JMIR Publications
The Leading eHealth Publisher

Reducing Faculty Burden with Question Creation: A randomized control trial comparing AI-generated versus expert-generated questions through student feedback and performance

Akshat Kumar, Katie Stinson, Litao Wang, Elizabeth Nugent, Yongsu Lee, Mark Hormann, LaTanya Jones Love, Toufeeq Syed

Submitted to: JMIR Medical Education
on: November 05, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Reducing Faculty Burden with Question Creation: A randomized control trial comparing AI-generated versus expert-generated questions through student feedback and performance

Akshat Kumar[1*] MBE, MS, MD; Katie Stinson[2*] MLIS; Litao Wang[3*] EdD, MEd; Elizabeth Nugent[3*] MD; Yongsu Lee[4*] PhD; Mark Hormann[3*] MD; LaTanya Jones Love[3*] MD; Toufeeq Syed[5*] MS, PhD

[1]University of California San Diego Health San Diego US
[2]The University of Texas Health Science Center at Houston McWilliams School of Biomedical Informatics Houston US
[3]The University of Texas Health Science Center at Houston McGovern Medical School Houston US
[4]The University of Texas Health Science Center at Houston School of Public Health Houston US
[5]The University of Texas Health Science Center at Houston McWilliams School of Biomedical Informatics McGovern Medical School Houston US
[*]these authors contributed equally

Corresponding Author:
Toufeeq Syed MS, PhD
The University of Texas Health Science Center at Houston
McWilliams School of Biomedical Informatics
McGovern Medical School
7000 Fannin St. #600
Houston
US

Abstract

Background: Developing high quality formative and summative assessment questions can take considerable time and effort, which in turn can contribute to burnout among medical school teaching faculty. Generative AI can decrease faculty burden by reducing the time needed to develop assessment materials.

Objective: This study compares students' performance on and subjective assessment of AI- versus expert-generated questions. We hypothesized that there would be no significant difference in student performance or subjective assessment of AI vs expert-generated questions.

Methods: We designed a single-center, prospective, double-blind randomized controlled trial where student participants randomly received one AI or expert-generated question per day over the course of four weeks.

Results: We found no significant difference between the distributions of the proportion of correct responses to AI- versus expert-generated questions (p=0.1809). Interestingly, the cumulative proportion of questions answered correctly over 28 days increased slightly for human-generated questions but stayed constant for AI-generated questions. There was a significant difference (p=0.0051) between the distribution of student difficulty ratings for AI- versus expert-generated questions.

Conclusions: Generative AI can reduce faculty burden by creating effective formative and summative assessments.

(JMIR Preprints 05/11/2024:68435)
DOI: https://doi.org/10.2196/preprints.68435

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.
2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Reducing Faculty Burden with Question Creation: A randomized control trial comparing AI-generated versus expert-generated questions through student feedback and performance

Authors

[1]Akshat Kumar, MD, MS, MBE
[2]Katie Stinson, MLIS
[3]Litao Wang, MEd, EdD
[3]Elizabeth Nugent, MD
[4]Yongsu Lee, PhD
[3]Mark Hormann, MD
[3]LaTanya Jones Love, MD
[2,3]Toufeeq Ahmed Syed, PhD, MS*

Affiliations

[1]Department of Internal Medicine, University of California San Diego Health, San Diego, CA, United States
[2]McWilliams School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, United States
[3]McGovern Medical School, University of Texas Health Science Center at Houston, Houston, TX, United States
[4]School of Public Health, University of Texas Health Science Center at Houston, Houston, TX, United States
*Corresponding Author

Abstract

Background: Developing high quality formative and summative assessment questions can take considerable time and effort, which in turn can contribute to burnout among medical school teaching faculty. Generative AI can decrease faculty burden by reducing the time needed to develop assessment materials.

Objective: This study compares students' performance on and subjective assessment of AI- versus expert-generated questions. We hypothesized that there would be no significant difference in student performance or subjective assessment of AI vs expert-generated questions.

Methods: We designed a single-center, prospective, double-blind randomized controlled trial where student participants randomly received one AI or expert-generated question per day over the course of four weeks.

Results: We found no significant difference between the distributions of the proportion of correct responses to AI- versus expert-generated questions (p=0.1809). Interestingly, the cumulative proportion of questions answered correctly over 28 days increased slightly for human-generated questions but stayed constant for AI-generated questions. There was a significant difference (p=0.0051) between the distribution of student difficulty ratings for AI- versus expert-generated questions.

Conclusions: Generative AI can reduce faculty burden by creating effective formative and summative assessments.

Keywords: Faculty burden; generative AI; medical education; formative assessment; summative assessment; multiple-choice questions

Introduction

Using AI to Decrease Faculty Burden

Instructors commonly use multiple choice questions (MCQs) to assess US medical students, but generating MCQs is laborious, time-consuming, and expensive, especially given that many faculty members have full-time competing commitments as basic scientists or physicians. Automated question generation may reduce faculty burnout by saving faculty significant time and also benefit medical students by increasing access to high-quality questions directly relevant to material taught in class, limiting the need for expensive third-party question banks. Here we test if large language models (LLMs) like GPT4 can generate questions of the same quality and difficulty as domain experts.

Large language models (LLMs) have recently emerged as an exciting new technology with massive potential in multiple disciplines. Such models are categorically different from traditional machine learning methods because they rely heavily on semi-supervised training tasks to build internal representations of various concepts and learn directly from the temporal, spatial, and/or syntactic structure of the underlying data itself (e.g., predicting masked words given some context in the case of LLMs). Applications of LLMs in medical education include on-demand interactive teaching (e.g., Socratic method of teaching) and production of educational materials. A recognized limitation of LLMs is their inclination to hallucinate (i.e., generate false information) in response to queries. While hallucinations limit the effectiveness of LLMs in multiple fields, we suspect they may allow LLMs to generate more effective MCQs since effective MCQs often require believable distractors.

Multiple prior studies have explored the possibility of using LLMs for MCQ generation in medical education, but these studies have shown mixed results. For example, Kıyak et. al incorporated two ChatGPT-generated MCQs into a fourth-year medical school clerkship exam and found that the questions were able to distinguish between high and low performing students [1]. Laupichler et. al. took this research further by comparing student performance on two sets of 25 questions where one set was generated by medical educators and the other was generated by ChatGPT [2]. Unlike the previous study, Laupichler et. al. found that there was, in fact, higher discriminatory power in expert vs LLM-generated questions (expert generated questions distinguished low from high performing students better than did LLM-generated questions). The researchers also found that students were able to correctly identify if a question was expert or AI-generated 57% of the time.

The competing results of the prior studies highlight the need for further studies comparing AI- and expert-generated questions, a gap which our work fills. Similar to previous work, our study also compares student performance on expert versus AI-generated questions. Unlike previous work, we test students longitudinally over a month to see if their performance on and ability to distinguish between expert versus AI-generated questions changes over time. We also assess student perceptions of using AI for medical education before and after exposure to AI-generated questions. Lastly, we utilize GPT4, a significantly more advanced LLM, as opposed to ChatGPT (which runs on the less powerful GPT-3.5) to generate our questions. We hypothesized there would be no statistically significant difference in quiz takers' perception of and performance on AI- versus versus expert-generated questions.

Methods

Purpose

We designed a single-center, prospective, double-blind randomized controlled trial where student participants received one randomly selected AI or expert-generated question per day over the course of one month (28 days). Participants were second-year medical students at the University of Texas Health Science Center McGovern Medical School. All questions in the study focused exclusively on the reproductive system and specifically covered the following topics: contraception, breast cancer and adnexal masses, sexually transmitted infections (STIs), and the effects of pregnancy on organ systems. Questions were delivered to students using QuizToo, an education technology platform developed for healthcare training and education. Participation in the study was voluntary, and the study was approved by the Committee for the Protection of Human Subjects (the Institutional Review Board at the University of Texas Health Science Center at Houston [UTHSC-H]) under protocol HSC-SBMI-23-1137.

QuizToo

QuizToo, a unique learning platform used in this study, was developed by Toufeeq Ahmed Syed, PhD, MS, Associate Professor and Assistant Dean of Education at the University of Texas Health Science Center at Houston [3]. QuizToo works by providing spaced micro-learning (i.e., daily question delivery) either as a stand-alone course or as a supplemental learning tool for another course. This platform delivers courses to learners via text message or email at set times and provides instant feedback and learning resources. In QuizToo, educators can upload content to be generated into questions and answers (multiple choice, yes/no, true/false, multiple select). Depending on the preference, educators can set the day and time for delivering questions. In addition to email, learners also have the option to respond via SMS or text messages. Learners are provided immediate feedback in the form of resources and explanations when they respond to questions. QuizToo also allows educators to track individual responses to questions, download response data, and create a Question/Answer bank to collaborate with other educators.

Study Procedures

Thirty medical students participated in the study. Complete participation included reviewing and acknowledging a Letter of Information, completing pre- and post-study assessments, enrolling in our QuizToo Reproductive Systems Module, and answering at least one question over the 28-day question period.

Questions were delivered in a double-blind randomized format. We created two sets of questions for the study: 1) AI-generated questions; and 2) expert-generated questions. We used GPT4 to create AI questions based on openly available content from Scholar-Rx Bricks as context for the questions [4]. The best AI questions were then manually selected for the study by our expert, the Reproductive Systems Course Director (EN). Expert-generated questions were sourced from the free and openly available MedQA US database [5]. Both sets of questions were then uploaded into QuizToo. To administer the question sets, we configured QuizToo to randomly divide the cohort (n=30) into two groups such that, every day, one group received an AI-generated question while the other group received an expert-generated question. QuizToo re-randomized cohorts daily.

Questions were delivered Monday through Sunday over a 28-day period. Notifications of question delivery were sent at 9 a.m. each day either via text message or email (participants previously selected their preferred method of delivery). Each question notification contained a hyperlink to access the question. Upon answering the question, the user received immediate feedback if they answered the question correctly. Participants were prompted for the following feedback after completing each question: 1) Please rate the difficulty of this question (response options: very easy; easy; neutral; difficult; very difficult); and 2) Do you think this question was AI-generated or human-generated (response options: AI-generated; human-generated).

Data

We collected the following data over the course of 28 days: participant responses to daily MCQs, participants' perception on if a daily question was AI- or expert-generated, participant feedback on the perceived difficulty of a question using a Likert scale (very easy, easy, neutral, difficult, very difficult), and participants' perception of the use of generative AI in education using a pre- and post-assessment survey at the beginning and end of the study, respectively. The specific questions asked in the pre- and post-assessment surveys are detailed in the Results section.

Collecting Expert-generated Questions

We used MedQA, an open-source medical question-answer dataset built by scraping several popular websites used by medical students (MedBullets, Amboss, and Lecturio), to obtain expert-generated questions [5]. The MedQA dataset contains questions from the US, mainland China, and Taiwan with the subset of US questions having the following four fields: question, answer, options (i.e., a list of all the available answer choices), and meta-information (if the question is from Step 1 or Step 2) [5]. We used the set of questions in the MedQA US database labeled as pertinent for the USMLE Step 1 exam in this study as all participants were preclinical students in their second year of medical school.

Since the MedQA database did not characterize questions beyond if they were from Step 1 or Step 2 and because MedQA contains over 12,000 English language questions, we first passed each question along with its answer choices through Llama2-70B and asked it to categorize the questions as either irrelevant or of one of the following four reproductive system topics: contraception, breast cancer and adnexal masses, STIs, and the effects of pregnancy on organ systems [6]. Seven questions from the first 30 questions in each category were then selected for the study by our expert, the Reproductive Systems Course Director (EN).

