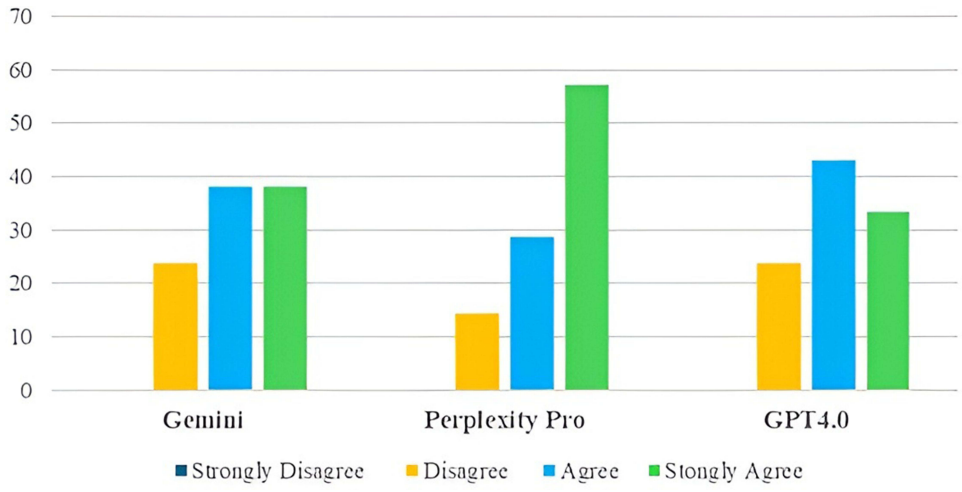


**Figure 3** Percentage of Score among 21 Anatomy MCQs Measuring Key Thinking and Problem-Solving Skills.


https://doi.org/10.2147/AMEP.S618711

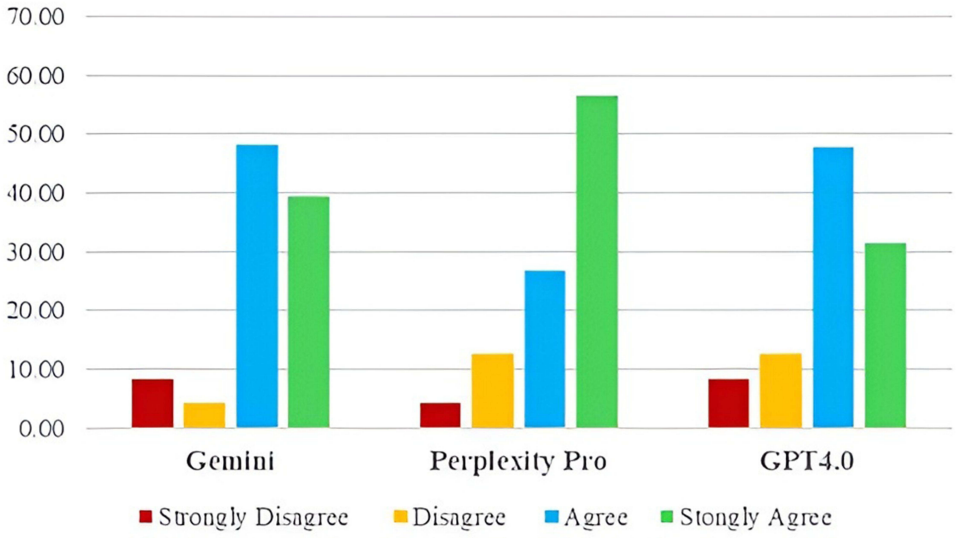


**Figure 4** Percentage of Score among 21 Pathology MCQs Measuring Key Thinking and Problem-Solving Skills.

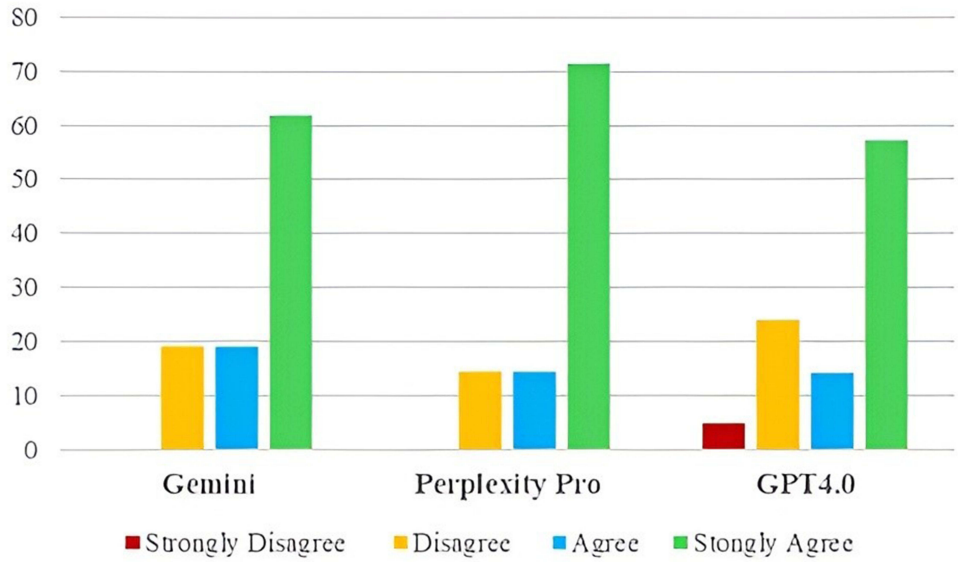


**Figure 5** Percentage of Score among 21 Family Medicine MCQs Measuring Key Thinking and Problem-Solving Skills.

mean rank, Chi-Square test statistics (X2), degrees of freedom (df), and p-values for the Anatomy discipline MCQ items, indicating the strength of differences in the ratings received by the MCQs created by the three AI tools. Analysis across the eight quality criteria showed that Perplexity Pro-created Anatomy MCQ items received higher mean scores and ranks on five of the eight criteria, with no statistically significant differences in mean scores and ranks among items created by three LLM tools.

Table 2 shows the mean score, mean rank, Chi-Square test statistics ($X^2$), degrees of freedom (df), and *p*-values for the Anatomy discipline MCQ items. Analysis across the eight quality criteria showed that Perplexity Pro-generated Anatomy MCQ items received higher mean scores and ranks for five of the eight criteria. Gemini Advanced-generated MCQ items received next-high mean scores and ranks for the remaining three of the eight criteria. However, no statistically significant differences were observed in the mean ranks of Anatomy MCQ items generated by Gemini Advanced, Perplexity Pro, and ChatGPT 4.0.


https://doi.org/10.2147/AMEP.S618711

**Table 2** Descriptive Statistics and Friedman Test results for Anatomy MCQ Items Generated from Three AI Tools (n=63). The Score for Each MCQ Item Was Provided by External Content Experts on a 4-Point Likert Scale (1=strongly Disagree; 4=strongly Agree) Based on the Quality Criteria

