Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Feedback and Reconstruction in Learning with Generative AI: An Integrative Review and Proposed Framework for Unaided Performance

Amar Shafiq

amarshafiq.scholar@yahoo.com

University of Electronic Science and Technology of China - Qingshuihe Campus
https://orcid.org/0000-0002-9030-2357

Systematic Review

Keywords: generative artificial intelligence, independent learning, metacognition, feedback, cognitive offloading, self-explanation

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11596206/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Generative artificial intelligence (GenAI) may help students produce higher-quality work; however, its effects on their future ability to work independently remain unclear. This integrative review examines this distinction by analysing research on feedback, metacognitive monitoring, cognitive offloading, and reconstructive learning activities. Targeted web searches and primary-source tracing informed a focused synthesis of adult learning studies and relevant foundational research. Three features may help explain variation in findings across studies: the information feedback conveys, the reasoning learners are expected to undertake and demonstrate, and the conditions under which learning is assessed. The proposed framework distinguishes between a regulatory pathway, in which feedback informs judgments and learning choices, and a practice pathway, in which reconstruction prompts relevant reasoning. This distinction helps explain why learners may become better calibrated without learning more, or learn more without becoming better calibrated. Feedback is expected to enhance the benefits of reconstruction when it improves engagement or reasoning quality, although its incremental benefit may be limited when the activity already provides comparable practice. Four conditional propositions describe these relationships and suggest comparisons that could help distinguish them from the effects of additional instruction, exposure, time, and motivation. The framework provides a foundation for testing when feedback and reconstruction support performance in the absence of assistance. The available evidence is not yet sufficient to establish the full causal pathways or lasting changes in capability, underscoring the value of delayed assessments, transfer tasks, and longitudinal research.

1. Introduction

A student may submit a persuasive analysis using generative artificial intelligence (GenAI) yet still struggle to explain the recommendation or apply the reasoning to a different case. This reflects a well-established principle in educational research: successful performance during an activity does not necessarily indicate what a learner has retained. Salomon et al. (1991) distinguished between the effects achieved with intelligent technologies and the changes that result from working with them. Similarly, Soderstrom and Bjork (2015) demonstrated why performance during practice should be distinguished from subsequent learning. GenAI brings these questions into everyday study. During a single conversation, a learner might request an explanation, ask the system to complete a solution, or seek feedback on an independent attempt. Each approach offers a distinct learning opportunity, even when the final answers appear equally strong.

The empirical findings underscore this variation. Fan et al. (2025) found that ChatGPT helped students revise essays but did not detect corresponding differences in knowledge acquisition or transfer. In a creativity task, Wong and Qiu (2026) reported stronger subsequent independent performance when students generated their own ideas before consulting ChatGPT. By contrast, Kestin et al. (2025) found substantial immediate learning benefits from a carefully designed AI tutor. As these studies examine different instructional approaches and assess distinct dimensions of learning, their findings suggest that

the educational consequences of GenAI depend on how learners use the assistance and how those effects are evaluated.

Feedback and reconstruction provide two intervention targets for examining this problem. Feedback may help learners recognize the limitations of their assisted performance, particularly by demonstrating what they can accomplish independently or how they used AI. Reconstruction enables learners to engage with reasoning that might otherwise remain implicit in a generated answer. Although these approaches overlap to some extent, each can also be applied independently. For example, learning that a test score fell below expectations does not explain how to address the gap, while producing a sound explanation does not necessarily correct an inaccurate assessment of one's abilities. By distinguishing these functions, we can better understand which information supports learners' decisions and which activities provide valuable practice.

Existing scholarship provides a robust conceptual foundation for these questions. Tankelevitch et al. (2024) examined the metacognitive demands of GenAI use, while recent reviews investigate relationships between generative systems and feedback, agency, and self-regulated learning (U. Maier et al., 2026; Shang, 2026; Wang et al., 2026; Xia et al., 2026). Nakamura's (2026) study of educational offloading is particularly relevant to the present argument. Collectively, this body of work underscores the importance of monitoring, verification, and sustained cognitive engagement. However, it remains less clear how to distinguish feedback's contribution from that of the subsequent learning activity, and what evidence would justify attributing an independent learning gain to either process.

This review examines how feedback and reconstructive activity can support adults' subsequent unaided performance when learning with GenAI. Its central premise is that recognizing a learning gap and practising the reasoning required to address it serve distinct purposes. Feedback can help learners make more informed judgments without changing their study practices, whereas reconstruction can strengthen independent performance without correcting those judgments. Accordingly, the framework identifies two pathways: a regulatory pathway, in which information shapes consequential learning choices, and a practice pathway, in which learners generate and evaluate relevant reasoning. The benefits of combining these pathways are expected to depend on whether feedback encourages learners to engage in that practice. Four conditional propositions describe these relationships and the comparisons required to distinguish their effects from those of additional instruction, exposure, time, and motivation.

2. Review Method

2.1 Scope and Search Approach

An integrative approach is appropriate because the relevant evidence spans educational psychology, human–computer interaction, and metacognition research. Although these traditions employ different terminology and methods, they address related questions about assistance and learning. Consistent

with guidance on integrative and theory-oriented reviews, this synthesis brings together and compares their concepts and explanations (Post et al., 2020; Snyder, 2019; Torraco, 2016). The substantive focus is on adult learning and skill practice involving text-generating AI. Earlier research provides definitions and mechanisms that help interpret recent studies. Its role is explanatory: findings from non-GenAI settings are not presented as direct evidence of an intervention's effectiveness with generative systems.

Literature searches and source verification continued through 2 November 2026. Web searches were supplemented by searches of publisher websites, institutional repositories, author manuscripts, and references cited in relevant reports. Search terms combined generative AI or ChatGPT with feedback, calibration, cognitive offloading, self-explanation, reconstruction, retention, transfer, and independent performance. Additional searches examined closely related frameworks, measurement research, and findings that provided alternative perspectives on the emerging argument. Sources were selected purposively. The search did not include comprehensive bibliographic exports from Scopus, Web of Science, PsycINFO, ERIC, or the ACM Digital Library, and screening and data extraction were not independently duplicated. Accordingly, this review presents a critical synthesis of selected evidence rather than an exhaustive review of the field.

The literature was evaluated based on what each source could substantiate. Studies involving adults who used GenAI and objective learning assessments provide the primary empirical evidence. Experiments in related contexts, qualitative studies, and measurement studies help clarify the underlying mechanisms and concepts. Previous reviews connect the findings to existing explanations and identify relevant primary research. Findings from school-based studies are included for comparison, while acknowledging that their populations differ from the adult-focused scope. Preprints and working papers are included when they directly inform the argument, with their provisional status noted. Studies of adoption intentions, AI-only performance, or assisted products alone cannot establish gains in independent capability. However, product measures may still be informative when considered alongside subsequent learning.

2.2 Evidence Appraisal and Synthesis

For the primary empirical comparisons, the synthesis closely examined the task and sample, the intervention and comparator, the feedback provided, the learning activities involved, and the assessment conditions. It also considered whether support remained available, the interval between practice and assessment, and whether learners were assessed on a repeated or novel task. Each conclusion was carefully evaluated in light of the strength of the study design. For example, randomisation can demonstrate the effect of an instructional package while leaving the contribution of each component unclear. Qualitative evidence may suggest a plausible process, but it cannot establish that an intervention caused an effect. When details regarding assessment access or implementation were unclear, conclusions were expressed with appropriate caution.

A pooled estimate would be challenging to interpret across interventions and outcomes that vary substantially; therefore, the review employs a critical narrative synthesis. Findings are compared in light

of study design and measurement, with null results and alternative explanations also considered. Statistical significance is not used to rank effects across studies. The review assigns greater weight to direct comparisons involving clearly described treatments and outcomes, while considering publication status and source accessibility when interpreting individual reports. Purposive selection, limited full-text access to some peripheral sources, and the absence of duplicate screening limit the review's coverage and reproducibility. These limitations are discussed further in the relevant section.

