Original Paper

Fine-Tuned Large Language Models for Generating Multiple-Choice Questions in Anesthesiology: Psychometric Comparison With Faculty-Written Items

Carlos Ramon Hölzing[1], MD; Charlotte Meynhardt[1]; Patrick Meybohm[1], MD; Sarah König[2], MD; Peter Kranke[1], MD

[1]Department of Anaesthesiology, Intensive Care, Emergency and Pain Medicine, University Hospital Würzburg, Würzburg, Germany
[2]Institute of Medical Teaching and Medical Education Research, University Hospital Würzburg, Würzburg, Germany

Corresponding Author:

Carlos Ramon Hölzing, MD
Department of Anaesthesiology, Intensive Care
Emergency and Pain Medicine, University Hospital Würzburg
Oberdürrbacher Str. 6
Würzburg 97080
Germany
Email: hoelzing_c@ukw.de

Abstract

Background: Multiple-choice examinations (MCQs) are widely used in medical education to ensure standardized and objective assessment. Developing high-quality items requires both subject expertise and methodological rigor. Large language models (LLMs) offer new opportunities for automated item generation. However, most evaluations rely on general-purpose prompting, and psychometric comparisons with faculty-written items remain scarce.

Objective: This study aimed to evaluate whether a fine-tuned LLM can generate MCQs (Type A) in anesthesiology with psychometric properties comparable to those written by expert faculty.

Methods: The study was embedded in the regular written anesthesiology examination of the eighth-semester medical curriculum with 157 students. The examination comprised 30 single best-answer MCQs, of which 15 were generated by senior faculty and 15 by a fine-tuned GPT-based model. A custom GPT-based (GPT-4) model was adapted with anesthesiology lecture slides, the National Competence-Based Learning Objectives Catalogue (NKLM 2.0), past examination questions, and faculty publications using supervised instruction-tuning with standardized prompt–response pairs. Item analysis followed established psychometric standards.

Results: In total, 29 items (14 expert, 15 LLM-generated) were analyzed. Expert-generated questions had a mean difficulty of 0.81 (SD 0.19), point-biserial correlation of 0.19 (SD 0.07), and discrimination index of 0.09 (SD 0.08). LLM-generated items had a mean difficulty of 0.79 (SD 0.18), point-biserial correlation of 0.17 (SD 0.04), and discrimination index of 0.08 (SD 0.11). Mann-Whitney *U* tests revealed no significant differences between expert- and LLM-generated items for difficulty (*P*=.38), point-biserial correlation coefficient (*P*=.96), or discrimination index (*P*=.59). Categorical analyses confirmed no significant group differences. Both sets, however, showed only modest psychometric quality.

Conclusions: Supervised fine-tuned LLMs are capable of generating MCQs with psychometric properties comparable to those written by experienced faculty. Given the limitations and cohort-dependency of psychometric indices, automated item generation should be considered a complement rather than a replacement for manual item writing. Further research with larger item sets and multi-institutional validation is needed to confirm generalizability and optimize integration of LLM-based tools into assessment development.

JMIR Form Res 2026;10:e84904; doi: 10.2196/84904

Keywords: medical education; multiple-choice questions; large language models; fine-tuning; psychometrics; assessment; anesthesiology; artificial intelligence

Introduction

Multiple-choice questions (MCQs) are fundamental to the objective assessment of medical students. They allow standardized testing across large cohorts and play a central role in evaluating foundational and applied knowledge [1]. However, the development of high-quality MCQs demands not only deep domain knowledge but also significant methodological and didactic expertise [2,3]. Effective items must balance appropriate difficulty, plausible distractors, minimal cueing, and strong discriminatory power to differentiate between varying levels of student performance [4].

Recent advances in artificial intelligence (AI), particularly large language models (LLMs), offer novel tools for automated question generation. For instance, efforts comparing ChatGPT-3.5–generated MCQs with expert-written items in neurophysiology revealed similar difficulty levels but lower discriminatory power in LLM-generated questions [5]. A systematic review of LLM use in medical MCQ generation found that while LLMs can produce examination-relevant items, many require additional modification due to quality issues [6]. Other studies highlight linguistic and structural shortcomings in automatically generated MCQs, particularly regarding distractor plausibility and alignment with instructional content [7,8].

