International Journal of Educational Technology in Higher Education

RESEARCH Open Access

# Human-centered GenAI feedback design in higher education: a multisite experiment on direct, reflective, and hybrid approaches to scientific argumentation

Huseyin Ates[1*]

*Correspondence:
Huseyin Ates
huseyin.ates@ahievran.edu.tr
[1]Kırşehir Ahi Evran University, Kırşehir, Turkey

**Abstract**

Generative artificial intelligence (GenAI) is increasingly used for feedback in higher education, yet evidence remains limited on how alternative human–AI feedback designs shape learning processes and durable outcomes. This study addresses that gap through a multisite, cluster-randomized, longitudinal field experiment comparing four feedback designs in introductory university science courses: peer feedback only, direct GenAI-supported feedback, reflective GenAI-supported feedback, and a hybrid design combining self-evaluation, peer feedback, and GenAI critique. The analytic sample comprised 1,176 first-year undergraduate students from 48 course sections across four universities and three science domains. Primary and secondary outcomes were argument-quality gain on four shared rubric dimensions—claim quality, evidence relevance and sufficiency, coherence of reasoning, and treatment of limitations or alternative explanations—conceptual learning, and delayed AI-free transfer; feedback uptake and self-regulated learning during revision were modeled as process mediators. Direct GenAI-supported feedback improved immediate argument-quality gain relative to peer feedback, whereas reflective and hybrid designs produced stronger feedback uptake and self-regulated learning. The hybrid condition yielded the highest adjusted mean for immediate argument-quality gain and showed the clearest advantage on conceptual learning; the reflective condition showed a positive but non-significant adjusted contrast on conceptual learning relative to direct GenAI-supported feedback. Both reflective and hybrid conditions outperformed direct GenAI-supported feedback on delayed AI-free transfer. Multilevel mediation analyses indicated that feedback uptake and self-regulated learning partially explained these advantages. By comparing four feedback designs, modeling revision processes, and assessing delayed AI-free transfer in a multisite field experiment, the findings suggest that the educational value of GenAI in higher education may depend less on AI access per se than on whether feedback environments preserve student agency, evaluative judgment, and ownership during revision.

**Keywords** Generative artificial intelligence, Feedback design, Higher education, Digital learning, Scientific argumentation, Self-regulated learning, Epistemic agency

## Introduction

Generative artificial intelligence (GenAI) is increasingly embedded in higher education, yet its educational significance remains unsettled (Wang & Fan, 2025; Xia et al., 2024). Evidence suggests that GenAI can support learning performance, perceptions, and higher-order thinking, but these benefits vary with the tool's role, the instructional design, and the task (Wang & Fan, 2025). In feedback contexts, this variation is crucial because feedback becomes educationally valuable not through comment delivery alone, but when learners interpret critique, compare it against criteria, judge its relevance, and use it to improve subsequent work (Carless & Boud, 2018; Hattie & Timperley, 2007; Nicol & Macfarlane-Dick, 2006). The central issue is therefore whether GenAI-supported feedback design sustains feedback uptake, evaluative judgment, and learner agency rather than merely automating critique (Buckingham Shum et al., 2023; Darvishi et al., 2024).

This issue is especially important in science learning. Scientific argumentation requires students to formulate claims, justify them with evidence and reasoning, consider alternatives, and revise explanations in response to critique (Driver et al., 2000; Osborne et al., 2004). Unlike generic writing, it requires disciplinary judgment about evidence adequacy, reasoning quality, and uncertainty, not only coherence or fluency (Osborne et al., 2004; Zhu et al., 2020). GenAI-supported feedback may support deeper reasoning when designed dialogically, yet it may also encourage passive uptake if students outsource evaluative judgment to the system (Darvishi et al., 2024; Tang & Putra, 2025; Zhu et al., 2020).

Recent reviews identify three gaps in the higher education feedback literature (Wang & Fan, 2025; Xia et al., 2024). First, disciplinary reasoning in university science courses has received less attention than general writing or broad feedback-source comparisons (Banihashem et al., 2024; Tang & Putra, 2025; Wang & Fan, 2025). Second, few studies compare alternative designs that sequence self-evaluation, peer input, and GenAI critique in different ways (Banihashem et al., 2024; Chang et al., 2026; Darvishi et al., 2024). Third, immediate product improvement is often examined without integrating feedback uptake, self-regulated learning, and later AI-free performance, making it difficult to determine whether gains reflect durable learning or temporary performance support (Wang & Fan, 2025; Xia et al., 2024; Zhu et al., 2020).

The present study addresses these gaps through a multisite, cluster-randomized, longitudinal field experiment in introductory university science courses. It compares four feedback designs—peer feedback only, direct GenAI-supported feedback, reflective GenAI-supported feedback, and a hybrid design combining self-evaluation, peer feedback, and GenAI critique—and examines immediate four-dimension argument-quality gain (improvement from initial to revised drafts on the four rubric dimensions scored at both stages: claim quality, evidence relevance and sufficiency, coherence of reasoning, and treatment of limitations or alternative explanations), conceptual learning, delayed AI-free transfer, feedback uptake, and self-regulated learning. By combining four feedback designs, process mediation, and delayed AI-free transfer, the study examines how

GenAI can be integrated into feedback processes in ways that amplify rather than displace students' evaluative work.

## Literature review and theoretical framework

### Feedback as uptake, judgment, and revision

The present study is grounded in a process-oriented understanding of feedback. Feedback is not simply information transmitted from a source to a learner; it becomes educationally meaningful when learners compare their work with criteria, interpret comments, and use those judgments to guide revision (Carless & Boud, 2018; Nicol, 2021; Nicol & Macfarlane-Dick, 2006). This view shifts attention from feedback provision to feedback use and foregrounds the learner processes through which comments are transformed into action.

This perspective is closely related to evaluative judgment, defined as the capacity to make decisions about the quality of one's own or others' work (Tai et al., 2018). Feedback is valuable not only when it is accurate or detailed, but when students are positioned to judge quality, compare alternatives, and decide how revision should proceed. Carless and Boud (2018) conceptualized this broader capability through feedback literacy, including students' capacity to appreciate feedback, make judgments, manage affective responses, and take action. Molloy et al. (2020) similarly argued that feedback-literate environments should support students' active role in interpreting and using feedback.

This distinction matters in technology-mediated feedback. Digital platforms and AI-enabled tools can make feedback faster and more scalable, but increased access to commentary does not automatically generate learning (Buckingham Shum et al., 2023). Nicol (2021) argued that feedback is most productive when it stimulates internal feedback through comparison between current performance, criteria, alternatives, and desired standards. In this study, feedback literacy refers to this broader learner capability, whereas feedback uptake refers to the task-specific process through which students attend to, interpret, select, and implement critique during a revision episode.

### Self-regulated learning and epistemic agency in AI-mediated feedback

Feedback uptake is closely related to self-regulated learning (SRL), which describes how students plan, monitor, evaluate, and adapt their cognition, motivation, and behavior in response to task demands (Nicol & Macfarlane-Dick, 2006; Panadero, 2017). In revision contexts, students must decide how to interpret comments, whether to trust them, which changes to prioritize, and how to integrate critique into future performance. Feedback therefore becomes educationally consequential when it supports learners' own regulatory activity.

This relationship can be understood through evaluative judgment and internal feedback. When students compare external comments with task criteria, prior understanding, and their own diagnosis of the work, they generate internal feedback that can inform planning, monitoring, and strategic adjustment (Nicol, 2021; Tai et al., 2018). Feedback uptake is therefore treated here as a proximal revision process that can support subsequent self-regulation.

AI-mediated feedback intensifies, this issue. GenAI may support SRL when it helps learners identify weaknesses, clarify criteria, compare alternatives, and plan revision moves (Buckingham Shum et al., 2023). However, it may also undermine SRL if students

accept suggestions uncritically or outsource difficult judgments to the system. Darvishi et al. (2024) showed that AI assistance in peer-feedback environments may increase dependence on the tool, raising questions about student agency. In this study, that concern is framed as epistemic agency: whether students remain responsible for judging the adequacy of claims, evidence, and reasoning, or whether those judgments are displaced by AI-generated critique.

Research on self-assessment provides a basis for reflective feedback design. Meta-analytic findings indicate that self-assessment can support self-regulated learning and self-efficacy because it requires learners to compare their work with explicit criteria and generate internal judgments before acting on external feedback (Panadero et al., 2017). In AI-supported revision, requiring students to articulate strengths, weaknesses, uncertainties, and revision priorities before viewing GenAI critique may therefore position AI feedback as a resource for comparison rather than as a substitute for judgment.

#### Scientific argumentation as disciplinary and epistemic work

Scientific argumentation provides a demanding context for studying AI-supported feedback because it differs from generic academic writing. Students must formulate claims, coordinate claims with evidence and reasoning, address limitations or alternative explanations, and revise explanations in light of critique (Driver et al., 2000; Osborne et al., 2004). Revision quality therefore depends not only on coherence or fluency, but also on disciplinary judgment about evidence, reasoning, uncertainty, and explanatory adequacy.

Feedback on scientific argumentation must address whether evidence is relevant and sufficient, whether reasoning justifies the claim, whether uncertainty or alternative explanations are acknowledged, and whether revisions improve the explanatory force of the argument. Zhu et al. (2020) showed that, in formative assessment of scientific argument writing, revision was associated with learning gains and contextualized feedback was more effective than generic feedback. Emerging work also suggests that GenAI can support disciplinary engagement when it functions as a dialogic partner rather than an answer source; Tang and Putra (2025), for example, found that a customized chatbot supported perspective-taking, reasoning, and argumentation in science learning.

For this reason, scientific argumentation is a critical case for examining whether GenAI-supported feedback amplifies or displaces students' evaluative work. If student judgment and ownership can be preserved in this context, the implications may extend to other higher education tasks that require evidence-based explanation, justification, and revision.

#### Why direct, reflective, and hybrid feedback designs may differ

The core theoretical claim of the present study is that alternative feedback designs may generate different learning processes because they distribute evaluative work differently across learners, peers, and AI-supported tools. From a human-centered digital learning perspective, the central issue is not AI access alone, but whether feedback design preserves learners' responsibility for interpreting critique, making judgments, and deciding how revision should proceed (Nicol, 2021; Tai et al., 2018).