3. Conceptual Foundations

3.1 Independent Performance and Evidence of Learning

Subsequent unaided performance refers to what learners can do when relevant AI assistance is unavailable. This focus reflects the central question: after working with generative AI, what can learners accomplish independently? Many educational and professional activities also require competent tool use; therefore, assessments that permit assistance are appropriate as well. Evaluating both objectives requires evidence from both types of assessment. Even when learners complete tasks without assistance, the nature of the task matters: identifying an answer on a multiple-choice test, explaining a principle, and analysing an unfamiliar case each demonstrate different capabilities. Any conclusions about learning should specify which capabilities the assessment was designed to measure.

Assessment independence, delay, and task novelty warrant separate consideration. A test administered immediately after removing the GenAI tool indicates how an individual performs without it at that point. To assess retention, testing must occur later; transfer, meanwhile, concerns performance beyond the material or situation encountered during practice. These features can vary independently. For example, a delayed test may repeat the original questions, whereas an immediate test may ask learners to apply what they have learned in a new way. A different surface context does not necessarily constitute far transfer if the underlying procedure remains unchanged (Barnett & Ceci, 2002). The distinction between learning and performance is therefore most useful when you consider the full pattern of practice and assessment (Soderstrom & Bjork, 2015). Table summarises the information needed to interpret that pattern.

**Table **

Information needed to interpret studies of learning with AI

Study feature	Information to report	Reason for reporting
Learning goal	Target operation and whether capability is expected with or without AI.	A strong assisted product and independent capability are distinct criteria.
Feedback information	Source, target, evidential basis, timing, dose, advice, and instructional content.	Actual unaided evidence, predicted scores, process descriptions, and task explanations test different routes.
Learner operation	Required output, source access, AI access, time, and observed uptake or accuracy.	An assigned explanation is an opportunity for processing; its occurrence and quality require evidence.
Assessment assistance	AI, source, peer, and other tool access, with actual enforcement.	Independence must be established rather than inferred from the word posttest.
Assessment timing	Interval after learning and any intervening practice.	Immediate independence does not establish persistence.
Task relationship	Repeated versus new material and the structural changes claimed as transfer.	Novelty and delay are separate axes; a new context need not constitute far transfer.
Claim and unit	Randomised ingredient or package, allocation unit, outcome unit, and uncertainty.	An intervention effect does not identify an unrandomised mediator.

Note. This reporting framework summarises the conceptual distinctions developed in the review; it has not been validated as an appraisal instrument.

3.2 What Confidence Refers To

Metacognitive monitoring involves judgments about cognition, whereas control concerns how those judgments guide subsequent actions (Nelson & Narens, 1990). When applying this distinction to GenAI, it is helpful to specify precisely what a confidence judgment concerns. A student may trust an AI-generated claim and expect an AI-assisted assignment to receive a high grade, yet still anticipate a low score on an unaided test. Another student may be sceptical of the system while overestimating their own understanding. These judgments concern different matters and therefore do not necessarily align. A general confidence measure may obscure the specific change that a feedback intervention is intended to produce.

Calibration describes how well a judgment corresponds to its relevant outcome. When an individual forecasts their performance without assistance, compare that forecast with their actual unaided result using the same scale and timeframe. Signed error indicates the direction of any difference, absolute error indicates its magnitude, and discrimination reflects whether confidence helps distinguish stronger responses from weaker ones (Schraw, 2009). It is useful to report the original forecast and performance alongside these measures. Signed error may remain unchanged even as achievement improves, and an average close to zero may conceal overconfidence among some learners and underconfidence among

others. Because short quizzes involve sampling uncertainty, drawing broad conclusions about an individual's calibration can be particularly challenging.

Research on internet search provides a useful historical comparison. Fisher et al. (2015) examined how access to external explanations may shape estimates of individuals' internal knowledge. A similar attribution challenge may arise with GenAI, although it requires evaluation using measures appropriate to that context. Fernandes et al. (2026) reported a discrepancy between performance and self-assessment in AI-assisted reasoning; however, their first study relied on a historical comparison rather than a concurrent randomized control. Because the judgments concerned assisted performance, these findings help clarify confidence during AI use, while leaving a distinct question unresolved: how accurately do learners anticipate what they can do once the assistance is removed?

3.3 Feedback Content and Evidential Basis

The source of feedback identifies who provided it, whereas its target concerns the information it conveys. For example, a large language model (LLM) might comment on an essay, while a teacher might describe a student's use of AI. Either source may provide feedback on the task, the learner, or the learning process. For the present review, we distinguish four informational targets, informed by established feedback research: the quality of a product or response, demonstrated unaided performance, the discrepancy between a forecast and an outcome, and observed learning behaviour (Hattie & Timperley, 2007; Shute, 2008). A single message may address multiple targets. Labels such as personalised feedback, on their own, do not provide enough information to distinguish their effects without knowing the message content, timing, and evidential basis.

This evidence is important. A score from a completed test taken without assistance provides different information from an algorithmic prediction of a future score, and each has distinct uncertainties. Similarly, a transcript may show that a student requested a complete solution, but it cannot indicate how carefully the student considered it. Identifying a forecast discrepancy requires both the learner's judgment and the corresponding result. If feedback also includes correct answers, explanations, or strategic guidance, these instructional elements may help explain subsequent improvement. How learners interpret and use this information is another important consideration, addressed in the literature on feedback literacy (Carless & Boud, 2018).

3.4 Offloading and Reconstructive Activity

Cognitive offloading involves using external resources or actions to alter the demands of a task (Risko & Gilbert, 2016). Its effects on learning depend on which aspects of the work the learner delegates. For example, a student learning to diagnose a problem may benefit from delegating transcription while retaining responsibility for the diagnostic judgment. Delegating that judgment may reduce opportunities to practise generating it, although evaluating or explaining the supplied reasoning may still support learning. Research on external memory also suggests that making information readily available can alter memory strategies without necessarily causing a general decline in ability (Sparrow et al., 2011). When

considering learning with GenAI, the key question is which cognitive operations learners retain or delegate in relation to their learning goals. Prompt frequency alone is a less useful measure.

Reconstruction refers to activities that prompt learners to make the reasoning underlying an answer explicit. They may self-explain a relationship, evaluate conflicting evidence, recover a rationale, or apply a principle under new conditions. The term encompasses activities grounded in research on self-explanation, constructive engagement, and retrieval (Chi et al., 1994; Chi & Wylie, 2014; Roediger & Karpicke, 2006), while acknowledging that these approaches differ in important respects. When learners explain while the source material is available, they can select and reorganize information; explaining from memory also involves retrieval. An intervention that changes both access to the source and the demands of producing an explanation draws on both processes. Its effects cannot be attributed to only one process simply because the activity is described as reconstruction.

4. Empirical Evidence on Assistance and Independent Learning

4.1 Product Quality and Subsequent Performance

An experiment involving 117 university students highlights the distinction between product improvement and support for independent learning (Fan et al., 2025). ChatGPT, human expert assistance, a learning-analytics checklist, and no additional support produced different revision processes and outcomes; however, the study found no differences in knowledge gain or transfer. This suggests that improvements in revisions alone may not fully capture the educational benefits. The absence of detected differences in learning also leaves some uncertainty: it does not establish that the conditions were equivalent or that students' capabilities declined. Differences in recorded activity may help explain the findings, although the authors did not examine these processes separately.

Other instructional approaches may also produce positive learning outcomes. Kestin et al. (2025) compared a research-informed AI tutor with active classroom learning among undergraduate physics students. The crossover study included 194 eligible students, assigned through peer-instruction groups, and found that the tutor produced higher scores on the immediate post-lesson assessment. This finding reflects the combined effects of the tutor's design, pacing, interaction, and setting; it does not demonstrate that open-ended chat alone was responsible. As assessment-access controls could not be fully verified in the materials reviewed, the result is best interpreted as evidence of immediate post-lesson learning. Whether it supports longer-term independent capability remains an open question.