Recent domain-specific efforts such as Hypnos [9], CDGen [10], and the Chinese Anesthesiology Benchmark [11] have demonstrated that LLMs can be effectively fine-tuned or benchmarked within anesthesiology. However, these studies primarily focus on domain adaptation and benchmark performance rather than psychometric validation of automatically generated examination items. To address this gap, a GPT-based model was adapted using anesthesia-specific teaching materials, the National Competence-Based Learning Objectives Catalogue in Medicine (NKLM 2.0), past examination items, and faculty publications [12]. Item development for both expert- and AI-generated questions was systematically mapped to the NKLM 2.0, Bloom's taxonomy, and the local examination blueprint to ensure comprehensive curricular coverage and to allow a fair psychometric comparison.

This study aimed to evaluate whether a fine-tuned LLM can generate MCQs (Type A) in anesthesiology with psychometric properties comparable to those written by expert faculty.

Methods

Overview

This study analyzed the performance of MCQs used in the regular written examinations of anesthesiology in the eighth semester of medical training with 157 students. The examination consisted of 30 items. Half of the items (n=15) were written by senior faculty members, and half (n=15) were generated by a fine-tuned LLM. Nine faculty members from the Department of Anesthesiology, each with at least 10 years of experience, participated in item creation. All had prior training in assessment design through institutional workshops on multiple-choice item writing. In addition, all items were independently reviewed by an educational specialist with a Master of Medical Education degree to ensure adherence to established item-writing principles. Faculty were aware of the study but blinded to the psychometric comparison during data collection.

Data analysis was performed fully anonymously. The participating students were regular medical students in their eighth semester. They were not informed about the origin of the examination questions and therefore did not know whether an item was generated by faculty or the LLM.

A customized GPT-based model was developed specifically for this study. The model was built as a domain-adapted instance of GPT-3.5-Turbo, configured to generate single-best-answer MCQs. Adaptation followed a supervised instruction-tuning approach: several hundred standardized prompt-response pairs were created using anesthesiology lecture slides, NKLM 2.0, past examination questions, and faculty publications. Faculty publications were included to capture authentic domain phrasing and ensure that the model reflected institution-specific conceptualizations of anesthetic procedures. Previous research shows that faculty development and the use of high-quality source material improve item validity and discrimination [13].

These materials were curated to align with Bloom's taxonomy and national curricular requirements [14]. The fine-tuning pipeline can be found in Multimedia Appendix 1.

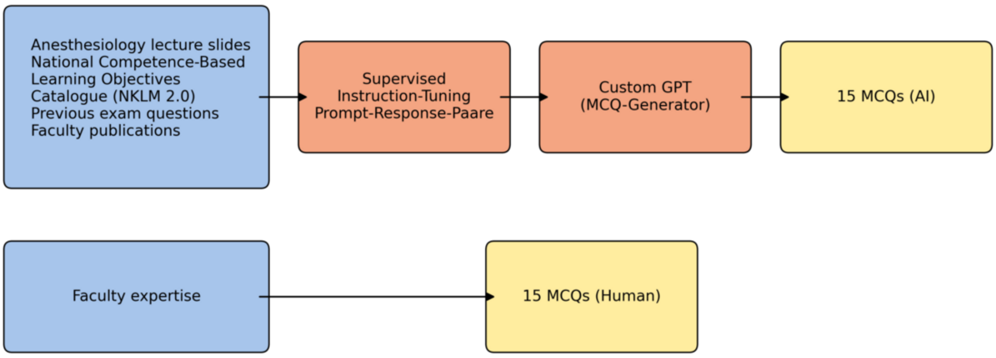
Item analysis followed established psychometric standards. Difficulty was defined as the mean proportion of correct responses (0-1). Values between 0.30 and 0.70 are generally considered optimal, those greater than 0.70 indicate easy items, and those less than 0.30 indicate difficult items [15,16]. The point-biserial correlation was classified as follows: negative correlation (r<0), very low correlation (0≤r<0.10), low correlation (0.10≤r≤0.20), and acceptable correlation (r≥0.20) [15]. The discrimination index (D) was calculated as the difference in difficulty between the upper and lower 27% performance groups, with values of 0.40 and above considered excellent; 0.30-0.39, good; 0.20-0.29, acceptable; and those less than 0.20, poor [15,16]. Statistical analysis was performed using SPSS Statistics version 27 (IBM Corp). Graphs were created with Prism 9 (GraphPad Software). Nominal variables were summarized as counts and percentages. The Shapiro-Wilk test was used to test for normal distribution. Group comparisons of categorical data were performed with the chi-square test or Fisher exact test if expected frequencies were less than 5. Continuous data were reported as mean and SD values and compared using the Mann-Whitney U test. A significance level of P≤.05 was applied.

