Open-Ended Questions Need Personalized Feedback: Analyzing LLM-Enabled Features with Student Data

Rachel Van Campenhout[1,*], Jeffrey S. Dittel[1], Bill Jerome[1], Michelle W. Clark[1] and Benny G. Johnson[1]

[1] *VitalSource Technologies, Raleigh, NC 27601, USA*

Abstract

Large language models (LLMs) offer new opportunities to support deeper learning through open-ended, formative practice. This paper investigates two novel types of automatically generated questions: compare-and-contrast prompts and student-authored exam questions. These question types are integrated into an ereader platform alongside conventional fill-in-the-blank items. To enable meaningful interaction with these open-ended tasks, an LLM is used to generate personalized feedback grounded in textbook content. A dataset of more than 90,000 student-question interactions is analyzed to evaluate how these new question types perform in terms of engagement, difficulty, persistence, and non-genuine responses, and how students interact with the LLM-generated feedback. Results are compared across contexts where questions were assigned as part of a course versus used voluntarily. Assigned usage dramatically increases engagement and improves performance across most metrics. To understand how students respond to the feedback itself, timing and textual overlap between the initial LLM-generated feedback and the student's second attempt are examined, revealing distinct patterns of reflection, revision, and potential feedback reuse. These results highlight both the promise and complexity of using LLMs to expand the cognitive scope of automated formative practice while maintaining pedagogical value at scale.

Keywords

automatic question generation, open-ended questions, personalized feedback, large language models, performance metrics

1. Introduction

Automatic question generation (AQG) has been a proficient area of research and development in the past decade, enabled by advancements in natural language processing tools, machine learning techniques, and artificial intelligence. Many approaches have been used to develop AQG pipelines and for equally varied use-cases. However, from their systematic review of 92 AQG studies, Kurdi et al. [1] found only one study that evaluated automatically generated (AG) questions using student data and called for quantitative evaluations of question performance metrics. The AQG system studied in this investigation is an expert-designed, rule-based system that uses textbook content as the corpus for natural language processing in order to select important sentences and key terms and transform them into formative practice questions for students to answer as they read. Formative practice significantly benefits all students, particularly those who struggle or are disadvantaged [2, 3], with integrated practice achieving six times the effect size compared to reading alone [4, 5]. Given this robust causal relationship [5, 6], leveraging AQG to scale formative practice widely was pragmatic. To support equity and access at scale, the AG formative practice was made available for free to any learner who uses textbooks containing it. Prior research on this AQG system in recent

EvalLAC '25: Second Workshop on Automated Evaluation of Learning and Assessment Content, July 22–26, 2025, Palermo, Italy

* Corresponding author.

Rachel.vancampenhout@vitalsource.com (R. Van Campenhout); jeff.dittel@vitalsource.com (J. S. Dittel); bill.jerome@vitalsource.com (B. Jerome); michelle.clark@vitalsource.com (M. Clark); benny.johnson@vitalsource.com (B. G. Johnson)

0000-0001-8404-6513 (R. Van Campenhout); 0000-0002-4913-4427 (J. S. Dittel); 0000-0002-4200-155X (B. Jerome); 0009-0002-1500-9166 (M. Clark); 0000-0003-4267-9608 (B. G. Johnson)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

CEUR-WS.org/Vol-4006/paper5.pdf

years has compared engagement, difficulty, persistence, and discrimination performance metrics of intermixed AG and human-authored questions within a courseware environment [7, 8], and evaluated these same performance metrics at scale with more than seven million student-question interactions [9]. Further research studied student learning behaviors via question interaction patterns [10], the most effective type of automatically generated feedback for student persistence [11], and performance metrics with faculty and student perceptions from classroom implementations [12]. Not only does studying AQ questions using student data provide valuable performance benchmarks for formative learning contexts, but this research leads to iterative improvement cycles of the AQG system itself—ultimately benefiting the learners who use it.

The AQG pipeline as originally created does not use large language models (LLMs) for question generation for two primary reasons: first, the LLMs were far less robust when the AQG pipeline was being developed and second, there is a possibility for factual inaccuracies from LLMs and the scale at which questions are being generated is far too great for human review—an ethical barrier. However, LLM technologies do have many advantages that could be applied to AQG pipelines if done so responsibly. Personalized feedback is one such opportunity. Generating open-ended questions is not challenging, but providing feedback is. Personalized, error-specific feedback is a hallmark feature of intelligent tutoring systems, well established for being the most effective computer-based learning environments [13, 14]. Once again, scale for this type of feedback was largely prohibitive, but text comparison is a strength of LLMs and could provide a solution to this challenge.

In the fall of 2024, two new open-ended question types were added to the existing AG practice question types: a glossary term compare and contrast (C&C) question and a write your own exam question. These two new question types were selected to engage students in higher-level cognitive process dimensions [15]. An LLM is then harnessed to compare the student's response to the relevant section of textbook content and provide constructive feedback. Although the rule-based AQG system was capable of generating such open-ended questions before, implementing them without providing feedback would have left students uncertain about their correctness, risking the perpetuation of misconceptions. Therefore, including these question types required the ability to provide accompanying personalized feedback.