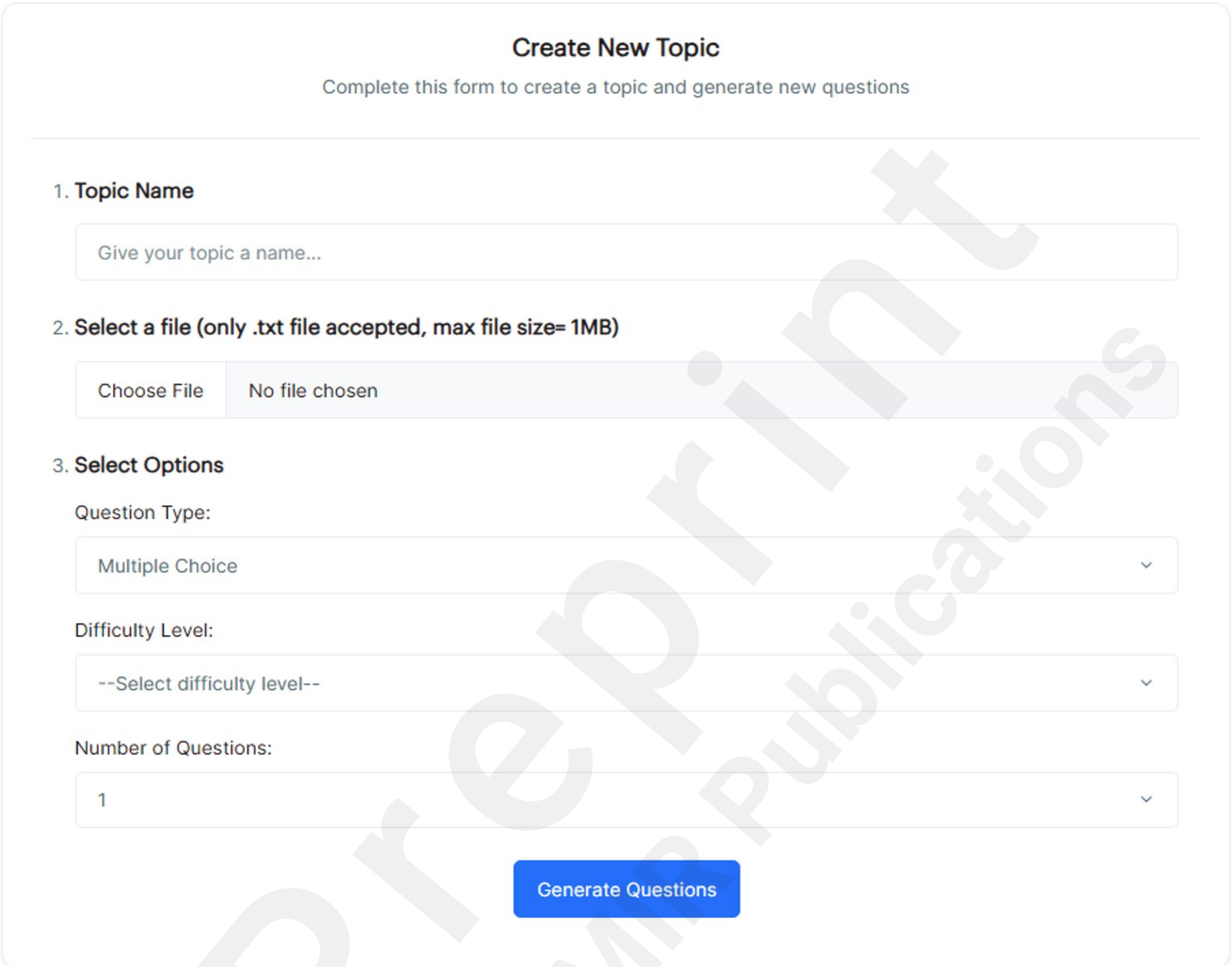
Creating AI-generated Questions

To create the AI-generated question set, we developed a custom Python script to access GPT4 through an API. GPT4 was asked to function as a quiz generator assistant when generating questions. We passed the following prompt into GPT4 to generate all questions: *"You are a quiz generator assistant. You help create MCQs based on the content. Questions are always strictly from the content and nowhere else. There is only one correct answer and the rest are distractors. Four answers max. Return in a minified JSONL format."* In addition to the previous prompt, we asked GPT4 to return the following information with each generated question:

- 1-sentence Lead-in
- Question
- Options
- Correct Answer
- 2-line Explanation
- Type
- Difficulty

- Link to Reference
- Keywords

Figure 1 shows an example interface we created to easily query GPT4 to generate a question.



Create New Topic

Complete this form to create a topic and generate new questions

1. Topic Name

Give your topic a name...

2. Select a file (only .txt file accepted, max file size= 1MB)

Choose File No file chosen

3. Select Options

Question Type: Multiple Choice

Difficulty Level: --Select difficulty level--

Number of Questions: 1

Figure 1: Interface used to provide GPT4 the information to create the AI-generated questions.

As before, seven questions in each category were selected for the study by our expert, the Reproductive Systems Course Director (EN), and uploaded into QuizToo.

Results

Participants

Of the 240 students in McGovern Medical School's second year class, 36 completed the pre-assessment, 30 enrolled in the study, 28 completed the study, and 22 completed the post-assessment. For the purposes of this study, we did not collect demographic information. A total of 28 participants answered at least one question. A total of 567 questions were answered throughout the study by the 28 participants.

Participant Responses to Daily MCQs

Figure 2 represents the proportion of participants who answered a question correctly. More specifically, the x-axis represents accuracy on different questions, with zero representing questions no one answered correctly and one representing questions everyone answered correctly. The y-axis represents the proportion of participants who answered all of the questions correctly. For example, Student A might have answered 62% of the expert-generated questions they submitted correctly and 88% of the AI-generated questions they submitted correctly, while Student B may have answered 90% of the expert-generated questions they submitted correctly and 70% of the AI-generated questions they submitted correctly. Figure 2 thus represents the proportion of students who performed at different total exam scores for both expert and AI-generated questions in the form of a histogram. Since these distributions are not normally distributed and may not necessarily have equal variance and spread, we used the Wilcoxon's Rank Sum Test, a non-parametric alternative to the more classic two-sample T-test, to assess if the two distributions were statistically significantly different from each other.

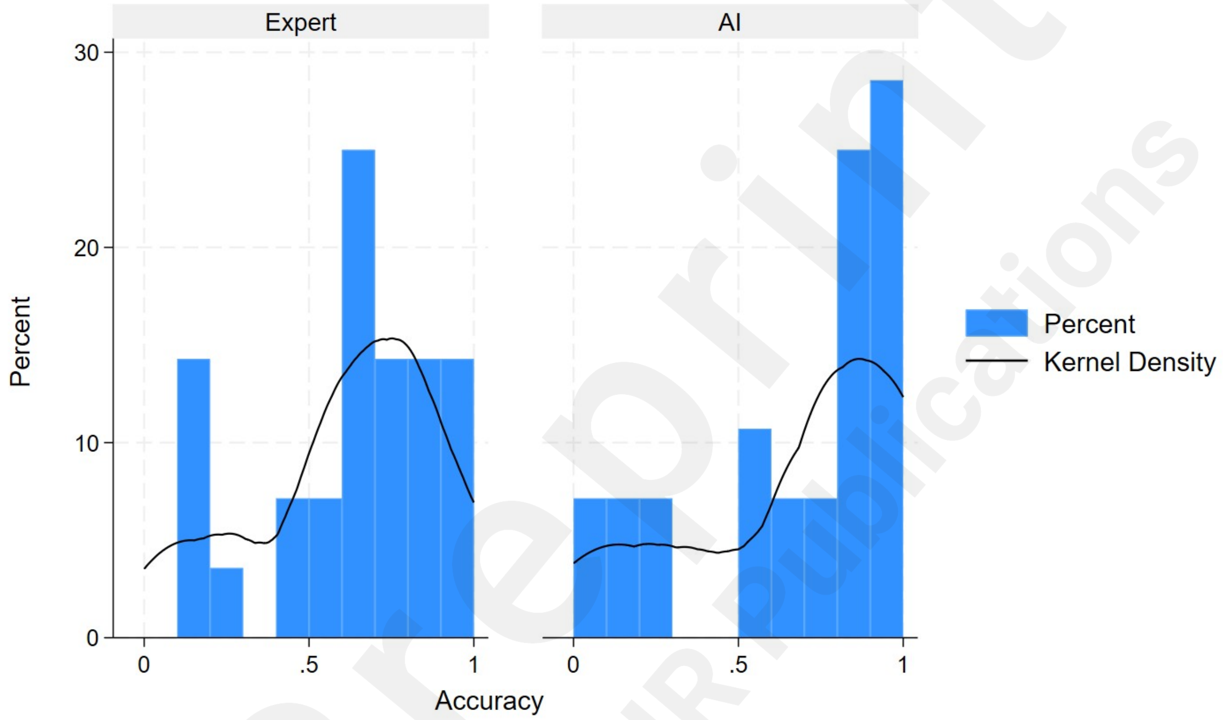


Figure 2: Proportion of students at various performance deciles for expert versus AI-generated questions.

In Figure 2, the x-axis of each histogram represents accuracy, with 0 being questions that were answered incorrectly by all participants and 1 being questions that were answered correctly by all participants. The y-axis represents the percentage of participants who answered questions correctly.

The histograms in Figure 2 represent distributions of accuracy for each question generator. The height of each bar shows the percentage of questions falling within a specific accuracy range. For example, the height of the first bar in the histogram for the AI-generated questions represents the percentage of questions with accuracy between 0.0 – 0.1.

Kernel density, or kdensity, is a statistical method used to estimate the probability distribution of a variable by replacing each datapoint with a commonly used probability density function (e.g., normal distribution) as a weight. We applied the "locally weighted scatterplot smoothing (LOWESS)" method, which helped us capture the overall distribution of a variable by reducing potential unnecessary noise and giving more weight to areas with a higher density of datapoints.

Between both sets of questions, we observed similar patterns in terms of accuracy. However, in the accuracy range of 80% to 100%, we found that more AI-generated questions were answered correctly. Overall, our p-value is relatively large (p=0.1809); therefore, there is no statistical evidence supporting a difference in accuracy between the expert-generated and AI-generated questions. The results of this assessment support the hypothesis that there is no significant difference in student performance of AI-generated questions compared to expert-generated questions.

Participant Responses to Daily MCQs Over Time

In addition to observing if the participants answered the questions correctly, we observed if the participants improved their ability to answer questions correctly over the duration of the study to assess learned behavior. The unique aspect of our study is that we were able to observe the participants' question-answering patterns over the 28-day period as opposed to the participants sitting for a 1-day test. For each question that was delivered to a participant, we tracked metadata including the day and time the question was delivered.

For each participant, we collected the correctly answered questions out of all delivered questions and sorted the data according to the day and time the question was delivered. Then, we calculated the cumulative proportions of correctly answered questions over time. See Figure 3 and Figure 4.

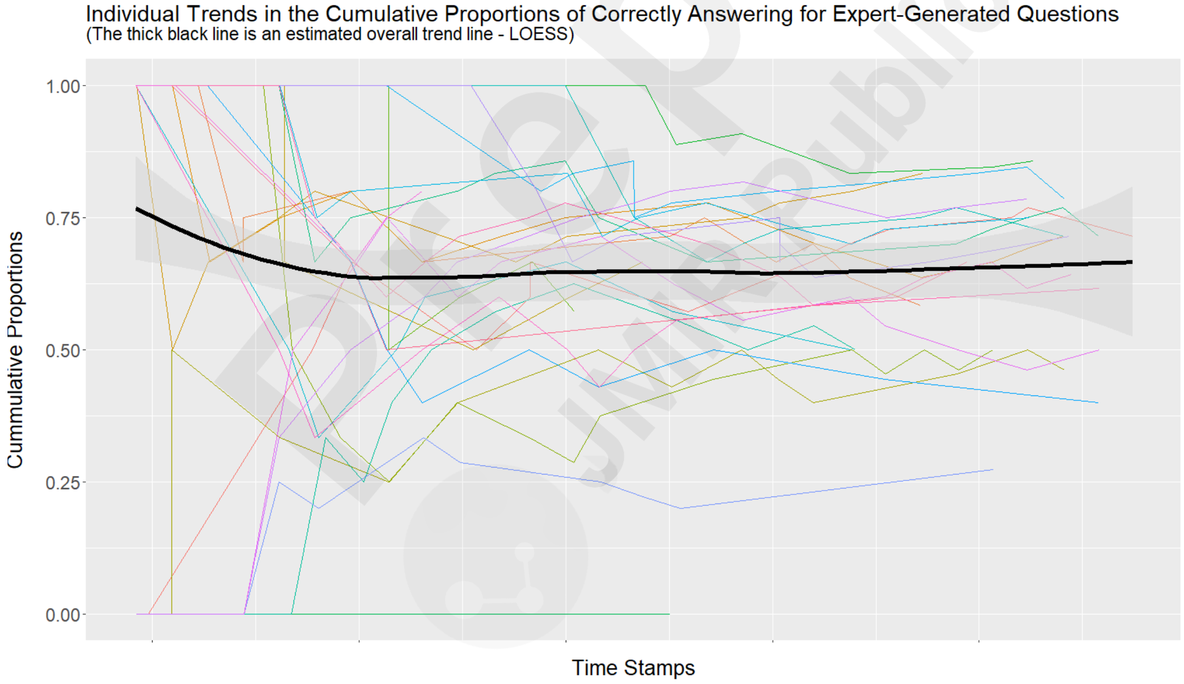


Figure 3: Individual trends in the cumulative proportions of correctly answering expert-generated questions. The thick black line is an estimated overall trend line.

