VitalSource®
Learning Science Group

AIED SEOUL 2026

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

iTextbooks Workshop 2026

Does Dynamic LLM Feedback Improve on Static?

LLMs lower the barrier to error-sensitive feedback

- Students get immediate feedback after an incorrect FITB attempt
- Static feedback is question-level: same response regardless of the student's answer
- Dynamic LLM feedback: conditioned on the student's actual wrong answer, more adaptive in principle

Does the added adaptivity translate to better outcomes?

Static Feedback: Three Types, One Clear Winner

Prior work compared three question-level feedback types for FITB practice:

Common Answer Feedback — strongest scaffold
A second cloze sentence with same target term removed; student retrieves the term again in a new context

Context Feedback
Extended textbook passage around the question: more context, no direct answer constraint

Outcome Feedback — weakest scaffold
Informs the student the response is incorrect — minimal, but always generatable

Common answer is the preferred static type when generatable: the benchmark for this study

Dynamic LLM Feedback

Unlike static feedback, dynamic feedback adapts to what the student actually entered

Inputs to the model: question stem · correct answer · student's incorrect answer

Output (2–3 sentences):

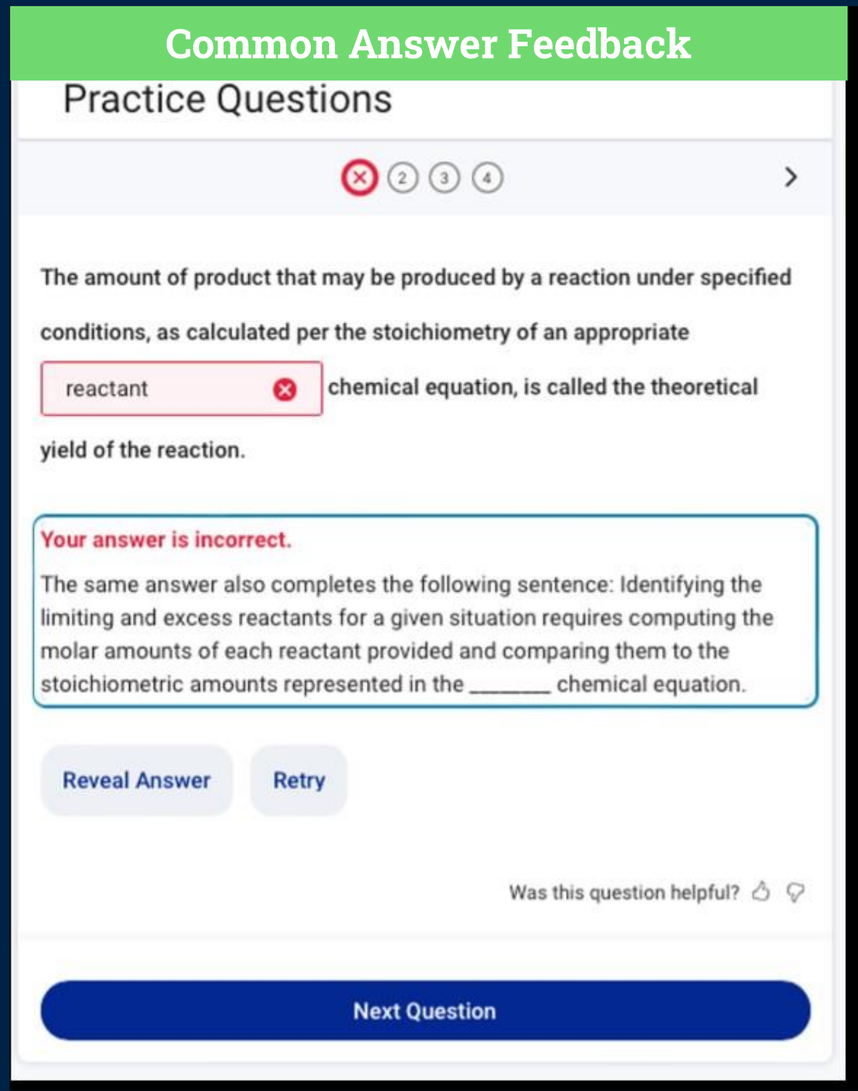
- Acknowledges the student's answer
- Explains why it doesn't fit
- Redirects without revealing the correct answer