In direct GenAI-supported feedback, students receive AI-generated critique without first externalizing their own diagnosis of the task. This design may support efficiency

and immediate argument-quality gain through rapid, criterion-referenced suggestions. At the same time, because much evaluative work is front-loaded onto the system, it may offer weaker conditions for internal comparison, self-generated judgment, feedback uptake, self-regulated revision, and delayed AI-free transfer.

In reflective GenAI-supported feedback, students first evaluate their own work against explicit criteria or identify uncertainty before engaging with AI-generated critique. This sequence may be beneficial because it requires learners to articulate an initial judgment and then compare that judgment with external feedback. It is consistent with the view that evaluative judgment develops through opportunities to make and test quality-related decisions, rather than merely receive recommendations about what to change (Nicol, 2021; Tai et al., 2018).

In hybrid feedback, self-evaluation, peer feedback, and GenAI critique are combined within the same instructional sequence. These sources may offer complementary affordances: self-evaluation foregrounds internal criteria and initial judgment, peer feedback highlights reader-oriented comprehensibility and ambiguity, and GenAI critique may identify rubric-aligned omissions or alternative revision moves (Banihashem et al., 2024; Chang et al., 2026). Together, they may create a strong comparison ecology for active judgment and revision because students must compare multiple forms of critique, select among them, and justify how revision decisions should proceed.

At the same time, reflective and hybrid designs also require additional evaluative activity. Any observed advantage should therefore be interpreted as an effect of the designed feedback process as a whole, rather than as an isolated effect of feedback source alone. This distinction is important because reflective and hybrid conditions may differ from direct GenAI-supported feedback not only in sequencing and source configuration, but also in time-on-task, reflective workload, and structured cognitive engagement.

Based on this framework, the study proposes a process model in which feedback design may affect learning outcomes both directly and indirectly. Direct effects may arise because alternative designs alter the timing, diversity, and content of critique available to learners. Indirect effects are expected to operate through feedback uptake, self-regulated learning during revision, and the serial pathway from uptake to self-regulated revision. Reflective and hybrid designs may therefore be better positioned to support argument-quality gain, conceptual learning, and delayed AI-free transfer because they require learners to compare, select, and justify revision decisions rather than merely implement external suggestions. The proposed process model is presented in Fig. 1.

## Research questions and hypotheses

The research questions and hypotheses were designed to examine both outcome differences across feedback designs and the revision processes that may explain those differences. To avoid conflating change in common argument quality with revised-product quality, strategic quality of revision was treated as a separate revised-draft indicator and used in a supplementary interpretive role rather than as part of the gain outcome.

**RQ1.** How do peer feedback, direct GenAI-supported feedback, reflective GenAI-supported feedback, and hybrid feedback differ in their effects on students' immediate four-dimension argument-quality gain in scientific argumentation tasks?

**RQ2.** How do these four feedback conditions differ in their effects on students' conceptual learning and delayed AI-free transfer in introductory university science courses?

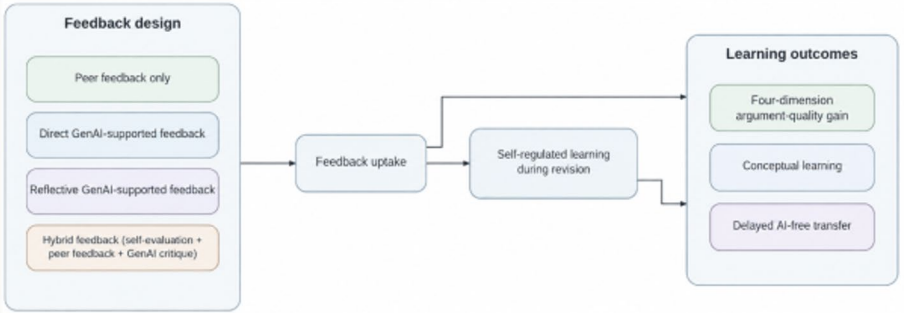


**Fig. 1** Process model of GenAI-supported feedback in scientific argumentation. The model proposes that feedback design affects four-dimension argument-quality gain, conceptual learning, and delayed AI-free transfer both directly and indirectly through feedback uptake and self-regulated learning during revision. Feedback uptake is modeled as a proximal process through which learners interpret, select, and implement critique, while self-regulated learning captures the planning, monitoring, and strategy adjustment that follow during revision. The model treats feedback design as a pedagogical configuration involving feedback source, sequencing, and structured evaluative activity. It is situated in introductory university science courses and scientific argumentation tasks

**RQ3.** How do these feedback conditions differ in terms of students' feedback uptake and self-regulated learning during the revision process?

**RQ4.** To what extent do feedback uptake and self-regulated learning mediate the relationship between feedback condition and students' four-dimension argument-quality gain, conceptual learning, and delayed AI-free transfer?

**H1a** *Reflective GenAI-supported feedback and hybrid feedback will yield higher levels of feedback uptake than direct GenAI-supported feedback because they require learners to make evaluative judgments before and during engagement with AI-generated critique.*

**H1b** *No directional hypothesis is specified between reflective GenAI-supported feedback and hybrid feedback, although the hybrid design may provide additional advantages through multi-source comparison.*

**H2a** *Reflective GenAI-supported feedback and hybrid feedback will yield higher levels of self-regulated learning during revision than direct GenAI-supported feedback.*

**H2b** *No directional hypothesis is specified between reflective GenAI-supported feedback and hybrid feedback for self-regulated learning during revision.*

**H3** *Reflective GenAI-supported feedback and hybrid feedback are expected to support stronger conceptual learning and delayed AI-free transfer than direct GenAI-supported feedback because they preserve learners' evaluative judgment and promote more agentic revision.*

**H4** *Direct GenAI-supported feedback may be expected to improve immediate four-dimension argument-quality gain relative to peer feedback by providing rapid, criterion-referenced critique; however, reflective and hybrid feedback designs are expected to yield comparable or stronger gains when deeper feedback uptake and self-regulated revision are supported.*

**H5a** *Feedback uptake will mediate the relationship between feedback condition and students' four-dimension argument-quality gain, conceptual learning, and delayed AI-free transfer.*

**H5b** *Self-regulated learning during revision will mediate the relationship between feedback condition and students' four-dimension argument-quality gain, conceptual learning, and delayed AI-free transfer.*

**H5c** *Feedback uptake and self-regulated learning will jointly mediate the relationship between feedback condition and students' four-dimension argument-quality gain, conceptual learning, and delayed AI-free transfer through a serial pathway from uptake to self-regulated revision.*

## Method

### Design

This study employed a multisite, cluster-randomized, longitudinal field experiment to examine how different feedback designs shaped students' scientific argumentation and learning in introductory university science courses. Cluster randomization was used because the intervention was implemented at the level of intact course sections, thereby reducing contamination across conditions during peer interaction and revision activities. The unit of randomization was the course section, whereas the primary unit of analysis was the student, with clustering handled through multilevel modeling. Four conditions were compared: peer feedback only, direct GenAI-supported feedback, reflective GenAI-supported feedback, and hybrid feedback combining self-evaluation, peer feedback, and GenAI critique.

The study followed a pre-test, post-test, and delayed post-test structure. The primary revision outcome captured change from initial draft to revised submission on the four rubric dimensions shared across draft stages; the full scoring procedure is described in Sect. "Four-dimension argument-quality gain". Strategic quality of revision was treated as a separate revised-draft indicator because it was scored only for revised submissions and was used as a concurrent supplementary indicator rather than as part of the gain outcome. Secondary outcomes were conceptual learning and delayed AI-free transfer. Two process variables—feedback uptake and self-regulated learning during revision—were specified a priori as mediators linking feedback design to learning outcomes.

Gain scores were retained because the instructional target was revision-related improvement from draft to final submission and because the outcome was preregistered in this form. This specification also matched the intervention logic: students produced an initial draft, received condition-specific feedback, and revised the same text. To reduce bias associated with baseline performance differences, models adjusted for baseline scientific argumentation, and baseline equivalence across conditions was examined before outcome modeling. Nevertheless, because students with stronger initial drafts may have had less room for improvement, ceiling-related constraints are addressed in the Limitations section and through a supplementary baseline-adjusted final-score sensitivity model.

The design, outcome definitions, focal contrasts, and analysis plan were preregistered on the Open Science Framework (OSF) prior to post-test analysis. The preregistered

focal contrasts emphasized comparisons of the more agentic feedback designs with direct GenAI-supported feedback while also retaining planned comparisons with the peer-feedback condition.

An a priori power analysis for a four-arm cluster-randomized design indicated that, assuming an intraclass correlation of .05, an average cluster size of 26 students, a two-sided alpha of .05, and a minimum detectable standardized effect of $d = .25$ for the planned between-condition contrasts, 48 sections would provide statistical power of .82. Power calculations were conducted in Optimal Design 3.01 under a balanced allocation scenario. On this basis, the target sample was set at approximately 1,200–1,300 students across four institutions.

### Setting and participants

The study was conducted across four universities in compulsory first-year biology, chemistry, and physics courses that required written scientific argumentation. The intervention was embedded in normal coursework rather than delivered as an external experimental activity. The initial sample comprised 1,248 students from 48 intact course sections.

Sections were randomly assigned to one of four feedback conditions, with the course section as the unit of randomization. Randomization was conducted by a researcher not involved in instruction, blocked by institution and discipline, and balanced by section meeting time. A common intervention schedule, shared task-design protocol, and common analytic rubric were used across sites to reduce institution-level heterogeneity.

Participation in instructional activities was part of normal coursework, whereas inclusion in the research dataset required informed consent for use of student work, survey responses, interview data, and platform logs. Students who did not consent completed the same class activities but were excluded from the analytic dataset. Additional exclusions and final analytic sample characteristics are reported in the Results section. Because research consent was voluntary, the analytic sample may reflect some consent-based selection.

### Instructional tasks and rubric

The intervention comprised three scientific argumentation cycles across 10 weeks. In each cycle, students produced an initial draft, received condition-specific feedback, revised the text, and, submitted a final version. Tasks included laboratory-based interpretive writing and short socio-scientific argumentation tasks, all designed to require coordination of claim, evidence, reasoning, and uncertainty. Tasks were developed through a shared design protocol and reviewed across institutions to support comparability across biology, chemistry, and physics.

Each cycle required an initial draft of approximately 350–500 words and a revised submission of 400–600 words. Students had 72 h to complete the first draft, 48 h to review assigned feedback, and 72 h to submit the revision. In the hybrid condition, the revision memo was limited to 100–150 words.

