LLM Feedback Isn't Automatically Better: Static Scaffolds Outperform Dynamic Feedback in Textbook-Embedded Practice

Benny G. Johnson*, Jeffrey S. Dittel, Oscar J. Ortiz, Rodrigo Bistolfi, Michelle W. Clark, Bill Jerome, Richard Benton, and Rachel Van Campenhout

VitalSource Technologies, 227 Fayetteville Street, Suite 400, Raleigh, North Carolina, 27601 United States

Abstract

Large language models (LLMs) lower the practical barrier to generating error-sensitive feedback conditioned on a student's actual incorrect answer, suggesting a plausible next step beyond question-level static feedback in textbook-embedded formative practice. In this study, we compare dynamic LLM-generated feedback to an existing static feedback approach for automatically generated fill-in-the-blank (FITB) cloze questions delivered alongside textbook content in an ereader platform to students learning in higher education contexts. During a randomized deployment, incorrect first attempts were assigned either to dynamic or static feedback, enabling intention-to-treat analysis. Secondary analyses examined complier average causal effects and realized feedback conditions. Results show that dynamic feedback did not outperform the static approach overall. Although dynamic feedback was associated with slightly lower answer reveal rates (a small persistence effect), it produced no detectable change in net target term recovery on the next action; the displaced sessions were absorbed primarily by incorrect retries rather than correct ones. Secondary decomposition analyses indicate that the strongest advantage over dynamic was observed for common answer feedback, which is associated with higher correct retry than dynamic at the feedback type level. We argue that this pattern is learning-science coherent: in textbook-embedded FITB practice designed to support reading comprehension and the doer effect, feedback that more directly constrains the answer space may be better aligned to the task than error-sensitive explanatory feedback.

Keywords

large language models, AI-generated feedback, formative practice, fill-in-the-blank questions, retrieval practice

1. Introduction

Feedback for formative practice and practice testing is a well-researched component of learning and is generally known to produce positive learning outcomes across varying educational contexts [1]. However, the effectiveness of feedback is dependent on learning context, learner characteristics, and feedback characteristics themselves [2, 3]. The immediacy of feedback matters, with immediate feedback found to reduce the amount of time students spent correcting mistakes, increase comprehension of the correct answer [4], and increase student satisfaction [5]. Feedback is commonly categorized into distinct types: knowledge of results (KR), knowledge of correct response (KCR), and elaborative feedback (EF). KR feedback is the least effective, with KCR only slightly more effective, while EF (which typically includes KR or KCR within the response) is the most effective strategy [5-8].

While there is conclusive evidence regarding types of feedback, theoretical and empirical research continues to show how the nuance of feedback content and context requires continued evaluation. For example, there is tension between the layered/sequenced hints with bottom-out

*Corresponding author.

iTextbooks'26: Seventh Workshop on Intelligent Textbooks, June 28, 2026, Seoul, Republic of Korea

Benny.Johnson@vitalsource.com; Jeff.Dittel@vitalsource.com; Oscar.Rey@vitalsource.com; Rodrigo.Bistolfi@vitalsource.com; Michelle.Clark@vitalsource.com; Bill.Jerome@vitalsource.com; Richard.Benton@vitalsource.com; Rachel.VanCampenhout@vitalsource.com

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

strategy common in intelligent tutoring systems [9] and benefits of singular elaborative feedback, with findings suggesting sequenced feedback boosted learner engagement and satisfaction but decreased learning outcomes [10]. Another example is the inclusion of KCR within EF; providing correct responses with longer explanations has been found to be effective [1], but also provides a bottom-out state that can lead to learners copying the answer with minimal processing [11, 12].

Despite the wealth of research on feedback theory, in their systematic review, Morris et al. [13] call for more large-scale empirical evaluations of formative practice and feedback in higher education. Quantitative research exploring the many nuances of effective feedback is especially necessary as feedback generated with artificial intelligence (AI) becomes more prominent in digital learning environments. In their 2020 systematic review of automatic question generation systems, Kurdi et al. [14] noted that only one of 92 studies included feedback. Since that time, many studies on automatically generated feedback have been published, in part due to the rise of large language models (LLMs). Research has included feedback for both cloze and open-ended question types [15-18]. Studies on LLM-generated personalized feedback evaluate the impact of erroneous feedback due to LLM hallucinations [19]. Research on AI-generated multimodal feedback found comparable learning gains and correctness compared to human-authored feedback, yet higher perceived clarity and satisfaction [17].

This study aims to extend current research on AI-generated feedback within the context of AI-generated formative practice in etextbooks. The formative practice, which includes matching, multiple choice, fill-in-the-blank (FITB), and open-ended question types, is generated through a rule-based automatic question generation (AQG) system [20-22]. These questions have been shown to be comparable to human-authored questions on engagement, difficulty, persistence, and discrimination [23, 24]. Classroom research has found that doing these questions improved exam scores and generated the doer effect [25]—the learning science principle that doing practice while learning is six times more effective than reading alone [26-28]. Supporting students in persisting after they answer a question incorrectly is an important part of the learning process, and feedback is a critical component of this [29]. A key reason this matters is that scaffolding is especially valuable when it helps students produce the correct answer themselves rather than simply being shown it [29]. After students answer a question incorrectly, as shown in Figure 1, students receive feedback as well as the option to retry, reveal the answer, and rate the question. In the present study, we focus on what happens next after an incorrect first attempt as a function of feedback type: whether students reveal the correct answer and whether their next action produces a correct response.