All AG questions are presented within a dedicated "CoachMe" panel alongside the ebook text, allowing learners to interact with questions while reading (Figure 1). Students may attempt each question as many times as desired, receiving immediate feedback indicating whether their response is correct or incorrect. The fill-in-the-blank (FITB) questions have contextual hints generated using related sentences of the same textbook section [11]. Students can revisit the textbook content for support before retrying, or when needed, can choose to reveal the correct answer. Additionally, students may rate the question after submitting a response using the thumbs up and down icon.

Each student interaction with the ereader platform generates microlevel clickstream data, and these "digital traces of student actions promise a more scalable and finer-grained understanding of learning processes" [16]. These high-quality data allow for investigation of learner behaviors as well as learning technologies, allowing for old research questions to be answered in new ways and new research questions to arise from novel data [17]. The clickstream data are stored with a student identifier, so no personally identifiable information is connected with engagement data. The platform does not capture any student demographic data. The analysis in this study includes all students who have answered these questions, with an interest in studying the difference between self-motivated usage and usage when assigned in a course context. Investigating the effectiveness of AI is required to ensure its application is beneficial and performing as intended for learners. The use of AI in educational technology should adhere to AI principles (such as accountability, transparency and explainability, responsibility and ethics, and efficacy) both during its conceptualization and development [18] as well as reporting efficacy findings and continuing to engage in iterative improvement [19].

Beyond evaluating the two new question types, this study contributes to the broader theoretical understanding of how generative AI can facilitate higher-order cognitive engagement [15] and

constructive learning activities [20]. Practically, our findings provide educators and educational technology designers concrete evidence supporting structured integration of cognitively demanding open-ended questions paired with AI-generated personalized feedback. By empirically demonstrating significant differences between assigned and unassigned contexts, this research underscores the critical role of instructional design in maximizing the benefits of automated formative practice at scale.

The primary research questions for this paper are:

1. What are the performance metrics for the new open-ended question types and how do they compare to the existing FITB questions as a benchmark?
2. How do the performance benchmarks differ between contexts where the questions are unassigned (students self-selecting to answer) and assigned (known classroom implementations)?
3. How does the LLM-generated personalized feedback perform?

2. Methods

2.1. Automatic Question Generation

A rule-based AQG pipeline underpins the generation of the standard FITB questions used for comparison in this investigation. While full implementation details can be found in earlier work (e.g., [9]), a brief overview is provided here for context. The pipeline uses spaCy [21] to perform syntactic and semantic analysis of textbook content and applies TextRank [22] to identify sentences deemed important. Very short (fewer than five words) or very long (more than 40 words) sentences are discarded. For each remaining sentence, a set of rule-based filters removes trivial or ambiguous terms (e.g., function words, overly predictable words [23], or list items), leaving only key terms as blank candidates. If multiple terms survive, each is turned into a separate FITB question. These items are placed at major subsection boundaries so that learners regularly encounter formative practice questions while reading the textbook.

2.1.1. Open-Ended Questions with LLM-Enabled Feedback

Building on the rule-based pipeline described above, two new question types extend CoachMe into more open-ended tasks designed to foster deeper cognitive engagement (shown in Figure 1). The existing questions (including the FITB questions used here for comparison) are primarily focused on basic comprehension and most closely align with lower-level recognition and recall cognitive processes in Bloom's taxonomy [15], and according to the ICAP framework [20], help maintain an "active" mode of engagement. Despite the seemingly modest cognitive demands, these question types have demonstrated effectiveness in supporting learning, as evidenced by the doer effect [5, 6]. While the standard items attend to essential knowledge-building, the newly introduced open-ended questions aim to elevate learners further along Bloom's taxonomy and into a more "constructive" mode within the ICAP framework.

The student-authored exam questions direct students to "Write a test question for the section '[Textbook Section Title]' as if you are the instructor preparing an exam." This templated prompt is placed at the end of each major section of the textbook. This aims to promote higher-order thinking by requiring students to reflect on and synthesize key concepts. Having students compose their own exam questions fosters metacognitive awareness and shifts them from simply receiving content to a more constructive level of engagement. Research has found this type of student question creation can increase engagement and significantly enhance comprehension and academic performance, particularly when feedback is provided [24, 25].

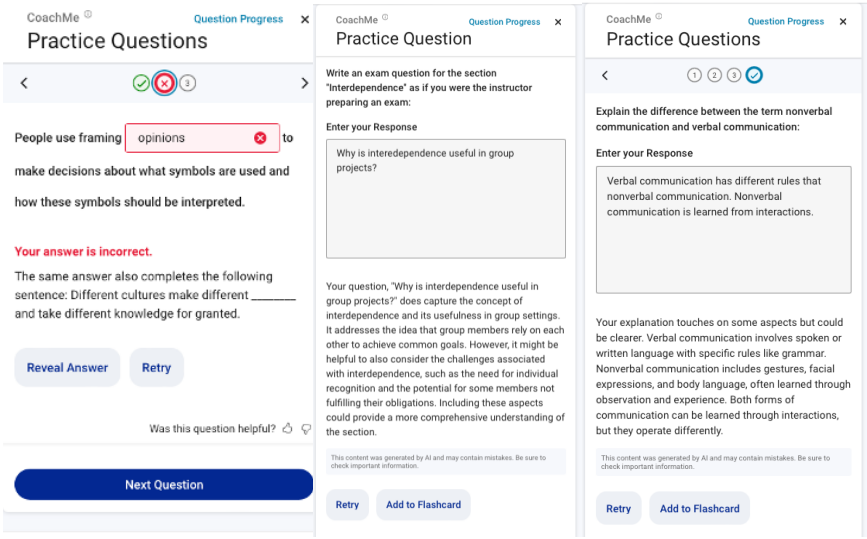

Figure 1: Examples of a FITB question, exam question, and C&C question.

