Journal of of Basic and Applied Pharmaceutical Science

A Pilot Study Comparing AI- and Instructor-Generated Practice Questions on Student Performance in Pharmacy Education

Parto Khansari[1], Rahul Nohria[2*], Leanne Coyne[1], and Douglas Ried[2]

[1]*West Coast University School of Pharmacy, 590 N Vermont Ave, Los Angeles, CA 90004, United States.*
[2]*Larkin University College of Pharmacy, 18301 N Miami Ave #1, Miami, FL 33169, United States.*

Article Details

Article Type: Research Article
Received date: 14th October, 2025
Accepted date: 03rd December, 2025
Published date: 05th December, 2025

***Corresponding Author:** Rahul Nohria, Associate Professor, Larkin University College of Pharmacy, 18301 N Miami Ave #1, Miami, FL 33169, United States.

Citation: Khansari, P., Nohria, R., Coyne, L., & Ried, D., (2025). A Pilot Study Comparing AI- and Instructor-Generated Practice Questions on Student Performance in Pharmacy Education. *J Basic Appl Pharm Sci, 3*(2): 111. doi: https://doi.org/10.33790/jbaps1100111.

Abstract

The integration of artificial intelligence (AI) into educational settings has expanded rapidly, yet empirical research on its practical applications remains limited. We conducted a pilot study to evaluate the use of AI-generated practice questions in a first-year pharmacy course and their impact on student performance in summative assessments. The objectives for this study were to see if 1) there was a difference in AI- and faculty- developed practice questions to prepare for summative assessment performance, and 2) practice questions led to better summative assessment performance. Students in an introductory pharmacology course received practice questions either created by AI or instructors. AI-generated practice questions were at least as effective as faculty generated practice questions, demonstrating marginally improved performance ($p = 0.06$), while requiring less faculty time to produce. Notably, students receiving practice questions from either AI or faculty generated demonstrated significantly higher summative assessment scores compared to a previous cohort that did not receive practice questions ($P<0.001$). These findings suggest that AI can be a valuable tool for maintaining or enhancing student outcomes without a perceived increase in faculty workload.

Keywords: ChatGPT, Practice Test, AI-generated, Faculty Burnout

Introduction

AI-based large language models (LLM), such as ChatGPT, are expanding rapidly. Since ChatGPT's launch in late 2022, there are over 9,000 publications (*search term: (Chat GPT) OR (ChatGPT)*) indexed in PubMed as of April 2025. Of these, approximately 2870 focus on education (*search term: (Chat GPT) OR (ChatGPT)) AND (education)*), and 705 specifically address education and healthcare (*search term: (Chat GPT) OR (ChatGPT)) AND (education) AND (healthcare)*). Much of this literature remains theoretical, with a focus on ChatGPT's potential applications. Empirical research is still emerging.

In a systematic review of 60 studies, Sallam [1] outlined key benefits of AI in healthcare, including improvements in scientific writing, data analysis, personalized learning, and workflow efficiency. However, the review also raised significant concerns regarding ethical issues, bias, plagiarism, and the potential spread of misinformation. The authors encouraged the development of ethical guidelines to ensure the responsible use of AI in healthcare settings [1].

Similarly, a review by Lo [2] analyzed 50 articles across educational disciplines, finding that ChatGPT's performance varied by subject. For example, ChatGPT excelled in economics but struggled in mathematics. ChatGPT was shown to assist in refining educational materials and assessments. However, Lo stressed the importance of verifying content to avoid inaccuracies and misinformation. The review also emphasized the need for updated institutional policies and comprehensive teacher training to effectively integrate AI into the classroom [2]. Although these articles highlight the need for oversight, they also clearly demonstrate the growing use of ChatGPT in education. One area that has received some attention but still lacks substantial empirical data is the use of ChatGPT for generating test questions.