Student work was scored using a five-dimension analytic scientific argumentation framework adapted from prior research on scientific argumentation and automated feedback in scientific writing (e.g., Osborne et al., 2004; Zhu et al., 2020). The first four dimensions—claim quality, evidence relevance and sufficiency, coherence of reasoning,

and treatment of limitations or alternative explanations—were scored for both initial and revised drafts and used to estimate four-dimension argument-quality gain. The fifth dimension, strategic quality of revision, was scored only for revised submissions. The same four common content dimensions were applied across intervention and transfer tasks, with topic-specific descriptors added where necessary. Full rubric descriptors, calibration procedures, and anchor materials are provided in the supplementary materials.

### Experimental conditions

An overview of the four intervention conditions is provided in Table 1. All conditions used the same scientific argumentation tasks, analytic rubric, revision deadlines, and final-submission requirements. The condition-specific sequences, student roles, AI roles, and intended pedagogical functions are summarized in the table; this section therefore highlights only the main design differences.

The peer-feedback condition used feedback from two peers through a common rubric-based template and included no AI support. The direct GenAI-supported condition provided immediate criterion-referenced GenAI feedback after draft submission, without requiring students to complete a prior self-evaluation. The reflective GenAI-supported condition required a brief rubric-aligned self-evaluation before students received GenAI feedback. The hybrid condition combined self-evaluation, peer feedback, GenAI critique, and a short revision memo in which students explained which suggestions they accepted, rejected, or modified.

The conditions therefore differed not only in feedback source and sequence but also in the amount of structured evaluative activity required before revision. This design choice was theoretically intentional because the study conceptualized feedback design as a process configuration rather than as feedback-source exposure alone. However, it

**Table 1** Overview of the four feedback conditions

| Condition | Feedback sources | Sequence | Student role | AI role | Intended pedagogical function |
|---|---|---|---|---|---|
| Peer feedback only | Peer feedback | Draft → peer feedback from two peers → revision → final submission | Reviewer and reviser | None | To support evaluative judgment through peer review and revision without AI assistance |
| Direct GenAI-supported feedback | GenAI feedback | Draft → GenAI-generated feedback → revision → final submission | Reviser | Provides criterion-referenced critique and revision suggestions | To provide immediate, scalable, rubric-aligned feedback with minimal preparatory judgment by the learner |
| Reflective GenAI-supported feedback | Self-evaluation + GenAI feedback | Draft → structured self-evaluation → GenAI-generated feedback → revision → final submission | Self-evaluator and reviser | Provides critique after the learner articulates an initial judgment | To strengthen feedback uptake by requiring prior evaluative judgment before AI-supported revision |
| Hybrid feedback | Self-evaluation + peer feedback + GenAI feedback | Draft → self-evaluation → peer feedback from two peers → GenAI-generated feedback → revision memo → final submission | Self-evaluator, reviewer, comparator, and reviser | Provides additional critique and revision suggestions after self- and peer-based judgment | To support structured comparison across self-evaluation, peer feedback, and AI-supported critique while requiring students to justify revision decisions |

All conditions used the same scientific argumentation tasks, analytic rubric, revision deadlines, and final-submission requirements. Conditions differed in feedback source, feedback sequence, and the amount of structured evaluative activity required before revision. Reflective and hybrid feedback therefore represent pedagogical feedback configurations rather than feedback-source exposure alone

also means that condition effects cannot be attributed exclusively to feedback source independent of time-on-task, reflective workload, or cognitive engagement. Access to condition-specific feedback resources was monitored through platform records, and potential contamination through non-assigned resources is addressed under implementation fidelity.

### GenAI system and prompting protocol

The AI-supported conditions used a secure, institutionally hosted interface based on OpenAI's GPT-5 model and integrated into the learning-management platform. The interface enabled interaction logs to be captured while maintaining institutional data-governance requirements, and students were instructed not to enter personally identifying information into the system.

The tool was configured to provide criterion-referenced feedback rather than replacement text. It identified strengths, weaknesses, missing evidence, unsupported claims, revision moves, and follow-up questions while prohibiting full-answer rewriting. To ensure consistency across institutions and cycles, all AI-supported conditions used the same prompt architecture, rubric anchors, fixed model parameters, and single-generation rule across the three intervention cycles. Full prompt materials, model settings, interface screenshots, and implementation rules are provided in Supplementary Appendix B.

### Procedure

The study ran for 10 instructional weeks (Fig. 2). Week 1 included prior knowledge, baseline scientific argumentation, background, prior GenAI experience, and baseline feedback literacy measures. Week 2 provided common orientation to scientific argumentation, the analytic rubric, and feedback principles; students in AI-supported conditions also received GenAI-use training. This additional training was necessary for procedural consistency but may have increased preparation time relative to the peer-feedback condition.

Weeks 3–8 comprised three argumentation cycles, each involving draft production, the assigned feedback condition, revision, and final submission. Week 9 included the conceptual-learning post-test and a novel AI-free transfer task. Week 10 included delayed AI-free transfer, the post-intervention questionnaire, and semi-structured interviews with a subsample.

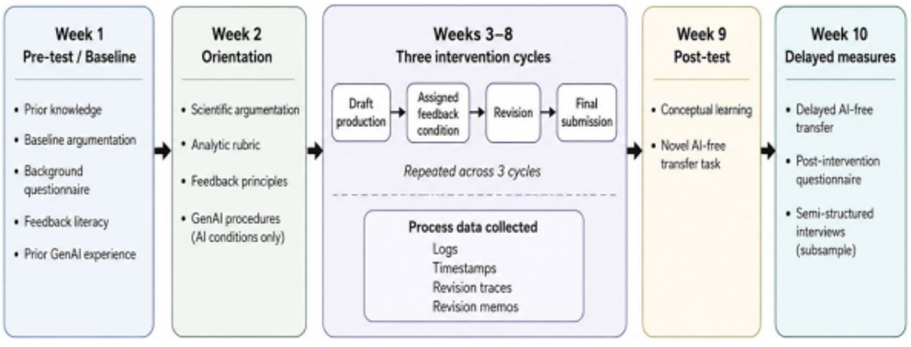


**Fig. 2** Study timeline and data collection

## Measures

A concise overview of all measures, data sources, timing, and scoring procedures is provided in Supplementary Table S7.

### *Four-dimension argument-quality gain*

Four-dimension argument-quality gain was calculated as the revised-draft common-content subtotal minus the initial-draft common-content subtotal and then averaged across the three intervention cycles. The common-content subtotal comprised the four rubric dimensions scored at both draft stages: claim quality, evidence relevance and sufficiency, coherence of reasoning, and treatment of limitations or alternative explanations. Each dimension was scored from 1 to 4, yielding comparable subtotals ranging from 4 to 16.

Strategic quality of revision was retained as a separate supplementary revised-draft indicator because it was scored only for revised submissions and required side-by-side comparison with the initial draft. Revision depth was coded separately as a supplementary process indicator on a three-point scale, from predominantly surface-level to predominantly substantive, to support interpretation of whether observed gains reflected substantive rather than primarily surface-level revision. Detailed rubric descriptors and revision-depth coding guidance are provided in Supplementary Appendix A.

### *Conceptual learning*

Conceptual learning was measured using discipline-specific pre/post assessments co-developed by instructors across the four institutions. Each assessment included 12–15 selected-response and short constructed-response items targeting the core concepts underlying the argumentation tasks. Scores were standardized within discipline before pooled modeling. Internal consistency at post-test was acceptable across disciplines, McDonald's $\omega = .79$ in biology, .82 in chemistry, and .80 in physics.

### *Delayed AI-free transfer*

Delayed AI-free transfer was assessed through a novel scientific argumentation task completed individually in a supervised setting without access to the study's GenAI tool or internet-enabled devices. The task differed in topic from the intervention activities and the Week 9 transfer task but required the same four common content dimensions. Because the delayed transfer task was a single independent performance rather than a paired draft sequence, strategic quality of revision was not applicable.

### *Feedback uptake*

Feedback uptake was treated as a four-indicator construct capturing the extent to which students attended to, interpreted, selected, and incorporated critique during revision. The indicators were students' post-revision rationales, coder-rated feedback–revision alignment, log-based feedback review time, and the proportion of substantive revisions traceable to feedback content.

In the hybrid condition, the 100–150 word revision memo served as the post-revision rationale. In the peer feedback, direct GenAI-supported feedback, and reflective GenAI-supported feedback conditions, students, submitted a shorter structured rationale with the final draft. This rationale asked students to identify the main revision they made, the

feedback point or self-evaluation judgment that informed it, and one suggestion they accepted, modified, or rejected.

Feedback–revision alignment was coded by comparing major implemented revisions with the feedback record available to the student in that cycle. A revision was coded as aligned when it addressed a specific issue raised in peer comments, GenAI-generated feedback, structured self-evaluation, or the revision memo, depending on condition.

The traceable-substantive-revision proportion was calculated as the number of substantive revisions linked to at least one received feedback point divided by the total number of substantive revisions identified in the draft comparison. Substantive revisions included changes to the claim, evidence selection or integration, explanatory reasoning, or treatment of uncertainty and alternative explanations; surface-level wording, grammar, punctuation, and formatting edits were not counted as substantive. Cycles with no substantive revision were coded as 0 for this indicator.

Feedback review time was derived from platform logs and represented the time students spent viewing assigned feedback resources before submitting the revised draft. Extreme log-duration values caused by inactive browser sessions were winsorized at the 99th percentile before standardization.

Reliability for the manually coded uptake indicators was assessed through independent double-coding of a stratified random 20% of coded revision episodes. The agreement coefficients are reported in Sect. "Preliminary analyses", and the full coding protocol is provided in Supplementary Appendix E. Each uptake indicator was averaged across cycles and z-standardized. A one-factor measurement model showed standardized loadings from .58 to .81, and the final uptake index was computed as the equal-weight mean of the four z-standardized indicators.

#### *Feedback literacy*

Baseline feedback literacy was assessed using the 22-item Student Feedback Literacy Instrument (SFLI), a higher-education measure of feedback attitudes and feedback practices (Weidlich et al., 2025; Woitt et al., 2025). Students responded on a 5-point Likert scale, and scores were calculated as the mean across items. Internal consistency in the present sample was high, McDonald's $\omega = .90$.

