Galore International Journal of Health Sciences and Research
[Indexed, Peer-Reviewed Journal]
Vol. 11; Issue: 2; April-June 2026
Website: www.gijhsr.com
Original Research Article *P-ISSN: 2456-9321*

The Pocket Tutor - Generating Competency-Based Clinical Vignettes with Artificial Intelligence for Medical Postgraduate Competitive Exam Preparation

Swapnil Banerjee[1], Manisha Agarwal[2]

[1]Final Year MBBS Student, Dr. KNS Memorial Institute of Medical Sciences, Barabanki, IND
[2]Department of Obstetrics & Gynaecology, Dr. KNS Memorial Institute of Medical Sciences, Barabanki, IND

Corresponding Author: Swapnil Banerjee

DOI: https://doi.org/10.52403/gijhsr.20260208

ABSTRACT

The transition of medical licensing examinations toward application-based clinical vignettes necessitates high-quality question banks. However, manual drafting of complex multiple-choice questions places a profound cognitive burden on educators. This study evaluates the efficacy of constrained Large Language Models to generate standardized, high-yield medical assessments. A cross-sectional, dual-cohort study was conducted involving undergraduate medical students (n=230) and senior medical faculty (n=32). A constrained prompt was engineered using Gemini 3 Pro to generate competency-based clinical vignettes. Participants evaluated the AI-generated content via digital surveys. A blinded Turing Test, embedding authentic past year questions among AI modules, assessed indistinguishability. Expert faculty rated clinical accuracy highly (Mean=4.38±0.71), with 96.9% certifying the content as clinically safe. The student cohort reported strong exam parity, with 81.3% finding the AI difficulty aligned with standard examinations. In the Turing Test, 75.0% of faculty and 47.0% of students could not distinguish AI-generated vignettes from human-authored questions (p=0.005). Furthermore, 90.0% of students desired to integrate the tool into exam preparation, while 93.8% of faculty considered it a viable drafting aid. Highly constrained artificial intelligence can successfully architect structurally sound, competency-based clinical assessments. By passing a clinical Turing Test among educators, this methodology serves as a reliable mock-testing tool for trainees and a time-saving resource for medical faculty.

Key Words: Artificial intelligence, Competency-based assessment, large language models, medical education, Undergraduate training.

INTRODUCTION

The integration of Artificial Intelligence (AI), particularly Large Language Models (LLMs), has initiated a paradigm shift in healthcare and medical education.[1,2] Following the public release of generative AI tools like Gemini, Grok, ChatGPT, these models have demonstrated remarkable natural language comprehension and reasoning capabilities, most notably evidenced by their ability to achieve passing-level performance on standardized, high-stakes examinations such as the United States Medical Licensing Examination (USMLE) without specialized, domain-specific training.[3-5] Consequently, the potential

applications of LLMs in medical training—ranging from virtual patient simulations to automated educational content generation—are being rapidly explored.[6]

Objective assessment through multiple-choice questions (MCQs), specifically those utilizing complex clinical vignettes, remains the cornerstone of evaluating medical students' competency and clinical reasoning.[7,8] However, authoring high-quality, competency-based MCQs that accurately reflect real-world clinical practice is a notoriously challenging, time-consuming, and resource-intensive process for medical educators. Recent audits indicate that drafting 100 new clinical MCQs can require approximately 96 person-hours of expert faculty time. While early studies suggest that generative AI can significantly reduce this workload and produce examination-relevant items,[7,8] the reliability, pedagogical value, and clinical safety of AI-generated questions demand rigorous evaluation.[9]

Although current literature indicates that LLMs possess the capability to generate medical MCQs, there is a consensus that these AI tools should serve as supplementary aids rather than autonomous creators; AI-generated outputs can occasionally be misleading or flawed, making moderation by human experts mandatory. Furthermore, there is a critical need to evaluate whether these AI-generated vignettes align with the specific difficulty, distractor quality, and formatting standards of competitive postgraduate entrance examinations, such as the NEET-PG and INI-CET. To address this gap, this study introduces and evaluates an AI-driven mock assessment pipeline designed to generate competency-based clinical vignettes across all four years of the Bachelor of Medicine, Bachelor of Surgery (MBBS) curriculum. By subjecting these AI-generated vignettes to a dual-evaluation framework—incorporating both large-scale student perception analytics and rigorous peer-review by senior medical faculty—this study aims to assess the clinical accuracy and educational utility of the tool. Additionally, we implemented a blinded "Turing Test" by embedding authentic Past Year Questions (PYQs) within the AI-generated assessment to determine if educators and learners could distinguish between human-authored and AI-generated clinical scenarios. Ultimately, this study explores the viability of a human-AI "co-pilot" model for streamlining high-yield medical assessment generation.