Yang and colleagues [3] provide psychological evidence in a review article supporting the use of tests as effective learning tools, particularly in enhancing knowledge retention [3]. Similarly, Augustin [4] emphasizes that active retrieval practice rather than passively reviewing content is a significantly more effective strategy for promoting learning and long-term retention, particularly among medical students [4]. In a study by Naujoks and colleagues [5], the authors investigated two studies examining the impact of practice tests on academic performance among undergraduate psychology students [5]. The findings of the study suggest that practice tests enhanced academic performances. Despite substantial evidence supporting the effectiveness of practice tests in enhancing student learning, a major drawback is the significant increase in faculty workload associated with generating multiple sets of questions. This added burden may discourage instructors from incorporating practice tests into their teaching.

In a study by Cheung and colleagues [6], researchers compared the quality and time required to generate 50 multiple-choice questions

(MCQs) using ChatGPT versus expert instructors on content from medical textbooks. The questions were randomized and evaluated for quality by five independent international reviewers. The results showed no significant difference in quality between the two sets of questions. However, ChatGPT generated the questions in just 20 minutes and 25 seconds, compared to 211 minutes and 33 seconds for the experts, more than ten times longer. Other studies have also noted that writing well-constructed exam questions is a time-consuming and challenging task for instructors [7].

To address this workload issue, we piloted a study comparing the impact of two sets of practice questions on subsequent exam performance: one set generated by ChatGPT based on course materials and another manually created by instructors. Our aim was to evaluate assessment outcomes with and without practice tests, and to compare the effectiveness of practice questions generated by instructors and ChatGPT.

Objective/Aims

1. To see if there was a difference in AI- and faculty- developed practice questions to prepare for summative assessment performance.
2. To see if practice questions led to better summative assessment performance.

Methods

Study Population

The study population consisted of first-year (P1) students enrolled in an introductory pharmacology course during 2024 at a private college of pharmacy on the west coast of the US. The course followed a structured Team-Based Learning (TBL) format, designed to enhance student engagement and collaborative learning [8]. The 2024 cohort was taught online as part of a hybrid curriculum and was exposed to both AI- and faculty- developed practice questions prior to summative assessments. For historical context, data was also examined in the 2022 cohort, where classes were similarly taught online due to COVID-19 but were not administered practice tests prior to summative assessments.

Under description of the Experimental and Control groups and Study Design

The 2024 cohort received weekly practice tests each Friday, consisting of 42 questions covering material from that week, in preparation for a weekly Monday assessment. There were a total of ten practice exams and ten summative assessments. To maintain alignment with TBL principles, students first completed an individual practice test, followed by a team-based test where teams worked collaboratively to answer the same questions without access to external resources. Practice test questions were either faculty-generated (FGQ) or AI-generated (AIQ), with AIQ questions created using ChatGPT. All AIQ questions were reviewed by instructors to ensure accuracy and alignment with course learning outcomes. A comprehensive description of the prompt employed to generate the AIQ is available in Appendix 1. FGQ and AIQ were administered to all students of the 2024 cohort at the same time (Figure 1).