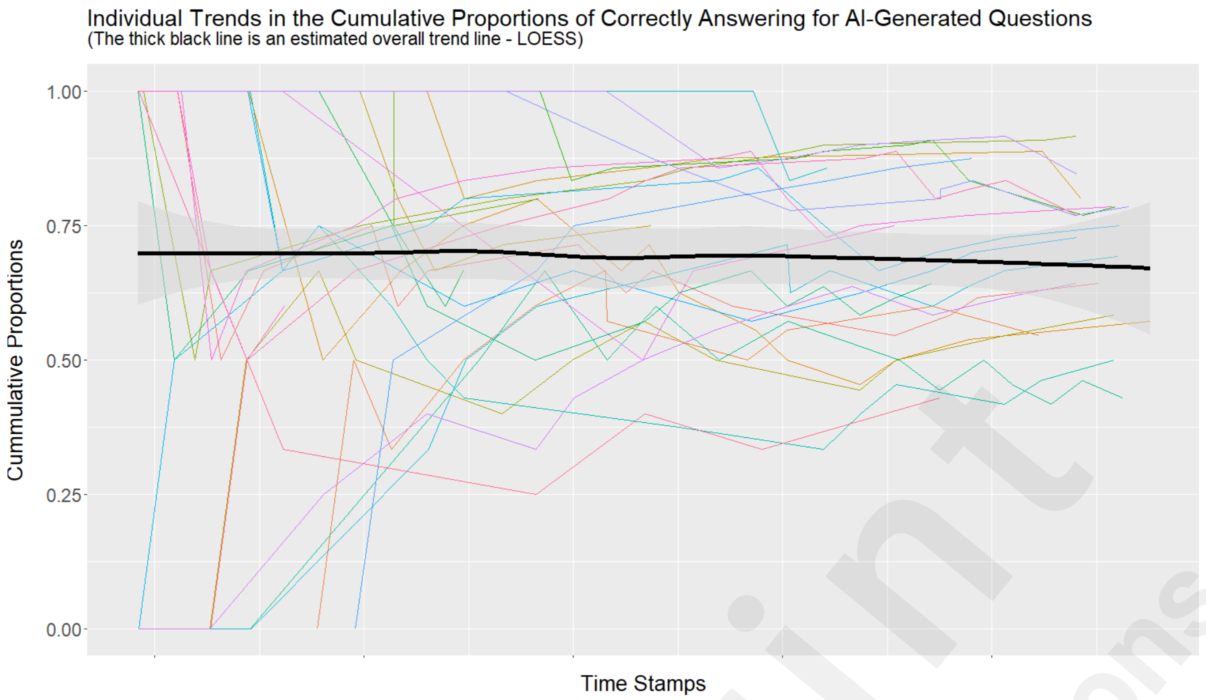


Figure 4: Individual trends in the cumulative proportions of correctly answering AI-generated questions. The thick black line is an estimated overall trend line.

Within these graphs, a black line on the 0.5 in the Cumulative Proportions axis indicates a pattern of random guessing. The results of each graph show that the overall trend is far above the random guess. Although there is a slight increase in correctly answering expert-generated questions and a slight decrease in correctly answering AI-generated questions, neither graph shows a drastic increasing nor decreasing trend between the question sets. The results of this analysis support that the participants did not learn patterns for answering each type of question over time. Thus, the results support the hypothesis that there is no significant difference in student performance of AI-generated questions compared to expert-generated questions, even when compared over time.

Participant Feedback: AI-Generated or Expert-Generated

One of the feedback questions each participant was asked to complete after each study question was whether they thought the question was expert-generated or AI-generated. Using the aforementioned day and time metadata for each question, we sorted the questions based on delivery day and time and calculated the cumulative proportions of correctly guessing a question's generator over time for each subject. However, due to some incomplete responses, the number of available data varied across the participants. See Figure 5 for the participant trends of guessing whether a question was expert-generated or AI-generated.

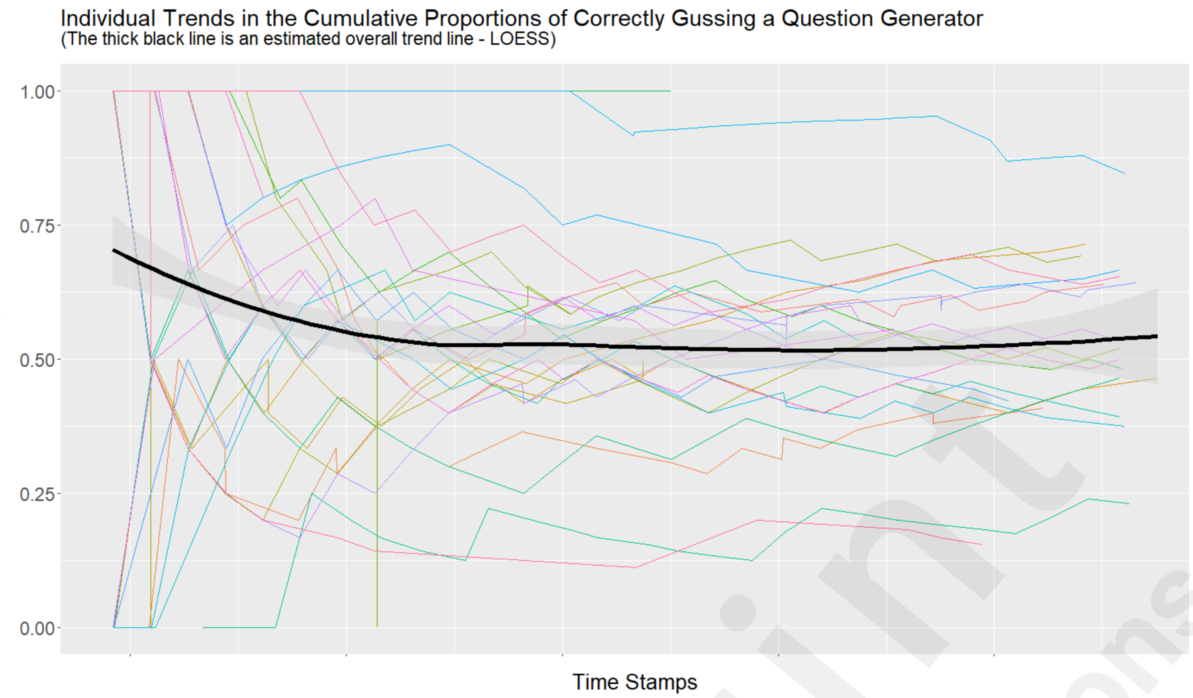


Figure 5: Individual trends in the cumulative proportions of correctly guessing whether a question was expert-generated or AI-generated. The thick black line is an estimated overall trend line.

Based on Figure 5, we can see the proportions approach the estimated overall value of around 0.5 (50% correct), which represents a random guess. Thus, as the number of feedback responses increased, the proportions approached a case of random guessing. Although we observed a slight increasing trend toward the end of the study, it is not significant enough to suggest a learned pattern from the participants. This data suggests that throughout the study, participants did not improve their assumption of whether a question was expert-generated or AI-generated. Thus, the results of this analysis support the hypothesis that there is no significant difference in the participants' ability to guess whether a question was AI-generated or human-generated.

Participant Feedback: Perceived Difficulty Level

To assist with determining the quality of the questions between expert-generated and AI-generated, we asked each participant to rate their perceived level of difficulty for each question using a Likert Scale of Difficulty (very easy, easy, neutral, difficult, very difficult). To conduct this analysis, we calculated the average difficulty of each question and then used the averages to conduct Wilcoxon's Rank-Sum Test. Figure 6 shows a histogram of the average difficulties across each type of question.

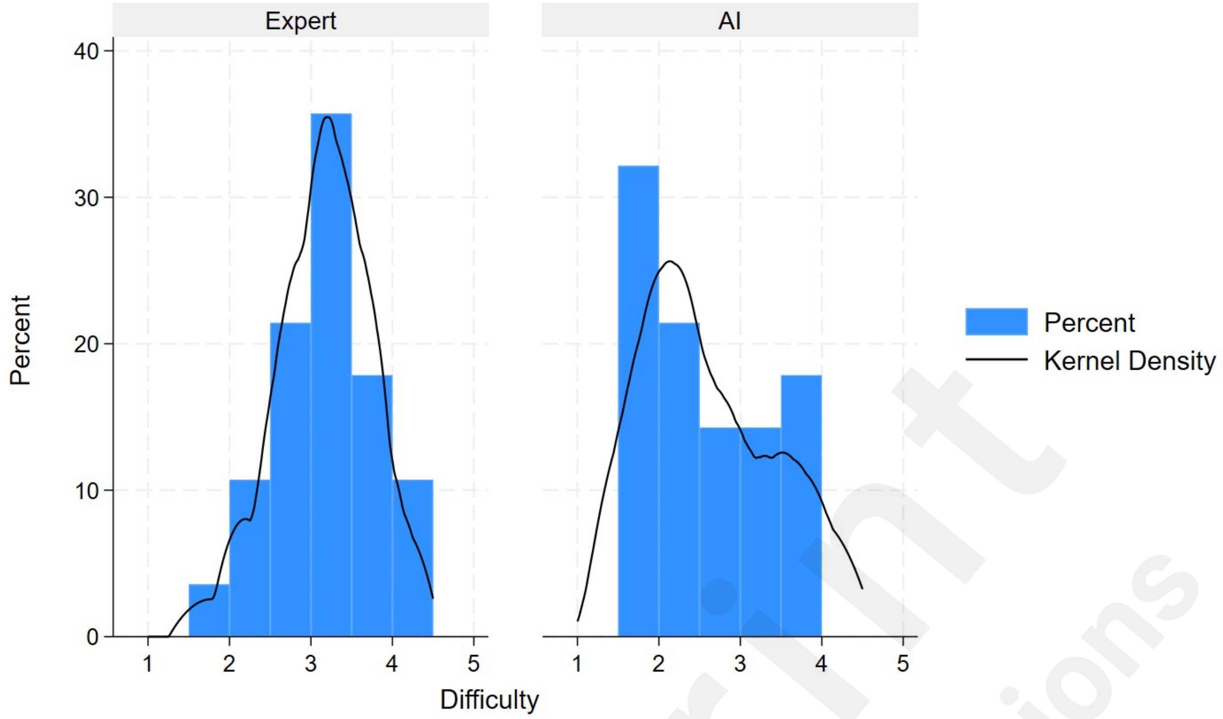


Figure 6: Histogram of Difficulty between expert-generated and AI-generated questions.

In Figure 6, the x-axis of each histogram represents perceived difficulty, with the lower numbers representing the questions perceived the least difficult and the higher numbers representing the questions perceived the most difficult. The y-axis represents the percentage of participants who selected their perceived difficulty across both question types (AI-generated and expert-generated).