Figure 1 summarizes the item-generation workflow, including consolidated inputs, supervised instruction-tuning, the custom GPT MCQ generator, and parallel faculty-written items converging into the examination.

Ethical Considerations

The study was submitted to the Ethics Committee of the University of Würzburg, which confirmed (reference number 2024--258-ka on November 11, 2024) that no formal review was required and that no ethical objections were raised.

Figure 1. Item-generation workflow. Consolidated inputs (anesthesiology lecture slides, NKLM 2.0, prior examination items, faculty publications) inform supervised instruction-tuning of a custom GPT configured for single-best-answer MCQs to produce 15 AI-generated questions. In parallel, faculty authored 15 questions. AI: artificial intelligence; MCQ: multiple-choice question.



Results

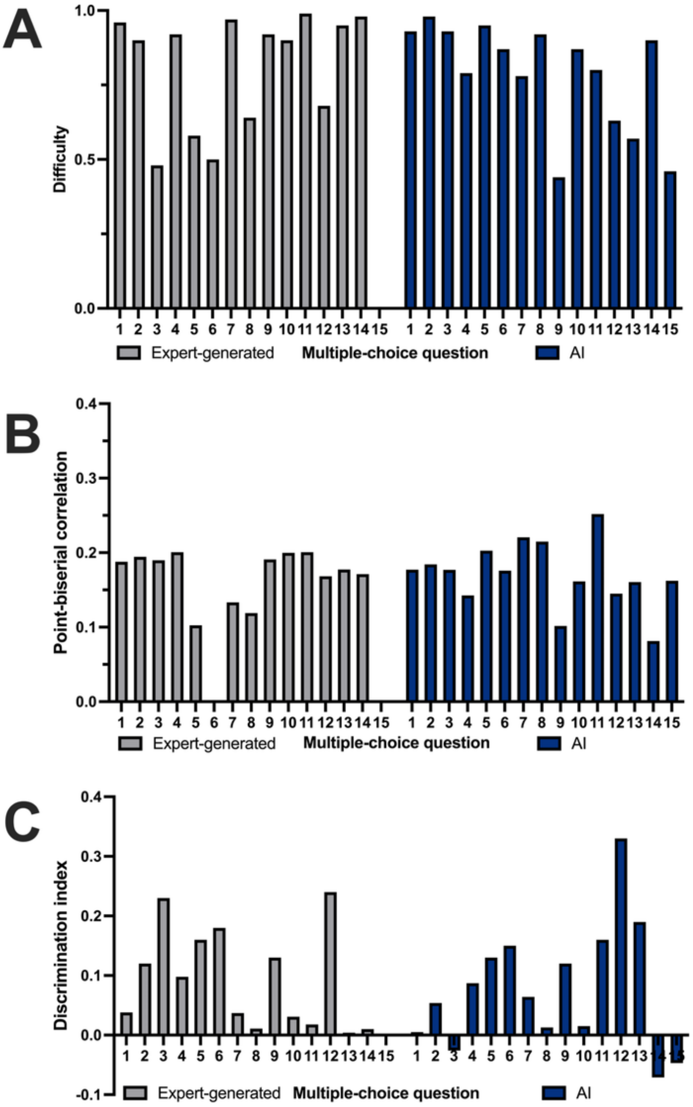
A total of 30 MCQs were analyzed. One expert-generated item was excluded from analysis due to a strongly negative discrimination index (–0.22) and negative point-biserial correlation (–0.20). Its difficulty (*P*=.86) indicated a ceiling effect, suggesting that most students answered it correctly despite unclear key wording.