This article was previously presented as a meeting abstract at the National Conference on Artificial Intelligence held at Dr. KNS Memorial Institute of Medical Sciences, Barabanki, on March 27-28, 2026.

The primary aim of this study is to systematically evaluate the clinical validity, educational efficacy, and user acceptability of an AI-driven, competency-based assessment generator ("The Pocket Tutor") designed for the undergraduate medical curriculum. The specific objectives were to evaluate clinical accuracy and safety by assessing the factual correctness, clinical relevance, and presence of any dangerous medical errors within AI-generated clinical vignettes and distractors, utilizing peer-review by senior medical faculty. Secondarily, to determine exam-standard alignment by measuring how effectively the AI-generated multiple-choice questions align with the difficulty and competency-based formatting of standardized postgraduate medical entrance examinations (e.g., NEET-PG, INI-CET). Furthermore, to conduct an educational "Turing Test" to determine the ability of both medical educators and medical students to distinguish between entirely AI-generated clinical scenarios and authentic, human-authored Past Year Questions (PYQs) in a blinded assessment. Finally, to assess educational utility and adoption feasibility to gauge student and faculty perceptions regarding the value of AI-generated rationales ("High-Yield Pearls") for rapid revision, and to explore the viability of adopting this AI tool as a supplementary "co-pilot" for routine exam drafting.

MATERIALS AND METHODS

Study Population

This cross-sectional study was conducted between January and March 2026. The study utilized a convenience sampling method, yielding a dual-cohort design of senior medical faculty (n=32) and undergraduate MBBS students (n=230).

Ethical Considerations

This study constituted an educational evaluation of an artificial intelligence tool utilizing anonymized, voluntary survey feedback. As the research involved no patient data, clinical interventions, biological specimens, or access to protected health information, it was classified as a minimal-risk educational assessment and was exempt from formal Institutional Ethics Committee (IEC) full-board review. Explicit digital informed consent was obtained from all participants prior to survey initiation, and all data were completely anonymized to ensure strict confidentiality.

Study Design and Setting

A cross-sectional, observational study was conducted to evaluate the efficacy, clinical safety, and user acceptability of an artificial intelligence-driven medical assessment generator, termed "The Pocket Tutor." The study was conducted at Dr. KNS Memorial Institute of Medical Sciences, Barabanki, targeting the current competency-based medical education (CBME) landscape and competitive examination standards (NEET-PG / INI-CET / USMLE / PLAB).

Study Population

The study utilized a dual-cohort design to ensure both expert validation and end-user applicability. Cohort A consisted of senior medical faculty members (n=32) representing various clinical and pre-clinical departments to provide subject-matter expert (SME) validation. Cohort B consisted of undergraduate MBBS students (n=230) across multiple years of study to evaluate the tool's practical utility and exam parity. Participation was voluntary. While participant identities were verified to ensure cohort validity, all collected data was kept strictly confidential. Personal identifiers were stripped and anonymized prior to statistical analysis to ensure participant privacy.

AI Engineering and Prompt Design

The Gemini 3 Pro Large Language Model (Google) was utilized to engineer the clinical vignettes. To prevent AI hallucination and ensure strict adherence to medical standards, a highly constrained "Master Prompt" framework was developed. The AI was instructed to generate high-order, multi-step clinical scenarios requiring diagnostic reasoning rather than simple fact-recall. Each generated module was mandated to contain a realistic patient demographic and primary clinical presentation, relevant vital signs, physical examination findings, preliminary laboratory/imaging data, four highly plausible multiple-choice options (one correct answer and three competitive distractors), and a "High-Yield Pearl," providing a concise, evidence-based rationale for the correct answer to facilitate rapid student revision.

The "Turing Test" Methodology

To objectively evaluate the quality and indistinguishability of the AI-generated questions, a blinded "Turing Test" was integrated into the study design. Within the mock assessment module provided to both cohorts, two authentic Past Year Questions (PYQs) from recent standard medical board exams were seamlessly embedded alongside the AI-authored questions. Participants were not informed which specific questions were AI-generated versus human-authored until after the assessment.

Data Collection Instruments

Two distinct, cohort-specific digital questionnaires were deployed. Prior to full-scale deployment, the survey instruments underwent face validity testing and a pilot run with a small subset of participants to

ensure clarity, relevance, and internal consistency.

The bespoke survey instruments demonstrated high internal consistency, yielding a Cronbach's alpha coefficient of 0.84 during the pilot testing phase.