| No | Criteria | Gemini | | Perplexity Pro | | GPT4.0 | | $X^2$(df) | *p*-value |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean Score [95% CI] | Mean Rank | Mean Score [95% CI] | Friedman Rank | Mean Score [95% CI] | Mean Rank | | |
| 1 | The question measures specific content, as outlined in the syllabus. | 3.09 [2.74, 3.44] | 1.93 | **3.28** [2.92, 3.64] | **2.21** | 3.00 [2.65, 3.35] | 1.86 | 1.66 (2) | 0.437 |
| 2 | The question measures key thinking and problem-solving skills. | 3.14 [2.78, 3.5] | 1.90 | **3.43** [3.09, 377] | **2.26** | 3.09 [2.74, 3.44] | 1.83 | 2.73 (2) | 0.255 |
| 3 | The question is carefully edited and formatted using correct grammar and spelling. | 3.38 [3.08, 3.68] | 1.86 | **3.76** [3.56, 3.96] | **2.26** | 3.43 [3.12, 3.74] | 1.88 | 3.57 (2) | 0.168 |
| 4 | The central idea is included in the stem, not the options. | **3.62** [3.35, 3.89] | **2.07** | 3.62 [3.35, 3.89] | 2.00 | 3.52 [3.21, 3.83] | 1.93 | 0.41 (2) | 0.815 |
| 5 | The stem of the question is worded positively and avoids negatives such as NOT and EXCEPT. | 3.86 [3.69, 4.02] | 2.00 | **3.95** [3.85, 4.05] | **2.14** | 3.76 [3.56, 3.96] | 1.86 | 3.43 (2) | 0.180 |
| 6 | Only 1 of the options is clearly correct. | **3.76** [3.56, 3.96] | **2.14** | 3.52 [3.18, 3.86] | 1.88 | 3.62 [3.39, 3.84] | 1.98 | 1.77(2) | 0.412 |
| 7 | The correct option is not cued by item writing errors | 3.38 [3.01, 3.75] | 2.10 | **3.52** [3.25, 3.79] | **2.14** | 3.14 [2.81, 3.47] | 1.76 | 2.62 (2) | 0.270 |
| 8 | All of the distractors are plausible | **3.62** [3.35, 3.89] | **2.07** | 3.48 [3.1, 3.85] | 2.02 | 3.48 [3.17, 3.78] | 1.90 | 0.61 (2) | 0.739 |

**Notes**: Mean score and ranks in bold show highest value among three AI tools; 95% CI (Confidence Interval); Test statistic (Chi-Square) (X2); Degrees of freedom (df); Significance level (p-value<0.05).


https://doi.org/10.2147/AMEP.S618711

Table 3 shows the mean score, mean rank, Chi-Square test statistics ($X^2$), degrees of freedom (df), and *p*-values for the Pathology discipline MCQ items. Analysis across the eight quality criteria showed that Perplexity Pro-generated Anatomy MCQ items received higher mean scores and ranks for four of the eight criteria. Gemini Advanced-generated MCQ items received next-high mean scores and ranks for the remaining three of the eight criteria. Both Gemini Advanced and ChatGPT 4.0 achieved a similar high mean score on the fifth criterion: the stem of the question being worded positively. However, no statistically significant differences were observed in the mean ranks of Pathology MCQ items generated by Gemini Advanced, Perplexity Pro, and ChatGPT 4.0.

Table 4 presents the mean score, mean rank, Chi-square test statistics (X2), degrees of freedom (df), and *p*-values for the Family Medicine discipline MCQ items. Analysis across the eight quality criteria showed that Perplexity Pro-generated Anatomy MCQ items received higher mean scores and ranks for four of the eight criteria. Gemini Advanced-generated MCQ items received next-high mean scores and ranks across the remaining four of the eight criteria. Only one criterion (the central idea is included in the stem) showed a significant difference in mean ranks [$X^2$ (2, n=21)=10.17, p = 0.006]. However, the Bonferroni-corrected p-value for pairwise comparison between the LLM tools was not statistically significant. The Gemini Advanced-generated MCQ item received a higher mean score and rank under this criterion. Across other criteria, no statistically significant differences were observed in the mean ranks of the Family Medicine MCQ items generated by the Gemini Advanced, Perplexity Pro, and ChatGPT 4.0 tools.

The comparative analysis of higher mean scores across the cumulative data for three disciplines of Anatomy, Pathology, and Family Medicine (8 criteria applied to 3 LLMs, totaling 24), as judged by external content experts, showed that Perplexity Pro achieved 13 higher mean scores, Gemini Advanced achieved 10, and both Gemini Advanced and ChatGPT 4.0 achieved only 1. (Table 5).