Wong and Qiu (2026) examined the sequence of learner and AI contributions more directly. In their experiment, 196 university students were assigned to one of three conditions: human-only work, unrestricted ChatGPT use, or a structured approach that required students to think first. This structured approach supported stronger independent creativity afterward. Although unrestricted AI use improved the assisted product, it did not result in a corresponding advantage in independent creativity. Because students completed the independent task as part of the study, the findings relate to subsequent

performance rather than long-term retention. Taken together, these comparisons underscore the value of reporting assisted-product quality and learning assessments separately. Table synthesizes these findings alongside studies on feedback and access arrangements.

**Table **

Selected adult GenAI studies and their learning assessments

Study	Focal contrast	Learning assessment	Interpretation
Fan et al. (2025)	Four support arrangements.	Revision, knowledge, and transfer.	Product advantage; no detected knowledge or transfer differences.
Kestin et al. (2025)	Structured tutor versus active class.	Immediate post-lesson assessment.	Combined instructional changes; assessment-access controls remain unclear.
Wong and Qiu (2026)	Think-first versus free AI or human-only.	Subsequent independent creativity.	No long-delay inference.
Fischer et al. (2025)*	Continuous, delayed, or no tutor access.	Immediate unaided economics test.	Continuous access exceeded delayed access.
Shen and Tamkin (2026)*	AI versus no AI during new-library practice.	Immediate skill assessment.	Less learning does not establish loss of established capability.
S. Maier et al. (2026)*	Process feedback crossed with reward; no-AI comparison.	Immediate unaided arithmetic.	One-sided tests; no detected calibration difference.
Dai (2026)	Feedback-literacy script crossed with calibration support.	Repeated writing assessments; withdrawal phase.	Outcome-definition and reporting ambiguities constrain interpretation.

Note. * Preprint or working paper. Assessment independence, delay, and transfer are considered separately.

4.2 Access Arrangements and the Choice of Comparison

Requiring students to work independently before using AI does not produce the same outcomes in every context. Fischer et al.'s (2025) working paper examined 334 university students preparing for an economics test. Access to AI improved students' subsequent performance without assistance, and continuous access yielded better results than access provided only after independent reading. This preliminary finding suggests that delaying assistance is not always appropriate. It does not necessarily contradict the creativity study: drafting an initial idea and completing a required reading period influence different aspects of learning. A useful explanation should consider what the restriction requires students to do, when assistance becomes beneficial, and which skills the task demands.

Reports of less favorable outcomes warrant the same careful consideration as reports of learning gains. In a preprint involving 52 experienced Python users learning an unfamiliar library, Shen and Tamkin (2026) found weaker immediate understanding and debugging performance following AI-assisted practice. This finding concerns learning new material; it does not show a loss of previously acquired programming ability. In high-school mathematics, Bastani et al. (2025) found that guardrails shaped the effects of AI-supported practice on subsequent unaided assessments. Although the students were not part of an adult-focused population, this comparison underscores the importance of instructional design. Collectively, these findings provide a more nuanced basis than the broad claim that AI use causes deskilling.

The comparator shapes the question a study can answer. Replacing active classroom instruction with a specialized tutor helps evaluate the full instructional package. Allowing AI where it was previously unavailable helps assess access within the task's specific incentives. Providing different feedback within the same AI environment tests the effects of the information supplied. Requiring students to reconstruct rather than review tests a production requirement, provided that time and information are comparable. These comparisons address distinct educational decisions. To understand their outcomes, it is important to preserve these distinctions rather than place every study on a single scale of greater or lesser AI dependence.

4.3 Evidence on Feedback and Calibration

Related experiments suggest that self-monitoring and aid selection can be modified, although their educational implications require further investigation. Ma et al. (2024) examined how accurately individuals calibrate their confidence when making AI-assisted decisions. In memory tasks, Ngai and Gilbert (2026) found that combining performance predictions with feedback improved strategic reminder use, whereas predictions alone were insufficient. Taken together, these findings connect individuals' judgments with their use of external aids. They also underscore why reducing offloading is not always the appropriate goal: when aids have clear costs and benefits, better regulation may involve using them more. To apply these findings to GenAI-assisted learning, further research should assess what learners retain and can use independently afterward.

Lee et al. (2025) extend calibration support to a more educational context. Their statistics-learning environment provided machine-learning predictions of eventual performance, and the reported associations linked learning behaviours to the intervention's effects. Rather than focusing on a completed, unaided test, the feedback emphasized a predicted outcome. It is also useful to distinguish this approach from generative conversational tutoring. The role of behaviour is consistent with a regulatory explanation, although a causal interpretation requires assumptions that random assignment to the overall intervention cannot establish on its own. The study offers valuable insights into potential links between monitoring and learning, while leaving the precise causal role of the intervening behaviour unresolved.

In a preprint, S. Maier et al. (2026) report an experiment testing process feedback among 704 adults practising arithmetic with an LLM assistant. The feedback helped clarify the division of work and its implications for practice. The authors reported reduced offloading and an immediate advantage on unaided tasks under prespecified one-sided tests, with no detected difference in calibration. The learning result was less conclusive under a conventional two-sided criterion, as the reported 95% confidence interval included the null. Nevertheless, this close precedent suggests that process feedback may influence behaviour and learning without a measurable change in calibration. Longer-term benefits, and whether calibration is necessary, remain open questions.

4.4 Learning Processes and Perceived Capability

Interaction records can indicate what students ask the system to do; however, the resulting activity is not always easy to interpret. A request for an explanation may encourage careful reasoning, or it may simply generate additional text to review. Stadler et al. (2024) reported reduced effort alongside less in-depth scientific inquiry with LLM support, underscoring the importance of considering ease alongside engagement quality. In related research on AI decision-making, Buçinca et al. (2021) found that cognitive forcing reduced overreliance, although it came at an experiential cost. The central educational question is what students gain from this added demand. Extra effort is worthwhile when it supports the intended learning or judgment.

Qualitative evidence provides a valuable complementary perspective on how learners navigate this tension. Gao and Zhang (2026) described accountable reconstruction in doctoral learning, including approaches for reconnecting AI-supported work with a learner's own reasoning. Zhu et al. (2026) distinguished between dependent and autonomous offloading in a multiwave study, highlighting different ways in which learners use assistance. However, their perceived-capability outcomes and temporal models cannot establish whether skills were retained or causally demonstrate mediation. This work helps clarify processes for future testing. It also reminds us that an observed action, a self-reported tendency, and an objectively demonstrated capability should not be treated as interchangeable indicators of the same construct.

5. How Feedback May Influence Learning

Feedback may address gaps in content knowledge, reshape learners' perceptions of their abilities, or highlight their learning strategies. A clear explanation may directly improve future performance, whereas a score provided without further guidance may identify a gap without supplying the knowledge needed to address it. Guidance on requesting complete answers may also encourage learners to approach the next task independently. These possibilities reflect established distinctions in feedback theory (Hattie & Timperley, 2007; Shute, 2008), and are particularly relevant when an external system can generate a substantial portion of the work being assessed. The central question is which information prompts which response and whether that response creates a stronger opportunity for learning.

Diagnostic feedback on capability is most useful when the forecast, assessment, and outcome correspond to the same clearly defined task. The learner independently predicts a score, completes the assessment under controlled conditions, and receives a result on the corresponding scale. Presenting the difference can clarify any discrepancy; however, the result remains only a sample of performance. A single brief quiz is insufficient to support a broad judgment that an individual is dependent or overconfident. Repeated assessments can provide learners with stronger evidence, although they also offer additional practice. Therefore, maintaining consistency across conditions is important when the study aims to measure the effect of feedback rather than differences in testing.