Faculty Instrument: Evaluated the AI modules for clinical safety (screening for dangerous medical errors or obsolete guidelines), factual accuracy, distractor quality, and alignment with the current NEET-PG clinical-vignette style using a 5-point Likert scale.

Student Instrument: Assessed the perceived difficulty level relative to standard Q-banks, the educational value of the AI's instant feedback rationale, and overall acceptability for regular exam preparation.

Statistical Analysis

Data extracted from the digital surveys were cleaned and analyzed using Python (version 3.10) utilizing the Pandas and SciPy statistical libraries. Descriptive statistics, including means, standard deviations (SD), and frequency distributions, were calculated for Likert-scale responses. Inferential biostatistics were applied to compare the perceptions of the faculty and student cohorts. A Pearson Chi-Square test of independence was utilized to analyze the categorical pass/fail rates of the blinded Turing Test between cohorts. A Mann-Whitney U test was employed to compare the continuous, non-parametric distributions of clinical accuracy ratings between faculty and students. A p-value of < 0.05 was considered statistically significant.

RESULTS

The study analyzed survey responses from a dual-cohort consisting of senior medical faculty (n=32) and undergraduate medical students (n=230). Data were analyzed utilizing descriptive statistics (means, standard deviations, and frequencies) and inferential biostatistics (Chi-Square and Mann-Whitney U tests) to establish statistical significance.

Faculty Validation of Clinical Accuracy and Safety

The faculty cohort (n=32) demonstrated exceptionally high acceptance of the AI-generated modules. When evaluating the clinical accuracy and factual correctness of the AI-generated scenarios, faculty awarded a mean score of 4.38 ± 0.71 (out of 5). Notably, 87.5% of senior educators rated the accuracy as either a 4 or 5.

Regarding clinical safety, 96.9% of the faculty certified that the AI's answers were clinically safe for student use, with 59.4% identifying "no errors" and 37.5% identifying only "minor phrasing issues" devoid of clinical risk. Only a single respondent (3.1%) flagged a major clinical concern. Furthermore, faculty rated the AI's alignment with current competency-based, clinical-vignette exam standards (NEET-PG/INI-CET) highly, yielding a mean score of 4.25 ± 0.72.

Table 1: Faculty Evaluation of AI-Generated Clinical Vignettes (n = 32)

Evaluation Metric	Mean Score (± SD) or Percentage
Clinical Accuracy & Factual Correctness (1-5 Scale)	4.38 (± 0.71)
Exam Alignment (NEET-PG/INI-CET) (1-5 Scale)	4.25 (± 0.72)
Clinical Safety (No dangerous errors/ contraindications)	96.9%
Workflow Integration (Desire to use as supplementary aid)	93.8%

• Note: Likert scale questions were rated from 1 (Highly Inaccurate/Poor) to 5 (Highly Accurate/Excellent).

Student Perception and Educational Efficacy

The undergraduate student cohort (n=230) similarly reported a positive reception, rating the overall clinical accuracy and relevance of the AI scenarios with a mean score of 3.93 ± 1.02.

When assessing exam parity, the majority of the students (62.2%) reported that the AI-generated difficulty "perfectly matched" the standard of the NEET-PG / INI-CET exams,

while an additional 19.1% found them "slightly harder." The educational value of the AI's immediate feedback mechanisms (the "High-Yield Pearls") was rated at a mean of 3.90 ± 0.98, indicating strong utility for rapid revision. Consequently, 90% of the student cohort expressed a desire to integrate this AI tool into their regular competitive exam preparation (40.4% "absolutely" and 49.6% as a "supplementary tool").

Table 2: Student Perception and Educational Efficacy (n = 230)

Evaluation Metric	Mean Score (± SD) or Percentage
Clinical Accuracy & Relevance (1-5 Scale)	3.93 (± 1.02)
Educational Value of "High-Yield Pearls" (1-5 Scale)	3.90 (± 0.98)
Difficulty Alignment (Matched or slightly harder than actual exams)	81.3%
Tool Acceptability (Desire to use for regular exam preparation)	90.0%

• Note: Likert scale questions were rated from 1 (Highly Inaccurate/Poor) to 5 (Highly Accurate/ Excellent).

The Clinical "Turing Test" & Comparative Biostatistics

A blinded "Turing Test" was conducted by embedding authentic human-authored Past Year Questions (PYQs) within the AI-generated module to evaluate indistinguishability.

Among the faculty, 24 out of 32 (75.0%) were entirely unable to distinguish the AI-generated questions from the human-made PYQs. Among the students, 108 out of 230 (47.0%) found them indistinguishable.