## Discussion

Summative assessment in professional higher education is vital for student learning and for decision-making about academic progress. When generating items using an LLM, it is essential to explicitly define the blueprint and outcome domain in the prompt to produce relevant items. In this study, by including specific topics and clear learning outcomes expressed using higher-order action verbs aligned with the assessment blueprint, we created high-quality MCQs that reflected the lesson outcomes and reduced the risk of bias and hallucination. Studies have shown that a well-constructed blueprint can facilitate the use of LLMs to create test items with minimal effort.[23] The knowledge tested by the item should be communicated to the LLMs in the prompt for item generation using at least two dimensions (topic and learning outcome domain).[24]

The recruitment of external content experts, blinded to the generation of MCQ items on Anatomy, Pathology, and Family Medicine CLOs in this study, enabled them to evaluate the MCQ items produced by three LLMs for item-writing flaws and for relevance to the specific subject content.[25] According to the criterion for measuring specific content, Perplexity Pro-generated Anatomy MCQ items in this study received a higher percentage of "strongly agree" ratings than those from Gemini Advanced and GPT 4.0. According to the criterion for assessing key thinking and problem-solving skills, Perplexity Pro-generated MCQ items across all three disciplines (Anatomy, Pathology, and Family Medicine) received a higher percentage of "strongly agree" ratings. A similar study comparing ChatGPT, Perplexity, and DeepSeek-generated MCQ items in Haematology found that expert validation yielded a 96% acceptance rate for Perplexity items, indicating its reliability for complex-reasoning, multi-response tasks.[26] CBME requires assessing judgment and reasoning rather than merely recalling facts. LLM-generated MCQs might encourage shallow learning, with trainees performing well on tests but struggling to apply the concepts in real-world, high-stress clinical settings.[27]

Current literature recommends that the vocabulary used in MCQs be simple enough for the weakest readers in the group to comprehend. MCQ options should be formatted vertically rather than horizontally to improve readability.[28] According to the criterion of careful editing and formatting using correct grammar, both Anatomy and Family Medicine MCQ items showed that Perplexity Pro had comparatively higher percentages of "strongly agree" ratings. For Pathology MCQ items, Gemini Advanced had comparatively higher percentages of "strongly agree" ratings. The existing literature proposes a few criteria to follow when constructing the stem of an MCQ. The stem should be meaningful and should highlight a specific problem, ensuring the item focuses on assessing the learning outcome. Irrelevant information in the stem reduces the test's reliability and validity.[29] The stem should be a question, because it helps the student focus on

**Table 3** Descriptive Statistics and Friedman Test Results for Pathology MCQ Items Generated from Three AI Tools (n=63). The Score for Each MCQ Item Was Provided by External Content Experts on a 4-Point Likert Scale (1=strongly Disagree; 4=strongly Agree) Based on the Quality Criteria

| No | Criteria | Gemini | | Perplexity Pro | | GPT4.0 | | $X^2$(df) | *p*-value |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean Score [95% CI] | Friedman Rank | Mean Score [95% CI] | Friedman Rank | Mean Score [95% CI] | Friedman Rank | | |
| 1 | The question measures specific content, as outlined in the syllabus. | **3.62** [3.28, 3.95] | **2.26** | 3.48 [3.17, 3.78] | 1.71 | 3.24 [2.89, 3.59] | 2.02 | 5.02 (2) | 0.081 |
| 2 | The question measures key thinking and problem-solving skills. | 3.15 [2.73, 3.56] | 2.00 | **3.33** [2.92, 3.75] | **2.19** | 2.95 [2.53, 3.37] | 1.81 | 2.37 (2) | 0.306 |
| 3 | The question is carefully edited and formatted using correct grammar and spelling. | **3.71** [3.5, 3.92] | **2.33** | 3.43 [3.09, 3.77] | 2.00 | 3.24 [2.99, 3.48] | 1.67 | 8.34 (2) | 0.15 |
| 4 | The central idea is included in the stem, not the options. | 3.38 [2.94, 3.82] | 2.10 | **3.381** [2.96, 3.8] | **2.14** | 3.14 [2.73, 3.56] | 1.76 | 3.53 (2) | 0.171 |
| 5 | The stem of the question is worded positively and avoids negatives such as NOT and EXCEPT. | **3.95** [3.85, 4.05] | **2.02** | 3.90 [3.77, 4.04] | 1.95 | **3.95** [3.85, 4.05] | **2.02** | 0.67 (2) | 0.717 |
| 6 | Only 1 of the options is clearly correct. | 3.00 [2.86, 3.14] | 1.95 | **3.09** [2.96, 3.23] | **2.05** | 3.05 [2.95, 3.15] | 2.00 | 0.80 (2) | 0.670 |
| 7 | The correct option is not cued by item writing errors | 3.43 [3.09, 3.77] | 1.98 | **3.48** [3.05, 3.89] | **2.14** | 3.24 [2.83, 3.64] | 1.88 | 1.41 (2) | 0.494 |
| 8 | All of the distractors are plausible | **3.76** [3.52, 4.0] | **2.07** | 3.62 [3.28, 3.95] | 1.95 | 3.71 [3.46, 3.97] | 1.98 | 0.56 (2) | 0.756 |