Process feedback is based on observations of activity. For example, a record may indicate that a student requested a completed recommendation before explaining; however, it cannot establish that no reasoning took place. A constructive message can describe the action and respectfully explain how it may affect practice while linking that consequence to the learning objective. This distinction is also important in experimental comparisons. If process feedback provides more extensive guidance, additional subject-matter knowledge, or more emotionally engaging language than score feedback, then the treatments differ in several respects. Any subsequent differences in performance may reflect content, attention, or motivation, in addition to the informational target.

Learners must also interpret and use the feedback they receive. Feedback literacy involves interpreting feedback and acting on its implications (Carless & Boud, 2018), while models of self-regulated learning distinguish between monitoring and the actions that follow (Panadero, 2017). A student may recognize a weakness but still lack an appropriate strategy or the background knowledge required to practice effectively. If an activity rewards only the immediate answer, continued delegation may remain a rational choice given the incentives. Therefore, a lack of improvement following diagnostic feedback may stem from several points in the process. Measuring how learners interpret feedback and what actions they choose to take next would help distinguish between feedback that was not informative and information that was understood but could not, or did not, lead to action.

These alternatives help explain why a change in calibration should not be assumed to account for every feedback effect. A message may encourage greater attention or practice without changing a forecast. More accurate forecasts may also result from improved learning rather than precede it. To study this process, measurements should follow the sequence suggested by the explanation: feedback first, the proposed change in judgment next, and relevant behaviors and outcomes afterward. The measurements themselves should avoid creating unequal practice opportunities or revealing the intended response. Even with this timing, unmeasured common causes may still influence an observed indirect association (Imai et al., 2010). It is therefore helpful to examine monitoring, allocation, and practice as separate links while leaving their causal status open to further testing.

6. Reconstruction and Practice in Reasoning

Reconstruction focuses on the reasoning learners are expected to develop. Research on self-explanation examines how learners connect ideas (Bisra et al., 2018; Chi et al., 1994). At the same time, the interactive, constructive, active, and passive (ICAP) framework distinguishes between generating new explanations and engaging with information already provided (Chi & Wylie, 2014). Together, these perspectives clarify what meaningful participation may involve following an AI response. For example, in a management case, students might connect evidence to a recommendation or explain why an alternative is less suitable. A brief explanation of their reasoning can reveal more about the intended task than the length of the preceding conversation.

The production requirement needs careful design. After receiving an AI-assisted analysis, a student might explain why the recommendation is reasonable, identify the assumption it depends on, and reconsider it when a constraint changes. These tasks require causal or conditional reasoning. Simply restating the answer in different words may involve little more than paraphrasing. Likewise, asserting that the student understands or takes ownership of the work provides little evidence of the reasoning involved. Reconstruction becomes easier to evaluate when the response makes the target reasoning visible and allows you to assess its accuracy. Otherwise, an activity that appears demanding may reinforce superficial or incorrect connections.

Whether sources remain available can substantially influence the activity. When learners have access to the case and the AI response, they can compare claims and reorganize the material. Removing these sources introduces an additional retrieval demand. Retrieval practice can be beneficial under appropriate conditions, and research on learning techniques distinguishes it from rereading and other study activities (Dunlosky et al., 2013; Roediger & Karpicke, 2006). Accordingly, a reconstruction study should clearly specify whether learners can access the sources. Benefits observed in a closed-source activity may reflect retrieval as well as explanation, whereas a source-available activity may primarily support integration. For novice learners, removing the material too early may also make successful reconstruction more difficult and allow errors to persist.

A review provides a useful comparison when described carefully. During rereading, students may privately test, explain, or compare ideas. An experiment requiring a written reconstruction therefore changes the obligation by requiring students to produce observable reasoning; it does not establish that the comparison group engages in no such mental activity. Providing both conditions with the same information and amount of time yields a more informative estimate of the effect of the production requirement. Outputs and process indicators can then reveal how closely participants followed the assigned activities. A benefit observed under these conditions supports the practical value of requiring reconstruction, while variation within either group remains relevant to understanding why it was beneficial.

Difficulty alone is not a sufficient reason to use the intervention. Learners' beliefs about effective study may differ from the strategies that support later performance (Bjork et al., 2013), and effort may sometimes be directed toward confusing or repetitive tasks. Reconstruction is expected to help when

the task elicits sufficiently accurate practice in relevant reasoning at a manageable level of challenge (Bisra et al., 2018). To assess its suitability, consider the quality of learners' responses and their subsequent performance, as well as the time required, errors made, and perceived effort. A slower activity may be worthwhile when it develops a necessary capability; it may be unnecessary when the goal is to complete a routine task with assistance.

7. An Integrative Framework

7.1 Proposed Relationships

The framework identifies two ways in which an intervention may support learning. In the regulatory pathway, feedback provides evidence about performance or learning behaviour, which learners can interpret and use to inform their subsequent choices. Monitoring and control are distinct processes; therefore, recognising a gap does not necessarily mean that a learner can or will respond effectively (Carless & Boud, 2018; Nelson & Narens, 1990). In the practice pathway, reconstruction involves learners generating and evaluating reasoning that supports the capability being developed, drawing on established research on self-explanation and constructive engagement (Chi et al., 1994; Chi & Wylie, 2014). This distinction clarifies where an intervention may fall short: feedback may be understood without changing practice, while reconstruction may prompt substantial effort that is inaccurate or unrelated to the intended learning outcome. Figure presents these relationships as hypotheses rather than as established causal pathways.

**Figure **

Proposed regulatory and practice pathways linking feedback and reconstruction to subsequent unaided performance

Note

All arrows denote hypothesised relationships. Metacognitive monitoring concerns judgments of unaided capability; metacognitive control concerns the allocation of cognitive work and the selection of learning strategies. The dashed arrow denotes the conditional influence of these choices on engagement in reconstructive activity. The dotted route represents instructional or motivational alternatives that do not need to operate through monitoring and control. Reconstruction may support reasoning practice without improved calibration. Boundary conditions delimit the proposed relationships. Assessment characteristics describe how the outcome is evaluated and are not causal stages. The framework is a conceptual synthesis, not a validated measurement model. The figure shows the principal proposed pathways. Feedback may also influence other learning activities, including independent practice and explanation seeking.

The key is to align the intervention, the process it is intended to support, and the outcome. If feedback concerns what a learner can do without AI, judgments about that learner's own unaided performance are

relevant to monitoring; trust in an AI-generated claim concerns a different target. If reconstruction involves explaining a trade-off, a factual recognition test may provide only limited evidence of its benefits. If the goal is to improve professional performance with assistance, an unaided memory test would capture only part of the picture. Defining these connections in advance makes the findings easier to interpret by clarifying what the intervention is expected to change and what evidence would support the proposed explanation.

Two patterns are particularly useful for understanding this framework. First, learners may become more accurate in judging what they can do without assistance, even if their actual ability remains unchanged. This may reflect successful updating without a corresponding change in effective practice. Second, learners may improve their unaided performance without becoming more accurate in assessing their abilities. This may indicate a benefit from practice that does not depend on improved monitoring. Neither pattern, on its own, identifies the cause. The first shows why improved calibration may be insufficient for learning, whereas the second challenges the claim that calibration improvement is necessary for every learning gain. To examine these possibilities, studies should separately measure learners' judgments, their subsequent choices, the quality of relevant practice, and independent performance. Carefully documenting cases in which no change occurs at a predicted point may also help clarify the limits of the proposed explanation.