Prior research on generated feedback types for the FITB revealed that the feedback type influences behavior and learning: how often they retry, reveal, and reach the correct answer on the

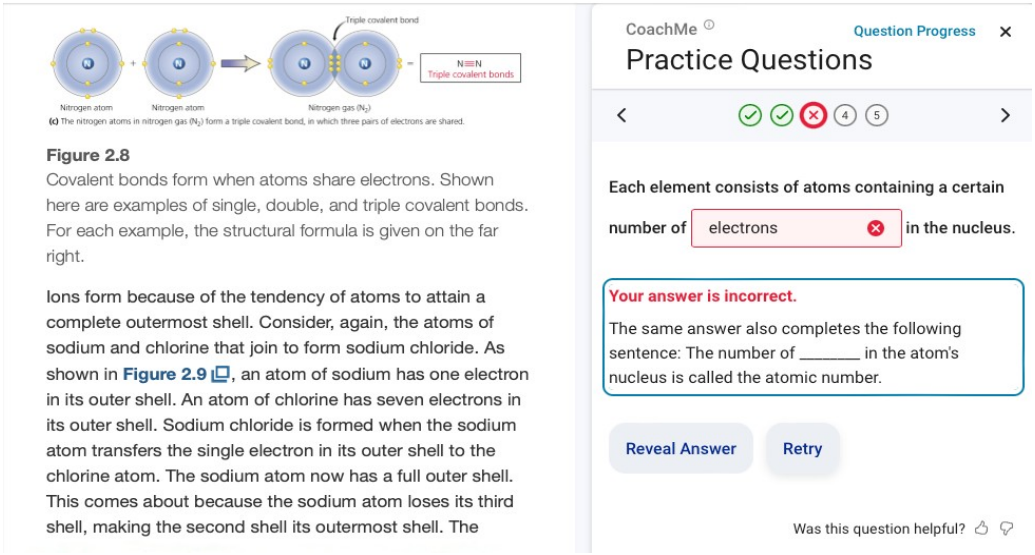

Figure 1: An example FITB formative practice question in a chemistry textbook.

next attempt [15]. Three types of generated feedback were randomly deployed for the FITB cloze questions: outcome feedback, context feedback, and common answer feedback. Outcome feedback merely informs the student that the response is incorrect. Context feedback provides an extended selection of the textbook passage surrounding the question sentence in order to give the student more contextual information for the next attempt. Common answer feedback presents a second sentence selected from nearby in the textbook content with the same target term removed, thereby giving the student another example, scaffolding the next attempt. These feedback types are static with respect to the student's answer, in the sense that they are generated at the question level rather than tailored to a student's specific incorrect answer. That study found that common answer feedback performed best on key behavioral outcomes, including lower answer reveal rates and higher success on retry attempts. As a result, common answer is the strongest existing static scaffold and the benchmark for any new feedback approach. In the current system, it is also the preferred static feedback type when it can be generated.

The advancement of LLMs has made it possible to scale dynamic, error-sensitive feedback conditioned on the student's incorrect answer, as shown in the research discussed. LLM-generated feedback was employed in this AQG system for certain open-ended questions with success [16, 30]. It seemed like a logical next step to extend this feedback approach to the FITB cloze questions. Unlike static feedback generated at the question level, an LLM can in principle respond differently to near-synonyms, valid-but-different answers, nonsense input, or other answer-specific situations. Given the longstanding effectiveness of error-sensitive feedback in tutoring and educational software [29], an answer-specific explanation seems like a more adaptive and effective form of feedback than a fixed response at the question level.

The present study compares dynamic (LLM-generated) feedback to the existing static feedback policy for FITB questions using a randomized deployment in the textbook ereader platform. The static policy selects from common answer, context, or outcome feedback at the question level rather than tailoring to a student's specific incorrect answer. The primary analysis asks whether dynamic feedback improves student behavior and target term recovery relative to the static approach, while a secondary analysis examines the feedback conditions actually shown to students in order to clarify which static feedback types appear most relevant to the observed differences.

This research further contributes to the nuances and tension in feedback literature by comparing two different AI-generated feedback types which both have valid theoretical underpinnings. In the context of textbook-based formative practice, and more specifically, FITB cloze questions that engage a recall cognitive process dimension [31], feedback may serve a task-specific role. The common answer static feedback provides a discrete scaffold in a consistent format, while dynamic feedback is personalized and potentially error-sensitive. The main goal of this study is to compare these approaches and determine which better serves students in this context. The study is organized around three research questions:

- **RQ1.** Under random assignment, does dynamic feedback improve outcomes relative to the existing static feedback approach?
- **RQ2.** Which feedback types actually shown to students appear most associated with the observed differences in outcomes?
- **RQ3.** What do the results imply about the relationship between feedback design and target term recovery in textbook-embedded formative practice?

The contribution of this investigation is threefold. Empirically, it provides a randomized evaluation of dynamic LLM feedback in a large-scale textbook setting drawn from natural learning contexts. Analytically, it combines the primary ITT comparison with secondary decomposition analyses. Theoretically, we examine how feedback type may differ in task alignment for textbook-embedded FITB practice questions.

2. Method

2.1. Static and dynamic feedback generation

In the current AQG system, the static feedback types are used in a preference hierarchy based on findings from prior research [15]. Common answer feedback is used whenever it can be generated. If it cannot, context feedback is used when available. Outcome feedback serves as the final static option because it can always be generated. Thus, although the static policy includes multiple feedback types, common answer is the most frequent static type in practice. Because common answer feedback serves as the strongest static benchmark, Figure 2 illustrates common answer feedback alongside the dynamic feedback studied here.

The dynamic condition uses an LLM to generate feedback based on the student's actual incorrect answer. The model used in this deployment was GPT-4.1 nano [32]. For each incorrect attempt, the model is given three inputs: the question stem, the correct answer, and the student's incorrect answer. The feedback generation prompt instructs the model to produce brief supportive feedback that acknowledges the student's answer, explains why it does not fit, redirects the student without revealing the correct answer, and encourages a retry. The design was intended to approximate a brief elaborative, answer-sensitive feedback style rather than a minimal hint, because elaborative feedback is often treated as a strong general-purpose feedback strategy [7, 8]. The generated response is constrained to two or three sentences. The full prompt, along with the dataset used in this study, will be made available in our open data repository [33].

A key instructional decision is that the feedback should not directly reveal the correct answer, given that immediately providing the correct response could allow learners to disengage from the retrieval process [11] (though a bottom-out option is provided in the reveal button shown in Figure 1). To enforce this, generated dynamic feedback is checked for the presence of the correct answer word. If the generated feedback contains the answer, it is rejected and the system falls back to the static feedback approach. This guardrail should be understood as a pedagogical design choice, not as evidence that the generated output was otherwise unusable.