No-leak guardrail: if generated response contains the answer → fall back to static
Model: `gpt-4.1-nano` (latency and cost feasible for live deployment)

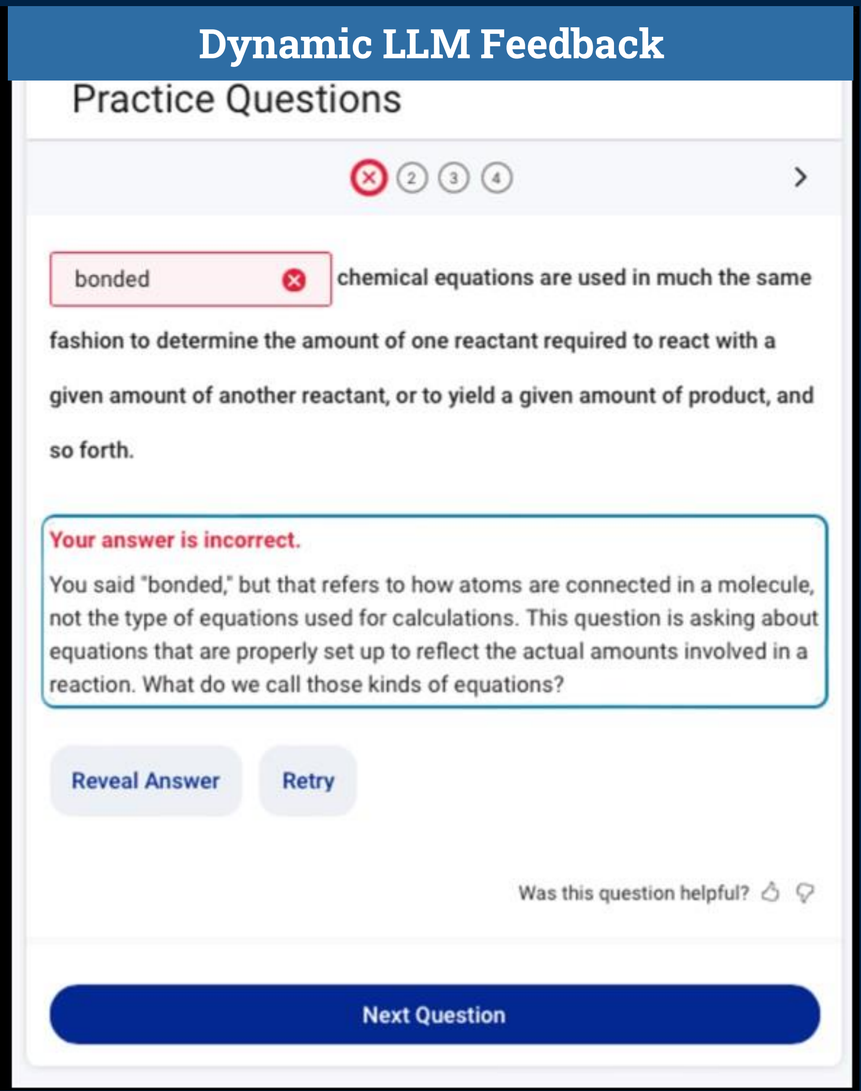
LLM feedback worked well for open-ended questions; FITB seemed like a logical next step

What Each Feedback Type Looks Like

Common Answer Feedback



Dynamic LLM Feedback



Three Research Questions

RQ1. Does dynamic feedback improve behavioral outcomes relative to the existing static approach?

RQ2. Which specific static feedback types appear most associated with the observed differences?

RQ3. What do results imply about the relationship between feedback design and retrieval practice?