The final dataset therefore included 14 expert-generated and 15 AI-generated items. Table 1 displays the descriptive metrics for expert- and AI-generated items. Expert-generated items showed a mean difficulty of 0.81 (SD 0.19), a mean point-biserial correlation of 0.16 (SD 0.07), and a mean discrimination index of 0.09 (SD 0.08). AI-generated items had a mean difficulty of 0.79 (SD 0.18), a mean point-biserial correlation of 0.17 (SD 0.04), and a mean discrimination index of 0.08 (SD 0.11). Mann-Whitney *U* tests indicated no significant differences between expert- and AI-generated items with respect to difficulty (*P*=.38), point-biserial correlation (*P*=.96), or discrimination index (*P*=.59; Figure 2).

Table 1. Overview of question metrics by expert and artificial intelligence (AI).

Question created by	Questions, n	Metrics		
		Minimum	Maximum	Difficulty, mean (SD)
Expert				
Difficulty	14	0.48	0.99	0.81 (0.19)
Point-biserial correlation	14	–0.02	0.20	0.16 (0.07)
Discrimination index	14	0.01	0.24	0.09 (0.08)
AI				
Difficulty	15	0.44	0.98	0.79 (0.18)
Point-biserial correlation	15	0.08	0.25	0.17 (0.04)
Discrimination index	15	–0.07	0.33	0.08 (0.11)

Figure 2. Psychometric characteristics of AI- and expert-generated multiple-choice items. (A) Item difficulty, (B) point-biserial correlation, and (C) discrimination index is displayed for each question. AI: artificial intelligence.



Reference ranges are as follows:

- Difficulty (P)=0.30-0.70 is considered desirable;
- Discrimination (r_{pb})≥0.20 is considered acceptable;
- r_{pb}<0 indicates flawed items [17].

The categorical distributions of item properties are summarized in Table 2.

Table 2. Psychometric characteristics of expert- and LLM[a]-generated items displayed side by side for difficulty, point-biserial correlation, and discrimination index.

	Questions, n (%)	
	Expert-generated (n=14)	LLM-generated (n=15)
Difficulty categories		
Extremely difficult (P≤0.25)	0 (0)	0 (0)
Tends to heavy (0.25<P≤0.4)	0 (0)	0 (0)
Optimal difficulty (0.4<P≤0.8)	5 (35.7)	6 (40.0)
Very simple (0.8<P≤0.9)	0 (0)	3 (20.0)
Extremely simple (P>0.9)	9 (64.3)	6 (40.0)

	Questions, n (%)	
	Expert-generated (n=14)	LLM-generated (n=15)
Point-biserial correlation categories		
Negative correlation (*r*<0)	1 (7.1)	0 (0)
Very low correlation (0≤*r*<0.10)	0 (0)	1 (6.7)
Low correlation (0.1≤r≤0.20)	8 (57.1)	10 (66.7)
Acceptable correlation (*r*≥0.20)	5 (35.7)	4 (26.7)
Discrimination index categories		
Urgent need for revision (D'<0)	0 (0)	3 (20.0)
Need for revision (0≤D'<0.2)	12 (85.7)	11 (73.3)
Check required (0.2≤D'<0.3)	2 (14.3)	0 (0)
Potential for improvement (0.3≤D'<0.4)	0 (0)	1 (6.7)
Item effectively distinguishes (D'≥0.4)	0 (0)	0 (0)

[a]LLM: large language model.