2.2. Randomized assignment design

In order to evaluate the feedback types fairly, randomized feedback assignment for the FITB questions was used. During the randomized phase, each incorrect attempt was assigned, when the incorrect response was submitted, to one of two study conditions: dynamic feedback or the existing static feedback approach. Assignment probability was equal for the two conditions.

As discussed previously, the static condition was not a single feedback type: depending on the question, it could yield common answer, context, or outcome feedback. For that reason, the

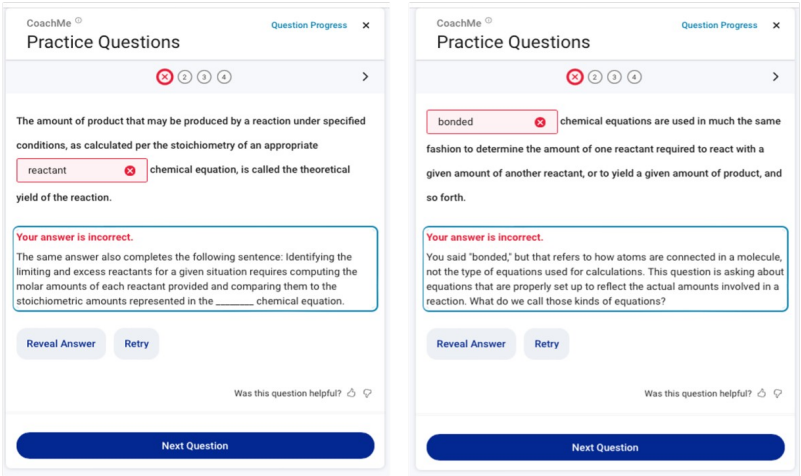

Figure 2: Examples of common answer and dynamic feedback for FITB questions.

primary randomized comparison in this study is assigned dynamic versus assigned static, not dynamic versus common answer in isolation.

A separate issue is that assignment and delivered feedback do not always coincide. Because of the no-leak pedagogical guardrail, some incorrect attempts assigned to dynamic fall back to static feedback. The primary estimand is therefore an intention-to-treat (ITT) [34] effect: the effect of assignment to the dynamic rather than the static condition. This creates a one-sided noncompliance structure in the randomized design: incorrect attempts assigned to dynamic may receive dynamic feedback or may fall back to static feedback, whereas incorrect attempts assigned to static never receive dynamic feedback.

We also report complier average causal effect (CACE) estimates as a secondary estimate alongside the ITT analysis [35]. Whereas ITT estimates the effect of assignment to the dynamic condition, CACE estimates the effect among cases in which assignment to dynamic results in dynamic delivery. Under this one-sided noncompliance structure, CACE is estimated on the risk difference scale by dividing the ITT difference by the compliance rate in the dynamic condition. Intuitively, if only part of the dynamic-assigned group actually receives dynamic feedback, CACE rescales the ITT effect to estimate what the effect would be among the subset for whom dynamic delivery is feasible. Accordingly, CACE should be interpreted as the effect for that subgroup rather than as the effect for all dynamic-assigned cases.

2.3. Observation window, outcomes, and analysis plan

The dataset consists of interactions recorded during the randomized deployment window from 9 April 2026 through 8 May 2026, using textbooks from eight publishers who granted permission for generative AI research. The unit of analysis is the student-question session. A session is defined as all interactions by a given student on a given question, ordered chronologically. Each row in the dataset corresponds to one student-question session anchored at an incorrect first attempt and followed through the subsequent recorded actions on that same question. Questions with no incorrect attempts during the observation window were excluded, as they do not contribute to the analysis. Sessions in which more than ten minutes elapsed between the student's initial incorrect answer and next action (if any) were removed to reduce the chance that the student had left the textbook between actions, potentially weakening the effect of feedback; this accounted for less than 0.5% of sessions. The final dataset contains 33,834 sessions across 23,148 questions, 5,596 students, and 1,363 textbooks. Using the Book Industry Standards and Communications major subject heading classification [36] available for most of the textbooks, the top subject domains as a percentage of session data were Psychology (22.7%), Social Science (19.1%), Political Science (15.7%), Business & Economics (12.2%), and Law (7.5%).

The outcomes examined are answer reveal and correct retry.

- Answer reveal was evaluated on all sessions anchored at an incorrect first attempt, coded as 1 if the student requested the correct answer immediately after that attempt and 0 otherwise. It is useful because it reflects whether the feedback helped keep the student working to produce the answer rather than leading directly to an answer request.
- Correct retry was also evaluated on all sessions anchored at an incorrect first attempt, coded as 1 if the student's next action produced a correct response and 0 otherwise. This captures whether the student was able to recover the target term after receiving feedback.

The primary analysis is an ITT comparison between assigned dynamic and assigned static. For interpretability, we first report mean-based ITT and CACE estimates for answer reveal and correct retry. Confidence intervals for these estimates were obtained using a student-level cluster bootstrap, which accounts for the non-independence of multiple sessions contributed by the same student. The reported intervals are based on 100,000 bootstrap samples, which produced stable estimates at the reported precision.

Mixed effects logistic regression models were then fit to account more fully for the clustered structure of the data. Each student typically engaged with multiple questions, and each question was also often answered by multiple students, so observations are non-independent within both student and question clusters. As in prior research [20], mixed effects logistic regression models were fit with random intercepts for both student and question. All regressions were performed using R version 4.4.1 [37], with package `glmmTMB` version 1.1.9 [38] for mixed effects models. The formula for the answer reveal model in R notation is:

```
glmmTMB(answer_reveal ~ assigned_condition
   + (1|student_id) + (1|question_id),
   family=binomial(link=logit), data=sessions)
```

An analogous model was fit for correct retry. These models estimate the primary ITT effect while accounting for repeated observations within students and clustering by question.

To clarify which delivered feedback types appear most relevant to the observed patterns, we also conduct secondary analyses using a realized condition factor, i.e., a factor indicating the feedback type actually shown to the student. This factor has seven levels: dynamic, common answer assigned or fallback, context assigned or fallback, and outcome assigned or fallback. Descriptive outcome rates and mixed effects logistic regression results are reported with dynamic feedback as the reference level. Because these realized condition analyses are shaped partly by fallback and are therefore post-random assignment, they are interpreted as descriptive and mechanism-oriented rather than as the primary causal result.