The compare-and-contrast (C&C) questions focus on conceptual clarity by having students compare related glossary terms. The system automatically identifies pairs of "coordinate" terms that share a common final word (e.g., "lactate threshold" and "ventilatory threshold") appearing close to each other in the same textbook section. It then inserts the standardized question stem, "Explain the difference between [Term 1] and [Term 2]." This task asks students to identify subtle distinctions, thereby engaging them in elaboration and deeper processing consistent with the "analyze" cognitive process dimension [15] and constructive engagement [20]. Research on C&C tasks suggests that recognizing similarities and differences and drawing comparisons improves conceptual clarity, facilitates retention beyond mere recall, supports the formation of conceptual categories, and aids in establishing meaningful links among ideas [26, 27, 28].

These new question types do not supplant the standard items; rather, they fulfill complementary roles. The standard questions help ensure students do not passively skim the text without active reflection of foundational content. The new open-ended questions require students to produce new representations of knowledge. This higher-order interaction can bridge connections between concepts more effectively and strengthen long-term retention.

Once a student submits an answer, the platform gathers that response along with relevant textbook passages or glossary entries and forwards them to an LLM-based evaluator. The evaluation process proceeds as follows:

1. An excerpt from the textbook or glossary is supplied to the LLM, ensuring feedback remains grounded in the source material and aligned with the textbook's terminology. This "textbook-centered" approach is designed to minimize hallucinations and maintain consistency with established vocabulary.
2. The LLM is instructed to gauge the accuracy, completeness, and clarity of the student's submission. In the case of a C&C question, for example, the LLM checks whether the student's explanation clearly differentiates the two related concepts. Because these question types are intrinsically open-ended, the system does not classify answers as strictly "correct" or "incorrect." Instead, the evaluator identifies strengths in the student's work and points out areas that might need additional clarification or elaboration.
3. The LLM produces a concise textual critique, which may include praise for effectively capturing key points, suggestions for further detail, or corrections if inaccuracies are evident. In the current implementation, GPT-4o [29] is used for feedback generation, with the temperature parameter set to 0 to decrease the likelihood of hallucinations. Since these question types often call for a higher level of cognitive engagement than the standard items,

the feedback is intended to encourage iterative refinement, allowing students to revise and resubmit their answers if they choose.

2.2. Data Collection

The dataset consists of all student-question interaction events for the LLM-enabled questions gathered from August 15, 2024, through February 9, 2025. The ereader platform stores the raw clickstream data with anonymous identifiers. Student consent for research and analytics is obtained through acceptance of the platform's terms of use and privacy policy. No student characteristics are collected and the learner context is in general not known, though the majority of data comes from higher education institutions in the United States. Data were grouped into student-question sessions, consisting of all actions of an individual student on a single question, ordered chronologically. A session may include multiple attempts on the question and, optionally, a thumbs up or down rating by the student (see Section 3.1.4).

This resulted in a dataset of 83,624 LLM-enabled question sessions (56,944 exam question and 26,680 C&C), encompassing 92,719 interaction events, 23,750 questions, 14,696 students, and 1,929 textbooks. (Because only 544 of these textbooks included a glossary, C&C questions could only be generated for those particular books.) For comparative purposes, data from the standard FITB questions were retrieved for the same textbooks and timeframe, resulting in 1,142,891 sessions spanning 236,511 questions. The datasets are made available in our open data repository [30].

These usage data reflect real-world learning contexts in which some courses assigned the questions as part of a participation grade, while in other cases the questions remained optional. Questions were categorized as either assigned or unassigned based on whether they were part of a known classroom implementation. Specifically, 21 course sections across four institutions were identified in which instructors explicitly required students to complete the practice questions; these constitute the assigned group. All other usage is considered unassigned and typically reflects voluntary student engagement, allowing for comparative analysis between contexts.

2.3. Analysis

2.3.1. Question Performance Metrics

Previous research on AI-generated questions has relied on several core metrics to characterize question performance, including engagement, difficulty, persistence, non-genuine response rates [9, 10], and student ratings [31]. The present study adopts these metrics to compare the performance of new LLM-enabled items and standard FITB questions (detailed in the Results and Discussion section). We adopt an exploratory approach for this study, using mean rates or proportions and quartiles for each metric. If notable differences emerge, future investigations may employ more advanced statistical approaches (e.g., mixed effects regression) to address variables such as subject domain or student-level factors.