Discussion

Principal Findings

In this study, we compared psychometric properties (difficulty, point-biserial correlation, and discrimination) of MCQs generated by a supervised fine-tuned LLM with those written by expert faculty in an undergraduate anesthesiology examination. Although no statistically significant differences were observed, the overall quality of both item sets remained moderate. The point-biserial correlations and discrimination indices suggest that neither set reliably distinguishes higher- from lower-performing students, a finding consistent with previous research indicating that even expert-authored items often underperform in psychometric analyses [18]. This pattern aligns with broader evidence in medical education, where cohort studies have demonstrated that AI-generated MCQs often achieve discrimination indices similar to expert-generated items but tend to be easier overall and still require expert review to ensure distractor plausibility and alignment with higher-order learning objectives [5,19-23].

Supervised adaptation with domain-specific materials likely contributed to the close alignment of psychometric indices between AI and faculty-written items. Other work shows that when AI-mediated question generation is guided by domain content, structured prompts, or instruction tuning, the output more closely resembles faculty items in both difficulty and discrimination [7,8,19,24]. Notably, neither item set in our study consistently achieved high point-biserial correlation or discrimination, confirming that generating functionally effective distractors remains a challenge for both experts and LLMs [25-30]. Prior studies have similarly identified that AI items often underperform in assessing higher cognitive levels or using plausible distractors without ambiguity [8].

The absence of psychometric superiority in either group suggests that AI-assisted question generation can produce items of comparable statistical quality to traditional item writing. However, psychometric analysis alone is insufficient for examination quality assurance; human oversight remains essential to safeguard content validity, blueprint alignment, and cognitive level coverage. Studies in high-stakes examination settings show that expert review reduces factual inaccuracies and improves alignment with assessment blueprints [31]. Importantly, in our study, the fine-tuned LLM generated all 15 candidate items within a few minutes. While we evaluated only a subset psychometrically, our study demonstrates that domain-adapted LLMs support rapid item drafting at scale. Automatic item generation methods have long promised efficiency gains by expanding item pools from templates rather than crafting each item manually [32]. Recent AI studies show that LLM-based MCQ generation can approach human performance while drastically reducing human effort [33]. In practice, educators may use LLM throughput to generate large candidate sets and then filter, refine, and align items to the blueprint and cognitive levels, shifting effort from generation toward qualitative review and validation.

Limitations and Future Work

Study limitations include a small item sample size, single-institution administration, and fine-tuning with primarily local teaching resources, which may reduce external validity. Cognitive level of items (eg, recall vs application) was not measured, although comparative studies indicate this is an important differentiator between AI- vs expert-generated MCQs [31]. Future work should involve larger item pools, multi-institutional validation, and systematic qualitative review of items, including stem clarity, distractor plausibility, and distractor efficiency, as well as cognitive demand. It would also be valuable to compare different fine-tuning or prompt-engineering strategies and to assess students' perceptions of AI-generated items [34].

Conclusion

This study demonstrates that a supervised fine-tuned LLM can generate MCQs with psychometric properties comparable to those created by experienced faculty. While neither approach consistently produced items with high point-biserial correlation or discrimination, the results indicate that automated question generation can complement traditional item writing in medical education.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Technical workflow and dataset construction for fine-tuned large language model–mediated item generation.
[PDF File (Adobe File), 165 KB-Multimedia Appendix 1]