3. Results and Discussion

3.1. Randomized comparison

We begin with the primary randomized comparison, which addresses RQ1 by testing whether dynamic feedback improves outcomes relative to the existing static approach. Table 1 presents the primary randomized comparison between the dynamic and static conditions for the answer reveal and correct retry outcomes across the full sample of 33,834 incorrect-first-attempt sessions. We first report outcome rates by assigned condition and then summarize both the ITT difference and the corresponding CACE estimate, with student-level cluster bootstrap confidence intervals reported for both.

Table 1
Randomized ITT and CACE estimates for answer reveal and correct retry.

Outcome	Dynamic Condition (%)	Static Condition (%)	ITT, 95% CI	CACE, 95% CI
Answer reveal	61.0	64.3	-3.3 [-4.4, -2.2]	-5.3 [-7.2, -3.6]
Correct retry	16.5	16.9	-0.4 [-1.3, 0.4]	-0.7 [-2.1, 0.7]

ITT and CACE are reported on the risk difference scale as percentage point differences (dynamic minus static).

For answer reveal, the dynamic condition had a slightly lower reveal rate than the static condition (61.0% vs. 64.3%), corresponding to an ITT difference of -3.3 percentage points. Thus, students assigned to dynamic feedback were modestly less likely to request the correct answer immediately after an incorrect first attempt. The corresponding CACE estimate was larger in magnitude (-5.3 percentage points), indicating that among cases where assignment to dynamic actually resulted in dynamic delivery, the reduction in answer reveal was somewhat stronger.

For correct retry, the difference between conditions was small. Students assigned to dynamic feedback recovered the correct answer at 16.5%, compared with 16.9% for those assigned to static feedback, corresponding to an ITT difference of -0.4 percentage points with a 95% confidence

interval that crosses zero. The corresponding CACE estimate was likewise small (-0.7 percentage points) and not distinguishable from zero.

Taken together, these randomized descriptive results suggest a nuanced pattern. Assignment to dynamic feedback modestly reduced immediate answer reveal, but produced no detectable change in correct retry relative to assignment to the existing static approach. The reduction in answer reveal did not translate into more students recovering the correct answer on the next action. Incorrect retry rates were correspondingly higher under dynamic (18.5%) than static (14.6%), consistent with the persistence shift being absorbed primarily by failed rather than successful retries. This pattern provides the first indication that the dynamic condition does not outperform the current static policy overall.

The CACE estimates point in the same direction as the ITT results and are larger in magnitude, as expected under one-sided noncompliance. Compliance with dynamic delivery was moderate rather than high (61.6%), reflecting fallback under the no-leak guardrail in a substantial minority of dynamic-assigned cases. The larger CACE estimates simply rescale the ITT effects to the dynamic-delivered subgroup; they do not change the qualitative conclusion that dynamic feedback reduced answer reveal but produced no detectable change in correct retry.

Table 2 reports the model-based ITT results from logistic mixed effects models with random intercepts for question and student. The answer reveal result was statistically significant; the correct retry result was not, consistent with the descriptive picture in Table 1. In other words, the model-based results sharpen the descriptive picture by showing that the same basic pattern persists after accounting for question- and student-level heterogeneity.

Table 2
Mixed effects logistic regression results for the randomized ITT comparison. The assigned dynamic condition is the reference level.

Outcome	Estimate (SE)	*p*	Odds Ratio
Answer reveal	0.305 (0.035)	< 2e-16 ***	1.36
Correct retry	0.073 (0.041)	.079	1.08

For answer reveal, assignment to the static condition yielded higher odds of requesting the correct answer than assignment to the dynamic condition. Thus, the model-based result supports the descriptive finding that dynamic feedback slightly reduced answer reveal relative to the static approach. The corresponding odds ratio was modest in magnitude (1.36). For correct retry, the model-based estimate did not reach statistical significance, consistent with the descriptive ITT and CACE estimates whose confidence intervals included zero. The model-based result thus confirms that dynamic and static feedback do not produce a detectable difference in net target term recovery.

Taken together, the ITT models reinforce the main randomized conclusion. Dynamic feedback does not outperform the existing static approach overall. Instead, the current evidence suggests a more nuanced pattern: dynamic feedback was associated with slightly lower immediate answer reveal, but with no detectable change in correct retry. Lower answer reveal is itself desirable because it reflects greater persistence in trying to produce the answer, but that persistence shift did not translate into more correct answers on the next action. These results strengthen the conclusion that dynamic feedback is not the better overall approach in this task.

3.2. Realized condition decomposition

We next examine the realized feedback conditions actually shown to students in order to address RQ2 and clarify which feedback types appear most responsible for the observed differences. Because the assigned dynamic condition combines dynamic-delivered and fallback-delivered cases, the realized condition rates reported below need not match the assigned condition rates reported

above. Table 3 presents descriptive outcome rates by realized feedback condition. These descriptives are not the primary causal result, but they are useful for understanding which delivered feedback states appear most relevant to the pattern observed in the randomized ITT analyses.

Table 3
Descriptive outcome rates by realized feedback condition.

Realized Condition	Answer Reveal (%)	Correct Retry (%)
Dynamic	63.6	14.5
Common assigned	63.1	18.4
Common fallback	55.5	21.3
Context assigned	63.2	16.2
Context fallback	57.7	17.8
Outcome assigned	70.4	12.9
Outcome fallback	60.9	16.9

Cell sizes differed across realized feedback conditions, ranging from 1,098 to 10,410. Common answer feedback accounted for 56.3% of static-delivered cases. Thus, although the static condition includes multiple feedback types, common answer constitutes the largest share of delivered static feedback in practice.

For answer reveal, the realized condition descriptives do not point to a simple common answer advantage over dynamic. The answer reveal rate for dynamic was 63.6%, compared with 63.1% for common assigned and 63.2% for context assigned, while outcome assigned was substantially higher at 70.4%. The fallback static conditions were lower still. Descriptively, then, answer reveal does not isolate a single static feedback type that clearly explains the broader pattern. Instead, the most distinctive feature of this outcome is the relatively unfavorable reveal rate for outcome-assigned feedback, whereas common answer and dynamic feedback appear broadly similar.