However, because the new question types lack predefined correct answers, an LLM-based approach was used to determine correctness (for C&C items) and to detect non-genuine responses (for both types). GPT-4o mini [32] was used to examine each C&C response and its accompanying feedback to decide whether a typical college instructor would reasonably consider it "complete and correct." Exam questions receive no correctness label but are checked for non-genuine attempts. Any submission that does not address the prompt meaningfully (e.g., random text, "idk") was flagged as non-genuine. To ensure reliability, the prompts were iteratively developed using a subset of responses, refining them until the LLM's outputs were consistent with typical college-level evaluation. It was verified that the LLM correctly identified non-genuine answers and assessed C&C accuracy in a way that reflected domain-reasonable expectations. After prompt refinement, spot checks were performed on additional cases in the full dataset to confirm the LLM was applying these criteria consistently. While not a formal validation study, this process ensured that the LLM's classifications were consistent with our instructional intent. Although sufficient for the present

analysis, we acknowledge that a more systematic validation, such as expert annotation of a sample set, would further strengthen the reliability of these measures. This remains an area for future work.

2.3.2. Feedback Usage

To examine how students might use the LLM-generated feedback, the time interval until a second attempt was computed. Specifically, if a student's initial attempt was incorrect or non-genuine and a follow-up attempt occurred, the elapsed time (in seconds) between the submissions was calculated. This is similar to prior work in intelligent tutoring systems, where response latency often serves as a proxy for reflection or cognitive engagement [13, 28]. Because these intervals tend to be skewed, we report the first quartile, median, and third quartile (Q_1–Q_3).

To assess whether LLM feedback fosters learning, the analysis focuses on sessions in which the first attempt was incorrect or non-genuine. The time interval data are stratified by the initial attempt category and the outcome of the second attempt (correct, incorrect, or non-genuine). This framework highlights pivotal transitions, such as moving from a non-genuine to a correct response, and establishes a basis for comparing revision times with textual overlap of the revised attempt with the feedback. Because the LLM's feedback can occasionally provide near-complete model answers, recognizing such overlap is relevant for distinguishing between independent construction of new text and reuse of provided material.

For this analysis, the LLM's feedback and the student's second answer were lowercased and stripped of punctuation to help mitigate superficial differences, then tokenized on whitespace. A token-level gestalt sequence-matching approach [33], implemented via Python's `difflib.SequenceMatcher`, produced a similarity percentage score, where 100% indicates a verbatim match. Reordered text reduces the similarity score, penalizing partial rearrangements. This method is intended to capture literal copying more effectively than simpler distance metrics, as it identifies matching subsequences across the entire submission. These findings are then related to the time interval results, exploring whether rapid resubmission coincides with higher textual overlap.

3. Results and Discussion

3.1. Performance Metrics

3.1.1. Engagement

Engagement measures whether students choose to attempt a given question upon encountering it. It serves as a proxy for how appealing or approachable a question is to students in a given context. Lower engagement may indicate a question type is perceived as more time-consuming, overly difficult, or less beneficial. Because engagement serves as a core driver of the doer effect in formative practice, it remains a critical baseline for understanding how new question types fare relative to standard items. In this analysis, engagement is measured as the number of students who answered each question, which provides a straightforward indicator of how often a question drew student participation when encountered.

Table 1 reports student engagement for each question type. For the assigned group, where engagement was more substantial, the mean and Q_1–Q_3 are reported. For the unassigned group, engagement was consistently low, so only the mean is reported. (In Tables 1–3, all cells represent over 1,000 sessions.) For unassigned questions, most were answered by only a few students. The mean number of students answering each exam question was 2.4, and 2.8 for C&C. The assigned courses show a very different pattern of behavior. A Mann–Whitney U test confirmed that significantly more LLM-enabled questions were answered in assigned contexts than in unassigned contexts ($U = 1.23 \times 10^6$, $p < .001$). The mean numbers of students answering the exam and C&C questions are very close (51.7 and 55.5, respectively), which seems reasonable given the similarity in effort involved. The FITB questions are considerably higher at 84.5, and indeed are answered at a much higher rate at each quartile.

Across the assigned group, FITB questions were answered at a substantially higher rate than exam questions at every quartile, with more than double the number of students answering at the 75th percentile. C&C questions showed stronger engagement than exam questions as well, with 75th percentile participation nearly 50% higher. These patterns suggest that students were more consistently willing to attempt FITB and C&C items when assigned. Faculty practices may contribute to this behavior; for example, many instructors assign participation credit for completing a portion (e.g., 80%) of the available questions, which could lead students to selectively skip certain question types.

Table 1
Number of students answering per question by type and assignment context. Assigned questions include mean and quartiles; unassigned questions include only the mean.

	Unassigned	Assigned			
	Mean	Mean	Q_1	Q_2	Q_3
Exam	2.4	51.7	4	21	62
C&C	2.8	55.5	3.8	28	90.5
FITB	3.9	84.5	24	60	143

3.1.2. Difficulty and Persistence

Difficulty is reflected by the percentage of correct first attempts (sometimes referred to as the difficulty index). While the open-ended exam questions are less amenable to an objective correctness classification, C&C responses can be more readily evaluated because they involve specific key distinctions. GPT-4o mini [32] was employed post hoc to analyze each student's submission together with its LLM-generated feedback, instructed to determine whether a typical college professor would regard it as "complete and correct." This offline classification does not affect the real-time feedback students receive, but rather serves as a means to compare overall difficulty of the C&C items to that of standard FITB questions. As shown in Table 2, the results for C&C and FITB confirm trends from prior research [9, 10] that the mean difficulties are higher when the practice is assigned, meaning students get the questions correct more frequently in a classroom context when they are assigned. Specifically, a chi-square test showed that the proportion of correct first attempts for C&C was significantly higher in the assigned context compared to unassigned ($\chi^2 = 207.87$, $p < .001$). The C&C questions had lower mean scores than the FITB questions, which is not unexpected given the higher level of cognitive effort and content comprehension required to answer the C&C compared to a single-term FITB. However, the difficulty index of 59.8 for the assigned C&C is within a reasonable range for such a complex question type.