The Figure 6 histograms represent distributions of difficulty for each question generator. The height of each bar shows the percentage of questions falling within a specific difficulty range. For example, the height of the first bar in the histogram for the Expert-generated questions represents the percentage of questions with difficulty between 1.5 – 2.0.

Within this histogram, we can see the expert-generated questions were perceived as more difficult by the participants. The p-value was very small (p=0.0051), indicating there is a significant difference in the perceived average difficulty between the expert-generated and AI-generated questions. For this analysis, the results support rejecting the hypothesis as there is a significant difference in perceived difficulty of the AI- versus expert-generated questions.

Perceptions of Generative AI

All participants were asked to complete a pre-assessment and a post-assessment as part of this study. These assessments asked questions pertaining to the participants' perception of generative AI before and after participating in the study. See Figures 7-10 for the response percentages to the pre- and post-assessments.

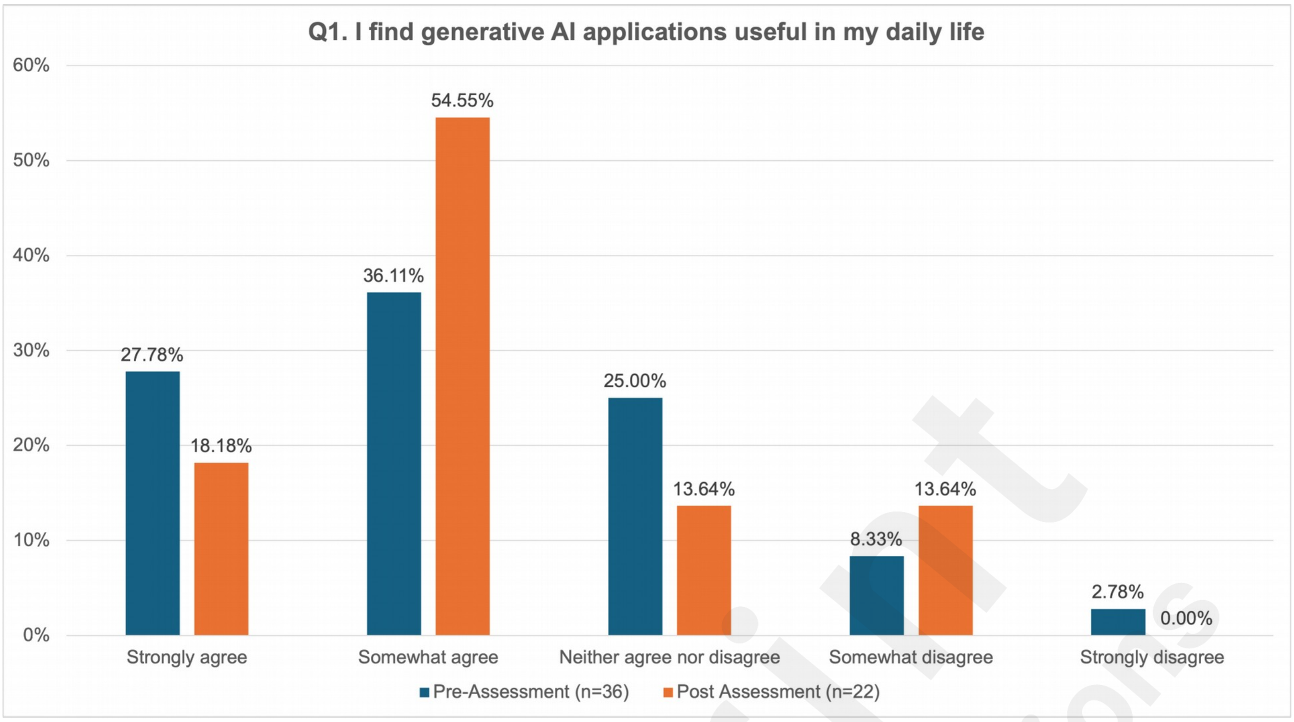


Figure 7: Responses to the question, “I find generative AI applications useful in my daily life,” by percentage.

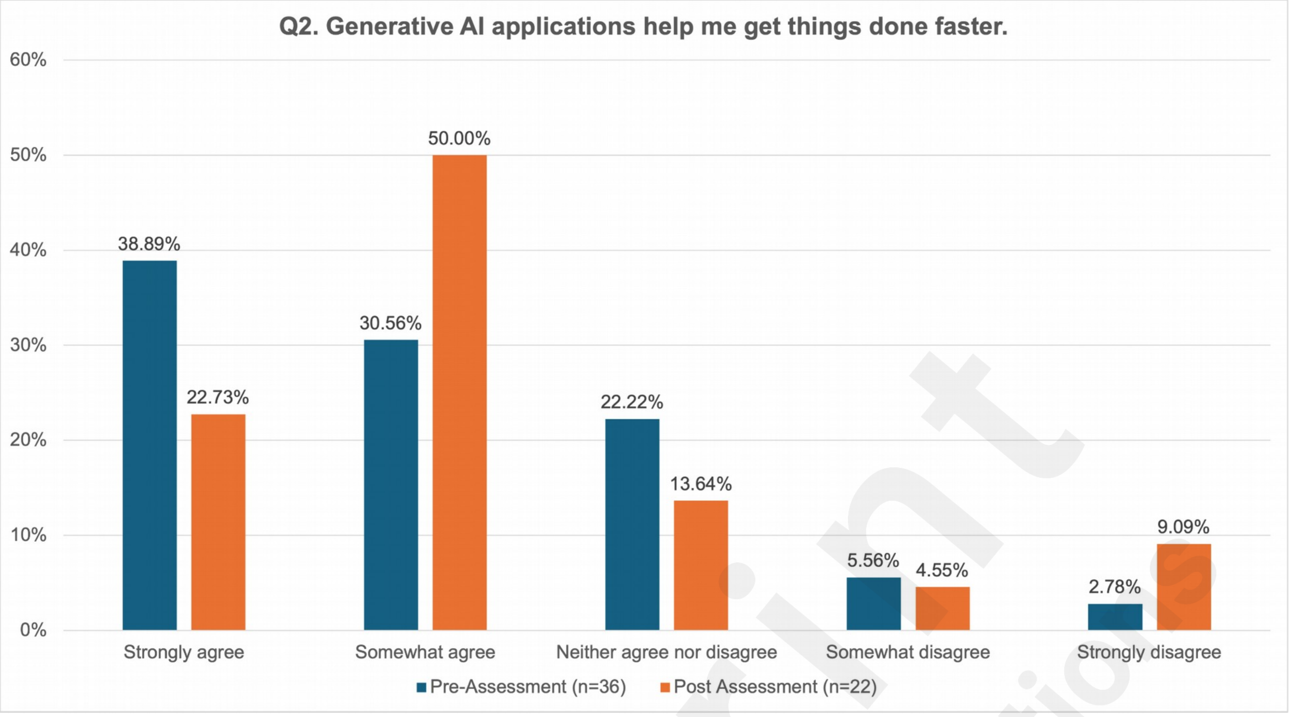


Figure 8: Responses to the question, “Generative AI applications help me get things done faster,” by percentage.

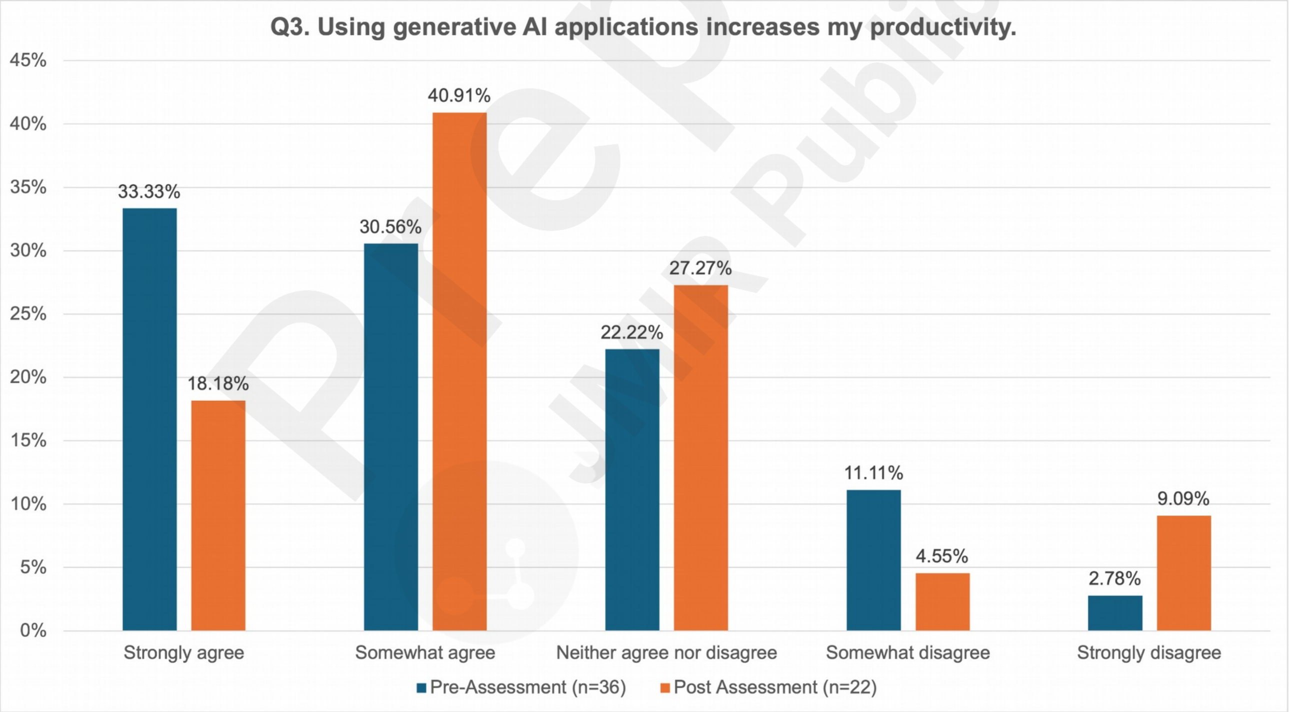


Figure 9: Responses to the question, “Using generative AI applications increases my productivity,” by percentage.

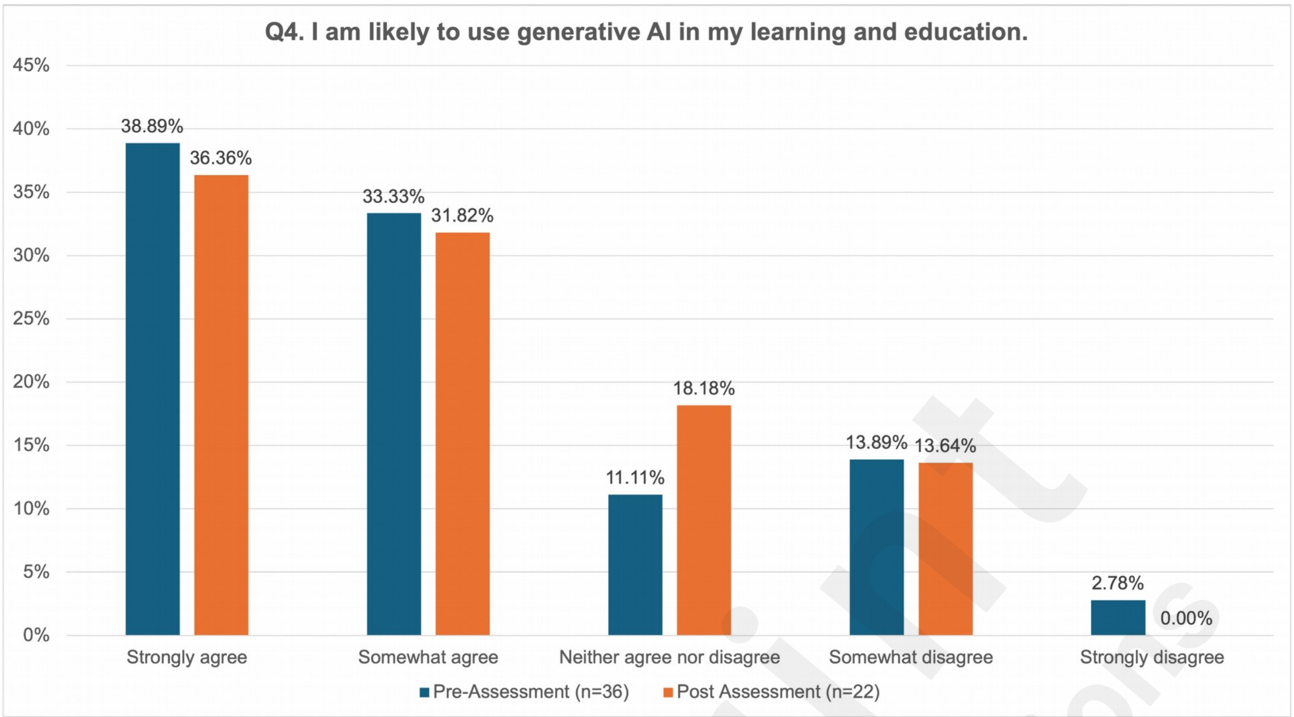


Figure 10: Responses to the question, “I am likely to use generative AI in my learning and education,” by percentage.