**Notes**: Mean score and ranks in bold show highest value among three AI tools; 95% CI (Confidence Interval); Test statistic (Chi-Square) (X2); Degrees of freedom (df); Significance level (p-value<0.05).


https://doi.org/10.2147/AMEP.S618711

**Table 4** Descriptive Statistics and Friedman Test Results for Family Medicine MCQ Items Generated from Three AI Tools (n=63). The Score for Each MCQ Item Was Provided by External Content Experts on a 4-Point Likert Scale (1=strongly Disagree; 4=strongly Agree) Based on the Quality Criteria

| No | Criteria | Gemini | | Perplexity Pro | | GPT4.0 | | $X^2$(df) | *p*-value |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean Score [95% CI] | Friedman Rank | Mean Score [95% CI] | Friedman Rank | Mean Score [95% CI] | Friedman Rank | | |
| 1 | The question measures specific content, as outlined in the syllabus. | **3.72** [3.39, 4.04] | **2.24** | 3.52 [3.18, 3.86] | 2.00 | 3.28 [2.9, 3.67] | 1.76 | 4.17 (2) | 0.125 |
| 2 | The question measures key thinking and problem-solving skills. | 3.43 [3.06, 3.79] | 2.05 | **3.57** [3.23, 3.91] | **2.07** | 3.24 [2.78, 3.69] | 1.88 | 0.86 (2) | 0.649 |
| 3 | The question is carefully edited and formatted using correct grammar and spelling. | 2.95 [2.59, 3.32] | 1.98 | **3.28** [2.9, 3.67] | **2.31** | 2.52 [2.08, 2.97] | 1.71 | 5.15 (2) | 0.076 |
| 4 | The central idea is included in the stem, not the options. | **3.81**[a] [3.5, 4.12] | **2.31** | 3.67 [3.4, 3.93] | 2.02 | 3.33 [2.97, 3.69] | 1.67 | 10.17(2) | 0.006 |
| 5 | The stem of the question is worded positively and avoids negatives such as NOT and EXCEPT. | **3.91** [3.77, 4.04] | **2.05** | 3.86 [3.69, 4.02] | 1.98 | 3.86 [3.69, 4.02] | 1.98 | 0.4 (2) | 0.819 |
| 6 | Only 1 of the options is clearly correct. | **3.48** [3.03, 3.92] | **2.05** | 3.38 [2.89, 3.87] | 2.04 | 3.33 [2.94, 3.72] | 1.90 | 0.75 (2) | 0.687 |
| 7 | The correct option is not cued by item writing errors | 3.00 [2.62, 3.38] | 2.00 | **3.19** [2.79, 3.59] | **2.17** | 2.95 [2.59, 3.32] | 1.83 | 1.96 (2) | 0.375 |
| 8 | All of the distractors are plausible | 3.09 [2.72, 3.47] | 2.00 | **3.24** [2.83, 3.64] | **2.05** | 3.19 [2.85, 3.53] | 1.95 | 0.15 (2) | 0.929 |

**Notes**: Mean score and ranks in bold show highest value among three AI tools; 95% CI (Confidence Interval); Test statistic (Chi-Square) (X2); Degrees of freedom (df); Significance level (p-value<0.05); [a] Bonferroni correction (adjusted p-value not significant).