7.2 Relationship to Existing Theory

Metacognitive theory distinguishes between monitoring and control, whereas self-regulated learning provides a cyclical framework of how these processes relate (Nelson & Narens, 1990; Panadero, 2017). Research on offloading examines the work learners transfer to external resources, while constructive-learning studies focus on the reasoning learners generate themselves (Chi & Wylie, 2014; Risko & Gilbert, 2016). Integrating these perspectives enables a more focused question: Does a particular type of feedback influence a decision in ways that shape relevant practice and subsequent learning? Each link must be supported by the study's design and measurement. Evidence of an overall intervention effect is informative but provides less insight than evidence identifying each of these individual links.

The distinction proposed here is more specific than the broader claim that metacognition and cognitive engagement are important when using GenAI. Metacognitive frameworks describe the demands of monitoring and guiding interactions with generative systems, while constructive-learning studies clarify why generating reasoning can support understanding (Chi & Wylie, 2014; Tankelevitch et al., 2024). This framework integrates these explanations through a shared set of intervention comparisons. It examines whether altering the feedback information affects subsequent practice when the activity requirement remains constant, and whether asking learners to reconstruct ideas improves learning when the feedback information remains constant. These comparisons let us assess each intervention's contribution without assuming the other is necessary. The framework's added value, therefore, depends on whether separating these factors provides more precise predictions than existing frameworks can make on their own.

There is considerable overlap with recent scholarship on generative AI. Nakamura (2026) connects calibrated monitoring, work allocation, verification, and reinternalisation. Shang (2026) examines the relationships among metacognition, trust, and co-regulation in foreign-language learning. Wang et al. (2026) investigate offloading and agency, while Xia et al. (2026) review generative AI-supported self-regulation. Building on these frameworks, the present framework distinguishes between the information conveyed through feedback and the activity expected of the learner. It links both to the assessment required to support a learning claim. Its contribution lies in the comparisons and predictions that follow, rather than in the general proposition that learners should remain cognitively involved.

Fan et al.'s (2026) SYNC model offers a closely related perspective by distinguishing metacognitive offloading from onloading and explaining how assisted performance may differ from learning. Together with Nakamura's (2026) augmentation–erosion framework, it underscores that supporting learner regulation and meaningful processing is already a central theoretical concern. This review focuses more specifically on the intervention question: What does feedback information contribute when the reconstruction requirement remains constant, and what does reconstruction contribute when the feedback information remains constant? The proposed contribution is to specify these comparisons and their conditional joint effects against clearly defined unaided outcomes.

Table situates this contribution within the related literature. Measurement research should also be considered before proposing a new capability construct. Existing scales for collaborative AI metacognition and autonomy provide useful starting points (Lu et al., 2026; Sidra & Mason, 2026). However, self-reported checking habits, demonstrated identification of a faulty explanation, and observed decisions to seek assistance provide different forms of evidence. Researchers should investigate their relationships before combining them into a total score or treating them as dimensions of a higher-order capability (Flake et al., 2017). The framework can guide task-based and behavioural assessment without presupposing a particular measurement structure.

**Table **

Relationship between the present framework and neighbouring literature

Existing contribution	Existing emphasis	Focus of the present review
Fan et al. (2026)	Metacognitive offloading and onloading; divergence between AI-assisted task performance and durable learning.	Separate feedback information from reconstruction requirements; specify comparisons testing their individual and conditional joint contributions to unaided performance.
Tankelevitch et al. (2024); Nakamura (2026)	Metacognitive demands, work allocation, verification, and reinternalisation.	Separate feedback information from required processing; specify contrasting interventions.
Shang (2026); Wang et al. (2026); Xia et al. (2026)	Co-regulation, agency, offloading, and self-regulated learning with GenAI.	Make outcome independence, delay, and task novelty explicit in each inference.
U. Maier et al. (2026); established feedback literature.	Feedback generation, quality, uptake, and instructional functions.	Distinguish the source of feedback from its informational target and evidential basis.
Sidra and Mason (2026); Lu et al. (2026)	Existing AI metacognition and collaborative-capability measures.	Distinguish self-report tendencies, demonstrated skills, and enacted choices before proposing new scores.

Note: This comparison situates the specific analytical contribution developed here within established accounts of metacognition, offloading, and constructive learning.

7.3 Conditional Propositions

Proposition 1

examines how learners revise their judgments after receiving diagnostic evidence. When an initial forecast is inaccurate, and a credible, unaided assessment measures the relevant capability, feedback reporting both the outcome and the discrepancy between the outcome and the forecast should prompt greater revision than a neutral progress update. We expect little change when the forecast is already accurate or when the assessment provides weak evidence. This prediction can be tested without assuming that learning improves immediately. Report forecasts and actual scores separately, as a smaller discrepancy could reflect revised confidence, improved performance, or both.

Proposition 2

connects this information to goal-directed control. Capability or process feedback may alter which cognitive operations learners retain or delegate when they understand its implications and have a feasible way to respond. The appropriate response depends on the nature of the gap and the task. For example, a student might complete additional problems independently, seek a clearer explanation, or obtain further AI support to address gaps in prerequisite knowledge. If only immediate output is rewarded, the same information may do little to promote independent practice. To evaluate this proposition, it is important to examine which operations learners retain or delegate, which strategies are

available to them, and which goals they pursue. A reduction in the number of prompts alone would not constitute sufficient evidence.

Proposition 3

examines the benefits of reconstruction practice. When learners have comparable information and time, prompting them to generate and verify relevant reasoning should help them perform the same operation independently later, provided their prior knowledge and the task's difficulty allow them to engage successfully. This benefit is expected to be greater for explanation and application than for recognizing familiar wording. The proposition would receive limited support if learners generated the intended reasoning accurately, yet a sensitive assessment produced an estimate precise enough to rule out a meaningful benefit. If the advantage disappeared when time or instructional content was held constant, this would instead suggest that the gains resulted from additional practice or information.

Proposition 4

examines whether feedback can enhance the benefits of reconstruction. When credible feedback about learners' capabilities identifies a relevant learning gap, and learners can choose how to respond, the framework predicts that the advantage of reconstruction over review will be larger under diagnostic feedback than under a neutral update. This prediction applies when reconstruction provides a practical means of addressing the identified gap; it does not apply when learners cannot interpret the feedback or complete the activities successfully. By contrast, if the production requirement already results in similarly accurate and complete reasoning across feedback conditions, the framework predicts that feedback will provide a smaller additional benefit through this pathway. Testing these predictions requires a prespecified outcome scale, sufficient precision to assess the interaction, and measures of reconstruction uptake and quality. Main effects alone would not demonstrate complementarity.

8. Research Agenda

8.1 Distinguishing Competing Explanations

Several explanations could account for an apparent benefit without supporting the proposed metacognitive process. For example, feedback may teach an answer or strategy that is not provided in the comparison condition. A reconstruction task may introduce examples, increase exposure, or provide an additional opportunity to retrieve information. A noticeable discrepancy may also prompt learners to pay closer attention or exert greater effort without improving calibration. Each explanation could indicate a beneficial educational outcome while suggesting a different reason for the intervention's effectiveness. The comparisons in Table are designed to help distinguish among these possibilities. In particular, matching the information and time helps determine whether learner production adds value beyond instruction and practice time.

**Table **

Comparisons needed to test the propositions and alternative explanations

Proposition	Critical comparison	Expected evidence	Alternative interpretation
1. Updating	Actual unaided score plus matching forecast discrepancy versus neutral progress; common probes.	Forecasts update toward credible evidence, especially when initially inaccurate.	An unchanged forecast with higher achievement supports a different route; noisy probes limit diagnosis.
2. Allocation	Cross feedback with learning versus output goals; measure retained and delegated operations.	Allocation changes in ways compatible with the stated goal and available strategies.	Lower prompt counts alone do not demonstrate adaptive regulation; an attention cue may explain change.
3. Practice	Required explanation or application versus review with matched information and time.	Accurate target-operation production and improved unaided performance on that operation later.	Benefits explained by extra content, exposure, or time weaken a distinct production account.
4. Joint effects	Capability feedback by reconstruction factorial; powered interaction; uptake measures.	Feedback modifies reconstruction uptake or quality when that link is possible.	A main effect is not synergy; an imprecise null interaction is not independence.
Note. These predictions concern future research. An observed indirect association is insufficient to establish causal mediation without additional assumptions or experimental evidence.			