#### *Self-regulated learning during revision*

Self-regulated learning during revision was measured using a 12-item task-specific scale adapted from the higher-education self-regulated learning tradition represented by the Motivated Strategies for Learning Questionnaire and subsequent SRL research emphasizing planning, monitoring, and strategic adaptation (Panadero, 2017; Pintrich et al., 1991). The scale sampled planning, monitoring, strategy adjustment, and evaluation, with three items per domain. Internal consistency of the self-report component was satisfactory, McDonald's $\omega = .88$.

To complement self-report data, the study incorporated learning-management-system trace indicators of feedback revisiting, spacing of revision activity, repeated rubric consultation, and sequencing of draft-comparison actions. The final SRL index was represented by the mean of the z-standardized self-report score and the z-standardized trace-based subindex.

#### *Covariates*

The models adjusted for prior knowledge, baseline scientific argumentation, prior achievement, prior GenAI experience, gender, and science domain. Prior achievement was operationalized as each student's standardized university-entry score; because entry metrics differed across institutions, scores were standardized within institution before pooled modeling. Institutional variation was handled through institution fixed effects, and exploratory moderation analyses additionally examined whether baseline feedback literacy conditioned the effects of the feedback designs.

### Fidelity of implementation

Implementation fidelity was monitored through four procedures: use of a common intervention schedule, rubric, and instructional materials across sites; standardized prompt architecture and fixed system settings in the AI-supported conditions; platform logs verifying that students accessed only the feedback resources assigned to their condition; and manual audits of a random 15% sample of student cases across sites. Fidelity was assessed using a predefined checklist covering sequence integrity, timing compliance, access restrictions, and completion of condition-specific activities. Audits were conducted by trained research assistants who were not involved in instruction, and 20% of audited cases were double-coded, with high agreement on checklist items (Cohen's $\kappa = .89$). Detailed audit criteria and procedures are reported in the supplementary materials.

### Data analysis

Because students were nested within course sections and only four institutions participated, the primary quantitative analyses were estimated using mixed-effects models with students at Level 1, sections at Level 2, and institution entered as a fixed effect. This approach was preferred to specifying institutions as a random third level because the number of higher-level units was too small to support stable estimation of institution-level random effects. Analyses were conducted in R 4.3.2.

For RQ1 and RQ2, separate mixed-effects models were estimated for four-dimension argument-quality gain, conceptual learning, and delayed AI-free transfer. Experimental condition was treated as the principal fixed effect, section was specified as a random intercept, and the models adjusted for prior knowledge, baseline scientific argumentation, prior achievement, baseline feedback literacy, prior GenAI experience, gender, science domain, and institution fixed effects. Planned pairwise contrasts were estimated from the fitted models using estimated marginal means and Holm-adjusted $p$ values. For outcomes collected across the intervention cycles, student-level means were computed across the three cycles before modeling; conceptual learning and delayed AI-free transfer were modeled at their respective post-test and delayed assessment points. The primary immediate outcome was modeled as a gain score, consistent with both the instructional design and the preregistered outcome definition, because the intervention was designed to examine revision-related improvement from initial draft to revised submission.

Because gain scores can be constrained for students whose initial drafts are already high quality, an additional baseline-adjusted final-score sensitivity model was estimated. In this model, the dependent variable was the revised-draft four-dimension common-content score averaged across the three intervention cycles, and the corresponding

initial-draft four-dimension common-content score was entered as a covariate. The model otherwise followed the same fixed- and random-effects structure as the primary outcome model, with feedback condition as the principal fixed effect, section as a random intercept, and adjustment for the same student-level covariates, science-domain indicators, and institution fixed effects. This supplementary analysis was used to examine whether the primary gain-score pattern was robust to an alternative specification less directly affected by ceiling-related limits on observable improvement.

For RQ3, between-condition differences in feedback uptake and self-regulated learning during revision were examined using the same fixed- and random-effects structure. Both process variables were averaged across the three intervention cycles before modeling, after their component indicators had been standardized and combined as described above.

For RQ4, 2–1-1 multilevel mediation models were estimated with feedback condition at the section level and both mediators and outcomes at the student level. The mediation analyses focused on the theoretically central contrasts comparing reflective GenAI-supported feedback and hybrid feedback with direct GenAI-supported feedback. Indirect effects were estimated separately through feedback uptake, through self-regulated learning during revision, and through the serial pathway from feedback uptake to self-regulated learning.

For mediation analyses, feedback uptake and self-regulated learning were treated as process indicators aggregated across the three revision cycles, whereas four-dimension argument-quality gain represented the corresponding average improvement across those same cycles. Conceptual learning and delayed AI-free transfer were measured after the intervention period, but the mediators reflected students' cumulative revision processes during the intervention. These models were therefore intended to test whether condition differences in outcomes were statistically associated with condition differences in revision processes, rather than to establish a strict within-cycle temporal causal sequence. The serial ordering from feedback uptake to self-regulated learning was theory-driven, reflecting the assumption that attending to, interpreting, and selecting feedback can support subsequent planning, monitoring, and strategic adjustment during revision.

All primary analyses followed an intention-to-treat approach. Missing data were handled using multiple imputation by chained equations with 20 imputations under a missing-at-random assumption. For mediation analyses, indirect effects were estimated using Monte Carlo confidence intervals based on 20,000 simulations. To assess robustness, the primary models were re-estimated using per-protocol analyses, models excluding low-fidelity sections, an alternative missing-data specification based on full-information maximum likelihood, and the supplementary baseline-adjusted final-score model described above. Additional analytic details and robustness results are reported in the supplementary materials.

#### *Qualitative analysis*

To complement the quantitative analyses, semi-structured interviews were conducted with a purposive subsample of 32 students and 12 instructors. The student subsample was balanced across conditions and approximately balanced across disciplines, while instructor interviews captured site-level variation in feasibility and implementation. Interviews focused on how participants interpreted feedback, made revision decisions,

and experienced ownership in AI-supported revision. All interviews were audio-recorded, transcribed verbatim, and analyzed using reflexive thematic analysis following Braun and Clarke (2006).

The qualitative component was interpretive rather than reliability-seeking. Themes were developed iteratively through memoing and team discussion, with reflexive attention to how researchers' assumptions informed coding. No inter-coder reliability coefficient was calculated, in keeping with the reflexive thematic analysis approach. The qualitative findings were used in an explanatory role to clarify how participants experienced the feedback designs and to interpret the mechanisms underlying the quantitative patterns.

#### *Ethical considerations*

Ethical approval was obtained from the institutional review boards of the participating universities before data collection. Participation in the research component was voluntary, and students were informed that non-participation would not affect course grades. All data were de-identified prior to analysis, and identifiable platform data were stored on institutionally approved encrypted servers with access restricted to the research team in accordance with local ethics procedures.

Participants were informed of their right to withdraw their research data up to the point at which the dataset had been de-identified and merged for analysis. Because the intervention involved GenAI tools, students received explicit guidance on ethical AI use, data-entry limits, and the requirement that, submitted work remain their own. Monitoring of AI use was limited to interactions within the study platform. To reduce the risk of unequal access across conditions, all groups received the same core instruction in scientific argumentation, rubric use, and feedback principles, and students in the non-AI condition were provided with access to the GenAI orientation materials and a demonstration version of the interface after the intervention.

## Results

### Preliminary analyses

The final analytic sample comprised 1,176 students nested within 48 course sections across four institutions. Of the 1,248 students initially enrolled, 72 were excluded because of non-consent for research use of their data, course withdrawal before the first intervention cycle, or absence from both the post-test and delayed transfer sessions. Final sample sizes were balanced across conditions, and attrition did not differ significantly by condition, $\chi 2(3) = 0.33$, $p = .954$. Missing data on the retained sample ranged from 0.6 to 3.4% across variables and were handled using multiple imputation with 20 imputations.

Baseline equivalence analyses indicated that the four conditions did not differ significantly on age, gender distribution, prior knowledge, baseline scientific argumentation, prior achievement, feedback literacy, or prior GenAI experience, all $p \geq .236$, suggesting that the randomization procedure produced broadly comparable groups prior to the intervention. Descriptive statistics and baseline-equivalence tests are reported in Table 2.

The internal consistency of the self-report measures was satisfactory, and inter-rater reliability for scientific argumentation scoring and revision-depth coding was high.

**Table 2** Sample characteristics and baseline equivalence across conditions

| Variable | Peer feedback only (n = 293) | Direct GenAI-supported feedback (n = 295) | Reflective GenAI-supported feedback (n = 294) | Hybrid feedback (n = 294) | Total (N = 1,176) | Test statistic | p |
|---|---|---|---|---|---|---|---|
| Age, years, M (SD) | 18.92 (1.31) | 18.88 (1.27) | 18.95 (1.29) | 18.90 (1.34) | 18.91 (1.30) | F(3, 1172) = 0.14 | .936 |
| Women, n (%) | 167 (57.0) | 171 (58.0) | 166 (56.5) | 169 (57.5) | 673 (57.2) | $\chi^2(3) = 0.18$ | .981 |
| Prior knowledge (0–100), M (SD) | 58.63 (10.84) | 59.11 (10.52) | 60.02 (10.41) | 59.38 (10.76) | 59.29 (10.63) | F(3, 1172) = 0.84 | .472 |
| Baseline scientific argumentation (4–16), M (SD) | 10.32 (2.08) | 10.51 (2.03) | 10.64 (2.10) | 10.57 (2.07) | 10.51 (2.07) | F(3, 1172) = 1.10 | .349 |
| Prior achievement, M (SD) | -0.05 (0.98) | 0.03 (1.01) | 0.04 (0.96) | -0.01 (1.00) | 0.00 (0.99) | F(3, 1172) = 0.71 | .548 |
| Feedback literacy (1–5), M (SD) | 3.41 (0.52) | 3.44 (0.50) | 3.47 (0.49) | 3.45 (0.51) | 3.44 (0.51) | F(3, 1172) = 0.67 | .571 |
| Prior GenAI experience (1–5), M (SD) | 2.78 (0.89) | 2.93 (0.92) | 2.88 (0.90) | 2.99 (0.91) | 2.89 (0.91) | F(3, 1172) = 1.42 | .236 |

Baseline-equivalence tests are based on the final analytic sample. Prior achievement scores were standardized within institution before pooled comparison. No significant between-condition differences were observed at baseline. Baseline scientific argumentation scores reflect the four common content dimensions of the analytic framework and therefore ranged from 4 to 16 rather than including the revised-draft strategic-quality-of-revision indicator

Implementation fidelity was also high across sites: based on instructor logs, platform records, and checklist-based audits of 176 randomly selected student cases, 95.8% of applicable instructional steps were delivered as intended, with high agreement on fidelity checklist items, Cohen's $\kappa = .89$.