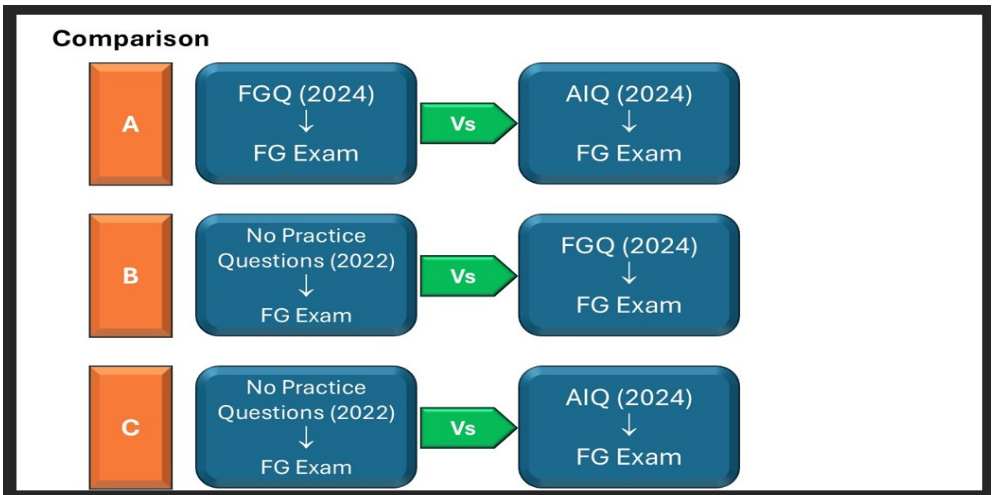


Fig 1. Experimental design of the study. Exam questions following faculty-generated practice questions (FGQ) were compared with exam questions following AI-generated practice questions (AIQ) in the 2024 cohort (comparison A). Exam questions for the 2022 cohort who did not receive practice questions were compared to FGQ (comparison B) and AIQ (comparison C).

Two instructors taught the course, each covering an equivalent number of topics. One instructor developed FGQ questions and the other created AIQ questions for the associated practice tests. Students were exposed to either FGQ or AIQ depending on which instructor delivered the content. Performance on the exam questions was linked back to the specific instructor. Questions were identified as AIQ or FGQ via a prefix in front of the question (e.g. AIQ #1, FGQ #2). In contrast, students in the 2022 cohort did not receive any practice questions.

The first objective was to determine whether there were differences in students' summative assessment outcomes depending on whether the preceding practice questions were FGQ or AIQ. For final summative assessments, all students were administered only faculty-generated questions.

A secondary objective of the study was to determine whether providing practice questions to students before the final assessment improved performance. Students in the 2022 cohort completed a midterm and a final exam during the course but did not receive either FGQ or AIQ questions beforehand.

Inclusion / Exclusion Criteria

Students enrolled in the required pharmacology course that completed all practice and summative assessments, either in 2022 (47 of 47 students) or 2024 (24 of 29 students) were included in the study. Four students were excluded from the 2024 cohort due to not participating in either practice tests or summative assessments.

Dependent Variable

Student performance was defined as the number of correct questions divided by the total number of questions on summative assessments.

Independent Variable

Two independent conditions were embedded in the intervention. The first was whether the practice tests contained FGQ or AIQ questions. FGQ were used for all final summative assessments for both cohorts. The second was whether students were administered a practice test before the summative assessment. The 2022 cohort did not receive a practice test, whereas the 2024 cohort did.

Statistical Analysis

Statistical comparisons of mean scores between FGQ and AIQ performance within the same cohort were conducted using paired t-tests. An independent t-test was used to compare the 2022 and 2024 cohorts. Given the small number of eligible participants, the p-value < 0.1. Statistical tests were conducted using Microsoft Excel, Microsoft Corporation, Version 16.0, 2019.

IRB

The protocol was reviewed and approved by the Institutional Review Board at West Coast University as an exempt protocol.

Results

Table 1 summarizes the data used to evaluate both study objectives. A direct comparison was made on the scores of the summative tests following AIQ and the FGQ used as practice assessments within the 2024 cohort (Figure 1, Comparison A). Students who completed the AIQ practice quizzes scored significantly higher (p < .06) on the summative assessment (mean = 80.6%) than when practicing with FGQs (mean = 77.7%).

Comparisons	Comparison A		Comparison B		Comparison C	
Study Groups	FGQ	AIQ	2022*	FGQ	2022*	AIQ
Percentage	77.7%	80.6%	64.5%	77.7%	64.5%	80.6%
Standard Deviation	.09%	.11%	.09%	.09%	.09%	.11%
t-test (df)	-1.98 (23)		6.36 (66)		5.49 (66)	
P-value	.06		< .001		< .001	

* Summative assessments in both years were faculty-generated, and the outcome measure is the mean score on the summative assessment. AIQ = Artificial intelligence generated practice questions; FGQ = Faculty Generated Questions. The p-value was set to 0.10.