For correct retry, the pattern is more informative. The correct retry rate for dynamic was 14.5%, compared with 18.4% for common assigned and 21.3% for common fallback. Context assigned was slightly above dynamic at 16.2%, and context fallback at 17.8%. By contrast, outcome assigned was actually below dynamic at 12.9%, with outcome fallback slightly above at 16.9%. Descriptively, then, the strongest advantage over dynamic is observed for common answer feedback, while context shows a smaller advantage and outcome shows no consistent advantage overall.

This descriptive decomposition already suggests a more specific interpretation than "static feedback beats dynamic" at the policy level: the within-static comparison to dynamic is non-uniform, with common answer showing the strongest advantage. The next analysis tests whether this pattern remains after accounting for student- and question-level variation in the realized condition models.

Table 4 reports the mixed effects logistic regression results for the realized condition decomposition, using dynamic feedback as the reference level. These models are intended as descriptive analyses rather than primary causal tests, but they are useful for clarifying which delivered feedback conditions appear most relevant to the overall pattern. Because delivered feedback states are not themselves randomized, the contrasts are best interpreted as suggestive rather than as identified causal effects of particular feedback types. Note that the model-based contrasts can shift relative to the descriptive rates in Table 3 because they adjust for student- and question-level variation.

Table 4
Mixed effects logistic regression results for the realized feedback condition decomposition. Dynamic feedback is the reference level.

Panel A. Answer reveal			
Condition	**Estimate (SE)**	***p***	**Odds Ratio**
Common assigned	0.165 (0.045)	2.79e-04 ***	1.18
Common fallback	-0.114 (0.060)	.058	0.89
Context assigned	0.308 (0.057)	8.05e-08 ***	1.36
Context fallback	0.092 (0.081)	.256	1.10
Outcome assigned	0.884 (0.073)	< 2e-16 ***	2.42
Outcome fallback	0.404 (0.101)	6.81e-05 ***	1.50

Panel B. Correct retry			
Condition	**Estimate (SE)**	***p***	**Odds Ratio**
Common assigned	0.338 (0.055)	7.27e-10 ***	1.40
Common fallback	0.311 (0.071)	1.14e-05 ***	1.36
Context assigned	-0.046 (0.070)	.508	0.95
Context fallback	-0.112 (0.098)	.249	0.89
Outcome assigned	-0.385 (0.088)	1.14e-05 ***	0.68
Outcome fallback	-0.177 (0.122)	.146	0.84

For answer reveal, the realized condition models do not suggest a simple common answer advantage over dynamic. Relative to dynamic, common assigned was associated with higher answer reveal odds, whereas common fallback trended lower but was not statistically reliable. Context assigned and both outcome conditions were also associated with higher answer reveal odds than dynamic, with the largest effect observed for outcome assigned. Thus, although the answer reveal descriptives are somewhat mixed, the model-based results do not isolate a clean common answer advantage on this outcome.

For correct retry, the decomposition is clearer. Relative to dynamic, both common assigned and common fallback were associated with higher odds of correct retry. Context feedback conditions did not differ statistically from dynamic. Outcome assigned was associated with lower odds of correct retry than dynamic, while outcome fallback did not differ statistically. The within-static comparison to dynamic is not uniform: common answer outperforms dynamic, context is not detectably different, and outcome when assigned directly is worse. This is where the learning science interpretation begins to come into view: the strongest comparator to dynamic is the static scaffold most directly associated with recovery of the target term.

On answer reveal, the descriptive pattern is mixed and the model-based contrasts do not identify a clean common answer advantage over dynamic. On correct retry, common answer clearly exceeds dynamic while outcome assigned underperforms. Together, these findings produce the policy level near-tie observed in the randomized analyses. This reinforces the importance of looking at specific feedback types rather than treating static feedback as a single category, and of separating persistence-related behavior from actual recovery of the correct answer.

Delivered feedback conditions were shaped by the no-leak guardrail. When the generated dynamic feedback contained the correct answer, it was replaced with static feedback. Fallback occurred in a substantial minority of dynamic-assigned cases and was non-random. Using edit distance between the student's answer and the correct answer as a proxy for response proximity, we found that fallback was more likely for responses closer to the correct answer, suggesting that leakage risk is greater when the student answer is already close to the target term. However, descriptive differences between conditions remained similar across edit distance strata, and adding

edit distance quartile as a covariate did not materially change the assigned condition effect in the mixed effects models. These supplementary checks did not materially alter the main randomized contrast or qualitative conclusions and are reported in the accompanying analysis notebook [33].

A concrete example helps clarify the role of fallback. Consider the following item from the textbook *Biological Psychology* [39]: "The ANS is essentially the collection of _____ that act as the manager of your internal organs." The student answered "neurons." The generated dynamic feedback was: "You answered 'neurons,' which are the cells that make up nerves, but the question is asking about the overall collection that acts as the manager of your internal organs. What do we call that collection of nerve fibers? Would you like to try again?" This is a coherent and plausible response that engages the student's answer and identifies the conceptual mismatch. However, it also comes close to defining the target term directly and therefore crosses the no-leak boundary. As a result, it is not shown. Instead, the system falls back to common answer feedback: "Some texts refer to a sub-set of the ANS known as the enteric nervous system, which refers to a fine network of _____ that are found only in the walls of the digestive tract and control the digestive process." The correct answer is "nerves." This example illustrates why fallback is best understood as a pedagogical boundary rather than as a response to low-quality dynamic output.

3.3. Interpreting the pattern

Finally, we turn to the broader interpretation of the results in order to address RQ3, namely what these findings imply about the relationship between feedback design and target term recovery in textbook-embedded formative practice. Viewed in hindsight, the overall pattern is learning-science coherent. Common answer feedback provides a second cloze sentence using the same target term, giving students another opportunity to retrieve that term while still requiring them to produce it themselves. Even when students do not explicitly compare the two cloze contexts as intended, the additional sentence still functions as a second retrieval cue.