Reliability was also assessed for the manually coded components of the feedback uptake index. A stratified random 20% of coded revision episodes was independently double-coded. Agreement was substantial for post-revision rationale specificity, weighted Cohen's $\kappa = .81$, feedback–revision alignment, Cohen's $\kappa = .84$, and traceable substantive revisions, Cohen's $\kappa = .82$. Disagreements were resolved through discussion before final indicator computation.

Cross-condition contamination was minimal. Taken together, these findings support the adequacy of the dataset and the integrity of the intervention. Section-level variance components and intraclass correlation coefficients are reported in Supplementary Table S1.

#### Effects of feedback condition on four-dimension argument-quality gain

To address RQ1, a linear mixed-effects model was estimated using students' mean four-dimension argument-quality gain across the three intervention cycles as the dependent variable, with section included as a random intercept and institution entered as a fixed effect. Feedback condition significantly predicted four-dimension argument-quality gain, $F(3, 43.8) = 18.76$, $p < .001$.

As shown in Fig. 3 and Table 3, direct GenAI-supported feedback outperformed peer feedback, and both reflective and hybrid feedback outperformed direct GenAI-supported feedback. The hybrid condition yielded the highest adjusted mean, but the hybrid–reflective contrast was not statistically significant. Supplementary revision-depth analyses showed a parallel directional pattern, indicating that stronger argument-quality

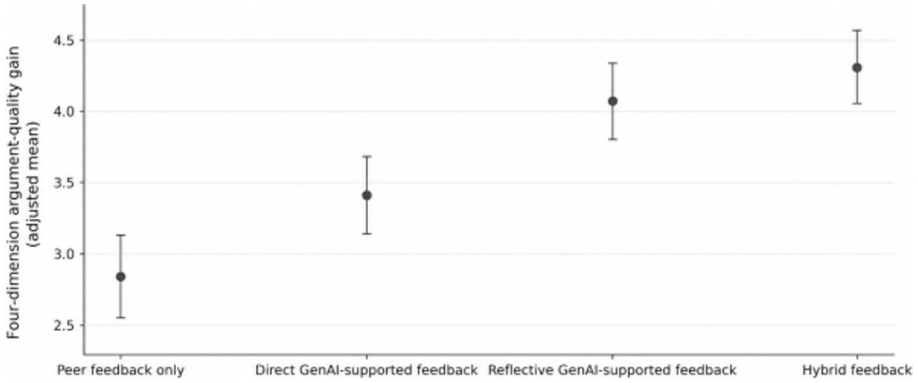


**Fig. 3** Estimated marginal means for four-dimension argument-quality gain across conditions. Error bars represent 95% confidence intervals

**Table 3** Multilevel model predicting four-dimension argument-quality gain

| Condition | | Adjusted M | | SE | 95% CI |
|---|---|---|---|---|---|
| *Panel A. Estimated marginal means by condition* | | | | | |
| Peer feedback only | | 2.84 | | 0.15 | [2.55, 3.13] |
| Direct GenAI-supported feedback | | 3.41 | | 0.14 | [3.14, 3.68] |
| Reflective GenAI-supported feedback | | 4.07 | | 0.14 | [3.80, 4.34] |
| Hybrid feedback | | 4.31 | | 0.13 | [4.05, 4.57] |
| **Contrast** | **b** | **SE** | **p** | **Cohen's d** | **95% CI for b** |
| *Panel B. Pairwise contrasts* | | | | | |
| Direct GenAI – Peer | 0.57 | 0.18 | .002 | 0.25 | [0.22, 0.92] |
| Reflective GenAI – Peer | 1.23 | 0.18 | < .001 | 0.54 | [0.88, 1.58] |
| Hybrid – Peer | 1.47 | 0.18 | < .001 | 0.64 | [1.12, 1.82] |
| Reflective GenAI – Direct GenAI | 0.66 | 0.17 | < .001 | 0.29 | [0.33, 0.99] |
| Hybrid – Direct GenAI | 0.90 | 0.17 | < .001 | 0.40 | [0.57, 1.23] |
| Hybrid – Reflective GenAI | 0.24 | 0.16 | .138 | 0.11 | [-0.08, 0.56] |
| **Effect** | | **Test statistic** | | | **p** |
| *Panel C. Overall model summary* | | | | | |
| Feedback condition | | F(3, 43.8) = 18.76 | | | < .001 |

The model was adjusted for prior knowledge, baseline scientific argumentation, prior achievement, feedback literacy, prior GenAI experience, gender, science domain, and institution fixed effects, with section included as a random intercept. Four-dimension argument-quality gain scores were computed as change scores from initial draft to revised submission on the four common content dimensions shared across draft stages and were averaged across the three intervention cycles. Higher scores indicate greater improvement in common-content scientific argument quality from draft to revision. The separate strategic-quality-of-revision indicator was not included in pre/post gain estimation. Holm-adjusted p values are reported for pairwise contrasts.

gains were accompanied by more substantive rather than primarily surface-level revision. Detailed revision-depth estimates are reported in Supplementary Table S2.

These findings support H4: direct GenAI-supported feedback improved immediate four-dimension argument-quality gain relative to peer feedback, while reflective and hybrid feedback produced stronger gains than direct GenAI-supported feedback.

### Effects on conceptual learning and delayed AI-free transfer

To address RQ2, separate linear mixed-effects models were estimated for conceptual learning and delayed AI-free transfer, with section included as a random intercept and institution entered as a fixed effect. Feedback condition significantly predicted both conceptual learning, $F(3, 43.5) = 7.62$, $p < .001$, and delayed AI-free transfer, $F(3, 43.9) = 12.03$, $p < .001$.

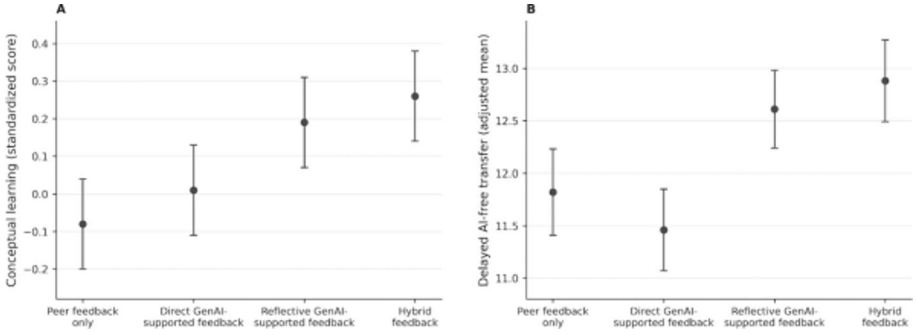


**Fig. 4** Estimated marginal means for conceptual learning and delayed AI-free transfer

**Table 4** Mixed-effects models predicting conceptual learning and delayed AI-free transfer

| Condition | | Adjusted M | | SE | 95% CI |
|---|---|---|---|---|---|
| *Panel A. Conceptual learning* | | | | | |
| Peer feedback only | | − 0.08 | | 0.06 | [− 0.20, 0.04] |
| Direct GenAI-supported feedback | | 0.01 | | 0.06 | [− 0.11, 0.13] |
| Reflective GenAI-supported feedback | | 0.19 | | 0.06 | [0.07, 0.31] |
| Hybrid feedback | | 0.26 | | 0.06 | [0.14, 0.38] |
| **Contrast** | **b** | **SE** | **p** | **Cohen's d** | **95% CI for b** |
| Direct GenAI – Peer | 0.09 | 0.09 | .314 | 0.08 | [-0.09, 0.27] |
| Reflective GenAI – Peer | 0.27 | 0.10 | .006 | 0.25 | [0.08, 0.46] |
| Hybrid – Peer | 0.34 | 0.09 | < .001 | 0.32 | [0.16, 0.52] |
| Reflective GenAI – Direct GenAI | 0.18 | 0.10 | .072 | 0.17 | [-0.02, 0.38] |
| Hybrid – Direct GenAI | 0.25 | 0.10 | .014 | 0.23 | [0.05, 0.45] |
| Hybrid – Reflective GenAI | 0.07 | 0.09 | .438 | 0.07 | [-0.11, 0.25] |
| **Effect** | | **Test statistic** | | | **p** |
| Feedback condition | | $F(3, 43.5) = 7.62$ | | | < .001 |
| **Condition** | | **Adjusted M** | | **SE** | **95% CI** |
| *Panel B. Delayed AI-free transfer* | | | | | |
| Peer feedback only | | 11.82 | | 0.21 | [11.41, 12.23] |
| Direct GenAI-supported feedback | | 11.46 | | 0.20 | [11.07, 11.85] |
| Reflective GenAI-supported feedback | | 12.61 | | 0.19 | [12.24, 12.98] |
| Hybrid feedback | | 12.88 | | 0.20 | [12.49, 13.27] |
| **Contrast** | **b** | **SE** | **p** | **Cohen's d** | **95% CI for b** |
| Direct GenAI – Peer | − 0.36 | 0.25 | .149 | − 0.12 | [− 0.85, 0.13] |
| Reflective GenAI – Peer | 0.79 | 0.26 | .003 | 0.27 | [0.28, 1.30] |
| Hybrid – Peer | 1.06 | 0.25 | < .001 | 0.36 | [0.57, 1.55] |
| Reflective GenAI – Direct GenAI | 1.15 | 0.26 | < .001 | 0.39 | [0.64, 1.66] |
| Hybrid – Direct GenAI | 1.42 | 0.27 | < .001 | 0.48 | [0.89, 1.95] |
| Hybrid – Reflective GenAI | 0.27 | 0.28 | .327 | 0.09 | [-0.28, 0.82] |
| **Effect** | | **Test statistic** | | | **p** |
| Feedback condition | | $F(3, 43.9) = 12.03$ | | | < .001 |

Models were adjusted for prior knowledge, baseline scientific argumentation, prior achievement, feedback literacy, prior GenAI experience, gender, science domain, and institution fixed effects, with section included as a random intercept. Conceptual learning scores were standardized within discipline before pooled analysis. Higher delayed-transfer scores indicate stronger AI-independent scientific argumentation on the novel task. Holm-adjusted p values are reported for pairwise contrasts.