Measurement can itself influence the process under investigation. Even an unaided quiz provides practice, whether or not the results are shared (Roediger & Karpicke, 2006). Predicting a score may prompt reflection, while asking learners to explain their reasoning may encourage them to reconstruct it. These experiences should be maintained consistently across conditions or evaluated as components of the intervention. Process coding raises a related issue: agreement among coders shows consistency, but the coded action must also match the proposed mental process. For example, requesting a complete answer reveals little about how the learner subsequently used the response. Combining selected interaction records with explanations, performance data, and carefully sampled retrospective evidence can support a more robust interpretation.

8.2 A Sequence of Experiments

An initial experiment could compare different feedback targets within a shared AI-assisted learning environment. One group would receive its actual score on an unaided probe, along with the difference between that score and its previously recorded forecast. A second group would receive factual feedback about its recorded AI use, while a control group would receive a neutral progress update. All groups would complete the same practice activities, make forecasts, and take unaided probes, with comparable time allocations and, where applicable, the same procedures for scoring assisted products and presenting that feedback. Correct answers and explanations would either be provided equally to all groups or consistently withheld during the key cycles. The contrast between capability and process

feedback may be particularly informative, as it compares two active forms of feedback while holding exposure to testing constant.

The primary outcome should assess how effectively students perform independently on new material after an appropriate and well-justified delay. For a business-case task, students could work through several cases with support and then, one week later, analyse new cases that require them to apply the same principles. An immediate assessment could provide a separate measure of learning before the delay. Forecasts should clearly specify which unaided test they address, and brief feedback probes should be treated as imperfect snapshots of ability. Clear scoring criteria, raters who are unaware of each student's assigned condition, and a plan for addressing missed follow-ups would strengthen the comparison. This design would help determine whether the information provided influences subsequent learning rather than merely improving the assisted submission.

A second experiment could examine capability feedback and reconstruction using a 2 × 2 design. Students would either receive diagnostic feedback or not, and would then either reconstruct the reasoning or review the same information for an equivalent duration. Access to AI would be unavailable during this activity. Sources would remain available in both conditions if the aim were to study explanation and organization without introducing unequal retrieval demands. The main effect of reconstruction would estimate the benefit of producing the reasoning, averaged across feedback conditions. A separately powered interaction would test whether diagnostic information changes that benefit. Measures of actual participation and accuracy would help clarify any differences between the assigned activity and the learning it produces.

Future research could build on these findings by replicating the most consequential contrast in a new sample and determining whether the effects persist or transfer. A longer interval tests persistence. Changes in task demands test transfer (Barnett & Ceci, 2002), whereas replication in another population or setting tests a further aspect of generalizability (Simons et al., 2017). Further evidence is needed to support claims about capability development or erosion during sustained AI use. This requires repeated exposure, comparable measures over time, and baseline and comparison trajectories that enable the interpretation of any observed changes. A lower post-test score following assisted practice may indicate that participants learned less from the experience; however, it does not demonstrate a decline in an established capability. Research can therefore progress from identifying an intervention effect to understanding its duration, scope, and implications for development.

8.3 Measurement and Analysis

The learning objective should guide the selection of measures. In case-based reasoning, assessments may examine how learners select evidence, explain causal relationships, evaluate alternatives, and revise recommendations. Interaction measures may track independent attempts, requests for explanations or completed answers, and verification of AI-generated claims. These behaviours are best interpreted using clear denominators and comparable opportunities: a learner who verifies two of three questionable claims has acted differently from one who verifies two of twenty. It is also useful to record the accuracy

of generated content, as identifying an error requires a different skill from interpreting sound advice. When describing the learning environment, include the model version, settings, access dates, and any significant changes made during data collection.

Performance and forecasting measures should provide sufficient range and reliability for their intended purposes. A four-item probe may be appropriate for brief feedback; however, it may not provide adequate precision for individual-difference or mediation analyses. Stronger claims about capability require adequately sampled tasks and scoring methods that reflect the reasoning being assessed. Confidence, trust, perceived understanding, and demonstrated competence are best treated as distinct constructs. Existing metacognition scales may offer useful evidence; however, adapting them for another language or population requires validation (Flake et al., 2017). For example, Lu et al. (2026) identify translation as an unresolved issue. Self-report and behavioural measures may be considered together without assuming that they constitute a single composite.

Analyses should focus on prespecified contrasts and the uncertainty surrounding the effects of educational interest. Sample-size justifications may be based on the smallest effect worth detecting, the expected precision, or a sensitivity analysis (Lakens, 2022). Interactions and indirect effects require their own planning. Missing follow-up data, noncompliance and deviations in AI access should also be reported. Primary analyses should reflect allocation, with sensitivity analyses clearly identified as such. A nonsignificant result may constrain a claim if its confidence interval rules out a meaningful benefit; however, a wide interval indicates greater uncertainty (Lakens, 2017). Exploratory process models may suggest explanations for future research, provided they remain clearly distinguished from comparisons specified before the outcomes were known.

9. Discussion

9.1 Theoretical Implications

The review shows how AI-assisted learning design connects to the evidence needed to understand its effects. Research on feedback clarifies what information learners receive. Explanations of monitoring and control help explain how this information may influence their decisions, while research on offloading and constructive learning examines which tasks learners continue to perform and practise. Assessment then determines which outcomes are observable. Considering these relationships collectively helps explain why improved performance with assistance may coincide with an unchanged learning score, or why increased use of assistance may sometimes support independent performance. To interpret either result meaningfully, describe the intervention, learner activity, and assessment criterion.

This approach also gives mixed findings a clearer and more substantive role in theory development. The benefits of structured tutoring challenge explanations that treat AI assistance solely as a substitute for learners' cognitive work, while the benefits of continuous access prompt reconsideration of the assumption that restrictions are invariably protective. The selected evidence does not establish

calibration as a necessary mediator of learning benefits; the precision, timing and coverage of monitoring measures also limit this inference. An experimentally time-matched comparison is needed to determine whether reconstruction offers benefits beyond additional learning time. A useful theory should accommodate these comparisons by specifying conditions in advance, thereby allowing the evidence to inform its predictions meaningfully.

The theoretical contribution offers a conditional explanation of when information about learning can be used in practice. The framework distinguishes among difficulties in interpreting feedback, acting on information once it is understood, and developing accurate, relevant reasoning. These challenges can lead to similar overall outcomes, but each may call for a different instructional response. More feedback may be insufficient when learners lack a workable strategy, and more demanding activities may be ineffective if they reinforce inaccurate reasoning. This distinction also helps explain why improved calibration does not always lead to better learning. The explanation remains provisional: it would be less convincing if distinguishing feedback information from reconstruction requirements added little explanatory value beyond differences in instructional content, exposure, time, or motivation.

9.2 Educational Implications

In educational practice, it helps to start by identifying the intended learning outcome. An activity may be designed to improve work completed with assistance, develop independent skills, or achieve both objectives. Students benefit from understanding this goal, and tasks should provide opportunities to practise and demonstrate the relevant skills. A single set of access restrictions is unlikely to be appropriate for all learners. A learner still developing prerequisite knowledge may benefit from an explanation before attempting a problem, while another may be ready to develop and defend a solution independently. When aligned with the learning objective, brief unaided assessments and focused reconstruction can help demonstrate progress, with support tailored to the task and the learner.