This matters because the present task is not a general tutoring exchange. It is textbook-embedded formative practice intended to support reading comprehension and the doer effect. For this task, the most useful feedback may not be the feedback that most directly engages the student's specific wrong answer, but the feedback that most effectively supports recovery of the correct target term. Put differently, error-sensitive feedback is not always the best instructional fit.

That perspective helps explain why dynamic feedback, despite being more adaptive in principle, did not outperform the static approach overall. The dynamic feedback often appears reasonable, supportive, and on topic, but it may still provide semantic activation without enough diagnostic constraint. It may help students think about the right region of meaning without narrowing the answer space as efficiently as common answer feedback. That kind of guidance may be sufficient to keep students engaged with the question, consistent with the lower answer reveal rates observed under dynamic assignment, while still being less effective than common answer feedback at helping them recover the correct answer themselves on the next action [11, 29].

A concrete example illustrates the distinction. Consider the item from the textbook *Employment Law for Paralegals* [40]: "The weight of the evidence indicated that a personal injury by _____ occurred while the worker was engaged in the performance of a work-related duty or engaged in an activity reasonably incidental to the employment." The student answered "negligence." The delivered dynamic feedback responded: "You said 'negligence,' but this term refers to a failure to take proper care, not an event like an injury. The question is asking about what actually caused the injury during work. What do we call an unexpected event that results in harm?" This feedback is reasonable and on topic, but the common answer feedback gives a tighter retrieval cue: "The worker was treated by a psychiatrist for approximately one year following the _____ and complained of dysphoria (general unhappiness), tearfulness, insomnia, poor concentration, and forgetfulness." The correct answer is "accident." Here, the dynamic response offers sensible guidance, but the common answer cue is the stronger scaffold for recovering the exact term.

This example makes the task alignment interpretation concrete, but the broader pattern could still admit alternative explanations. Several merit consideration, but none overturns this interpretation. One possibility is that common answer feedback improves retry success only by making the task too easy. Yet this feedback is not merely a shallow cue that inflates retry success while bypassing meaningful retrieval. It does not reveal the answer, but provides another cloze prompt that still requires the student to produce the target term. If common answer were simply functioning as a giveaway, correct retry rates would approach ceiling; instead, they remain far lower. The results are also difficult to reconcile with a simple verbosity explanation. Dynamic feedback was associated with lower, not higher, answer reveal, suggesting that students were not merely rejecting it as too cumbersome to use.

Fallback patterns were discussed in the preceding section: the edit-distance robustness check showed that adjusting for response proximity does not materially alter the assigned-condition effect, so fallback does not provide an alternative explanation for the within-static common answer advantage. Another possibility is that a larger model or better prompting strategy might have produced stronger dynamic feedback than the current implementation. However, the study still provides a meaningful result for a realistic deployment under latency- and cost-feasible constraints. Moreover, the observed pattern is not simply a uniform advantage of static over dynamic. For correct retry, the largest and most consistent advantage over dynamic was observed for common answer feedback, with context not detectably different from dynamic and outcome assigned actually worse. That suggests that the central issue is not simply model or prompt quality, but also alignment between feedback structure and the cognitive demands of the task.

These findings are also consistent with other recent research on AI-generated feedback. Zhao et al. [17] found that engagement patterns varied between multiple choice and open-ended multimodal feedback, highlighting the importance of tailoring feedback strategies by question type. While personalized, error-sensitive feedback was beneficial in prior analyses of open-ended questions in this environment, applying that same feedback approach for a FITB question, which engages different cognitive processes, may not be the most effective strategy for that type compared to the common answer feedback previously applied.

In this setting, then, the central issue is not whether dynamic feedback can produce reasonable responses, but whether it is the best fit for the task. For textbook-embedded FITB practice, feedback that more directly constrains the answer space may be better aligned to correct recovery of the target term than feedback that is more adaptive in principle but less retrieval-focused in practice.

3.4. Limitations and future directions

Several limitations should be kept in mind when interpreting these results. First, the dynamic condition reflects one realistic deployment configuration, and we do not claim that other implementations would necessarily yield the same pattern. Dynamic feedback is not a single design space, and more retrieval-focused, less explanatory prompting strategies may perform differently. The present result is therefore best understood as evidence about this task under a practical deployment configuration rather than as a universal statement about LLM-based feedback systems.

Second, the primary randomized comparison is between dynamic feedback and the static approach, not dynamic and common answer in isolation. The role of common answer is clarified through the secondary realized condition analyses rather than isolated directly in the primary ITT estimand. Because the realized condition decomposition is shaped partly by fallback, those analyses are interpreted as descriptive rather than as the primary causal result. Although edit distance checks suggested that the main randomized contrast was not driven by response proximity, fallback remains one reason the realized condition analyses should be interpreted cautiously.

Third, the present findings come from a particular mix of books and courses during a bounded observation window, and feature use was largely voluntary rather than assigned. Accordingly, the results characterize behavior in a naturalistic, self-directed usage context and may differ in settings where practice is required or more tightly integrated into course activity.

Fourth, randomization occurred at the incorrect attempt level, so students may experience multiple feedback conditions across different questions; the mixed effects models account for repeated observations within students but do not directly address possible carryover effects.

Several directions follow naturally from these findings. A first priority is to evaluate whether dynamic feedback can be redesigned to provide stronger narrowing cues without directly revealing the answer. The main challenge suggested by the present results may not be the use of an LLM itself, but the difficulty of producing feedback that is both non-leaking and sufficiently constraining for this retrieval task. That points toward prompt and system designs that are less focused on elaborating the student's incorrect answer and more focused on guiding the student toward the correct term through retrieval-supportive cues. One promising direction is to combine these strengths: future dynamic feedback could remain sensitive to the student's specific error while also incorporating stronger retrieval cues to support target term recovery. This is a promising direction given recent successes in multi-agent systems for feedback refinement [41].

A second direction is investigating hybrid approaches, such as selecting different feedback types depending on the nature of the student's error. Some responses may benefit from more retrieval-constraining cues, whereas others may benefit from answer-sensitive guidance that still preserves the student's need to produce the target term. This would align with the present interpretation that feedback effectiveness depends on matching support to the cognitive demands of the task.