Persistence occurs when a learner continues after an initial incorrect attempt until they eventually arrive at a correct response. As with difficulty, persistence applies only to question types where correctness is defined (C&C and FITB). Although the system's generative feedback focuses on iterative improvement rather than binary correctness, persistence nevertheless provides insight into how willing students are to revise more demanding items. The persistence data are a subset of the difficulty dataset, as it is only the students who were incorrect on their first attempt. Also consistent with prior research [9, 10], persistence increases when questions are assigned. For C&C questions, a chi-square test indicated that persistence was significantly higher in assigned contexts ($\chi^2 = 204.21$, $p < .001$). Persistence for C&C is much lower than for FITB. This could be related to two factors. First, the effort to answer C&C questions is much higher than for FITB, so it is not unexpected students would be less inclined to attempt them more than once. Second, the post-submission experience differs considerably: FITB initially provides correctness without revealing answers, prompting retries or answer reveals, whereas incorrect C&C responses immediately receive comprehensive LLM-generated corrective feedback, reducing incentives to retry. Given this, students who persist may show added effort to rephrase the correct response on their own, but students who don't persist have still received personalized corrective feedback—both beneficial learning experiences.

Table 2
Difficulty and persistence rates by question type and assignment context. Difficulty is defined as the percentage of first attempts marked correct; persistence is the percentage of initially incorrect attempts that were ultimately followed by a correct one.

	Difficulty		Persistence	
	Unassigned	Assigned	Unassigned	Assigned
C&C	50.6	59.8	8.2	16.9
FITB	65.9	79.9	61.2	94.7

3.1.3. Non-Genuine Responses

Non-genuine answers are those that do not constitute a legitimate effort. For FITB items, a rule-based filter detects obviously invalid submissions (e.g., single character, "idk"). As previously discussed, an LLM was used to evaluate whether a response substantively engaged with the C&C terms or proposed a meaningful exam question; if not, the answer is flagged as non-genuine. Non-genuine responses are lower for students in the assigned group for the open-ended questions: exam questions 11.7% assigned versus 16.8% unassigned and C&C questions 15.2% assigned compared to 19.1% unassigned. Chi-square tests confirm these differences are statistically significant for both exam questions ($\chi^2 = 200.01$, $p < .001$) and C&C questions ($\chi^2 = 63.86$, $p < .001$). The FITB questions have 6.6% non-genuine responses for assigned versus 3.9% unassigned. C&C questions had the highest non-genuine response rate for both groups. Given the cognitive demand combined with the need for understanding of two domain-specific terms, this is perhaps not surprising.

3.1.4. Student Ratings

Student "thumbs up/down" ratings (Figure 1) provide a mechanism for detecting problematic questions. Students could give a rating after submitting an answer, with one rating opportunity per question session. Table 3 shows higher overall rating frequency for unassigned questions. This initially seems counter-intuitive given the engagement is much lower for unassigned. We attribute this finding to rating fatigue [31]; students are more willing to rate early questions, but decline as they continue to answer. The students in the assigned group answer dramatically more questions, driving down their rating frequency. We also see an inverse relationship between the groups. The unassigned group has more thumbs up than thumbs down ratings while the assigned group has more thumbs down ratings. This could be attributed to students in the assigned group becoming more selective in motivation for rating, letting questions they like go by and negatively rating ones they liked less. These findings are consistent with prior research analyzing aggregate ratings [10, 34]. Exam questions have the highest thumbs up and down ratings for both groups. However, because the exam questions use only a templated prompt, they are not susceptible to some of the reasons FITB questions often get thumbs down, such as coming from an example or content students consider less helpful. Therefore, the thumbs down reasoning for exam questions is more likely related to not liking the question type itself.

Table 3
Thumbs up and thumbs down ratings per 1,000 student-question sessions, by question type and assignment context.

	Thumbs Up		Thumbs Down	
	Unassigned	Assigned	Unassigned	Assigned
Exam	3.67	0.57	2.08	0.64
C&C	2.22	0.21	0.93	0.42
FITB	2.30	0.06	1.45	0.09

3.2. Feedback Usage

To investigate how students engaged with the personalized feedback, we examined both how quickly they revised their answers and how extensively they incorporated the LLM's feedback text. Short intervals may indicate minimal attention to the feedback, whereas longer intervals could suggest more deliberate review. This also facilitates assessing whether rapid resubmissions align with potential "copy-paste" behavior.