A Pearson chi-square test of independence was performed to evaluate the difference in Turing Test pass rates between faculty and students. The results were statistically significant (χ^2 (1) = 7.75, p = 0.005), revealing a fascinating paradigm: senior medical experts were significantly more likely to view the AI-generated questions as indistinguishable from standard human questions than the students were.

Furthermore, a Mann-Whitney U test compared the clinical accuracy ratings between the two cohorts. Faculty ratings (Mean = 4.38) were significantly higher than student ratings (Mean = 3.93) (U = 4542.0, p = 0.024). This data suggests that while students remain somewhat cautious of AI-generated content, senior subject-matter experts objectively validate its high clinical fidelity and standard.

Table 3: Comparative Analysis of AI Indistinguishability and Accuracy

Evaluation Parameter	Faculty Cohort (n = 32)
Clinical Accuracy Rating (Mean ± SD)	4.38 ± 0.71
The "Turing Test" (% indistinguishable from human PYQs)	75.0%

*Indicates statistical significance (p < 0.05).

DISCUSSION

The integration of Large Language Models (LLMs) into medical education has predominantly been evaluated through the lens of AI as a test-taker, with landmark studies by Jaleel et al. [1] and Gilson et al. [3] demonstrating the ability of models to pass standardized examinations such as the United States Medical Licensing Examination (USMLE). However, this study shifts the paradigm, evaluating the efficacy of Generative AI not as a candidate, but as a test-creator. Our findings demonstrate that when constrained by rigorous prompt engineering, as explored by Al Shuraiqi et al.,[7] AI can successfully generate highly accurate, competency-based clinical vignettes that are overwhelmingly accepted by both senior faculty and undergraduate medical students.

A paramount concern regarding the use of Generative AI in medical education is the risk of clinical "hallucination"—the generation of plausible but medically incorrect or dangerous information. Our study heavily mitigates this concern. With 96.9% of the senior faculty cohort (n=32) certifying the AI-generated correct answers and distractors as clinically safe and devoid of dangerous contraindications, the data

aligns with recent literature by Takahashi et al. [9] suggesting that carefully bounded LLMs can produce highly reliable medical text. Furthermore, 93.8% of the faculty expressed a desire to integrate this tool into their assessment-drafting workflow. This presents a scalable solution to a well-documented challenge in medical academia: the significant time and cognitive burden placed on faculty to author high-quality, clinical-vignette Multiple Choice Questions (MCQs) for Competency-Based Medical Education (CBME) curriculums.

The most compelling and statistically significant finding of this study is the inverse relationship observed in the blinded "Turing Test" between the expert and novice cohorts ($\chi^2(1) = 7.75$, $p = 0.005$). While one might hypothesize that senior faculty would be more adept at identifying AI-generated content due to their extensive clinical and academic experience, the data revealed the opposite. A significant majority of faculty (75.0%) found the AI questions indistinguishable from authentic Past Year Questions (PYQs), compared to only 47.0% of students. We postulate that this discrepancy highlights a fundamental difference in how experts and students evaluate assessment material. Faculty members likely judged the questions based on core clinical structure, syllabus alignment, and diagnostic logic—areas where the AI excelled. Conversely, students, who are deeply immersed in the repetitive patterns of standard preparation Q-banks, may have been hyper-sensitive to minor deviations in phrasing, stylistic nuance, or standard commercial formatting, leading them to flag the AI questions as "different." Despite this, the student cohort reported high educational efficacy, with 90% expressing the desire to utilize the tool for regular NEET-PG/INI-CET preparation, reinforcing the acceptability of AI as a supplementary medical tutor, a trend consistent with the multicenter survey findings of Tran et al. [10]

Limitations

While the sample size of 262 total participants is robust, it utilized a convenience sampling method without an a priori formal power analysis. Furthermore, this study relies on the perceived clinical validity and educational efficacy reported by participants via subjective Likert scales, rather than objective psychometric item analysis under strict examination conditions. Additionally, it remains a single-center cross-sectional analysis evaluating a specific LLM iteration; as models update rapidly, outputs may vary. Future multi-centric studies are required to evaluate the long-term impact on actual standardized exam scores.

CONCLUSION

"The Pocket Tutor" validates the immense potential of Generative AI to democratize and scale medical assessment. By successfully passing a clinical Turing Test among senior medical educators, the AI demonstrated its capability to author safe, highly accurate, and exam-aligned clinical vignettes. As medical board examinations worldwide continue to prioritize complex clinical reasoning, AI-driven assessment generators offer a dual benefit: providing medical students with limitless, high-yield mock testing, while simultaneously alleviating the profound MCQ-authoring burden on academic faculty. The future of medical education will heavily rely on human-AI collaboration, maintaining continuous faculty oversight while leveraging AI for unprecedented scalability.