Table 1. Comparison of Percentage Score between artificial intelligence generated questions and 2024 and 2022 control groups.

A statistically significant difference was observed in mean scores on the summative assessment between the 2022 and 2024 student cohorts (Figure 1, Comparisons B and C). The 2022 cohort, which did not receive weekly Friday practice quizzes, achieved a mean score of 64.5%, whereas the 2024 cohort, following FGQs (Figure 1, Comparison B), attained a mean score of 77.7% (p < .001).

Under the third condition, the 2024 cohort, which completed AIQ as practice quizzes prior to the faculty-developed summative assessment, also demonstrated significantly higher performance compared to the 2022 cohort, which received faculty-developed questions but no practice questions (Figure 1, Comparison C). Students in the 2024 AIQ group achieved a mean score of 80.6%, compared to 64.5% in the 2022 group (p < .001). Thus, both comparisons, AIQ versus no practice questions, and AIQ versus FGQ, indicated that AI-generated practice questions were associated with improved assessment performance.

Discussion

Our findings support previous research showing that practice testing is a worthwhile and consistently effective learning strategy, especially in higher education and professional training. For the 2024 cohort, following the FGQ, there was a 13.2% improvement in summative scores. Following the AIQ, the 2024 cohort achieved a 16.1% improvement in summative scores. Findings from the current study indicate that practice questions aided in student learning. AIQ questions performed just as effectively as FGQ on summative assessment performance.

Incorporating practice exam questions into the curriculum offers several benefits, particularly before comprehensive or high stakes assessments. Among various learning strategies, practice exams have been identified as one of the most effective methods for enhancing learning and retention [9]. Exposure to practice questions enhances knowledge retention through repeated retrieval of information [10]. Furthermore, practice testing provides opportunities for immediate feedback, encourages peer engagement, and promotes the development of critical thinking and metacognitive skills. It also supports self-assessment and helps students refine their study strategies [11].

Despite documented benefits, writing exam questions is time-consuming for faculty, which may deter the use of practice tests. In this study, students completed 10 practice tests, one at the end of each week. Creating well-constructed questions that assessed understanding across different cognitive levels, including higher-order and critical thinking, required significant instructor time. For each exam, instructors wrote 42 questions. Depending on the complexity and intended cognitive level, drafting each question took approximately 5 to 15 minutes. In contrast, using AI with a well-designed prompt, the entire process of generating, reviewing, and revising all 42 questions to meet instructional goals took less than 30 minutes, without compromising the quality of the exam questions. To mitigate potential burnout among faculty members, AI systems could provide immediate answer keys and rationales. These tools are particularly valuable for formative assessments and practice tests, enabling students to receive instant feedback, without additional burden on instructors.

These results are consistent with other studies in medical and health sciences education that demonstrated practice exams can predict performance on summative assessments [12, 13]. Similarly, a study with pharmacy students found that repeated testing on drug knowledge and calculation improves learning outcomes [14]. Burnout among professors in higher education is a growing concern, compounded by the pressure to produce high-quality teaching materials [15]. Developing exam questions that accurately assess student learning while aligning with course objectives is an intellectually demanding and time-consuming challenge. AI tools have the potential to reduce instructors' workloads by generating diverse, high-quality exam questions, suggesting multiple-choice distractors, and even tailoring questions to varying levels of difficulty [16].

An anecdotal finding from this study was that the quality of AIQ questions is highly dependent on the design of the input prompt. In the initial phase of the study, the quality of the AIQ questions was suboptimal. For instance, only 23 out of the 42 questions were considered suitable for the exam. Correct answers were often easily identifiable due to inconsistent formatting. For example, the correct options were significantly longer than the others or included a question indicator absent from the incorrect choices. Additionally, most questions reflected low-level cognition, focusing primarily on recall.