https://doi.org/10.2147/AMEP.S618711

**Table 5** Highest Mean Scores Received by Generative AI Tools Across Anatomy, Pathology and Family Medicine MCQ Items Across 24 Quality Criteria

| No | Generative AI Tool | Total No. of Highest Mean Scores Across | Quality Criteria |
|---|---|---|---|
| 1 | Gemini Advanced | 10 | The central idea is included in the stem, not the options.<br>Only 1 of the options is clearly correct.<br>All of the distractors are plausible.<br>The question measures specific content, as outlined in the syllabus<br>The question is carefully edited and formatted using correct grammar and spelling,<br>The stem of the question is worded positively and avoids negatives.<br>All of the distractors are plausible.<br>The central idea is included in the stem, not the options. |
| 2 | Perplexity Pro | 13 | The question measures specific content, as outlined in the syllabus.<br>The question measures key thinking and problem-solving skills.<br>The question is carefully edited and formatted using correct grammar and spelling.<br>The stem of the question is worded positively and avoids negatives.<br>The correct option is not cued by item writing errors.<br>The central idea is included in the stem, not the options.<br>Only 1 of the options is clearly correct.<br>The correct option is not cued by item writing errors.<br>All of the distractors are plausible |
| 3 | GPT 4.0<br>(along with Gemini Advanced) | 1 | The stem of the question is worded positively and avoids negatives. |

answering the question.[30] According to the criterion, the central idea is included in the stem and not in the options; both Anatomy and Family Medicine MCQ items showed that Gemini advanced had comparatively higher percentages of "strongly agree" ratings. Pathology MCQ items showed that the Perplexity Pro had comparatively higher percentages of "strongly agree" ratings.

The one-best-choice MCQ items often fail to measure learning outcomes due to a lack of effective distractors. When constructing high-quality scenario-based MCQ items, as practiced in this study, distractor options must be well written. A previous study has shown that educators spend a significant amount of time creating the stem, with less time spent creating plausible distractor options.[31] According to the criterion, all of the distractors are plausible; both Anatomy and pathology MCQ items showed that Gemini advanced had comparatively higher percentages of "strongly agree" ratings. Family medicine MCQ items showed that the Perplexity Pro had comparatively higher percentages of "strongly agree" ratings. A review of 734 single-best-choice MCQ items, encompassing 2,936 options, revealed that the difficulty indices rose—indicating easier items—as the number of non-functioning distractors increased.[32]

While previous studies have examined LLM-generated MCQs and compared their performance, this study focused on their use in pre-clinical and clinical modules of an MBBS program and employed evidence-based quality criteria for MCQ construction. The study included the topics and learning outcomes outlined in the module's assessment blueprint. It incorporated them into the prompt to generate 63 higher-order clinical-scenario-based MCQ items in the Anatomy and Pathology disciplines of the pre-clinical semester 4 program and in the Family Medicine disciplines of the semester 6 program, to orient LLMs toward clinical integration.

## Limitations and Future Directions

This study provided valuable insights into the quality of three LLMs for generating higher-order, scenario-based MCQs in the medical field, though it had some limitations. Other LLMs could not be used because of institutional subscriptions. The number of MCQ items per discipline was small (63), limiting the power of the data analysis and the generalizability of the

findings. The sample size could not be determined because the project needed to follow the medical program module LOs to determine the number of MCQs. The sample size was insufficient to support robust statistical inference for the eight quality criteria. Consequently, none of the comparisons reached statistical significance. During the initial validation of the quality criteria, inter-rater reliability could not be calculated due to the subject-specific assignment of the content experts. However, instrument validation was assessed by testing internal consistency (Cronbach's alpha). Although two subject disciplines (Anatomy and Pathology) were from the pre-clinical program, one (Family Medicine) was from the clinical program. The quality of MCQ item construction, as assessed on a Likert scale by external content experts, could not be corroborated by item analysis indices such as difficulty and discrimination indices. Future studies should include the use of generated MCQ items in actual assessments and corroborate the scores using item-analysis indices.