The purpose of these questions was to ascertain the participants’ perceptions of using generative AI in education and if their perceptions changed after participating in the study. In response to the first assessment question, “I find generative AI applications useful in my daily life,” the largest change in the post-assessment was an increased 18.44 percentage shift to “somewhat agree,” indicating a positive change in perception of generative AI. For the second assessment question, “Generative AI applications help me get things done faster,” the largest change in the post-assessment was a 19.44 percentage increase to “somewhat agree,” demonstrating that the participants perceived generative AI as beneficial to their efficiency. In the third assessment question, “Using generative AI applications increases my productivity,” the “strongly agree” decreased by 15.15 percent while “somewhat agree” increased by 10.35 percent and “neither agree nor disagree” increased by 5.05 percent. Finally, the responses to the last question, “I am likely to use generative AI in my learning and education,” remained predominantly flat with an increase of 7.07 percent in the response “neither agree nor disagree”. This indicates that there was little to no change in the participants’ likelihood of using generative AI as a result of participating in this study. Overall, this data suggests that the participants in this study are receptive to the use of generative AI in their medical education.

Discussion

Perceptions of Generative AI

A challenge for any new technology is adoption. While the public and students have been quick to adopt new content generation technologies such as ChatGPT for meme creation and homework, there are many uses for generative AI technologies that can also help educators. This study focused on using GPT4 to alleviate some of the faculty burden associated with developing content and formative and summative assessment materials in medical schools. To determine if medical students would be amenable to studying and learning from generative AI-created content, we asked participants their perceptions of this technology. The results from the pre- and post-assessments suggest that the participants either strongly agreed or somewhat agreed with having positive perceptions of generative AI. While our study and other similar studies are testing students’ abilities to respond to AI-generated questions and finding positive results, it will be important for future studies to ascertain the ease of use, time needed, and feelings of adoption from various educators.

Additionally, our study used a new course delivery platform, QuizToo, to administer the MCQs to the participants [3]. Micro-learning platforms have yielded positive results of retention and can be used to deliver just-in-time courses or supplement a class [7, 8]. While QuizToo can be used to deliver courses on any topic, it was designed to facilitate learning for busy physicians, nurses, and other healthcare professionals who don’t have time in their schedules to stop and dedicate time to an hour-long course. Instead, this platform delivers a daily (or weekly, depending how the course is set up) question to the learner’s smartphone or email (learner’s choice). Additionally, depending on the course, learners can select the time they wish for the question to be delivered, enabling flexibility based on individual schedules.

Similarities and Differences between Expert-generated and AI-generated MCQs

In addition to studying perceptions of AI, we also tested for accuracy and difficulty, both in aggregate and viewed over the duration of the study, for expert-generated and AI-generated questions. While there were some marginal differences between the expert-generated and AI-generated MCQs, such as the expert-generated questions being perceived as more difficult, the results from each of the statistical tests were predominately similar. However, a surprising result was that the participants were largely guessing whether a question was expert-generated or AI-generated. To combat cheating, there are now multiple guides circulating for how to tell when content is AI generated and some obvious signs, such as sentences written oddly as though they were translated from a foreign language. Thus, we anticipated that participants would be able to learn the GPT4 writing pattern over time. When generating the questions, we had an expert review the questions for accuracy and to ensure they made sense, but we did not make any grammatical changes to the questions or answer choices. Thus, we found it impressive that over the 28-day period, participants were unable to improve their guess of the origin of the question over time.

There was also no statistical significance of the correct answers between the expert-generated and AI-generated questions for the participants. However, in the 80-100 percentile range of correctness, participants were more likely to answer the AI-generated questions correctly. Alternatively, although there was no dramatic difference, over time the participants slightly improved at answering the expert-generated questions over the AI-generated questions. Further research of the individual questions with a larger population is needed to understand the nuanced fluctuations in this data. While these fluctuations exist, they are very minimal and do not show an indication of great differences between the expert-generated and AI-generated questions.

AI Can Generate Expert-level Questions

Overall, for faculty burden to be alleviated by using generative AI to assist with question generation, these tests needed to show that the AI-generated questions had the same look and feel and were of the same quality as expert-generated questions. The results of the tests we conducted on the data from this study support our hypothesis that generative AI can create questions as effectively as a human expert of the subject matter. It is important to note that the questions were not downloaded directly from GPT4 and uploaded into QuizToo without oversight. At this point in generative AI technology development, there is still a critical need for experts to review the generated content to check for accuracy, quality, and readability. However, we believe that it is still remarkably easier to upload content, generate questions, and review than it is to develop unique questions, correct answers, and especially distractors. As generative AI improves, this process will continue to simplify and expedite the creation of formative and summative assessments and educational content.

Future Directions

The results of this study show great promise for the use of AI to assist educators when creating learning materials and assessments for medical students. However, further testing needs to occur to ensure the results are transferable across the broader population. This study's cohort was small and conducted within a single institution. Future studies should encompass a larger body of medical students across multiple medical schools with a variety of curriculum structures to discover how generative AI question creation can be adapted to multiple settings and scenarios. Additionally, future studies should also scale to incorporate the education needs of higher-level medical students, such as those in their third and fourth years who are working through their clinical rotations.

In future studies, we will also scale our evaluation to generate a better understanding of the participant's perception of difficulty. During evaluation of the questions, we will collect more data including the Likert Scales of Difficulty and incorporate Bloom's Taxonomy to discover why questions are considered difficult and how they influenced a participant's ability to learn the material [9-12].

Conclusion

Generative AI has quickly become instrumental in many use cases across the educational landscape, many of which can alleviate the content and assessment development burden for instructors and educators. While generative AI can be used across all types of education, our data supports that generative AI can be leveraged to create formative and summative assessment questions to reduce faculty burden at medical schools. As this technology develops, it will be essential to continue testing its usability and efficacy to facilitate medical education.

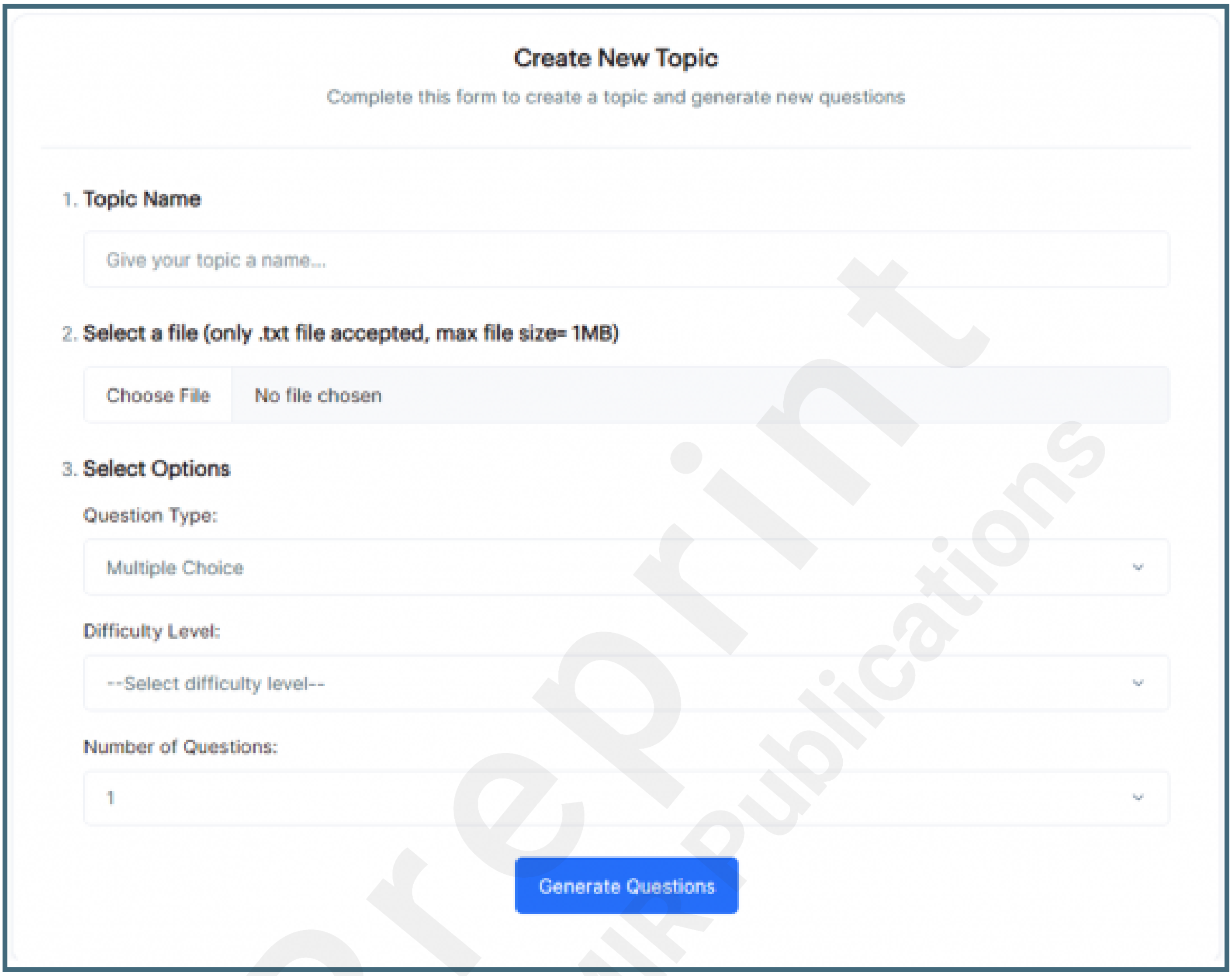
References

1. Kıyak YS, Coşkun Ö, Budakoğlu Iİ, Uluoğlu C. ChatGPT for generating multiple-choice questions: Evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam. *Eur J Clin Pharmacol*. 2024;80(5):729-735. doi:10.1007/s00228-024-03649-x
2. 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*. 2024;99(5):508-512. doi:10.1097/ACM.0000000000005626
3. *QuizToo*. (n.d.). https://quiztoo.com/
4. ScholarRx. (2024, August 26). *Medical school curriculum & assessment Resources | ScholarRX*. https://scholarrx.com/rx-bricks/
5. Openmedlab. (n.d.). *Awesome-Medical-Dataset/resources/MedQA.md at main · openmedlab/Awesome-Medical-Dataset*. GitHub. https://github.com/openmedlab/Awesome-Medical-Dataset/blob/main/resources/MedQA.md
6. *meta-llama/Llama-2-70b-chat-hf · Hugging Face*. (n.d.). https://huggingface.co/meta-llama/Llama-2-70b-chat-hf
7. Ahmed, T., Stinson, K., Johnson, J., & Latif, Z. (2024). QuizTime: Innovative Learning Platform to Support Just-In-Time Asynchronous Quizzes to Improve Health Outcomes. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, *2023*, 253–260.
8. Jape, D., Zhou, J., & Bullock, S. (2022). A spaced-repetition approach to enhance medical student learning and engagement in medical pharmacology. *BMC medical education*, *22*(1), 337. https://doi.org/10.1186/s12909-022-03324-8
9. *Sample Likert Scales // Division of Belonging and Student Affairs // Marquette University*. (n.d.). https://www.marquette.edu/student-affairs/assessment-likert-scales.php
10. Adams N. E. (2015). Bloom's taxonomy of cognitive learning objectives. *Journal of the Medical Library Association: JMLA*, *103*(3), 152–153. https://doi.org/10.3163/1536-5050.103.3.010
11. McDaniel, R. (2010, June 10). Bloom’s Taxonomy. *Vanderbilt University*. https://cft.vanderbilt.edu/guides-sub-pages/blooms-taxonomy/
12. Morton, D. A., & Colbert-Getz, J. M. (2017). Measuring the impact of the flipped anatomy classroom: The importance of categorizing an assessment by Bloom's taxonomy. *Anatomical sciences education*, *10*(2), 170–175. https://doi.org/10.1002/ase.1635

Supplementary Files

Figures

Interface used to provide GPT4 the information to create the AI-generated questions.