As shown in Fig. 4 and Table 4, the conceptual-learning pattern favored the more agentic feedback designs. Hybrid feedback significantly outperformed direct GenAI-supported feedback, whereas the reflective–direct contrast was positive but did not reach statistical significance after adjustment. Reflective and hybrid feedback did not differ significantly from each other.

For delayed AI-free transfer, both reflective and hybrid feedback significantly outperformed direct GenAI-supported feedback, while the hybrid–reflective contrast was not statistically significant. These findings provide partial support for H3 for conceptual learning and stronger support for H3 for delayed AI-free transfer. Detailed estimates and pairwise contrasts are reported in Table 4.

### Effects on feedback uptake and self-regulated learning

To address RQ3, separate linear mixed-effects models were estimated for feedback uptake and self-regulated learning during revision, with section included as a random intercept and institution entered as a fixed effect. Feedback condition significantly predicted both feedback uptake, $F(3, 43.6) = 16.28$, $p < .001$, and self-regulated learning, $F(3, 43.2) = 11.91$, $p < .001$.

As shown in Table 5, reflective and hybrid feedback yielded higher feedback uptake and self-regulated learning than direct GenAI-supported feedback, whereas reflective and hybrid feedback did not differ significantly from each other on either process

**Table 5** Mixed-effects models predicting feedback uptake and self-regulated learning during revision

| Condition | Adjusted M | SE | 95% CI |
|---|---|---|---|
| *Panel A. Feedback uptake* | | | |
| Peer feedback only | − 0.06 | 0.05 | [− 0.16, 0.04] |
| Direct GenAI-supported feedback | − 0.02 | 0.05 | [− 0.12, 0.08] |
| Reflective GenAI-supported feedback | 0.24 | 0.05 | [0.14, 0.34] |
| Hybrid feedback | 0.31 | 0.05 | [0.21, 0.41] |
| **Contrast** | **b** | **SE** | **p** |
| Direct GenAI – Peer | 0.04 | 0.07 | .563 |
| Reflective GenAI – Peer | 0.30 | 0.07 | < .001 |
| Hybrid – Peer | 0.37 | 0.07 | < .001 |
| Reflective GenAI – Direct GenAI | 0.26 | 0.07 | < .001 |
| Hybrid – Direct GenAI | 0.33 | 0.07 | < .001 |
| Hybrid – Reflective GenAI | 0.07 | 0.07 | .314 |
| **Effect** | **Test statistic** | | **p** |
| Feedback condition | F(3, 43.6) = 16.28 | | < .001 |
| **Condition** | **Adjusted M** | **SE** | **95% CI** |
| *Panel B. Self-regulated learning during revision* | | | |
| Peer feedback only | − 0.03 | 0.05 | [− 0.13, 0.07] |
| Direct GenAI-supported feedback | − 0.09 | 0.05 | [− 0.19, 0.01] |
| Reflective GenAI-supported feedback | 0.17 | 0.05 | [0.07, 0.27] |
| Hybrid feedback | 0.26 | 0.05 | [0.16, 0.36] |
| **Contrast** | **b** | **SE** | **p** |
| Direct GenAI – Peer | − 0.06 | 0.07 | .401 |
| Reflective GenAI – Peer | 0.20 | 0.08 | .015 |
| Hybrid – Peer | 0.29 | 0.08 | < .001 |
| Reflective GenAI – Direct GenAI | 0.26 | 0.08 | .002 |
| Hybrid – Direct GenAI | 0.35 | 0.08 | < .001 |
| Hybrid – Reflective GenAI | 0.09 | 0.08 | .226 |
| **Effect** | **Test statistic** | | **p** |
| Feedback condition | F(3, 43.2) = 11.91 | | < .001 |

Both process variables were standardized composite indices. Models were adjusted for prior knowledge, baseline scientific argumentation, prior achievement, feedback literacy, prior GenAI experience, gender, science domain, and institution fixed effects, with section included as a random intercept. Higher scores indicate stronger engagement with feedback and greater self-regulation during revision. Holm-adjusted p values are reported for pairwise contrasts.

variable. These findings support H1a and H2a and are consistent with the nondirectional expectations in H1b and H2b. Full estimates are presented in Table 5.

### Mediation analyses

To address RQ4, 2–1-1 multilevel mediation models were estimated to examine whether feedback uptake and self-regulated learning during revision mediated the relationship between feedback condition and four-dimension argument-quality gain, conceptual learning, and delayed AI-free transfer. Because the theoretically focal contrasts concerned the more agentic feedback designs, the mediation analyses compared the reflective and hybrid conditions primarily against direct GenAI-supported feedback.

Across outcomes, mediation analyses indicated that reflective and hybrid feedback were associated with stronger feedback uptake and self-regulated learning relative to direct GenAI-supported feedback. For four-dimension argument-quality gain and delayed AI-free transfer, indirect effects were significant through feedback uptake, through self-regulated learning, and through the serial uptake → self-regulated learning pathway. For conceptual learning, indirect effects were strongest through self-regulated learning and the serial pathway, whereas the uptake-only pathway did not reach significance. These findings support H5a with qualification and support H5b and H5c. Detailed direct, indirect, and total effects are reported in Table 6.

Table 7 provides a concise summary of the empirical support for each hypothesis across the outcome, process, and mediation analyses.

### Qualitative findings on feedback use and epistemic agency

The qualitative findings were used to clarify how students and instructors experienced the feedback designs and to help interpret the quantitative patterns. Four themes were

**Table 6** Direct, indirect, and total effects of feedback condition on outcomes

| Contrast | Outcome | Direct effect, β [95% CI] | Total indirect effect, β [95% CI] | Via feedback uptake, β [95% CI] | Via SRL, β [95% CI] | Via uptake → SRL, β [95% CI] | Total effect, β [95% CI] |
|---|---|---|---|---|---|---|---|
| Reflective GenAI vs. Direct GenAI | Four-dimension argument-quality gain | 0.36 [0.11, 0.61] | 0.30 [0.18, 0.44] | 0.18 [0.08, 0.31] | 0.07 [0.02, 0.14] | 0.05 [0.01, 0.10] | 0.66 [0.41, 0.91] |
| Hybrid vs. Direct GenAI | Four-dimension argument-quality gain | 0.52 [0.27, 0.77] | 0.38 [0.25, 0.52] | 0.22 [0.11, 0.35] | 0.09 [0.03, 0.16] | 0.07 [0.02, 0.13] | 0.90 [0.65, 1.15] |
| Reflective GenAI vs. Direct GenAI | Conceptual learning | 0.05 [– 0.04, 0.14] | 0.13 [0.06, 0.21] | 0.03 [– 0.01, 0.08] | 0.06 [0.02, 0.12] | 0.04 [0.01, 0.08] | 0.18 [0.08, 0.28] |
| Hybrid vs. Direct GenAI | Conceptual learning | 0.08 [– 0.02, 0.18] | 0.17 [0.09, 0.25] | 0.04 [– 0.00, 0.09] | 0.08 [0.03, 0.14] | 0.05 [0.02, 0.10] | 0.25 [0.15, 0.35] |
| Reflective GenAI vs. Direct GenAI | Delayed AI-free transfer | 0.65 [0.30, 1.00] | 0.50 [0.31, 0.72] | 0.21 [0.09, 0.36] | 0.18 [0.07, 0.32] | 0.11 [0.04, 0.19] | 1.15 [0.80, 1.50] |
| Hybrid vs. Direct GenAI | Delayed AI-free transfer | 0.80 [0.44, 1.16] | 0.62 [0.42, 0.84] | 0.26 [0.12, 0.41] | 0.22 [0.10, 0.36] | 0.14 [0.06, 0.23] | 1.42 [1.06, 1.78] |

Standardized coefficients are reported for all focal contrasts. Indirect-effect confidence intervals were derived from Monte Carlo simulation; direct- and total-effect confidence intervals were derived from model-based standard errors. Confidence intervals that do not cross zero indicate statistically reliable effects. For the revision outcome, mediation models used four-dimension argument-quality gain rather than the separate revised-draft strategic-quality-of-revision indicator, because gain estimation required directly comparable scoring across draft stages

**Table 7** Summary of hypotheses and empirical support

| Hypothesis | Related outcome or process | Main finding | Interpretation |
|---|---|---|---|
| H1a | Feedback uptake | Reflective and hybrid feedback yielded higher feedback uptake than direct GenAI-supported feedback | Supported |
| H1b | Reflective vs hybrid feedback uptake | No significant difference emerged between reflective and hybrid feedback on feedback uptake | Consistent with nondirectional expectation |
| H2a | Self-regulated learning during revision | Reflective and hybrid feedback yielded higher self-regulated learning than direct GenAI-supported feedback | Supported |
| H2b | Reflective vs hybrid self-regulated learning | No significant difference emerged between reflective and hybrid feedback on self-regulated learning | Consistent with nondirectional expectation |
| H3 | Conceptual learning and delayed AI-free transfer | Hybrid feedback outperformed direct GenAI-supported feedback on conceptual learning; reflective feedback showed a positive but non-significant contrast. Both reflective and hybrid feedback outperformed direct GenAI-supported feedback on delayed AI-free transfer | Partially supported for conceptual learning and more strongly supported for delayed AI-free transfer |
| H4 | Four-dimension argument-quality gain | Direct GenAI-supported feedback outperformed peer feedback; reflective and hybrid feedback outperformed direct GenAI-supported feedback | Supported |
| H5a | Mediation through feedback uptake | Feedback uptake mediated effects on argument-quality gain and delayed AI-free transfer; the uptake-only pathway did not reach significance for conceptual learning | Supported with qualification |
| H5b | Mediation through self-regulated learning | Self-regulated learning mediated the effects of reflective and hybrid feedback on the focal outcomes | Supported |
| H5c | Serial mediation through feedback uptake → self-regulated learning | The serial pathway from feedback uptake to self-regulated learning was significant across the focal contrasts | Supported |

identified. First, students described GenAI feedback as fast and useful but sometimes over-directive, especially in the direct GenAI-supported condition. Second, students in the reflective condition reported that prior self-evaluation helped them treat AI feedback as a comparison point rather than an answer key. Third, students in the hybrid condition described self-, peer-, and AI-based feedback as complementary perspectives that supported more deliberate revision choices. Fourth, students and instructors noted a tension between efficiency and ownership: AI feedback reduced the effort required to generate revision ideas but could also weaken students' sense of authorship. Together, these themes support the interpretation that reflective and hybrid designs promoted stronger uptake and self-regulated learning by requiring comparison, selection, and justification rather than simple implementation of external critique. Representative quotations are provided in Supplementary Table S8.