Study Design

Each incorrect FITB attempt was independently randomized at submission time

- Dynamic condition: LLM generates feedback conditioned on the student's specific wrong answer
- Static condition: existing policy: common answer if available, then context, then outcome

Primary analysis: ITT (intention-to-treat)

Effect of assignment to dynamic vs. static, not of delivery
One-sided noncompliance: dynamic-assigned may fall back to static (no-leak guardrail); static-assigned always receive static

Secondary analysis: realized-condition decomposition

Which delivered feedback types drive the pattern?

Randomized Deployment, Real Students

33,834
sessions

5,596
students

23,148
unique questions

1,363
textbooks

8
publishers

Measuring What Happens After Feedback

Incorrect first attempt
on an FITB question

→

Feedback shown
dynamic or static

→

Student's next action
four possibilities

Correct retry | Incorrect retry | **Answer reveal** | Abandon

Outcome 1: Correct Retry

Student answers correctly on a subsequent attempt
Unconditional: includes all students regardless of reveal

Outcome 2: Answer Reveal

Student clicks to reveal the answer rather than retry
Persistence-without-recovery pattern: reveals ↑, retries flat

Primary Results: Dynamic Doesn’t Outperform Static

Outcome	Dynamic	Static	ITT difference (dynamic − static)
Answer Reveal	**61.0%**	**64.3%**	**−3.3pp [−4.4, −2.2]** CI excludes zero ✓
Correct Retry	**16.5%**	**16.9%**	**−0.4pp [−1.3, +0.4]** CI crosses zero (not detectable)

Dynamic feedback reduced answer reveals slightly, but produced no detectable gain in correct retries.

ITT · 33,834 sessions · 95% CI from 100K student-level cluster bootstrap replicates

Persistence Without Recovery

Where did the −3.3pp in answer reveals go? Broken down by next action:

	Dynamic	Static
Correct retry	**16.5%**	**16.9%**
Incorrect retry	**18.5%**	**14.6%**
Answer reveal	**61.0%**	**64.3%**

The 3.3pp persistence gain went to failed retries (+3.9pp), not correct ones (−0.4pp).

Mixed Effects Models Confirm the Pattern

Mixed effects logistic regression models with random intercepts for student and question:

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

Answer Reveal — significant

Static condition OR = 1.36 vs. dynamic ($p < .001$)

Static-assigned students were more likely to request the answer, confirming dynamic reduced reveals

Correct Retry — not significant

Static condition OR = 1.08 vs. dynamic ($p = .079$)

No detectable difference after adjusting for student- and question- level heterogeneity

Decomposition: “Static” Is Not One Thing

The static condition is a hierarchical policy, not a single feedback type:

- Common answer feedback: if generatable from nearby text (56.3% of static-delivered cases)
- Context feedback: extended passage surrounding the question sentence
- Outcome feedback: only informs student their answer is incorrect

Dynamic-assigned cases may fall back to static (no-leak guardrail):

- Dynamic delivered (assigned dynamic, received dynamic): 61.6%
- Fallback cases (assigned dynamic, received static): labeled “common fallback,” “context fallback,” “outcome fallback”

Secondary question: Which delivered feedback states account for the pattern?

- Not causal; realized conditions are not randomized, but informative for interpretation

Correct Retry by Realized Feedback Condition

Realized Condition	Correct Retry	vs. Dynamic	OR (model) significance
Dynamic (baseline)	**14.5%**	reference	
Common assigned	**18.4%**	**+3.9pp**	OR 1.40 ***
Common fallback	**21.3%**	**+6.8pp**	OR 1.36 ***
Outcome assigned	**12.9%**	**−1.6pp**	OR 0.68 ***

Not “Static Beats Dynamic” — It’s More Nuanced

Common Answer
18–21%
correct retry

significantly outperforms dynamic (OR 1.36–1.40 ***)

Dynamic
14.5%
correct retry

context feedback: comparable (OR ~0.95, ns)

Outcome
12.9%
correct retry

significantly underperforms dynamic (OR 0.68 ***)

The advantage is concentrated at the extremes of the retrieval-cue spectrum.