Feedback is most valuable when grounded in available evidence. A recorded action can document how AI was used, while a test score can indicate performance on the assessed tasks. Neither, on its own, supports a broad judgment about a student’s cognitive dependence. Reconstruction can make a student’s reasoning easier to evaluate, but it is useful only when the resulting work is relevant and accurate. Novices may also benefit from examples or continued access to sources as they work. These considerations favour learning activities with clear, observable purposes and access arrangements tailored to the task and learner.

9.3 Limitations and Further Research

The review has several limitations related to its search and selection processes. Web-based searches tend to favor sources that are readily accessible and well indexed. In contrast, the selected search terms and the focus on English-language publications may have excluded relevant work described using different terminology or published in other languages. The focus on text-generating systems also limits the findings' applicability to other modalities. Some recent results remain provisional, and for several peripheral sources, claims relied on primary abstracts or publication records because information was

limited. The review did not include comprehensive database coverage, duplicate screening, or a formal risk-of-bias assessment. Accordingly, the synthesis presents a focused conceptual argument, while the completeness and balance of the underlying sources remain uncertain.

This framework focuses on adult learning contexts in which learners are expected to perform independently of AI at a later stage. Much of the educational evidence reviewed is drawn from university students completing short, structured tasks; therefore, further research is needed to assess its applicability to experienced professionals, workplace learning, other languages, and sustained AI use (Simons et al., 2017). In these contexts, prior knowledge, incentives, and the consequences of errors may influence how feedback and reconstruction function.

The empirical literature still leaves several relationships open to further investigation. Few selected studies examine the target of feedback, measure the reasoning learners actually employ, and assess delayed, unaided transfer within a single design. Many interventions combine additional instruction with changes in access or activity, while brief tasks and measures of perceived capability provide only limited insight into development. Dai's (2026) factorial writing study offers a close precedent for integrating feedback-literacy and calibration support. However, the effective-adoption values reported in Table and section 4.4 require clarification before detailed process estimates can be interpreted. These overlaps matter for the present framework: combining metacognitive supports or using a factorial design alone is insufficient to establish originality.

A broader synthesis could help assess how much the present conclusions depend on source selection. Conducting database searches, identifying multiple reports from the same study, and independently verifying critical data extraction would provide a stronger evidentiary foundation. It would be particularly valuable to examine provisional publications, immediate outcomes, and interventions that include additional instructional content separately. The framework offers distinctions for these comparisons, although the findings may also suggest ways to refine it. Further longitudinal research is needed to determine whether repeated patterns of assistance use result in lasting changes in capability beyond the brief learning episodes represented in much of the selected literature.

10. Conclusion

Learning with GenAI depends, in part, on how its assistance shapes learners' work. Feedback can help learners assess their capabilities or reconsider which tasks to delegate, while reconstructing an answer can provide valuable practice in reasoning that a generated response might otherwise displace. Neither contribution can be assessed solely on the finished work. By linking feedback and learner activity to clearly defined assessments, the proposed framework specifies comparisons for testing which approaches support unaided performance and through which processes. To determine whether short-term differences lead to lasting improvements in independent capability, evidence of persistence and transfer will also be necessary.

References

1. Barnett SM, Ceci SJ (2002) When and where do we apply what we learn? A taxonomy for far transfer. Psychol Bull 128(4):612–637
2. Bastani H, Bastani O, Sungu A, Ge H, Kabakcı Ö, Mariman R (2025) Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, *122*(26), e2422633122
3. Bisra K, Liu Q, Nesbit JC, Salimi F, Winne PH (2018) Inducing self-explanation: A meta-analysis. Educational Psychol Rev 30(3):703–725
4. Bjork RA, Dunlosky J, Kornell N (2013) Self-regulated learning: Beliefs, techniques, and illusions. Ann Rev Psychol 64(1):417–444
5. Buçinca Z, Malaya MB, Gajos KZ (2021) To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, *5*(CSCW1), 1–21, Article 188
6. Carless D, Boud D (2018) The development of student feedback literacy: Enabling uptake of feedback. Assess evaluation High Educ 43(8):1315–1325
7. Chi MT, De Leeuw N, Chiu M-H, LaVancher C (1994) Eliciting self-explanations improves understanding. Cogn Sci 18(3):439–477
8. Chi MT, Wylie R (2014) The ICAP framework: Linking cognitive engagement to active learning outcomes. Educational Psychol 49(4):219–243
9. Dai Z (2026) Rethinking AI-assisted writing instruction: feedback literacy scripts, calibration training, and student writing development. Front Psychol 17:1829268
10. Dunlosky J, Rawson KA, Marsh EJ, Nathan MJ, Willingham DT (2013) Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. Psychol Sci Public interest 14(1):4–58
11. Fan Y, Tang L, Le H, Shen K, Tan S, Zhao Y, Shen Y, Li X, Gašević D (2025) Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. Br J Edu Technol 56(2):489–530
12. Fan Y, Yan L, Molenaar I, Greiff S, Gasevic D (2026) A metacognitive approach to learning and performance in human-AI interaction. Commun Psychol 4:150
13. Fernandes D, Villa S, Nicholls S, Haavisto O, Buschek D, Schmidt A, Kosch T, Shen C, Welsch R (2026) AI makes you smarter but none the wiser: The disconnect between performance and metacognition. Comput Hum Behav 175:108779
14. Fischer M, Rau HA, Rilke RM (2025) *AI tutoring enhances student learning without crowding out reading effort (IZA Discussion Paper No. 18338)*. IZA Institute of Labor Economics
15. Fisher M, Goddu MK, Keil FC (2015) Searching for explanations: How the Internet inflates estimates of internal knowledge. J Exp Psychol Gen 144(3):674–687

16. Flake JK, Pek J, Hehman E (2017) Construct validation in social and personality research: Current practice and recommendations. Social Psychol personality Sci 8(4):370–378
17. Gao L, Zhang F (2026) Cognitive offloading and metacognitive calibration in generative AI-mediated doctoral learning: A grounded theory study of accountable reconstruction. Front Psychol 17:1920947
18. Hattie J, Timperley H (2007) The power of feedback. Rev Educ Res 77(1):81–112
19. Imai K, Keele L, Tingley D (2010) A general approach to causal mediation analysis. Psychol Methods 15(4):309–334
20. Kestin G, Miller K, Klales A, Milbourne T, Ponti G (2025) AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting. Sci Rep 15(1):17458
21. Lakens D (2017) Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. Social Psychol personality Sci 8(4):355–362
22. Lakens D (2022) Sample size justification. Collabra: Psychol 8(1):33267
23. Lee H, Stinar F, Zong R, Valdiviejas H, Wang D, Bosch N (2025) *Learning behaviors mediate the effect of AI-powered support for metacognitive calibration on learning outcomes.* Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems
24. Lu H, Chen R, Wei R (2026) Human-Centered AI Metacognition: Development and Validation of a Scale for Preserving Cognitive Autonomy in Human–AI Collaboration. J Intell 14(10):270
25. Ma S, Wang X, Lei Y, Shi C, Yin M, Ma X (2024) *Are you really sure? Understanding the effects of human self-confidence calibration in AI-assisted decision making* Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems
26. Maier S, Schwabe K, Schneider M, Feuerriegel S (2026) Designing Against Deskilling: Metacognitive Feedback Reduces Cognitive Offloading to LLM Assistants. *arXiv preprint arXiv:2609.20143*
27. Maier U, Seibold M, Klotz C (2026) LLM-generated formative feedback in education: A qualitative systematic literature review. Computers Educ Open 10:100374
28. Nakamura D (2026) Augmentation or erosion? Regulating cognitive offloading in generative AI-supported STEM learning. Educational Psychol Rev 38(1):124
29. Nelson TO, Narens L (1990) Metamemory: A Theoretical Framework and New Findings. In: Bower GH (ed) Psychology of Learning and Motivation, vol 26. Academic, pp 125–173
30. Ngai C, Gilbert SJ (2026) Metacognitive training facilitates optimal cognitive offloading. Cogn Research: Principles Implications 11(1):21
31. Panadero E (2017) A review of self-regulated learning: Six models and four directions for research. Front Psychol 8:422
32. Post C, Sarala R, Gatrell C, Prescott JE (2020) Advancing theory with review articles. J Manage Stud 57(2):351–376
33. Risko EF, Gilbert SJ (2016) Cognitive offloading. Trends Cogn Sci 20(9):676–688