Create New Topic

Complete this form to create a topic and generate new questions

1. Topic Name

Give your topic a name...

2. Select a file (only .txt file accepted, max file size= 1MB)

Choose File No file chosen

3. Select Options

Question Type:

Multiple Choice

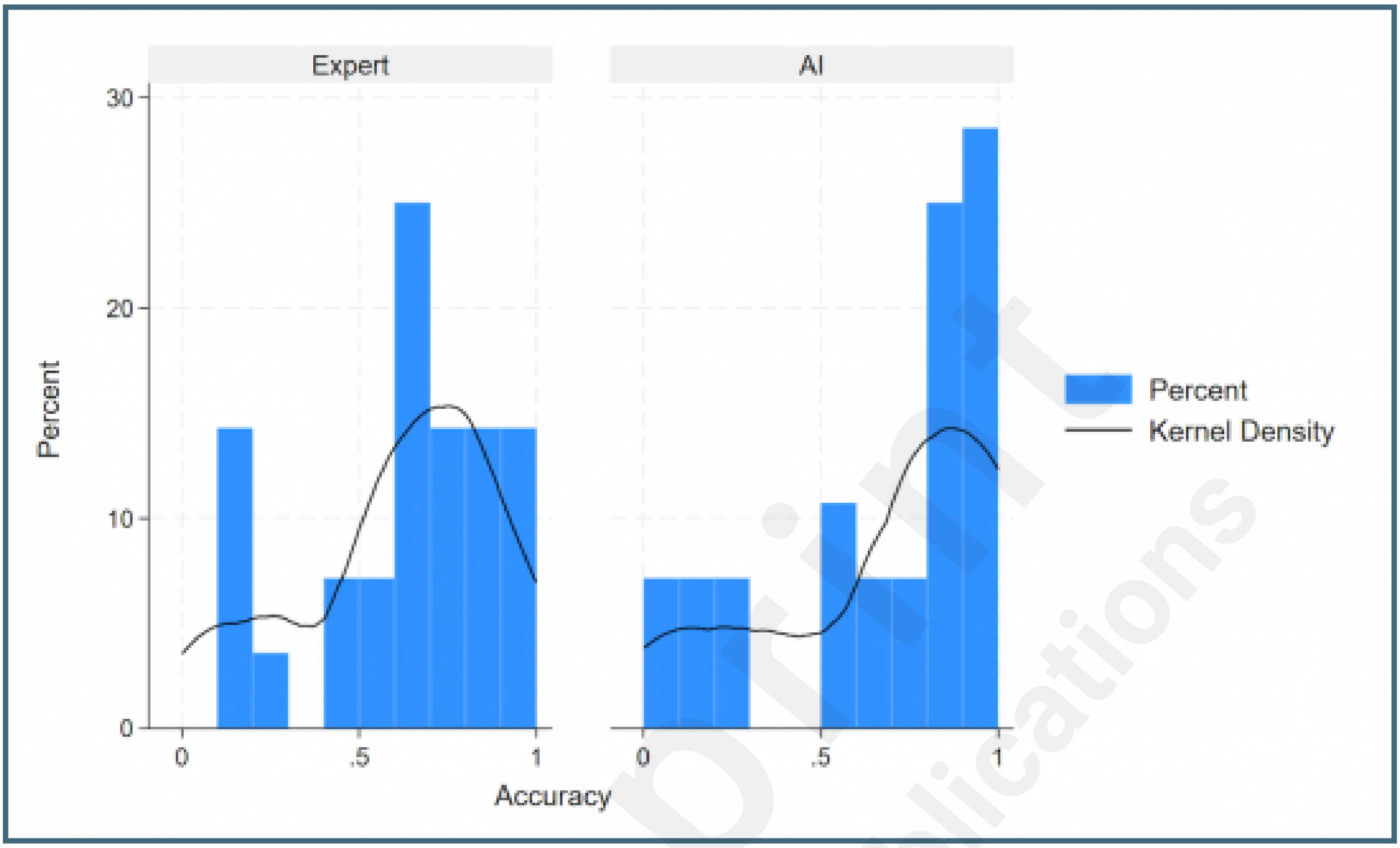
Difficulty Level:

--Select difficulty level--

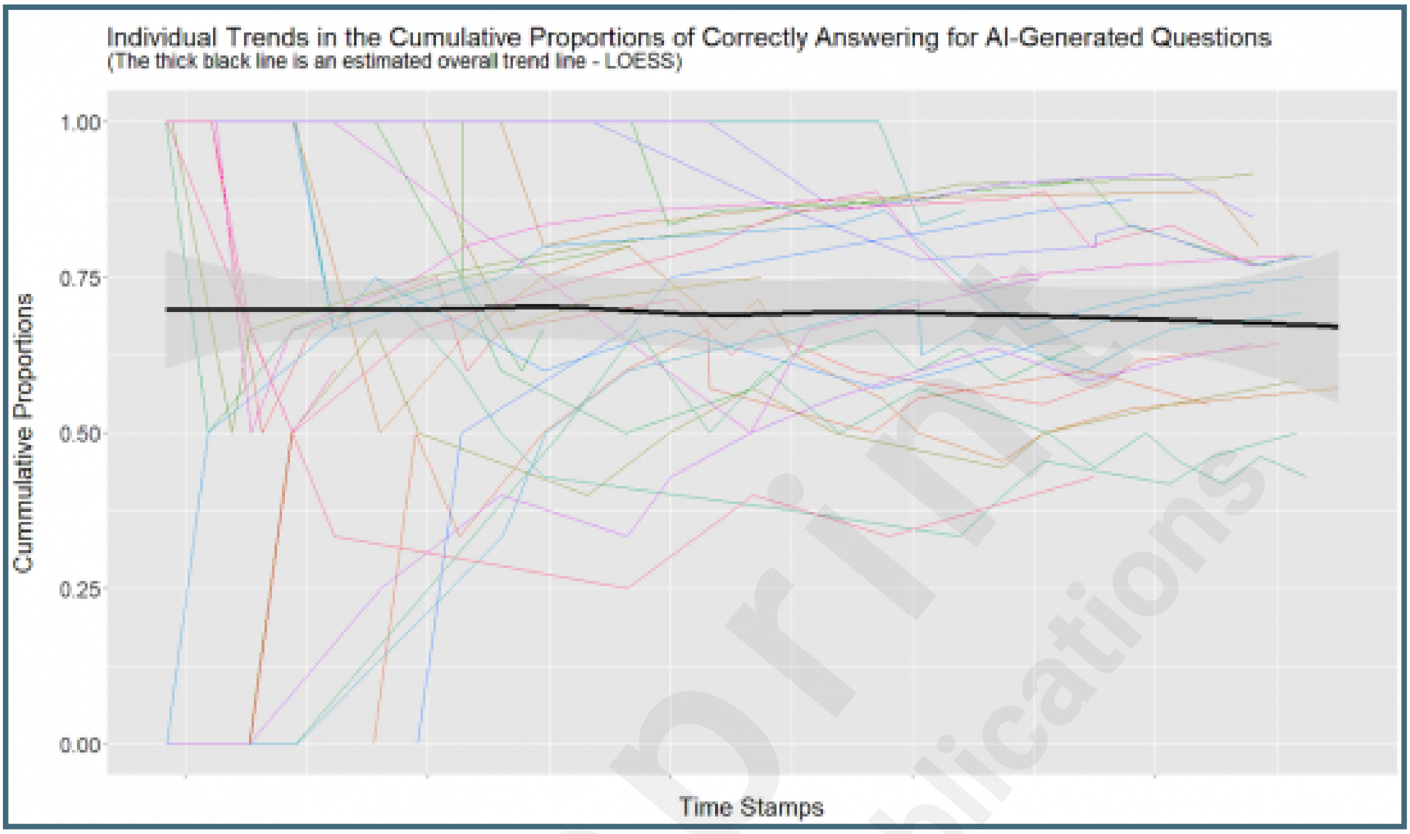
Number of Questions:

1

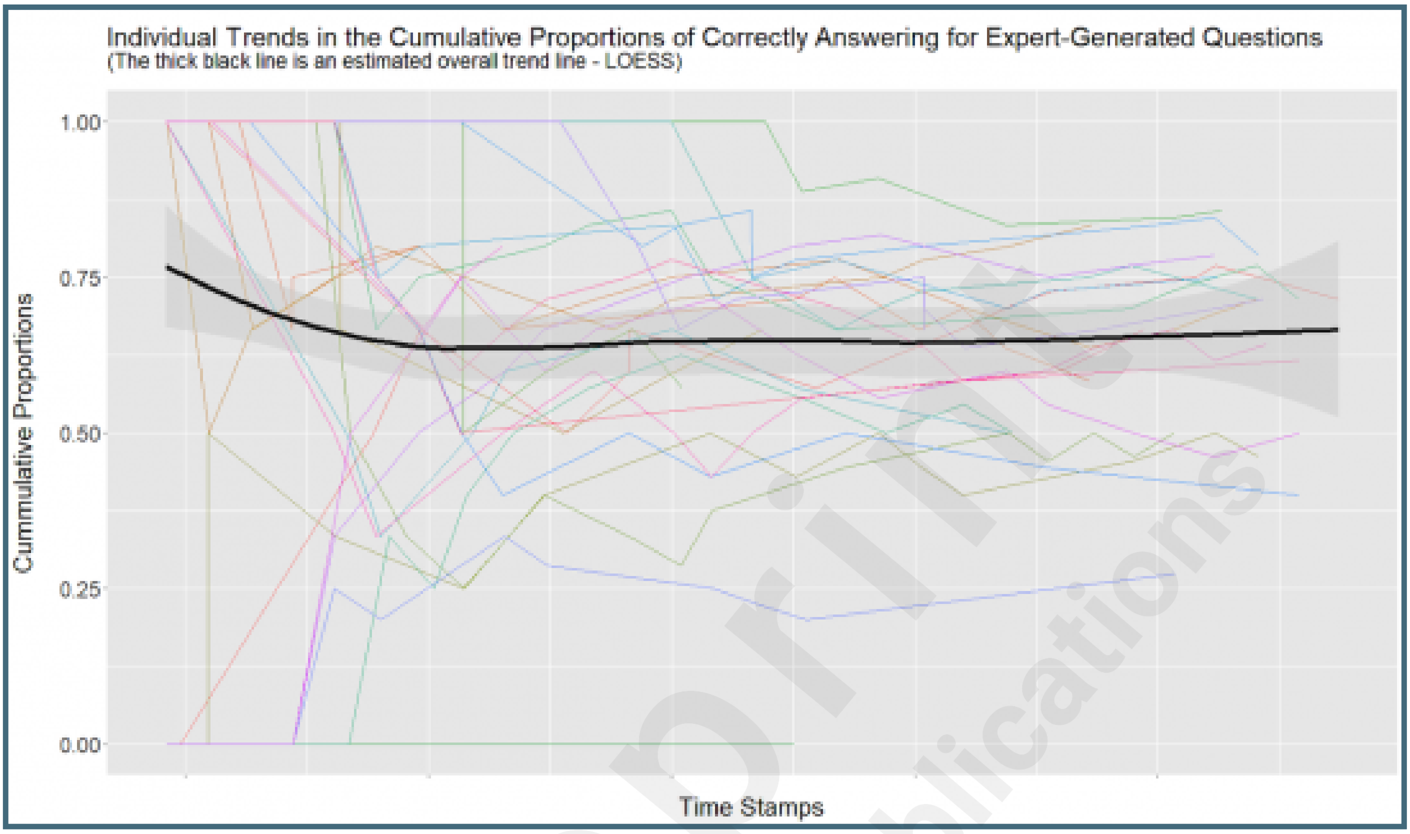
Proportion of students at various performance deciles for expert versus AI-generated questions.