## Conclusion

The study compared the ability of three LLM tools (Gemini Advanced, Perplexity Pro, and ChatGPT 4.0) to generate high-quality, clinical-scenario-based MCQ items across three disciplines in a medical program, using an 8-point quality rubric. Perplexity Pro achieved comparatively higher percentages of "strongly agree" ratings across 13 of 24 criteria in the combined data from three subject disciplines. Gemini Advanced achieved comparatively higher percentages of "strongly agree" ratings across 10 of 24 criteria in the combined data from three subject disciplines. Both Gemini Advanced and ChatGPT4.0 achieved the highest percentage of "strongly agree" ratings for only 1 criterion. According to the 8-point criteria for assessing the quality of generated MCQ items, there is no significant difference in the mean scores and ranks of the three LLMs across the quality criteria. Evaluating LLMs will provide educators with practical, evidence-based guidance for developing AI-supported MCQ items.

## Data Sharing Statement

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

## Acknowledgments

The authors express their sincere gratitude to the IMU University Joint Committee on Research and Ethics for approving this study.

## Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

## Disclosure

The authors report no conflicts of interest in this work. There is no financial interest to be disclosed.

## References

1. Sadaf S, Khan S, Ali SK. Tips for developing a valid and reliable bank of multiple choice questions (MCQs). *Educ Health*. 2012;25(3):195–197. doi:10.4103/1357-6283.109786
2. Bridge PD, Musial J, Frank R, Roe T, Sawilowsky S. Measurement practices: methods for developing content-valid student examinations. *Med Teach*. 2003;25(4):414–421. doi:10.1080/0142159031000100337
3. Adkoli B, Deepak K. *Blueprinting in Assessment: Principles of Assessment in Medical Education*. 1st ed. New Delhi, India: Jaypee Brothers Medical Publishers Ltd; 2012:205–213.
4. Patil SY, Gosavi M, Bannur HB, Ratnakar A. Blueprinting in assessment: a tool to increase the validity of undergraduate written examinations in pathology. *Int J Appl Basic Med Res*. 2015;5(Suppl 1):S76–79. doi:10.4103/2229-516X.162286
5. Haruna-Cooper L, Rashid MA. GPT-4: the future of artificial intelligence in medical school assessments. *J R Soc Med*. 2023;116(6):218–219. doi:10.1177/01410768231181251
6. Kraakevik JA, Beck Dallaghan GL, Byerley JS, et al. Managing expansions in medical students' clinical placements caused by curricular transformation: perspectives from four medical schools. *Med Educ Online*. 2021;26(1):1857322. doi:10.1080/10872981.2020.1857322