34. Roediger HL, Karpicke JD (2006) Test-enhanced learning: Taking memory tests improves long-term retention. Psychol Sci 17(3):249–255
35. Salomon G, Perkins DN, Globerson T (1991) Partners in cognition: Extending human intelligence with intelligent technologies. Educational researcher 20(3):2–9
36. Schraw G (2009) A conceptual analysis of five measures of metacognitive monitoring. Metacognition Learn 4(1):33–45
37. Shang X (2026) The psychology of learning with generative AI in foreign language education: building a theory of metacognition, trust, and human–AI co-regulation through a systematic review. Front Psychol 17:1974520
38. Shen JH, Tamkin A (2026) How AI impacts skill formation. *arXiv preprint arXiv:2601.20245*
39. Shute VJ (2008) Focus on formative feedback. Rev Educ Res 78(1):153–189
40. Sidra S, Mason C (2026) Generative AI in human-AI collaboration: validation of the collaborative AI literacy and collaborative AI metacognition scales for effective use. Int J Human–Computer Interact 42(7):5084–5108
41. Simons DJ, Shoda Y, Lindsay DS (2017) Constraints on generality (COG): A proposed addition to all empirical papers. Perspect Psychol Sci 12(6):1123–1128
42. Snyder H (2019) Literature review as a research methodology: An overview and guidelines. J Bus Res 104:333–339. https://doi.org/10.1016/j.jbusres.2019.07.039
43. Soderstrom NC, Bjork RA (2015) Learning versus performance: An integrative review. Perspect Psychol Sci 10(2):176–199
44. Sparrow B, Liu J, Wegner DM (2011) Google effects on memory: Cognitive consequences of having information at our fingertips. *science*, *333*(6043), 776–778
45. Stadler M, Bannert M, Sailer M (2024) Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. Comput Hum Behav 160:108386
46. Tankelevitch L, Kewenig V, Simkute A, Scott AE, Sarkar A, Sellen A, Rintel S (2024) *The metacognitive demands and opportunities of generative AI.* Proceedings of the 2024 CHI conference on human factors in computing systems
47. Torraco RJ (2016) Writing integrative literature reviews: Using the past and present to explore the future. Hum Resour Dev Rev 15(4):404–428
48. Wang G, Wang W, Yang D, Ren J (2026) Generative AI, Cognitive Offloading, and Learner Agency in Higher Education: A Scoping Review. Behav Sci 16(7):1150
49. Wong SSH, Qiu SX (2026) Think first, ChatGPT later: Guiding human–AI collaboration for learning gains in independent human creativity. Educational Psychol Rev 38(1):45
50. Xia Q, Liu Q, Tlili A, Chiu TKF (2026) A systematic literature review on designing self-regulated learning using generative artificial intelligence and its future research directions. Comput Educ 240:105465. https://doi.org/10.1016/j.compedu.2025.105465

51. Zhu Q, Li X, Dong Y, Chang P, Fan M (2026) Not all cognitive offloading is equal: distinguishing dependent and autonomous offloading to generative AI. Front Psychol 17:1878629

Unsectioned Paragraphs

Note. This reporting framework summarises the conceptual distinctions developed in the review; it has not been validated as an appraisal instrument.

Note. * Preprint or working paper. Assessment independence, delay, and transfer are considered separately.

Note: This comparison situates the specific analytical contribution developed here within established accounts of metacognition, offloading, and constructive learning.

Note. These predictions concern future research. An observed indirect association is insufficient to establish causal mediation without additional assumptions or experimental evidence.

Figures

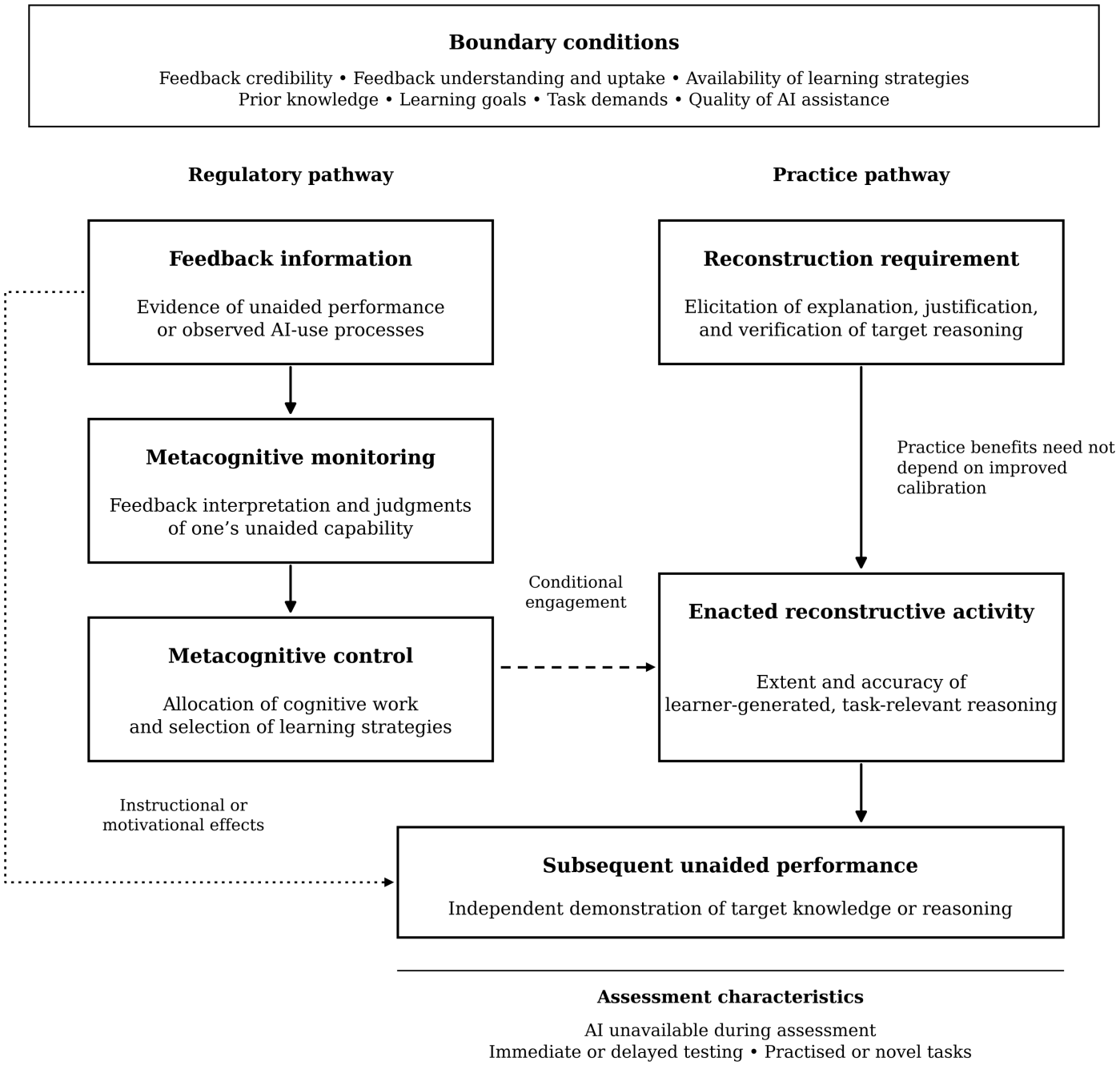


Figure 1

Proposed regulatory and practice pathways linking feedback and reconstruction to subsequent unaided performance