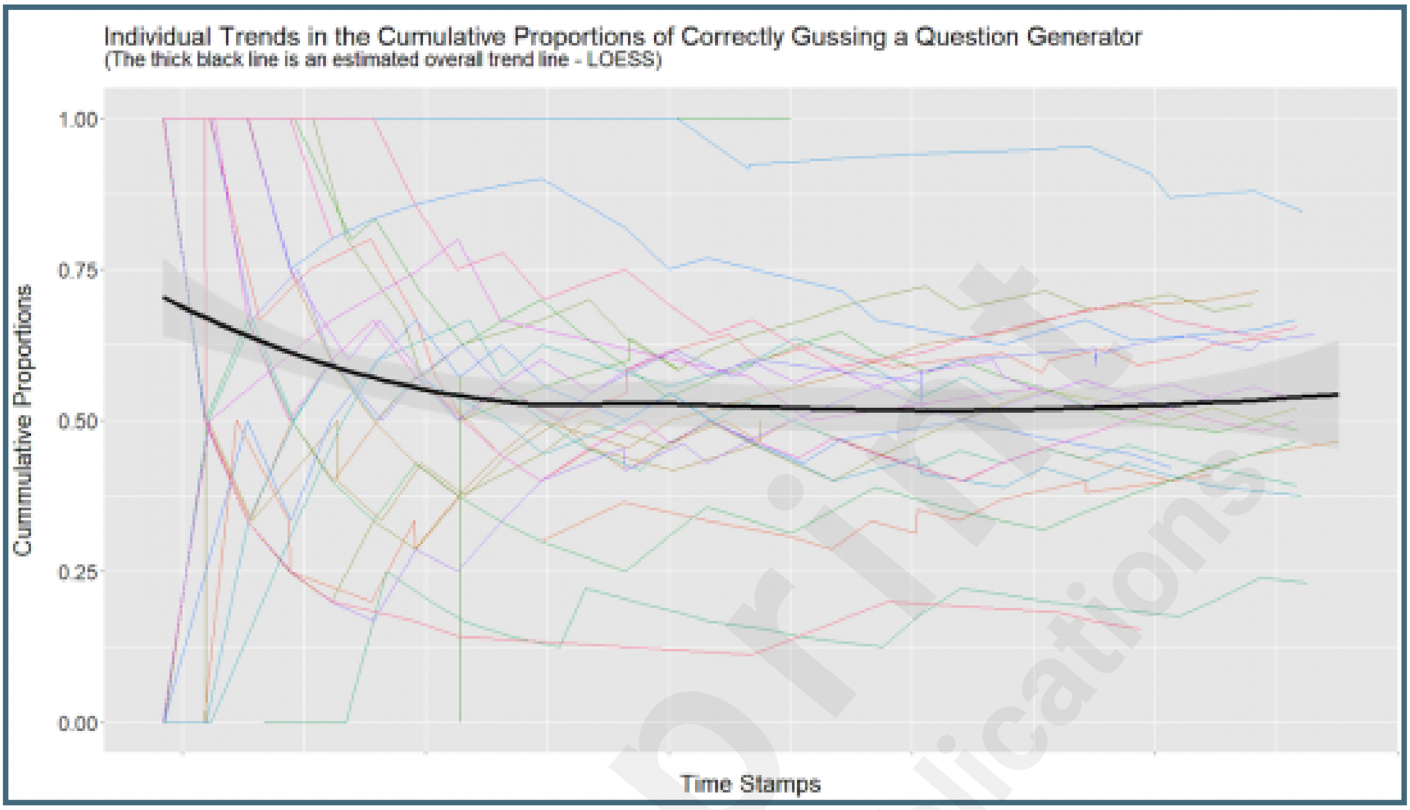
Individual trends in the cumulative proportions of correctly answering expert-generated questions. The thick black line is an estimated overall trend line.



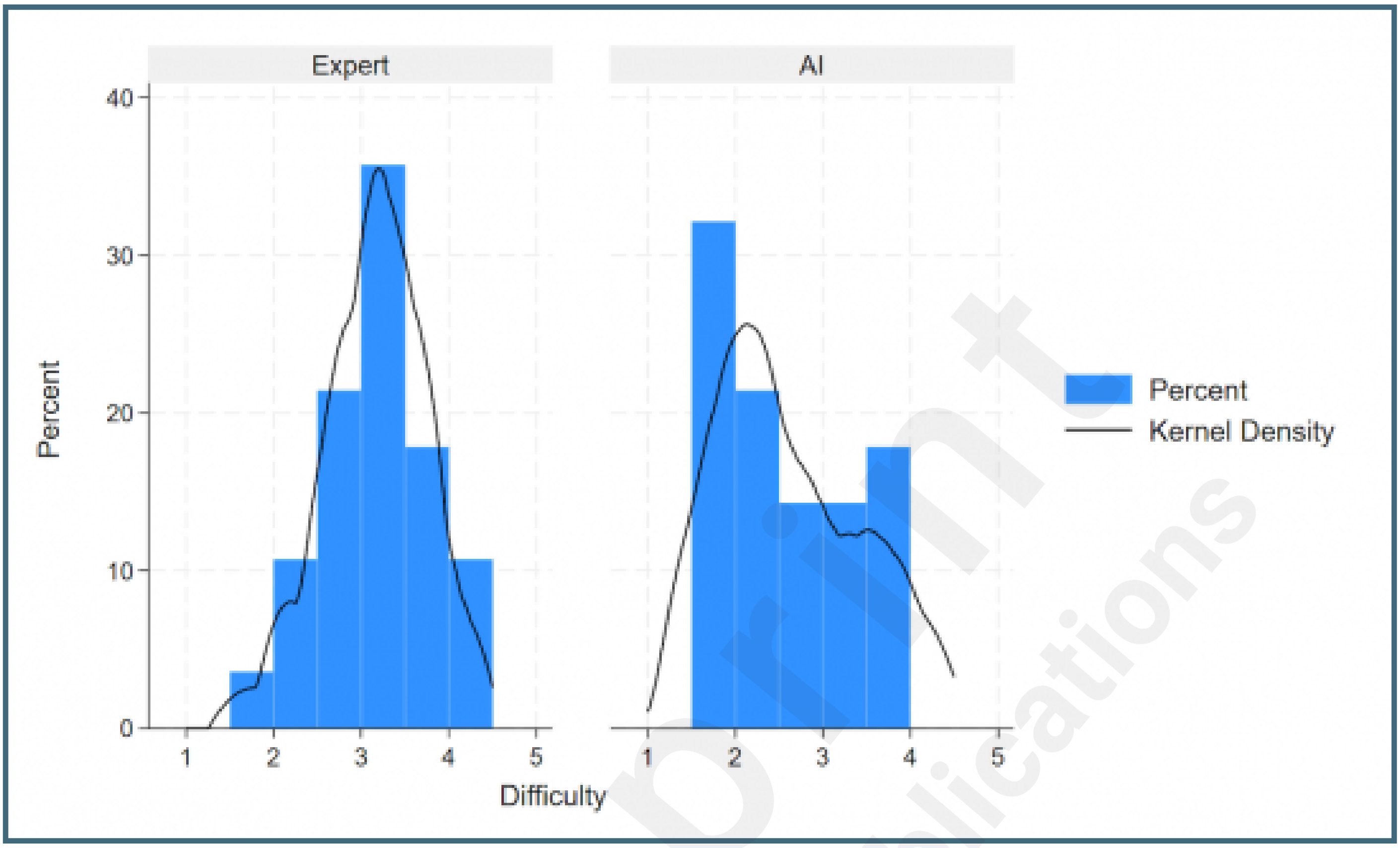
Individual trends in the cumulative proportions of correctly answering AI-generated questions. The thick black line is an estimated overall trend line.



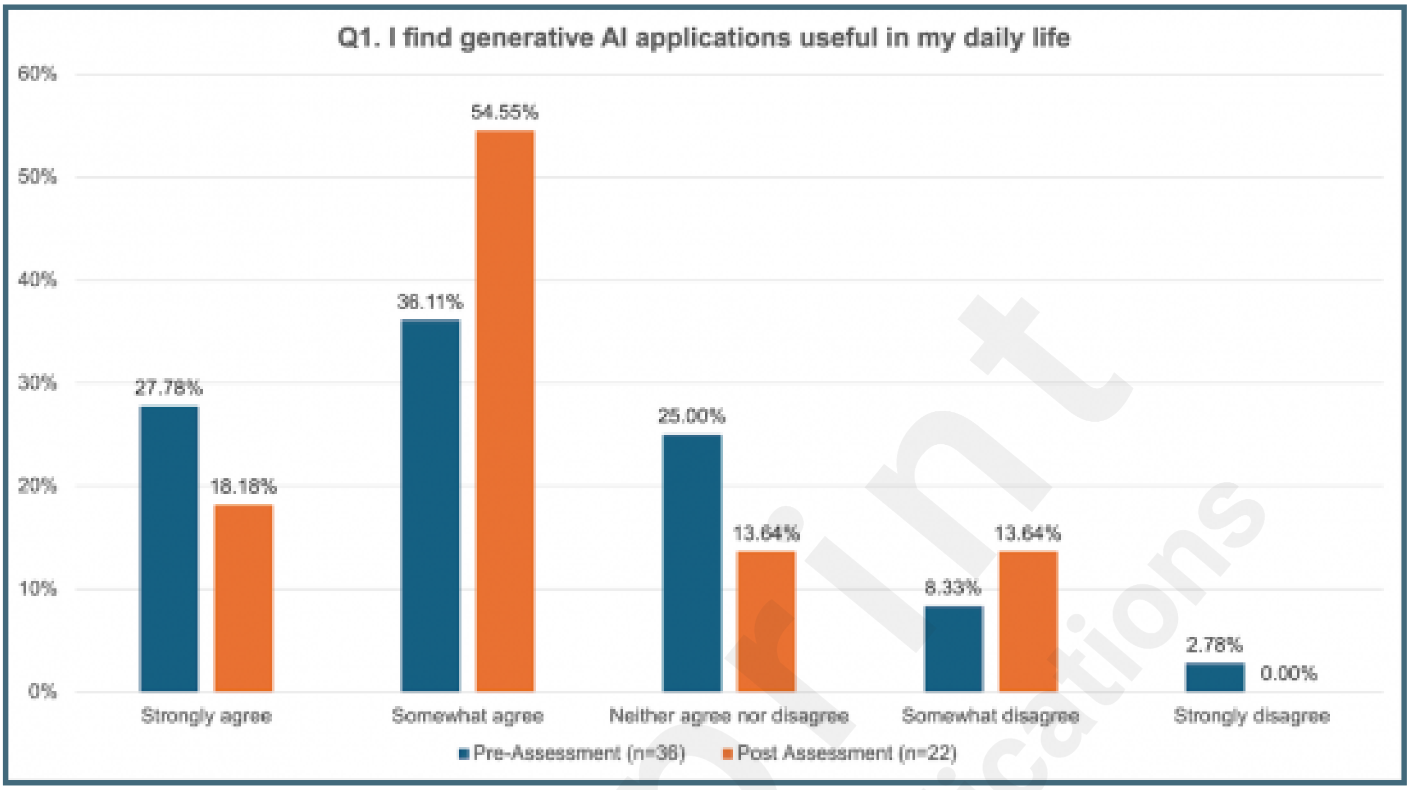
Individual trends in the cumulative proportions of correctly guessing whether a question was expert-generated or AI-generated. The thick black line is an estimated overall trend line.