The analysis focuses on cases in which the first attempt was incorrect (C&C 28.4%) or non-genuine (exam question 15.5%, C&C 17.7%). Although FITB items show high persistence (61.2% unassigned, 94.7% assigned), only 18.2% of exam-question sessions and 13.2% of C&C sessions with a non-correct first attempt proceeded to a second attempt. In Tables 4 and 5, which group data by first and second answer attempt categories, every cell comprises more than 100 sessions unless otherwise noted in the corresponding discussion. Tables 4 and 5 present the elapsed time between first and second attempts and the overlap score between each second attempt and the LLM feedback, disaggregated by question type, answer pattern (e.g., incorrect → correct), and assignment context. Second attempts on exam questions are classified only as genuine or non-genuine. All cells represent more than 100 sessions, except in the case of most incorrect or non-genuine second attempts (37–107 sessions), and two specific cells, incorrect → non-genuine for unassigned and assigned, contain just 9 and 11 sessions, respectively.

Table 4
Time interval (s) between first and second student attempts, by question type, answer pattern (e.g., incorrect → correct), and assignment context.

			Unassigned			Assigned		
	First Attempt	Second Attempt	Q_1	Q_2	Q_3	Q_1	Q_2	Q_3
Exam	Non-Genuine	Genuine	15	24	56	11	14	28
		Non-Genuine	16.25	40.5	114	9	21	52
C&C	Incorrect	Correct	39	73	125.5	18	29	59
		Incorrect	51	79	157	36.75	49	83.5
		Non-Genuine	15	58	162	24.5	55	115
	Non-Genuine	Correct	17	26.5	53	12	15	22
		Incorrect	50	77	112	28	60	91
		Non-Genuine	11	18.5	44.25	7	11	21.5

Table 5
Token-level textual overlap (percentage) between initial LLM-generated feedback and student second attempt, by question type, answer pattern (e.g., incorrect → correct), and assignment context.

			Unassigned			Assigned		
	First Attempt	Second Attempt	Q_1	Q_2	Q_3	Q_1	Q_2	Q_3
Exam	Non-Genuine	Genuine	28.0	79.2	84.5	59.6	77.6	83.1
		Non-Genuine	0.0	3.9	9.2	0.0	4.5	10.4
C&C	Incorrect	Correct	25.1	39.4	78.7	35.4	63.9	84.8
		Incorrect	16.0	23.3	29.2	15.7	19.3	22.5
		Non-Genuine	0.0	0.0	13.8	4.2	12.0	20.4
	Non-Genuine	Correct	69.9	85.2	95.7	71.2	81.8	95.1
		Incorrect	15.4	20.3	29.1	18.8	28.1	38.0
		Non-Genuine	0.0	0.0	8.0	0.0	0.0	5.1

There are several overall patterns noticeable from the time intervals. The first is that for every response pattern—across each quartile except one—the assigned group took less time to respond the second time. In many cases, they took roughly half the time, as seen in the C&C incorrect → correct response pattern. At first this seemed counterintuitive, as it could be assumed that students in the unassigned group would put less effort (i.e., time) than their assigned peers. However, when we consider the number of questions students in these groups answer, that changes the interpretation. Students in the unassigned group only answer a mean of 2.4 exam questions and 2.8 C&C, while the students in the assigned group answer a mean of 51.7 exam questions and 55.5 C&C. Students in the assigned group, familiar with expectations, respond faster, whereas the unassigned group likely requires additional time due to limited experience with the question type.

Another intriguing finding is how similar the elapsed times are each quartile for both assigned and unassigned for the exam question non-genuine → genuine and C&C non-genuine → correct response patterns. Prior research established that a percentage of students who input non-genuine responses for FITB follow it up with the correct response, indicating a strategy to reveal feedback as scaffolding [9, 10]. The similarity of elapsed times for the non-genuine to genuine/correct response patterns suggests a similar strategy is being employed here.

The overlap for non-genuine → non-genuine responses for both question types for both assigned and unassigned groups was 10.4% or less. Students who continued to enter non-genuine responses after receiving feedback did not appear to be considering the feedback or attempting to enter it back in. For C&C questions when students were incorrect on both attempts, they had among the highest time interval across all quartiles, yet low overlap (≤ 29.2%). This may reflect prolonged struggle or repeated guesswork.

However, overlap scores for exam questions (non-genuine → genuine) and C&C questions (incorrect → correct) reveal a wider range. Although the upper end of the overlap range still suggests significant reliance on the LLM's explanation, the lower overlap scores and longer time intervals may imply more genuine reflection and partial rewriting or paraphrasing rather than copying verbatim. The literal reuse of feedback does not necessarily impede learning—some learners may paraphrase or synthesize the feedback effectively—yet identifying instances of minimal revision can clarify the extent of students' engagement with the system's feedback.

4. Conclusion

Incorporating generative AI into educational technology should maintain focus on research-based methods that benefit student learning and adhere to responsible AI principles. The addition of open-ended questions that engage higher-order cognitive processes combined with personalized feedback to an existing AQG system provides students with a robust formative learning tool. This large-scale investigation of open-ended question types with LLM-enabled feedback provides a valuable comparison of performance metrics to established AG practice benchmarks. Addressing our second research question, we find that assigning questions has a profound effect on engagement, with clear impact on the rest of the performance metrics, indicating that structured classroom use encourages students to invest more effort in tackling these cognitively demanding tasks. Identifying effective strategies to encourage engagement in unassigned contexts remains an important direction for future research. Regarding research question one, assigned contexts showed higher difficulty and lower persistence for the new question types compared to FITB items, as expected given the greater effort required. The exam questions had notably less engagement, and combined with the thumbs down ratings, indicate a need for further consideration regarding the frequency of their appearance in the textbook.