Why Common Answer Wins: Task Alignment

Question stem (Employment Law for Paralegals):
"The weight of the evidence indicated that a personal injury by _____ occurred while the worker was engaged in a work-related duty." Student answered: **"negligence"**

Dynamic LLM Feedback

"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?"

Reasonable and on-topic. Engages the wrong answer, explains the mismatch, redirects. But doesn't give a second cloze cue; student still must recall "accident" unaided.

Common Answer Feedback

"The worker was treated by a psychiatrist for approximately one year following the _____ and complained of dysphoria, tearfulness, insomnia, and poor concentration."

A second cloze sentence using the same target term. Doesn't explain; constrains the answer space. Still requires the student to produce "accident."

Employment Law for Paralegals · correct answer: "accident"

Semantic Activation vs. Diagnostic Constraint

Dynamic feedback
Semantic activation

Engages the student's specific wrong answer
Explains why it doesn't fit
Redirects toward the right semantic region

Keeps students engaged (lower reveal rates)
But doesn't narrow the answer space as efficiently;
student still must recall the target term unaided

Common answer feedback
Diagnostic constraint

A second cloze sentence with the same target term
Another chance to retrieve the same term in context
Doesn't explain — constrains

Student still must produce the answer themselves
Better aligned to what retrieval practice requires

Dynamic feedback provides semantic activation. Common answer provides diagnostic constraint. For FITB retrieval practice, constraint wins.

Implications: Match Feedback to the Task

The broader principle

Feedback effectiveness is not absolute — it depends on
- What cognitive operation the task demands
- Whether feedback structure matches that demand

FITB retrieval practice asks students to produce
- Elaboration on the wrong answer is off-task
- A second retrieval cue is on-task

Alternative explanations considered

"Common = giveaway" — No: student must still produce the answer
"Dynamic too verbose" — Length confounded; outcome feedback also fails
"Model quality" — The structural advantage holds across subtypes

Design guidance

For FITB practice:
- ✓ Common answer
- ≈ Context feedback
- ✗ Outcome explanation

For open-ended tasks:
- Dynamic may excel
- Error-specific reasoning fits

The question is not "is LLM feedback good?" but "good for what?"

Limitation: single platform, FITB cloze questions · generalization to other formats requires further study

Summary of Contributions

01 · ITT at scale

Static common-answer feedback reduces answer-reveal rate by 3.3pp vs. dynamic LLM feedback (33K sessions)

Correct-retry rate unchanged (−0.4pp, ns); persistence went to failed retries, not recovery

02 · Within-static decomposition

Common answer (56% of questions) drives the benefit: OR 1.36–1.40 vs. dynamic

Outcome feedback significantly underperforms dynamic (OR 0.68 ***)

03 · Mechanism

Task alignment: a second cloze sentence is a retrieval cue, not a giveaway

Dynamic's semantic activation lacks the diagnostic constraint the task requires

For FITB retrieval practice: a carefully authored scaffold beats a dynamically generated one. Not because LLMs are bad, but because the task demands diagnostic constraint, not elaboration.

Practical Implications

Instructional designers

Don’t assume adaptive = better for retrieval tasks
Invest in authoring common-answer scaffolds for FITB banks
If using LLMs, target context-level feedback over outcome explanations

Platform teams & publishers

LLM feedback ROI depends on existing scaffold coverage
Target LLMs at questions without scaffolds; don’t replace scaffolds that already work
Prioritize LLM prompt design around cloze constraints, not error diagnosis

Open questions for future work:

Does task-alignment hold for open-ended constructed-response or essay questions?
Can LLMs generate common-answer-style cues automatically?
Do effects persist to downstream assessment?

VitalSource®

Learning Science Group

The question is not

“Is LLM feedback better?”

The question is

“Better for what task, and why?”

For textbook-embedded FITB retrieval practice, static scaffolds win, not despite being simple, but **because** they are.

Open dataset

Thank you · Questions welcome