To address these issues, teaching materials, such as instructor-prepared handouts and lecture slides, were uploaded into ChatGPT. A refined prompt was developed with clear criteria for constructing effective MCQ emphasizing consistent formatting and targeting a spectrum of cognitive complexity, from basic comprehension to higher-order critical thinking. The prompt also indicated the use of brief clinical scenarios and answer choices that required the integration of multiple concepts to arrive at the correct response. As a result, the quality, relevance, and difficulty level of the AIQ questions improved significantly, although expert review remains essential to ensure accuracy and appropriateness.

Consistent with our findings, Ahmed and colleagues [17], in their study of medical school exams comparing expert versus ChatGPT generated questions, also emphasized that prompt refinement is a crucial step in producing high-quality questions that align with learning objectives [17]. In another study by Law and colleagues [18], the authors noted that the AI-generated items lacked the depth and complexity of those written by experts, again emphasizing the importance of human oversight to maintain content quality. Future research on the quality and alignment of the prompt with academic content is necessary to continue to assess the quality of AIQ questions. As AI becomes increasingly integrated into education, it is essential for educators to assess its potential to enhance student learning outcomes, and to reduce faculty burnout.

One of the main limitations of this study is the sample size. In this pilot study, the intervention was assessed in a single cohort. Additionally, while both the 2022 and 2024 cohorts were taught online, the format of the courses differed. The 2022 cohort included traditional midterms and finals, whereas the 2024 cohort included weekly exams. More frequent testing may have promoted continuous engagement and improved short-term knowledge retention. Furthermore, the close timing of practice tests to the summative exams could have resulted in short-term performance gains rather than deep learning and long-term retention.

Despite these limitations, evidence supports practice questions as a learning tool and in this study, AIQ questions were at least as effective as questions created by instructors. Further research is needed to evaluate the effectiveness of AI-generated practice tests in supporting long-term learning.

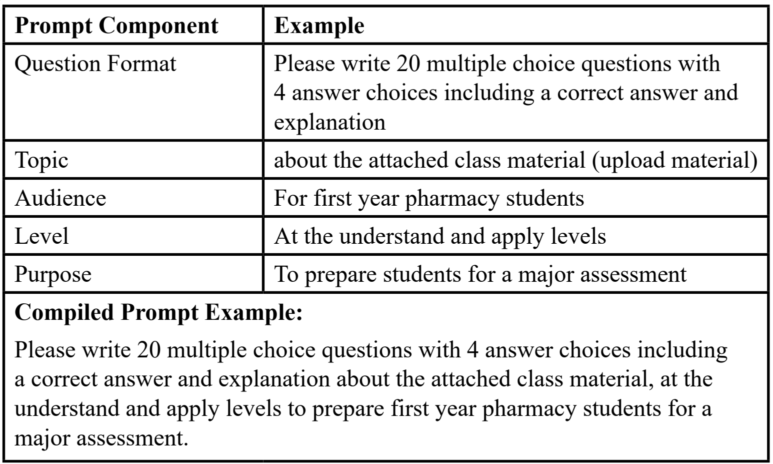
Conclusion

This study demonstrated that practice tests incorporating AI-generated questions were just as effective as those created by faculty in improving student outcomes. Though further research is needed, AI may help instructors incorporate practice tests without the typical burden of writing additional test questions.

Competing Interests: The authors declare that they have no competing interests.