Studying the use of feedback (research question three) by using both time intervals between first and second attempts as well as text overlap percentages between the feedback and student responses revealed several patterns in student behaviors. The time interval between responses was shorter in assigned contexts, suggesting experience answering more open-ended questions decreased the time it took students to craft a second attempt. The overlap analysis shows that many learners who had

incorrect or non-genuine first responses incorporate moderate to large portions of the LLM's feedback into their next correct submission. However, this approach of rephrasing or revising after a copy-paste does not necessarily preclude learning. A promising area for future research is further analyzing subgroups of student responses, including non-genuine responses, to reveal additional ways LLM-enabled feedback could scaffold learners.

As more domain-level, student-level, and other factors emerge from continued usage data, future work may employ more rigorous statistical modeling (e.g., mixed effects regression) to examine these factors in greater depth. In addition, because correctness and non-genuine responses were determined via an LLM-based evaluator, there is a possibility of classification errors or biases. Future analyses can consider sampling student responses for expert review and refining LLM prompts if necessary. Overall, these findings highlight both the promise and complexity of leveraging LLM technology to expand the cognitive range of automated practice. As generative AI continues to advance, maintaining rigorous analyses of usage patterns and performance metrics will remain crucial for ensuring that new capabilities genuinely advance student learning rather than merely accelerating the completion of tasks. In this case, we are satisfied that this first investigation shows a valid application of LLM abilities to provide the personalized feedback required by open-ended questions to support learning.

Declaration on Generative AI

During the preparation of this work, the authors used OpenAI o3 and GPT-4.5 for: refining draft content; paraphrasing and rewording; grammar and spelling checks. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. International Journal of Artificial Intelligence in Education, 30(1), 121–204. https://doi.org/10.1007/s40593-019-00186-y

[2] Black, P., & Wiliam, D. (2010). Inside the black box: raising standards through classroom assessment. Phi Delta Kappan, 92(1), 81–90. https://doi.org/10.1177/003172171009200119

[3] Theobald, E. J., Hill, M. J., Tran, E., Agrawal, S., Arroyo, E. N., Behling, S., Chambwe, N., Cintr, D. L., Cooper, J. D., Dunster, G., Grummer, J. A., Hennessey, K., Hsiao, J., Ira-non, N., Jones, L., Jordt, H., Keller, M., Lacey, M. E., Littlefield, C. E., … Freeman, S. (2020). Active learning narrows achievement gaps for underrepresented students in undergraduate science, technology, engineering, and math. Proceedings of the National Academy of Sciences, 117(12), 6476–6483. https://doi.org/10.1073/pnas.1916903117

[4] Koedinger, K. R., Kim, J., Jia, J., McLaughlin, E., & Bier, N. (2015). Learning is not a spectator sport: Doing is better than watching for learning from a MOOC. In Proceedings of the Second ACM Conference on Learning@Scale (pp. 111–120). https://doi.org/10.1145/2724660.2724681

[5] Koedinger, K. R., McLaughlin, E. A., Jia, J. Z., & Bier, N. L. (2016). Is the doer effect a causal relationship? How can we tell and why it's important. In Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (pp. 388–397). http://dx.doi.org/10.1145/2883851.2883957

[6] Van Campenhout, R., Jerome, B., Dittel, J. S., & Johnson, B. G. (2023). The doer effect at scale: Investigating correlation and causation across seven courses. In Proceedings of the 13th International Learning Analytics and Knowledge Conference (LAK 2023) (pp. 357–365). https://doi.org/10.1145/3576050.3576103

[7] Van Campenhout, R., Dittel, J. S., Jerome, B., & Johnson, B. G. (2021). Transforming textbooks into learning by doing environments: An evaluation of textbook-based automatic question generation. In Third Workshop on Intelligent Textbooks at the 22nd International Conference

on Artificial Intelligence in Education CEUR Workshop Proceedings (pp. 1–12). https://ceur-ws.org/Vol-2895/paper06.pdf

[8] Johnson, B. G., Dittel, J. S., Van Campenhout, R., & Jerome, B. (2022). Discrimination of automatically generated questions used as formative practice. In Proceedings of the Ninth ACM Conference on Learning@Scale (pp. 325–329). https://doi.org/10.1145/3491140.3528323

[9] Van Campenhout, R., Clark, M., Jerome, B., Dittel, J. S., & Johnson, B. G. (2023). Advancing intelligent textbooks with automatically generated practice: A large-scale analysis of student data. 5th Workshop on Intelligent Textbooks. The 24th International Conference on Artificial Intelligence in Education (pp. 15–28). https://intextbooks.science.uu.nl/workshop2023/files/itb23_s1p2.pdf