Data Availability Statement

The data that support the findings of this study, as well as the exact generative AI prompts utilized for the clinical vignettes, are available from the corresponding author upon reasonable request.

Declaration by Authors

Ethical Approval: Approved

Acknowledgement: The first author would like to extend his heartfelt thanks to Dr. Anupama, Dr. Alok for their massive help making this paper successful and the entire senior medical faculty at Dr. KNS

Memorial Institute of Medical Sciences. Their willingness to take time out of their busy schedules to critically evaluate this tool, grant permission for data collection, and wholeheartedly back a student-led academic initiative provided the crucial clinical validation this study required.

Finally, profound appreciation is given to every senior, junior, and batchmate who voluntarily participated in the survey. Their time, feedback, and peer support were instrumental in bringing "The Pocket Tutor" to life and making this study a massive success.

Source of Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest: The authors declare no conflict of interest.

REFERENCES

1. Jaleel A, Aziz U, Farid G, et al. Evaluating the Potential and Accuracy of ChatGPT-3.5 and 4.0 in Medical Licensing and In-Training Examinations: Systematic Review and Meta-Analysis. JMIR Med Educ. 2025; 11: e68070.
2. Maaß L, Grab-Kroll C, Koerner J, et al. Artificial Intelligence and ChatGPT in Medical Education: A Cross-Sectional Questionnaire on students' Competence. J CME. 2024; 14(1): 2437293.
3. Gilson A, Safranek CW, Huang T, et al. Correction: How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2024; 10: e57594.
4. Kang J, Ahn J. Technologies, opportunities, challenges, and future directions for integrating generative artificial intelligence into medical education: a narrative review. Ewha Med J. 2025; 48(4): e53.
5. Rabbani SA, El-Tanani M, Sharma S, et al. Generative Artificial Intelligence in Healthcare: Applications, Implementation Challenges, and Future Directions. BioMedInformatics. 2025; 5(3): 37.
6. Peralta Ramirez AA, Trujillo López S, Navarro Armendariz GA, et al. Clinical Simulation with ChatGPT: A Revolution in Medical Education? J CME. 2025; 14(1): 2525615.
7. Al Shuraiqi S, AlZaabi A, Aal Abdulsalam A. Prompt Engineering Strategies for Generating Medical Case-Based MCQs with Large Language Models: A Multi-Model Comparative Study. Machine Learning and Knowledge Extraction. 2026; 8(2): 41.
8. Hölzing CR, Meynhardt C, Meybohm P, et al. Fine-Tuned Large Language Models for Generating Multiple-Choice Questions in Anesthesiology: Psychometric Comparison With Faculty-Written Items. JMIR Form Res. 2026; 10: e84904.
9. Takahashi H, Shikino K, Kondo T, et al. Educational Utility of Clinical Vignettes Generated in Japanese by ChatGPT-4: Mixed Methods Study. JMIR Med Educ. 2024; 10: e59133.
10. Tran C, Hryciw BN, Moore SW, et al. Perceptions and Use of Generative Artificial Intelligence in Medical Students: A Multicenter Survey. J Med Educ Curric Dev. 2025; 12: 23821205251391969.

Appendix A

The "Pocket Tutor" Master Prompt

The following highly constrained prompt was utilized in Gemini 3 Pro to generate the clinical modules:

"Act as an expert medical examiner for the Indian NEET-PG and INI-CET board exams. Generate a high-yield, competency-based clinical vignette Multiple Choice Question (MCQ). You must strictly adhere to the following constraints: 1) Provide a realistic patient demographic, chief complaint, vital signs, and relevant lab/imaging findings. 2) The clinical scenario must require multi-step diagnostic reasoning, not simple fact-recall. 3) Provide 4 plausible options (A, B, C, D) with only one unambiguously correct answer and three highly competitive distractors. 4) Conclude with a 'High-Yield Pearl' consisting of 2-3 sentences explaining the exact evidence-based rationale for the correct answer to facilitate rapid student revision. Ensure absolute clinical safety and adhere to the latest medical guidelines."

How to cite this article: Swapnil Banerjee, Manisha Agarwal. The pocket tutor - generating competency-based clinical vignettes with artificial intelligence for medical postgraduate competitive exam preparation. *Gal Int J Health Sci Res*. 2026; 11(2): 69-75. *DOI: https://doi.org/10.52403/gijhsr.20260208*

******