7. Ellsworth RA, Dunnell P, Duell OK. Multiple-choice test items: what are textbook authors telling teachers? *J Educ Res*. 1990;83(5):289–293. doi:10.1080/00220671.1990.10885972
8. Kaya M, Sonmez E, Halici A, Yildirim H, Coskun A. Comparison of AI-generated and clinician-designed multiple-choice questions in emergency medicine exam: a psychometric analysis. *BMC Med Educ*. 2025;25(1):949. doi:10.1186/s12909-025-07528-6
9. 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. doi:10.1186/s12909-025-06796-6
10. Özer NE, Balcı Y, Bölükbaşı G, Ilhan B, Güneri P. Examining the role of artificial intelligence in assessment: a comparative study of chatgpt and educator-generated multiple-choice questions in a dental exam. *Eur J Dent Educ*. 2025;30:881–895. doi:10.1111/eje.70034
11. Lotto C, Sheppard SC, Anschuetz W, et al. ChatGPT generated otorhinolaryngology multiple-choice questions: quality, psychometric properties, and suitability for assessments. *OTO Open*. 2024;8(3):e70018. doi:10.1002/oto2.70018
12. Klang E, Portugez S, Gross R, et al. Advantages and pitfalls in utilising artificial intelligence for crafting medical examinations: a medical education pilot study with GPT-4. *BMC Medic Educ*. 2023;23(1):772. doi:10.1186/s12909-023-04752-w
13. Zhang Q, Huang Z, Huang Y, et al. Generative AI in medical education: feasibility and educational value of LLM-generated clinical cases with MCQs. *BMC Med Educ*. 2025;25(1):1502. doi:10.1186/s12909-025-08085-8
14. Elzayyat M, Mohammad JN, Zaqout S. Assessing LLM-generated vs expert-created clinical anatomy MCQs: a student perception-based comparative study in medical education. *Med Educ Online*. 2025;30(1):2554678. doi:10.1080/10872981.2025.2554678
15. Karahan BN, Emekli E. Comparison of applicability, difficulty, and discrimination indices of multiple-choice questions on medical imaging generated by different AI-based chatbots. *Radiography*. 2025;31(5):103087. doi:10.1016/j.radi.2025.103087
16. van Uhm J, van Haelst MM, Jansen PR. AI-powered test question generation in medical education: the DailyMed approach. *medRxiv*. 2024. doi:10.1101/2024.11.11.24317087
17. Borowitz MJ, Blackford AL, Nagelia S, Hruban RH. Large language models can generate high-quality pathology multiple-choice questions comparable with questions written by a human expert. *Mod Pathol*. 2026;39(1):100940. doi:10.1016/j.modpat.2025.100940
18. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ*. 2015;49(6):560–575. doi:10.1111/medu.12678.
19. Riehm L, Nanji K, Lakhani M, Pankiv E, Hasanee D, Pfeifer W. The use of large language models in generating multiple choice questions for health professions education: a systematic review and network meta-analysis. *PLoS One*. 2026;21(1):e0340277. doi:10.1371/journal.pone.0340277
20. Naseer MA, Nasir Y, Tabassum A, Ali S. ChatGPT-4 versus human generated multiple choice questions - A study from a medical college in Pakistan. *J Shalamar Med Dent Coll - JSHMDC*. 2024;5(2):58–64. doi:10.53685/jshmdc.v5i2.253
21. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. *Appl Meas Educ*. 2002;15(3):309–333. doi:10.1207/S15324818AME1503_5
22. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale NJ: Lawrence Erlbaum Associates Publishers; 1988.
23. Stadler M, Horrer A, Fischer MR. Crafting medical MCQs with generative AI: a how-to guide on leveraging ChatGPT. *GMS J Med Educ*. 2024;41(2):Doc20. doi:10.3205/zma001675
24. Magzoub ME, Zafar I, Munshi F, Shersad F. Ten tips to harnessing generative AI for high-quality MCQS in medical education assessment. *Med Educ Online*. 2025;30(1):2532682. doi:10.1080/10872981.2025.2532682
25. Smeby SS, Lillebo B, Gynnild V, et al. Improving assessment quality in professional higher education: could external peer review of items be the answer? *Cogent Med*. 2019;6(1):1659746. doi:10.1080/2331205X.2019.1659746
26. Boufrikha W, Sallem A, Laabidi B, et al. Evaluation of three artificial intelligence chatbots for generating clinical haematology multiple-choice questions for medical students. *Sci Rep*. 2026;16(1):5802. doi:10.1038/s41598-026-36839-x
27. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. *Med Teach*. 2024;46(4):446–470. doi:10.1080/0142159X.2024.2314198
28. Haladyna TM. *Developing and Validating Multiple-Choice Test. items*.3rd ed. UK: Routledge; 2004. DOI:10.4324/9780203825945
29. Haladyna TM, Downing SM. A taxonomy of multiple-choice item-writing rules. *App Meas Educ*. 1989;2(1):37–50. doi:10.1207/S15324818AME0201_3
30. Statman S. Ask a clear question and get a clear answer: an enquiry into the question/answer and sentence completion formats of multiple choice items. *System*. 1988;16(3):367–376. doi:10.1016/0346-251X(88)90079-6
31. Gierl MJ, Bulut O, Guo Q, Zhang X. Developing,analysing, and using distractors for multiple-choice tests in education: a comprehensive review. *Rev Educ Res*. 2017;87(6):1082–1116. doi:10.3102/0034654317726529
32. Shakurnia A, Ghafourian M, Khodadadi A, et al. Evaluating functional and non-functional distractors and their relationship with difficulty and discrimination indices in four-option multiple-choice questions. *Educ Med J*. 2022;14(4):55–62. doi:10.21315/eimj2022.14.4.5