#### Exploratory moderation by baseline feedback literacy

Exploratory moderation analyses examined whether feedback-condition effects varied by baseline feedback literacy and science domain. These analyses were not preregistered confirmatory tests and should therefore be interpreted cautiously.

The only statistically significant interaction was condition × baseline feedback literacy for delayed AI-free transfer, $F(3, 1158) = 2.83$, $p = .037$. The disadvantage of direct GenAI-supported feedback relative to reflective and hybrid feedback was more pronounced among students with lower initial feedback literacy, suggesting that structured self-evaluation and comparison may be especially useful for learners with weaker feedback-literacy resources. Condition × science domain interactions were not significant for any primary or secondary outcome, all $p > .10$. Detailed moderation results are reported in the Supplementary Materials.

#### Sensitivity analyses

Sensitivity analyses supported the robustness of the main findings. The pattern of results remained substantively unchanged in per-protocol analyses excluding students who completed fewer than two intervention cycles ($n = 1{,}102$), models excluding the three lowest-fidelity sections, and alternative missing-data specifications using full-information maximum likelihood.

To address possible ceiling-related constraints in gain scores, a baseline-adjusted final-score model used revised-draft four-dimension common-content score as the outcome and the corresponding initial-draft score as a covariate. Initial-draft score strongly predicted revised-draft performance, $b = 0.72$, SE = 0.04, 95% CI [0.64, 0.80], $p < .001$, and feedback condition remained significant, $F(3, 43.7) = 17.21$, $p < .001$. The condition pattern was consistent with the primary gain-score model: direct GenAI-supported feedback exceeded peer feedback, reflective and hybrid feedback exceeded direct GenAI-supported feedback, and the hybrid–reflective contrast remained non-significant. These results indicate that the primary findings were not driven solely by ceiling-related limits on possible gain. Full estimates are reported in Supplementary Table S6.

## Discussion

### Interpretation of the principal findings

The present study examined how alternative GenAI-supported feedback designs shaped revision, learning, and transfer in introductory university science courses. Direct GenAI-supported feedback improved immediate four-dimension argument-quality gain relative to peer feedback, whereas reflective and hybrid feedback designs produced stronger feedback uptake, self-regulated learning, and delayed AI-free transfer. Conceptual-learning benefits were clearest for the hybrid condition, while the reflective condition showed a positive but non-significant adjusted contrast relative to direct GenAI-supported feedback. The hybrid condition yielded the highest adjusted mean for four-dimension argument-quality gain, whereas both reflective and hybrid designs outperformed direct GenAI-supported feedback on delayed AI-free transfer. These findings suggest that GenAI can support rapid improvement in common-content argument quality, but more agentic feedback configurations appear more useful when the goal is durable learning and AI-independent transfer.

This pattern differentiates short-term draft improvement from broader educational value. Direct GenAI-supported feedback appears useful when students need rapid, criterion-referenced critique to strengthen a draft in the moment. However, the stronger outcomes associated with reflective and hybrid designs suggest that conceptual learning and transfer are more likely when feedback processes require students to engage in their own

evaluative work. Supplementary indicators support this interpretation: the more agentic conditions were associated not only with larger four-dimension argument-quality gains, but also with more substantive revision and stronger revised-draft strategic quality. This suggests that the advantages of reflective and hybrid feedback were linked to purposeful revision rather than surface-level editing alone.

At the same time, the observed advantages of reflective and hybrid designs should not be interpreted as evidence that feedback sequencing alone caused the differences. These conditions also required additional self-evaluation, comparison, and, in the hybrid condition, revision justification. The benefits may therefore reflect the combined effect of feedback source, sequencing, time-on-task, reflective workload, and structured cognitive engagement. This interpretation is consistent with the study's design logic: the intervention tested feedback designs as pedagogical configurations, not isolated feedback sources.

The absence of statistically significant differences between reflective and hybrid conditions on several outcomes should also be interpreted cautiously. Rather than demonstrating equivalence, the pattern suggests that requiring learners to externalize an initial evaluative judgment before receiving AI critique may be a particularly important ingredient of effective GenAI-supported feedback. The additional components of the hybrid condition may still offer pedagogical benefits, but these advantages were not always statistically separable from reflective GenAI-supported feedback alone in the present sample.

#### Theoretical and empirical implications

These findings support the claim that feedback design is a more meaningful analytic unit than feedback source alone. Direct GenAI-supported feedback provided rapid, criterion-referenced critique, but it did so without requiring learners to externalize an initial judgment about the quality of their own work. By contrast, reflective and hybrid designs required learners to compare, interpret, and justify revision decisions before or alongside engagement with AI critique. This pattern is consistent with process-oriented accounts of feedback, which emphasize that learning depends not merely on receiving comments, but on learners' active interpretation and use of them (Carless & Boud, 2018; Nicol & Macfarlane-Dick, 2006).

The findings also extend work on self-regulated learning in AI-mediated environments. The mediation analyses suggest that reflective and hybrid designs outperformed direct GenAI-supported feedback partly because they strengthened feedback uptake and self-regulated learning during revision. Their advantages therefore appear to reflect a more productive organization of evaluative activity rather than short-term score improvement alone (Buckingham Shum et al., 2023; Carless & Boud, 2018; Nicol & Macfarlane-Dick, 2006; Panadero, 2017).

Because the mediation variables and intervention-cycle gain scores were aggregated across the same revision period, the mediation findings should be interpreted as evidence of process-linked explanatory pathways rather than as definitive proof of temporally ordered causal mechanisms. The serial pathway from feedback uptake to self-regulated learning was theoretically grounded in the assumption that attending to and interpreting feedback supports subsequent planning, monitoring, and strategic

adjustment during revision. However, the present design does not fully isolate within-cycle temporal ordering between uptake, regulation, and improvement.

The exploratory moderation pattern connects the study's theoretical emphasis on feedback literacy to the empirical findings. Students with lower initial feedback literacy may be especially vulnerable to the limitations of direct AI feedback because they may have fewer resources for judging critique, selecting among suggestions, and translating feedback into independent performance. Structured reflective prompts before AI critique may therefore function as compensatory scaffolds, although this interpretation remains exploratory and should be tested directly in future research.

The study also contributes to discussions of epistemic agency in AI-supported learning. The findings are consistent with concerns that AI assistance may weaken student agency when critique is accepted too readily or when difficult judgments are outsourced to the system (Buckingham Shum et al., 2023; Darvishi et al., 2024). They also align with prior work showing that AI and peer feedback offer complementary affordances, that AI-supported peer feedback benefits from scaffolding, and that GenAI is most educationally useful when positioned as a dialogic partner rather than an authoritative answer source (Banihashem et al., 2024; Chang et al., 2026; Tang & Putra, 2025). More broadly, the findings suggest that concerns about authorship, evidence of learning, and learner responsibility are not only matters of policy or integrity, but also matters of feedback design (Wang & Fan, 2025; Xia et al., 2024).

#### Practical implications for higher education digital learning

The findings have several practical implications for higher education digital learning. When the goal is quick draft improvement, direct GenAI-supported feedback may be useful because it provides rapid, criterion-referenced critique with relatively low orchestration demands. However, when the goal is deeper learning, stronger reasoning, or AI-independent transfer, instructors should not present AI critique as the first or only evaluative input. The results instead suggest that AI critique should be preceded by brief, structured self-evaluation in which students identify strengths, weaknesses, uncertainties, and revision priorities before viewing external feedback.

Feedback configurations that combine multiple perspectives with structured comparison and revision justification may be useful for complex reasoning tasks in which students must weigh evidence, evaluate alternatives, and explain revision decisions. In the present study, the hybrid condition included not only self-, peer-, and AI-based feedback, but also structured evaluative activity, additional comparison work, and a revision memo. Its practical value should therefore be understood as arising from this broader configuration rather than from the mere addition of multiple feedback sources. More broadly, GenAI feedback systems should be designed around comparison, selection, and justification rather than passive implementation of suggestions.

A further practical implication concerns cost-effectiveness and orchestration burden. Although the hybrid condition produced the highest adjusted means on several outcomes, the reflective GenAI-supported design achieved a broadly similar pattern on feedback uptake, self-regulated learning, and delayed transfer while requiring substantially less coordination. For large enrolments or resource-constrained settings, structured self-evaluation followed by constrained GenAI feedback may therefore represent a sustainable default design. More resource-intensive hybrid configurations may be most

appropriate when tasks are high-stakes, conceptually complex, or explicitly intended to develop comparison across self, peer, and AI perspectives.

At the institutional level, decisions about GenAI adoption should not be made only in terms of scalability or efficiency. They should also consider whether digital-learning designs preserve student agency, evaluative judgment, and ownership through reflective scaffolding, rubric visibility, revision accountability, and appropriate data-governance procedures.

### Limitations

Several limitations should be considered when interpreting the findings. First, the study intentionally compared pedagogical feedback designs rather than equal-duration exposure conditions. The reflective and hybrid conditions involved additional guided evaluative activity, including self-evaluation, comparison, and, in the hybrid condition, revision justification. As a result, the benefits of reflective and hybrid designs cannot be fully disentangled from additional time-on-task, reflective workload, or structured cognitive engagement. The conceptual-learning advantage observed for the hybrid condition may therefore partly reflect greater exposure time, additional structured engagement, or increased opportunities to rehearse evaluative judgments, rather than the intrinsic superiority of the hybrid feedback sequence alone.

Second, students in AI-supported conditions received additional GenAI-use training. Although this training was necessary to support ethical, consistent, and procedurally comparable use of the interface, it may have increased preparation time relative to the peer-feedback condition and may have influenced students' familiarity with feedback resources or confidence in navigating the intervention. Future equal-time designs are needed to separate the effects of feedback configuration from the effects of additional training and preparation.

Third, gain-score models may underrepresent improvement among students whose initial drafts were already strong, because such students had less available room for improvement on the four-dimension scale. Although baseline scientific argumentation was included as a covariate, baseline equivalence was established, and the supplementary baseline-adjusted final-score sensitivity model produced a substantively similar pattern, ceiling-related constraints remain a possible influence on the magnitude of observed gain differences.