Finally, future work should continue to evaluate these feedback approaches in authentic settings while extending beyond immediate next action behavior. The current study focuses on answer reveal and correct retry on the same question, which are appropriate first outcomes for understanding the retrieval process. A next step is to examine whether feedback differences observed here translate into broader gains in comprehension, retention, or course performance, and whether similar patterns emerge in tasks beyond textbook-embedded FITB retrieval practice.

4. Conclusion

This study examined whether dynamic LLM-generated feedback could outperform an existing static feedback approach for automatically generated FITB retrieval practice embedded in textbooks. In randomized analyses, dynamic feedback did not outperform the static approach overall. Although dynamic feedback was associated with slightly lower answer reveal rates, it produced no detectable change in correct retry. Secondary decomposition analyses further suggested that common answer feedback, the scaffold most directly aligned to target term recovery, was the static type with the most consistent advantage over dynamic for correct retry.

The broader lesson is that LLM-based feedback is not automatically the best option. In textbook-embedded FITB formative practice designed to support reading comprehension and the doer effect, feedback that more directly constrains the answer space may be better aligned to the task than feedback that is more adaptive in principle but less retrieval-focused in practice. More generally, the results caution against an overly broad inductive leap: because LLMs can generate fluent, personalized, and often reasonable responses, it does not follow that they are the best form of support for a given instructional task. This result identifies an important case in which a static scaffold specifically aligned to retrieval outperforms dynamic explanatory feedback for correct retry, and suggests that future progress may depend less on making feedback more elaborate than on better aligning it to the cognitive demands of the task.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-5.4 for drafting content, paraphrasing and rewording, and grammar and spelling checks; Claude Sonnet for citation management; GPT-5.5, Gemini, and Claude Opus for simulated peer review; and Claude Opus for revision assistance. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Wisniewski, B., Zierer, K., Hattie, J.: The power of feedback revisited: A meta-analysis of educational feedback research. Front. Psychol. 10, 487662 (2020). https://doi.org/10.3389/fpsyg.2019.03087

[2] Heckler, A.F., Mikula, B.D.: Factors affecting learning of vector math from computer-based practice: Feedback complexity and prior knowledge. Phys. Rev. Phys. Educ. Res. 12, 010134 (2016). https://doi.org/10.1103/PhysRevPhysEducRes.12.010134

[3] Ramadan Elbaioumi Shaddad, A., Jember, B.: A step toward effective language learning: An insight into the impacts of feedback-supported tasks and peer-work activities on learners' engagement, self-esteem, and language growth. Asian-Pac. J. Second Foreign Lang. Educ. 9, 39 (2024). https://doi.org/10.1186/s40862-024-00261-5

[4] Anderson, J.R., Corbett, A.T., Conrad, F.: Skill acquisition and the LISP tutor. Cogn. Sci. 13, 467 –506 (1989).

[5] Schaeffer, L.M., Margulieux, L.E., Chen, D., Catrambone, R.: Feedback via educational technology. In: Lin, L., Atkinson, R. (eds.) Educational Technologies: Challenges, Applications, and Learning Outcomes, pp. 59–72. Nova Science Publishers (2016).

[6] Huang, K., Chen, C.H., Wu, W.S., Chen, W.Y.: Interactivity of question prompts and feedback on secondary students' science knowledge acquisition and cognitive load. Educ. Technol. Soc. 18(4), 159–171 (2015).

[7] Shute, V.J.: Focus on formative feedback. Rev. Educ. Res. 78, 153–189 (2008). https://doi.org/10.3102/0034654307313795

[8] Van der Kleij, F.M., Feskens, R.W., Eggen, T.M.: Effects of feedback in a computer-based learning environment on students' learning outcomes. Rev. Educ. Res. 85(4), 475–511 (2015). https://doi.org/10.3102/0034654314564881

[9] VanLehn, K.: The behavior of tutoring systems. Int. J. Artif. Intell. Educ. 16, 227–265 (2006). https://dl.acm.org/doi/10.5555/1435351.1435353

[10] Cao, J., Zhao, C.Q., Schunn, C., McLaughlin, E.A., Lin, J., Koedinger, K.R.: Assessing the impact and underlying pathways of sequenced AI feedback on student learning. arXiv preprint arXiv:2604.07469 (2026). https://arxiv.org/abs/2604.07469

[11] Kulhavy, R.W.: Feedback in written instruction. Rev. Educ. Res. 47, 211–232 (1977). https://www.jstor.org/stable/1170128

[12] Aleven, V., Koedinger, K.R.: Limitations of student control: Do students know when they need help? In: International Conference on Intelligent Tutoring Systems, pp. 292–303. Springer (2000). https://dl.acm.org/doi/10.5555/648030.745996

[13] Morris, R., Perry, T., Wardle, L.: Formative assessment and feedback for learning in higher education: A systematic review. Rev. Educ. 9(3), 1–26 (2021). https://doi.org/10.1002/rev3.3292

[14] Kurdi, G., Leo, J., Parsia, B., Sattler, U., Al-Emari, S.: A systematic review of automatic question generation for educational purposes. Int. J. Artif. Intell. Educ. 30(1), 121–204 (2020). https://doi.org/10.1007/s40593-019-00186-y

[15] Van Campenhout, R., Kimball, M., Clark, M., Dittel, J.S., Jerome, B., Johnson, B.G.: An investigation of automatically generated feedback on student behavior and learning. In: Proceedings of LAK '24: 14th International Learning Analytics and Knowledge Conference, pp. 850–856 (2024). https://doi.org/10.1145/3636555.3636901

[16] Van Campenhout, R., Dittel, J.S., Jerome, B., Clark, M.W., Johnson, B.G.: Open-ended questions need personalized feedback: Analyzing LLM-enabled features with student data. In: Proceedings of the Second Workshop on Automated Evaluation of Learning and Assessment Content (EvalLAC '25). CEUR Workshop Proceedings (2025). https://ceur-ws.org/Vol-4006/paper5.pdf

[17] Zhao, C.Q., Cao, J., Lin, J., Koedinger, K.R.: LLM-based multimodal feedback produces equivalent learning and better student perceptions than educator feedback. In: Proceedings of