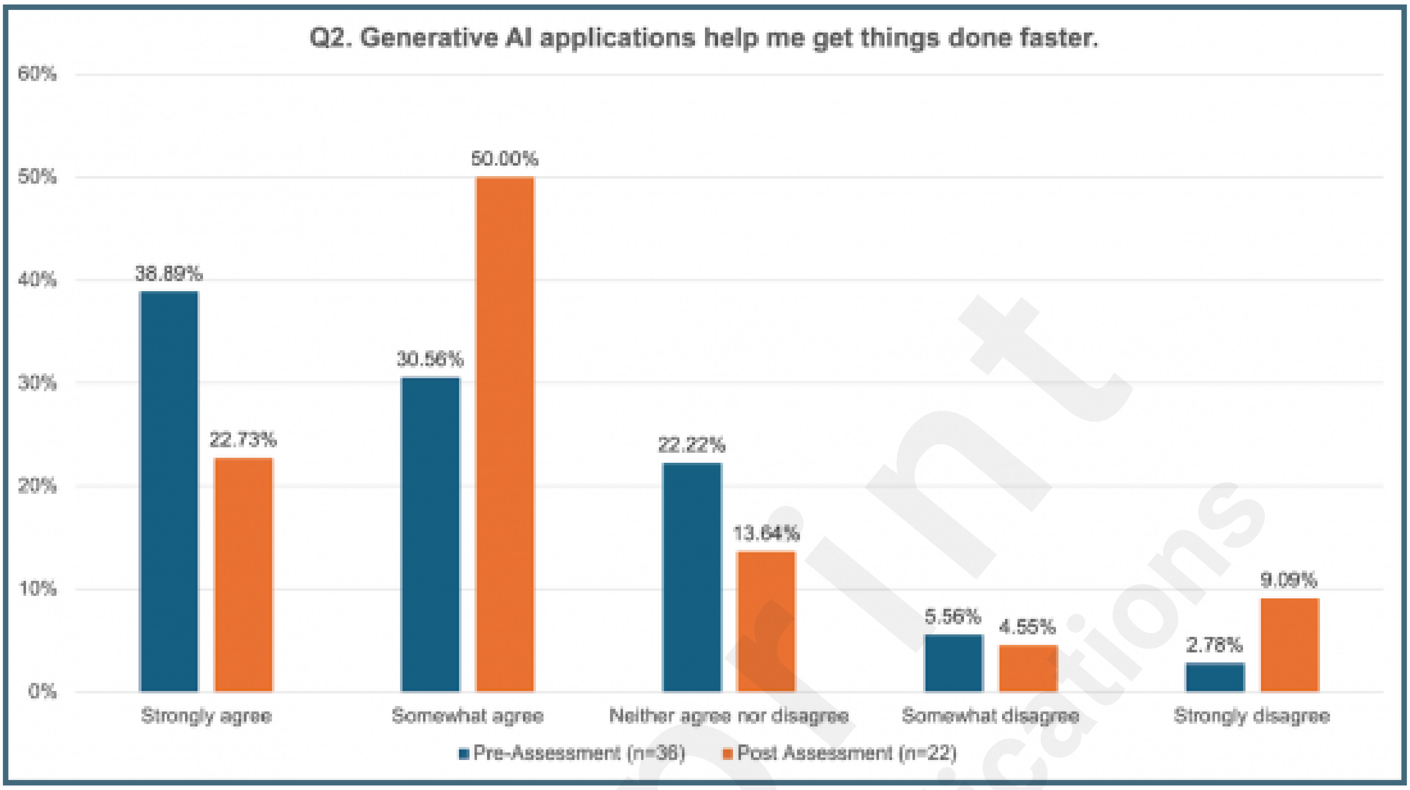
Histogram of Difficulty between expert-generated and AI-generated questions.



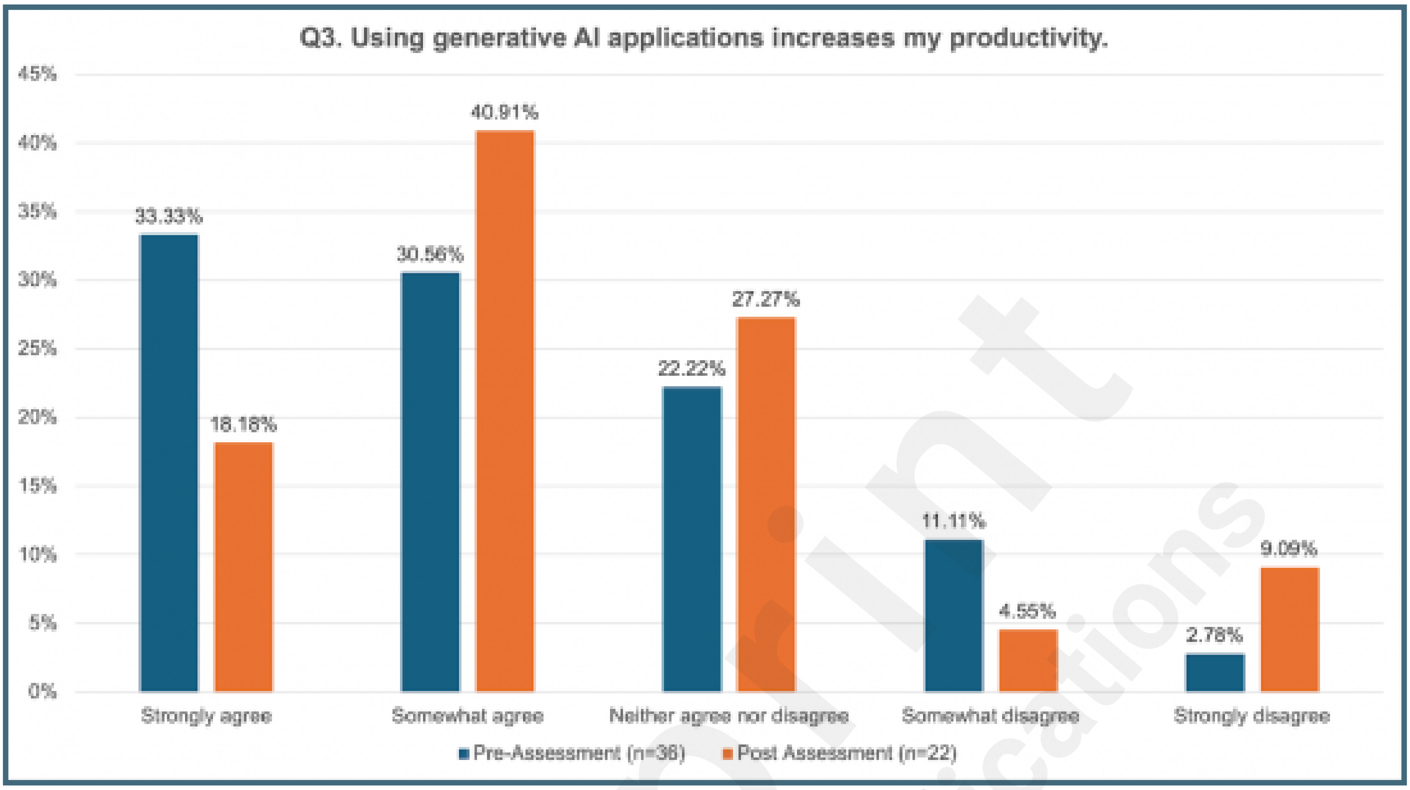
Responses to the question, “I find generative AI applications useful in my daily life,” by percentage.



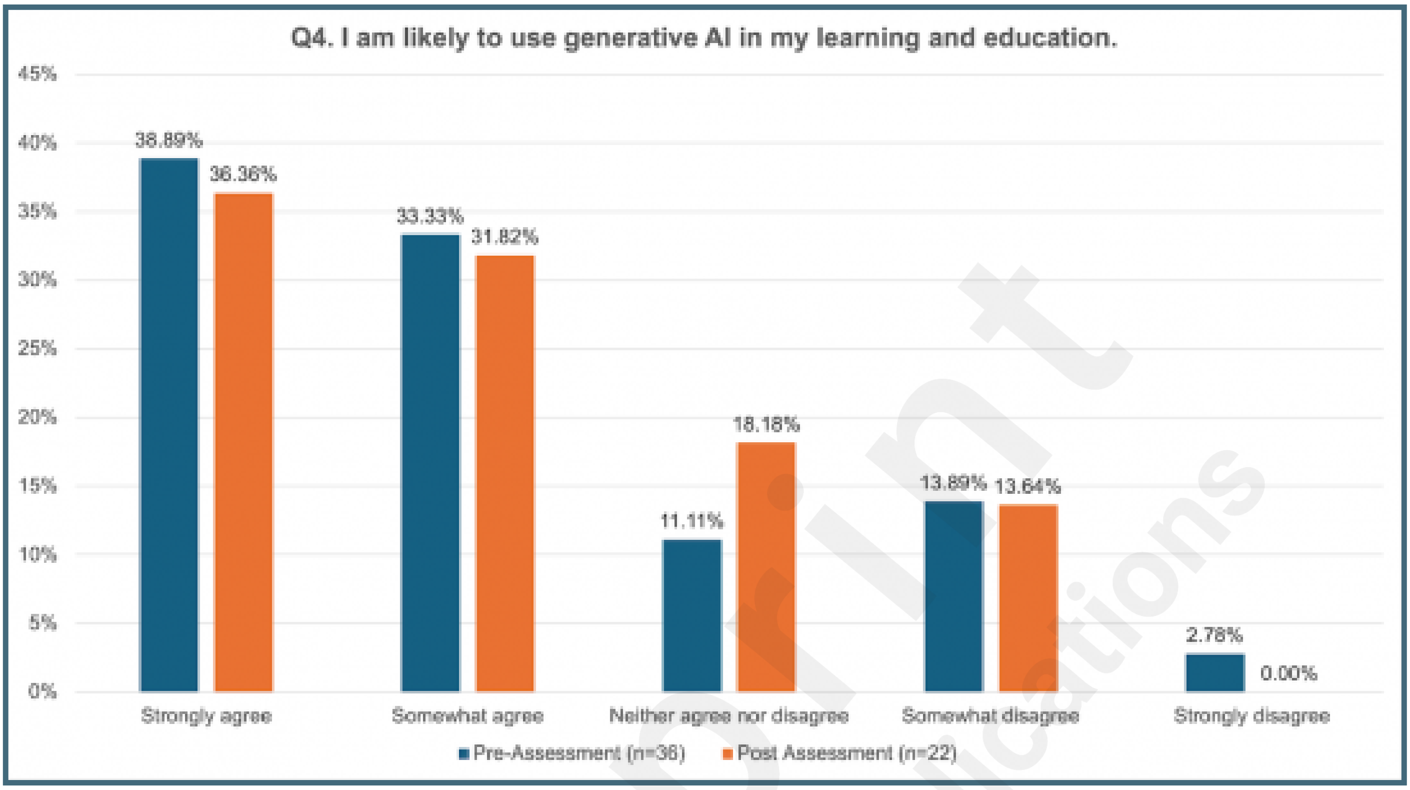
Responses to the question, “Generative AI applications help me get things done faster,” by percentage.



Responses to the question, “Using generative AI applications increases my productivity,” by percentage.



Responses to the question, “I am likely to use generative AI in my learning and education,” by percentage.


Q4. I am likely to use generative AI in my learning and education.
Strongly agree
Somewhat agree
Neither agree nor disagree
Somewhat disagree
Strongly disagree
■ Pre-Assessment (n=36)
■ Post Assessment (n=22)