References

1. St-Onge C, Young M, Renaud JS, Cummings BA, Drescher O, Varpio L. Sound practices: An exploratory study of building and monitoring multiple-choice exams at Canadian undergraduate medical education programs. Acad Med. Feb 1, 2021;96(2):271-277. [doi: 10.1097/ACM.0000000000003659] [Medline: 32769474]
2. Sideris GA, Singh A, Catanzano T. Writing High-Quality Multiple-Choice Questions. In: Image-Based Teaching. Springer; 2022:123-146. [doi: 10.1007/978-3-031-11890-6_9]
3. Collins J. Education techniques for lifelong learning: writing multiple-choice questions for continuing medical education activities and self-assessment modules. Radiographics. 2006;26(2):543-551. [doi: 10.1148/rg.262055145] [Medline: 16549616]
4. Caldwell DJ, Pate AN. Effects of question formats on student and item performance. Am J Pharm Educ. May 13, 2013;77(4):71. [doi: 10.5688/ajpe77471] [Medline: 23716739]
5. Laupichler MC, Rother JF, Grunwald Kadow IC, Ahmadi S, Raupach T. Large language models in medical education: Comparing ChatGPT- to human-generated exam questions. Acad Med. May 1, 2024;99(5):508-512. [doi: 10.1097/ACM.0000000000005626] [Medline: 38166323]
6. Artsi Y, Sorin V, Konen E, Glicksberg BS, Nadkarni G, Klang E. Large language models for generating medical examinations: systematic review. BMC Med Educ. Mar 29, 2024;24(1):354. [doi: 10.1186/s12909-024-05239-y] [Medline: 38553693]
7. Grévisse C, Pavlou MAS, Schneider JG. Docimological quality analysis of LLM-generated multiple choice questions in computer science and medicine. SN Comput Sci. 2024;5(5):636. [doi: 10.1007/s42979-024-02963-6]
8. Al Shuraiqi S, Aal Abdulsalam A, Masters K, Zidoum H, AlZaabi A. Automatic generation of medical case-based multiple-choice questions (MCQs): A review of methodologies, applications, evaluation, and future directions. BDCC. 2024;8(10):139. [doi: 10.3390/bdcc8100139]
9. Wang Z, Jiang J, Zhan Y, et al. Hypnos: A domain-specific large language model for anesthesiology. Neurocomputing. Apr 2025;624:129389. [doi: 10.1016/j.neucom.2025.129389]
10. Li Y, Zhan Y, Yu B, et al. Fine-tuning LLMs for anesthesiology via compositional data generation. IEEE Trans Emerg Top Comput Intell. 2025;9(6):4051-4065. [doi: 10.1109/TETCI.2025.3567602]
11. Zhou B, Zhan Y, Wang Z, et al. Benchmarking medical LLMs on anesthesiology: A comprehensive dataset in Chinese. IEEE Trans Emerg Top Comput Intell. 2025;9(4):3057-3071. [doi: 10.1109/TETCI.2024.3502465]
12. Fakultätentag M. Nationaler Kompetenzbasierter Lernzielkatalog Medizin. 2025. URL: https://nklm.de/menu [Accessed 2026-02-10]
13. Naeem N, van der Vleuten C, Alfaris EA. Faculty development on item writing substantially improves item quality. Adv Health Sci Educ Theory Pract. Aug 2012;17(3):369-376. [doi: 10.1007/s10459-011-9315-2] [Medline: 21837548]
14. Bloom BS, et al. Taxonomy of Educational Objectives. Vol 2. Longmans; 1964.
15. Escudero EB, Reyna NL, Morales MR. The level of difficulty and discrimination power of the Basic Knowledge and Skills Examination (EXHCOBA). Revista Electrónica de Investigación Educativa. 2000;2(1):2. URL: https://redie.uabc.mx/redie/article/download/15/27/75 [Accessed 2026-02-10]
16. Möltner A, Schellberg D, Jünger J. Grundlegende quantitative analysen medizinischer prüfungen. GMS Zeitschrift für Medizinische Ausbildung; 2006.
17. Moran V. Item and Exam Analysis, in Item Writing for Nurse Educators. Springer International Publishing; 2023:55-64.
18. Rush BR, Rankin DC, White BJ. The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. BMC Med Educ. Sep 29, 2016;16(1):250. [doi: 10.1186/s12909-016-0773-3] [Medline: 27681933]
19. Cheung BHH, Lau GKK, Wong GTC, et al. ChatGPT versus human in generating medical graduate exam multiple choice questions-A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). PLOS ONE. 2023;18(8):e0290691. [doi: 10.1371/journal.pone.0290691] [Medline: 37643186]