LAK '26: 16th International Learning Analytics and Knowledge Conference, pp. 632–642 (2026). https://doi.org/10.1145/3785022.3785124
[18] Dai, W., Tsai, Y.S., Lin, J., Aldino, A., Jin, H., Li, T., Gašević, D., Chen, G.: Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. Comput. Educ. Artif. Intell. 7 (2024). https://doi.org/10.1016/j.caeai.2024.100299
[19] Steinbach, M., Bhandari, S., Meyer, J., Pardos, Z.A.: When LLMs hallucinate: Examining the effects of erroneous feedback in math tutoring systems. In: Proceedings of the 12th ACM Conference on Learning @ Scale, pp. 139–150. ACM (2025). https://doi.org/10.1145/3698205.3729555
[20] Van Campenhout, R., Dittel, J.S., Jerome, B., Johnson, B.G.: 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 (2021). https://ceur-ws.org/Vol-2895/paper06.pdf
[21] Van Campenhout, R., Clark, M., Jerome, B., Dittel, J.S., Johnson, B.G.: Advancing intelligent textbooks with automatically generated practice: A large-scale analysis of student data. In: Fifth Workshop on Intelligent Textbooks at the 24th International Conference on Artificial Intelligence in Education. CEUR Workshop Proceedings, pp. 15–26 (2023). https://ceur-ws.org/Vol-3444/itb23_s1p2.pdf
[22] Van Campenhout, R., Clark, M.W., Dittel, J.S., Jerome, B., Brown, N., Johnson, B.G.: AI-generated formative practice and feedback: Performance benchmarks and applications in higher education. In: Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con), pp. 337–344. National Council on Measurement in Education (2025). https://aclanthology.org/2025.aimecon-main.36/
[23] Van Campenhout, R., Brown, N., Jerome, B., Dittel, J.S., Johnson, B.G.: Toward effective courseware at scale: Investigating automatically generated questions as formative practice. In: Proceedings of Learning @ Scale, pp. 295–298. ACM (2021). https://doi.org/10.1145/3430895.3460162
[24] Johnson, B.G., Dittel, J.S., Van Campenhout, R., Jerome, B.: Discrimination of automatically generated questions used as formative practice. In: Proceedings of the 9th ACM Conference on Learning @ Scale, pp. 325–329. ACM (2022). https://doi.org/10.1145/3491140.3528323
[25] Van Campenhout, R., Autry, K., Clark, M.W., Johnson, B.G.: Scaling the doer effect: A replication analysis using AI-generated questions. In: Proceedings of the 12th ACM Conference on Learning @ Scale, pp. 24–33. ACM (2025). https://doi.org/10.1145/3698205.3729545
[26] Koedinger, K., Kim, J., Jia, J., McLaughlin, E., Bier, N.: Learning is not a spectator sport: Doing is better than watching for learning from a MOOC. In: Proceedings of the 2nd ACM Conference on Learning @ Scale, pp. 111–120. ACM (2015). https://doi.org/10.1145/2724660.2724681
[27] Koedinger, K.R., McLaughlin, E.A., Jia, J.Z., Bier, N.L.: Is the doer effect a causal relationship? How can we tell and why it's important. In: Proceedings of LAK '16: 6th International Learning Analytics and Knowledge Conference, pp. 388–397 (2016). https://doi.org/10.1145/2883851.2883957
[28] Koedinger, K.R., Scheines, R., Schaldenbrand, P.: Is the doer effect robust across multiple data sets? In: Proceedings of the 11th International Conference on Educational Data Mining (2018).
[29] VanLehn, K.: The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educ. Psychol. 46(4), 197–221 (2011). https://doi.org/10.1080/00461520.2011.611369
[30] Van Campenhout, R., Dittel, J.S., Johnson, B.G.: Scaling effective characteristics of ITSs: A preliminary analysis of LLM-based personalized feedback. In: Graf, S., Markos, A. (eds.) Generative Systems and Intelligent Tutoring Systems. ITS 2025. Lecture Notes in Computer

Science, vol. 15723, pp. 171–181. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-98281-1_13
[31] Anderson, L.W., Krathwohl, D.R., Airasian, P.W., Cruikshank, K.A., Mayer, R.E., Pintrich, P.R., Raths, J., Wittrock, M.C.: A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. Longman (2001).
[32] OpenAI: GPT-4.1 nano: Technical specifications and system card. Microsoft Azure AI Model Catalog (2025). https://ai.azure.com/catalog/models/gpt-4.1-nano
[33] VitalSource Supplemental Data Repository (2026). https://github.com/vitalsource/data
[34] Hollis, S., Campbell, F.: What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ 319(7211), 670–674 (1999). https://doi.org/10.1136/bmj.319.7211.670
[35] Angrist, J.D., Imbens, G.W., Rubin, D.B.: Identification of causal effects using instrumental variables. J. Am. Stat. Assoc. 91(434), 444–455 (1996). https://doi.org/10.2307/2291629
[36] Book Industry Study Group: Complete BISAC subject headings list (2025). https://www.bisg.org/complete-bisac-subject-headings-list
[37] R Core Team: R: A language and environment for statistical computing, version 4.4.1 [Computer software]. R Foundation for Statistical Computing (2024). https://www.R-project.org/
[38] Brooks, M.E., Kristensen, K., Benthem, K.J., Magnusson, A., Berg, C.W., Nielsen, A., Skaug, H.J., Maechler, M., Bolker, B.M.: glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. R J. 9(2), 378–400 (2017). https://doi.org/10.32614/RJ-2017-066
[39] Lyons, M., Harrison, N., Brewer, G., Robinson, S., Sanders, R.: Biological Psychology, 1st edn. Learning Matters (2014).
[40] Romano, N.: Employment Law for Paralegals, 3rd edn. Emond Publishing (2026).
[41] Cao, J., Zhao, C.Q., Chen, X., Wang, S., Schunn, C., Koedinger, K.R., Lin, J.: From first draft to final insight: A multi-agent approach for feedback generation. arXiv preprint arXiv:2505.04869 (2025). https://arxiv.org/abs/2505.04869