References

1. Sallam, M. (2023). ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare 11*, no. 6: 887. https://doi.org/10.3390/healthcare11060887.
2. Lo, C.K. (2023). What Is the impact of ChatGPT on education? A rapid review of the Literature. *Educ Sci. 13*(4):410. https://doi.org/10.3390/educsci13040410.
3. Yang, B.W., Razo, J., Persky, A.M. (2019). Using testing as a learning tool. *Am J Pharm Educ. 83*(9):7324. doi: 10.5688/ajpe7324. PMID: 31871352; PMCID: PMC6920642.
4. Augustin, M. (2014). How to learn effectively in medical school: test yourself, learn actively, and repeat in intervals. *Yale J Biol Med. 87*(2):207-12. PMID: 24910566; PMCID: PMC4031794.
5. Naujoks, N., Harder, B., Händel, M. (2022). Testing pays off twice: Potentials of practice tests and feedback regarding exam performance and judgment accuracy. *Metacogn Learn. 17*(2): 479–498. https://doi.org/10.1007/s11409-022-09295-x
6. Cheung, B. H. H., Lau, G. K. K., Wong, G. T. C., Lee, E. Y. P., Kulkarni, D., Seow, C. S., Wong, R., & Co, M. T. (2023). ChatGPT versus human in generating medical graduate exam multiple choice questions-A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). *PloS one, 18*(8), e0290691. https://doi.org/10.1371/journal.pone.0290691
7. Rush, B.R., Rankin, D.C. & White, B.J. (2016) The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC Med Educ. 16*, 250. https://doi.org/10.1186/s12909-016-0773-3
8. Michaelsen, L., Sweet, M. (2008). The essential elements of team-based learning. *New Dir Teach Learn. 116*:7 - 27. 10.1002/tl.330.
9. Dunlosky, J., Rawson, K.A., Marsh, E.J., Nathan, M.J., Willingham, D.T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychol Sci Public Interest. 14*(1) 4-58. PMID (PubMed ID): 26173288. DOI: 10.1177/1529100612453266.
10. Roediger, H.L., Butler, A.C. (2011). The critical role of retrieval practice in long-term retention, *Trends Cogn Sci. 15*(1):20-27. ISSN 1364-6613, https://doi.org/10.1016/j.tics.2010.09.003.
11. Nicol, D.J., Macfarlane-Dick D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Stud High Educ. 31*(2):199–218. https://doi.org/10.1080/03075070600572090.
12. Kibble, J. (2007). Use of unsupervised online quizzes as formative assessment in a medical physiology course: effects of incentives on student participation and performance. *Adv. Physiol. Educ. 31*, 253–260. doi: 10.1152/advan.00027.
13. Jeyaraju, M., Linford, H., Mendes, T.B., Caufield-Noll, C., Tackett, S. (2023). Factors leading to successful performance on U.S. national licensure exams for medical students: A scoping review. *Acad Med. 98*(1):136–148. doi:10.1097/ACM.0000000000004877.
14. Coker, A.O., Lusk, K.A., Maize, D.F., Ramsinghani. S., Tabor, R.A., Yablonski, E.A., et al. (2018). The effect of repeated testing of pharmacy calculations and drug knowledge to improve knowledge retention in pharmacy students. *Curr Pharm Teach Learn. 10*(12), 1609–1615. https://doi.org/10.1016/j.cptl.2018.08.019
15. Sabagh, Z., Hall, N.C., Saroyan, A. (2018). Antecedents, correlates and consequences of faculty burnout. *Educ Res. 60*(2):131–156.https://doi.org/10.1080/00131881.2018.1461573.
16. Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International, 61*(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148
17. Ahmed, A., Kerr, E., O'Malley. A. (2025). Quality assurance and validity of AI-generated single best answer questions. *BMC Med Educ. 25*, Article: 300. https://doi.org/10.1186/s12909-025-06881-w
18. Law, A.K., So, J., Lui, C.T., Choi, Y.F., Cheung, K.H., Kei-Ching, H.K. et al. (2025). AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination. *BMC Med Educ. 25*(1):208. doi: 10.1186/s12909-025-06796-6. PMID: 39923067; PMCID: PMC11806894.



Prompt Component	**Example**
Question Format	Please write 20 multiple choice questions with 4 answer choices including a correct answer and explanation
Topic	about the attached class material (upload material)
Audience	For first year pharmacy students
Level	At the understand and apply levels
Purpose	To prepare students for a major assessment
Compiled Prompt Example: Please write 20 multiple choice questions with 4 answer choices including a correct answer and explanation about the attached class material, at the understand and apply levels to prepare first year pharmacy students for a major assessment.	

Appendix 1