20. E K, S P, R G, et al. Advantages and pitfalls in utilizing artificial intelligence for crafting medical examinations: A medical education pilot study with GPT-4. BMC Med Educ. Oct 17, 2023;23(1):772. [doi: 10.1186/s12909-023-04752-w] [Medline: 37848913]
21. Ayub I, Hamann D, Hamann CR, Davis MJ. Exploring the potential and limitations of Chat Generative Pre-trained Transformer (ChatGPT) in generating board-style dermatology questions: A qualitative analysis. Cureus. Aug 2023;15(8):e43717. [doi: 10.7759/cureus.43717] [Medline: 37638266]
22. 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. Jul 1, 2025;25(1):949. [doi: 10.1186/s12909-025-07528-6] [Medline: 40597998]
23. 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. Dec 2025;30(1):2554678. [doi: 10.1080/10872981.2025.2554678] [Medline: 40884796]
24. Emekli E, Karahan BN. AI in radiography education: Evaluating multiple-choice questions difficulty and discrimination. J Med Imaging Radiat Sci. Jul 2025;56(4):101896. [doi: 10.1016/j.jmir.2025.101896] [Medline: 40157013]
25. Bitew SK, et al. Distractor generation for multiple-choice questions with predictive prompting and large language models. Presented at: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases; Sep 18-22, 2023; Torino, Italy. [doi: 10.1007/978-3-031-74627-7_4]
26. Baldwin P, Mee J, Yaneva V, et al. A natural-language-processing-based procedure for generating distractors for multiple-choice questions. Eval Health Prof. Dec 2022;45(4):327-340. [doi: 10.1177/01632787211046981] [Medline: 34753326]
27. Wang R, et al. High-quality distractors generation for human exam based on reinforcement learning from preference feedback. In: Natural Language Processing and Chinese Computing. Springer Nature Singapore; 2025. [doi: 10.1007/978-981-97-9440-9_8]
28. Zhang L, VanLehn K. Evaluation of auto-generated distractors in multiple choice questions from a semantic network. Interactive Learning Environments. Aug 18, 2021;29(6):1019-1036. [doi: 10.1080/10494820.2019.1619586]
29. Abdulghani H, Ahmad F, Aldrees A, Khalil M, Ponnamperuma G. The relationship between non-functioning distractors and item difficulty of multiple choice questions: A descriptive analysis. J Health Spec. 2014;2(4):148. [doi: 10.4103/1658-600X.142784]
30. Rezigalla AA, Eleragi A, Elhussein AB, et al. Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. BMC Med Educ. Apr 24, 2024;24(1):445. [doi: 10.1186/s12909-024-05433-y] [Medline: 38658912]
31. 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. Feb 8, 2025;25(1):208. [doi: 10.1186/s12909-025-06796-6] [Medline: 39923067]
32. Embretson SE, Kingston NM. Automatic item generation: A more efficient process for developing mathematics achievement items? J Educational Measurement. Mar 2018;55(1):112-131. URL: https://onlinelibrary.wiley.com/toc/17453984/55/1 [Accessed 2026-02-10] [doi: 10.1111/jedm.12166]
33. Olney AM. Generating Multiple Choice Questions from a Textbook: LLMs Match Human Performance on Most Metrics. Grantee Submission; 2023.
34. Ng O, et al. Student Perspective Matters for GenAI in Question Setting in Medical Education. Medical Science Educator; 2025. [doi: 10.1007/s40670-025-02396-7]

Abbreviations

AI: artificial intelligence
LLM: large language model
MCQ: multiple-choice question
NKLM 2.0: National Competence-Based Learning Objectives Catalogue in Medicine

Edited by Amaryllis Mavragani; peer-reviewed by Assad Ali Rezigalla, Liang Ding; submitted 26.Sep.2025; final revised version received 18.Dec.2025; accepted 24.Dec.2025; published 18.Feb.2026

Please cite as:
Hölzing CR, Meynhardt C, Meybohm P, König S, Kranke P
Fine-Tuned Large Language Models for Generating Multiple-Choice Questions in Anesthesiology: Psychometric Comparison With Faculty-Written Items
JMIR Form Res 2026;10:e84904

URL: https://formative.jmir.org/2026/1/e84904
doi: 10.2196/84904