[10] Van Campenhout, R., Clark, M., Dittel, J. S., Brown, N., Benton, R., & Johnson, B. G. (2023). Exploring student persistence with automatically generated practice using interaction patterns. 2023 International Conference on Software, Telecommunications and Computer Networks (SoftCOM) (pp. 1–6). https://doi.org/10.23919/SoftCOM58365.2023.10271578

[11] Van Campenhout, R., Kimball, M., Clark, M., Dittel, J. S., Jerome, B., & Johnson, B. G. (2024). An investigation of automatically generated feedback on student behavior and learning. In Proceedings of LAK24: 14th International Learning Analytics and Knowledge Conference (pp. 850–856). https://doi.org/10.1145/3636555.3636901

[12] Van Campenhout, R., Clark, M., Johnson, B. G., Deininger, M., Harper, S., Odenweller, K., & Wilgenbusch, E. (2024). Automatically generated practice in the classroom: Exploring performance and impact across courses. In Proceedings of the 32nd International Conference on Software, Telecommunications and Computer Networks (SoftCOM 2024) (pp. 1–6). https://doi.org/10.23919/SoftCOM62040.2024.10721828

[13] VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46(4), 197–221. https://doi.org/10.1080/00461520.2011.611369

[14] Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. Review of Educational Research, 86(1), 42–78. https://doi.org/10.3102/0034654315581420

[15] Anderson, L. W. (Ed.), Krathwohl, D. R. (Ed.), Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C. (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives (Complete edition). Longman.

[16] Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., Slater, S., Baker, R., & Warschauer, M. (2020). Mining big data in education: Affordances and challenges. Review of Research in Education, 44(1), 130–160. https://doi.org/10.3102/0091732X20903304

[17] McFarland, D. A., Khanna, S., Domingue, B. W., & Pardos, Z. A. (2021). Education data science: Past, present, future. AERA Open, 7(1), 1–12. https://doi.org/10.1177/23328584211052055

[18] Van Campenhout, R., Soto-Karlin, D., Selinger, M., & Jerome, B. (2025). Learning engineering in practice: A case study on developing LLM-based educational tools. In R. A. Sottilare & J. Schwarz (Eds.), Adaptive instructional systems. HCII 2025. Lecture Notes in Computer Science (Vol. 15813, pp. 132–150). Springer. https://doi.org/10.1007/978-3-031-92970-0_10

[19] Van Campenhout, R., Brown, N., & Johnson, B. G. (2025). AI principles in practice with a learning engineering framework. In Proceedings of the 17th International Conference on Computer Supported Education (pp. 312–318). https://doi.org/10.5220/0013358600003932

[20] Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. Educational Psychologist, 49(4), 219–243.

[21] Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength natural language processing in Python. https://doi.org/10.5281/zenodo.1212303

[22] Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into text. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (pp. 404–411). https://aclanthology.org/W04-3252

[23] Evert, S. (2009). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), Corpus linguistics: An international handbook (Vol. 2, pp. 1212–1248). Mouton de Gruyter. https://doi.org/10.1515/9783110213881.2.1212

[24] Yu, F.-Y., & Pan, C.-C. (2014). The effects of student question-generation with online prompts on learning. Educational Technology & Society, 17(3), 267–279. https://www.jstor.org/stable/jeductechsoci.17.3.267

[25] Rosenshine, B., Meister, C., & Chapman, S. (1996). Teaching students to generate questions: A review of the intervention studies. Review of Educational Research, 66(2), 181–221. https://doi.org/10.3102/00346543066002181

[26] Alfieri, L., Nokes-Malach, T. J., & Schunn, C. D. (2013). Learning through case comparisons: A meta-analytic review. Educational Psychologist, 48(2), 87–113. https://doi.org/10.1080/00461520.2013.775712

[27] Gentner, D., & Namy, L. L. (1999). Comparison in the development of categories. Cognitive Development, 14(4), 487–513. https://doi.org/10.1016/S0885-2014(99)00016-7

[28] Aleven, V., & Koedinger, K. (2000). Limitations of student control: Do students know when they need help? In Proceedings of the 5th International Conference on Intelligent Tutoring Systems (pp. 292–303).

[29] OpenAI. (2024, August 8). *GPT-4o system card*. https://openai.com/index/gpt-4o-system-card/

[30] VitalSource Supplemental Data Repository. (2025). https://github.com/vitalsource/data

[31] Johnson, B. G., Dittel, J., & Van Campenhout, R. (2024). Investigating student ratings with features of automatically generated questions: A large-scale analysis using data from natural learning contexts. In Proceedings of the 17th International Conference on Educational Data Mining (pp. 194–202). https://doi.org/10.5281/zenodo.12729796

[32] OpenAI. (2024, July 18). *GPT-4o mini: advancing cost-efficient intelligence*. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[33] Ratcliff, J. W., & Metzener, D. E. (1988). Pattern matching: The gestalt approach. Dr. Dobb's Journal, 13(7), 46–51.

[34] Jerome, B., Van Campenhout, R., Dittel, J. S., Benton, R., & Johnson, B. G. (2023). Iterative improvement of automatically generated practice with the Content Improvement Service. In R. Sottilare & J. Schwarz (Eds.), Adaptive Instructional Systems. HCII 2023. Lecture Notes in Computer Science (pp. 312–324). Springer, Cham. https://doi.org/10.1007/978-3-031-34735-1_22