Fourth, peer-comment quality was not independently scored as a separate explanatory variable. Although peer feedback was structured through a common rubric-based template, variation in the specificity, accuracy, and actionability of peer comments may have introduced noise into the peer and hybrid conditions. This limitation is especially relevant for interpreting comparisons between peer-only and AI-supported feedback designs, because some observed differences may partly reflect variability in the quality of peer input rather than feedback source alone.

Fifth, the mediation models used aggregated process and outcome indicators across the intervention period. Therefore, the indirect effects should be interpreted as process-consistent explanatory associations rather than definitive evidence of a temporally sequenced causal chain. Although the theoretical ordering from feedback uptake to self-regulated learning is consistent with the study's framework, stronger causal claims

would require more fine-grained temporal designs, such as cycle-level lagged mediation or within-cycle process tracing.

Sixth, another limitation concerns the averaging of four-dimension argument-quality gain across the three intervention cycles. This approach matched the preregistered outcome definition and provided a stable estimate of overall revision-related improvement, but it may have obscured differences in learning trajectories across cycles. For example, students in the reflective and hybrid conditions may have shown increasing gains over time as evaluative judgment, feedback internalization, and self-regulated revision developed across repeated cycles. The delayed AI-free transfer task partly addresses this concern by providing a post-intervention measure of cumulative, AI-independent argumentation performance. Nevertheless, future research should use cycle-level growth models or lagged analyses to examine whether feedback designs differ not only in average gain but also in the rate and timing of improvement.

Seventh, the study was conducted in first-year university science courses, and the extent to which the findings generalize to other disciplines, levels of study, or institutional contexts remains uncertain. Scientific argumentation is a demanding context for examining evidence-based reasoning and revision, but different patterns may emerge in disciplines with different epistemic norms, writing practices, or assessment structures.

Finally, although AI-free transfer tasks were completed without access to the study's GenAI system in supervised settings, the design cannot eliminate broader concerns about students' external familiarity with AI beyond the intervention context. The analytic sample may also reflect some consent-based selection because instructional participation was compulsory whereas research participation was voluntary. In addition, the participating institutions were modeled as fixed effects because only four institutions were included, and the moderation analyses were exploratory and should therefore be interpreted with caution.

#### Future research directions

Future research should extend this work in several directions. First, equal-time comparison studies are needed to determine whether the advantages of reflective and hybrid feedback persist when time-on-task, GenAI-use training, and structured engagement are held constant across conditions. Such designs would help clarify whether observed differences are attributable primarily to feedback sequencing, reflective workload, cognitive engagement, or their combined effect.

Second, future studies should measure peer-feedback quality directly. Peer-comment specificity, accuracy, actionability, and alignment with rubric criteria should be modeled as explanatory variables, especially in studies comparing peer-only, AI-supported, and hybrid feedback designs. This would clarify when peer feedback provides productive contrast with AI critique and when variability in peer input limits its effectiveness.

Third, future work should use ceiling-sensitive analytic approaches to complement gain-score models. Latent growth models, residualized change models, and baseline-adjusted final-outcome models may help clarify how feedback designs affect students who begin at different levels of argumentation quality.

Fourth, more temporally fine-grained designs are needed to examine learning trajectories across revision cycles. Cycle-level growth models could test whether reflective and hybrid feedback designs support increasing gains over time as students develop

evaluative judgment, feedback internalization, and self-regulated revision. Lagged mediation designs and within-cycle process tracing could also provide stronger tests of the uptake → self-regulated learning → transfer pathway.

Fifth, future research should examine the minimum effective amount of structured self-evaluation required before AI critique. The present findings suggest that prior self-evaluation is important, but they do not establish whether shorter or lighter reflective prompts would produce comparable benefits in large-enrolment or resource-constrained courses.

Finally, future studies should examine cost-effectiveness and implementation burden. Studies comparing learning gain, delayed transfer, instructor workload, platform complexity, and student time investment would help institutions decide when reflective GenAI-supported feedback is a scalable default and when more resource-intensive hybrid configurations are worth the additional orchestration.

## Conclusion

The findings suggest that the pedagogical value of GenAI feedback in higher education depends less on the availability of AI-generated critique than on how feedback processes are designed. In the present study, direct GenAI-supported feedback supported immediate improvement in four-dimension argument-quality gain, whereas reflective and hybrid designs were more consistently associated with stronger feedback uptake, self-regulated learning during revision, and AI-independent transfer. Conceptual-learning benefits were strongest for the hybrid condition and more tentative for the reflective condition, whose adjusted contrast relative to direct GenAI-supported feedback was positive but not statistically significant. These findings reinforce the view that feedback is educationally valuable not simply when it is delivered efficiently, but when learners are positioned to interpret, compare, and act on critique as active evaluative agents (Carless & Boud, 2018; Nicol & Macfarlane-Dick, 2006).

More broadly, the findings suggest that the benefits of GenAI in higher education are highly design-sensitive. When AI feedback is organized primarily around speed and efficiency, it may support short-term improvement without necessarily strengthening the broader learning processes that underlie durable understanding and transfer. By contrast, when AI-supported feedback is embedded in reflective and comparative designs that preserve student judgment, self-regulation, and ownership, it is more likely to support meaningful learning (Buckingham Shum et al., 2023; Darvishi et al., 2024; Wang & Fan, 2025). At the same time, these benefits should be interpreted as effects of feedback designs as pedagogical configurations, including feedback source, sequencing, structured evaluative activity, and cognitive engagement, rather than as effects of AI access or feedback source alone. For higher education digital learning, the central implication is therefore not whether students can access GenAI feedback, but how GenAI feedback is pedagogically organized so that human evaluative work remains at the center (Buckingham Shum et al., 2023; Xia et al., 2024).

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s41239-026-00614-9.

Supplementary Material 1.

**Acknowledgements**
Not applicable.

**Author contributions**
H.A conceived and designed the study; developed the methodology and intervention materials; conducted the investigation; collected, curated, and analyzed the data; interpreted the findings; prepared the tables and figures; wrote the original draft of the manuscript; reviewed and revised the manuscript; and approved the final version.

**Funding**
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

**Data availability**
The datasets generated and/or analyzed during the current study are not publicly available because they include de-identified student performance, survey, interview, and platform-trace data collected under institutional ethics approvals and data-governance restrictions. However, de-identified data are available from the author on reasonable request, subject to institutional and ethical requirements.

## Declarations

**Competing interests**
The authors declare no competing interests.

Received: 12 April 2026 / Accepted: 6 July 2026
Published online: 16 July 2026

**References**

Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education, 21*, Article 23. https://doi.org/10.1186/s41239-024-00455-4

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Buckingham Shum, S., Lim, L.-A., Boud, D., Bearman, M., & Dawson, P. (2023). A comparative analysis of the skilled use of automated feedback tools through the lens of teacher feedback literacy. *International Journal of Educational Technology in Higher Education, 20*, 40. https://doi.org/10.1186/s41239-023-00410-9

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education, 43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Chang, Y., Liu, Q., Lu, Y., & Miao, E. (2026). Leveraging generative AI to facilitate peer feedback in collaborative argumentation learning. *International Journal of Educational Technology in Higher Education, 23*, Article Article 10. https://doi.org/10.1186/s41239-026-00586-w

Darvishi, A., Khosravi, H., Sadiq, S. W., Gašević, D., & Siemens, G. (2024). Impact of AI assistance on student agency. *Computers & Education, 210*, 104967. https://doi.org/10.1016/j.compedu.2023.104967

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education, 84*(3), 287–312. https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1098-237X(200005)84:3%3C287::AID-SCE1%3E3.0.CO;2-A

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Molloy, E., Boud, D., & Henderson, M. (2020). Developing a learning-centred framework for feedback literacy. *Assessment & Evaluation in Higher Education, 45*(4), 527–540. https://doi.org/10.1080/02602938.2019.1667955

Nicol, D. (2021). The power of internal feedback: Exploiting natural comparison processes. *Assessment & Evaluation in Higher Education, 46*(5), 756–778. https://doi.org/10.1080/02602938.2020.1823314

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199–218. https://doi.org/10.1080/03075070600572090

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching, 41*(10), 994–1020. https://doi.org/10.1002/tea.20035

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology, 8*, Article Article 422. https://doi.org/10.3389/fpsyg.2017.00422

Panadero, E., Jonsson, A., & Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review, 22*, 74–98. https://doi.org/10.1016/j.edurev.2017.08.004

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1991). *A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ)* (NCRIPTAL Technical Report No. 91-B-004). National Center for Research to Improve Postsecondary Teaching and Learning, University of Michigan. https://files.eric.ed.gov/fulltext/ED338122.pdf

Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement: Enabling students to make decisions about the quality of work. *Higher Education, 76*(3), 467–481. https://doi.org/10.1007/s10734-017-0220-3

Tang, K.-S., & Putra, G. B. S. (2025). Generative AI as a dialogic partner: Enhancing multiple perspectives, reasoning, and argumentation in science education with customized chatbots. *Journal of Science Education and Technology, 35*, 128–140. https://doi.org/10.1007/s10956-025-10240-1

Wang, J., & Fan, W. (2025). The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. *Humanities and Social Sciences Communications, 12*, 621. https://doi.org/10.1057/s41599-025-04787-y

Weidlich, J., Jivet, I., Woitt, S., Orhan Göksün, D., Kraus, J., & Drachsler, H. (2025). The student feedback literacy instrument (SFLI): Multilingual validation and introduction of a short-form version. *Assessment & Evaluation in Higher Education, 50*(5), 677–693. https://doi.org/10.1080/02602938.2025.2451729

Woitt, S., Weidlich, J., Jivet, I., Orhan Göksün, D., Drachsler, H., & Kalz, M. (2025). Students' feedback literacy in higher education: An initial scale validation study. *Teaching in Higher Education, 30*(1), 257–276. https://doi.org/10.1080/13562517.2023.2263838

Xia, Q., Weng, X., Ouyang, F., Lin, T. J., & Chiu, T. K. F. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. *International Journal of Educational Technology in Higher Education, 21*, 40. https://doi.org/10.1186/s41239-024-00468-z

Zhu, M., Liu, O. L., & Lee, H.-S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education, 143*, 103668. https://doi.org/10.1016/j.compedu.2019.103668

## Publisher's Note