

From AI-Assisted Performance to Independent Thinking: A Framework for Developing Higher-Order Thinking with Generative AI in Higher Education

Chen Wang^{1*}

¹ Jiangxi Fenglin College of Foreign Economy and Trade, Jiujiang, Jiangxi, China

Accepted

2026-10-15

Keywords

Generative artificial intelligence;
Higher-order thinking; Human–AI
collaboration; Scaffolding;
Self-regulated learning; Transfer of
learning

Corresponding Author

Chen Wang

Copyright 2026 by author(s)

This work is licensed under the
CC BY 4.0



<https://doi.org/10.70693/itphss.v3i5.701>

Abstract

In an AI-supported task, a student may explain a weak comparison once it has been identified, yet fail to notice a similar problem without a reminder. A decision about further help therefore needs evidence about what the student did without being prompted. We examine such decisions in English-medium university argumentation and evidence evaluation. The framework brings together English-language work on scaffolding, learner-generated reasoning, metacognitive calibration, and cognitive offloading. We distinguish help that supplies a reason from help that initiates a check. Readiness to work without the first does not establish readiness to work without the second. Coding examples and boundary cases make evidence about these functions interpretable, while six conditional propositions connect reconstruction and regulation to adaptation and transfer. A proposed experiment gives every group comparable diagnostic practice but varies whether teachers use its results to adjust assistance. This comparison isolates the value of using the evidence from learning through the diagnostic task itself. We relate the account to four GenAI approaches and illustrate it with a hypothetical sociology lesson. Help may be retained or restored when a new task requires it. The framework and coding rules need validation, particularly beyond the language and disciplinary settings considered here.

1. Introduction

Imagine two undergraduates writing reports on a university learning-support programme with generative artificial intelligence (GenAI). Both recommend expanding the programme and present a persuasive case. The teacher then asks them to reconsider the recommendation in light of different starting scores for participants and non-participants. One student revises the causal argument; the other repeats the original explanation. This hypothetical contrast captures the problem addressed here: similar assisted reports may conceal quite different capacities to reason with the evidence.

The distinction between performance during practice and learning that lasts or transfers is well established (Soderstrom & Bjork, 2015). GenAI makes it especially relevant to coursework because it can supply the argument as well as the information used to build it. Assessing the finished report is therefore only part of the task. We also need to know whether its author can later examine the reasoning and change it when the evidence changes.

Fan et al. (2025) illustrate this assessment problem. In a laboratory writing study with 117 university students, ChatGPT assistance produced greater essay improvement, but differences between conditions in knowledge gain and transfer were not statistically significant. The product and the independent measures told different stories. These results establish neither an absence of learning nor equivalence between conditions, and the bounded study cannot show whether the pattern would persist over a semester or in another discipline.

A different picture emerges in university physics. Kestin et al. (2025) reported better learning with a research-designed AI tutor than with in-class active learning. This was a different tutor and teaching arrangement from the writing study, so the result cannot answer whether reasoning gains endure across domains. What the comparisons do clarify is why assistance design and the outcome chosen for assessment both matter.

The move from practice to unaided work changed the result in Bastani et al. (2025). Over four mathematics sessions at one Turkish high school, unrestricted AI assistance improved practice performance; students subsequently performed worse without it. A guarded tutor mitigated the adverse effect. For our argument, the useful contrast is between help that leaves a step to be practised and help that completes it. These were short-term school assessments, not tests of our mechanisms or of lasting university argumentation.

Shared control need not end for learning to continue, as Molenaar's (2022) account of human–AI learning makes clear. Bearman et al. (2024) focus on students' evaluative judgement, while Vendrell and Johnston (2026) address pedagogical support for critical thinking. We build on these concerns at the next instructional decision: what reasoning can this student now undertake, and what evidence warrants a change in help? Section 7.1 compares that emphasis with four GenAI approaches.

AI may contribute both to an assignment and to the reflection offered as evidence of understanding. That reflection can be a legitimate collaborative product while telling us little about what the student could do independently. The difficulty is one of instructional inference, not necessarily misconduct. Kane (2013) argues that assessment evidence must warrant its proposed interpretation and use. A report that demonstrates successful collaboration may still be insufficient grounds for withdrawing reasoning support.

We ask how assisted performance differs from independent higher-order thinking, how cognitive and regulatory activity might connect them, and what should guide changes in support. The framework follows cognitive responsibility, learner reconstruction, support adjustment, and transfer across tasks. Its six propositions specify how evidence about particular learner functions could inform those decisions; they remain proposals for investigation.

2. Conceptual Approach and Scope

2.1. Theoretical synthesis and model development

We develop the account through focused theoretical synthesis (Jaakkola, 2020). Scaffolding provides a way to connect present success with help to later use of the reasoning. Generative processing and self-regulated learning explain what students might do within that support. Cognitive offloading identifies a possible obstacle: the activity handed to AI may be the one being

learned. Together, these perspectives pose an assessment question: what evidence is sufficient to justify changing assistance?

The model builds on contingent scaffolding and learner-generated reasoning rather than proposing new versions of either. Its contribution is to connect evidence about a particular cognitive function to decisions about content and regulatory support. The proposed test holds diagnostic practice constant so that using the evidence, rather than simply collecting it, becomes the contrast. Longer study, extra retrieval practice, and better information access remain competing explanations for any benefit.

Our literature selection was organised around this explanation rather than a comprehensive search. We prioritised 2023–2026 publications, retaining earlier work where it supplied a theoretical foundation. Checks of publication records, available abstracts, and relevant passages ran through 7 October 2026. GenAI studies helped establish the limits of application. The four approaches discussed in Section 7.1 were chosen for their relevance to judgement, teaching, or assessment in higher education; they locate the proposal without representing the entire field.

This is a focused synthesis, not a systematic review. It reports no comprehensive search, independent screening, study-quality scores, or new empirical sample.

2.2. Language, population, and task boundaries

The language of the literature and the language of instruction are separate boundaries. We draw on English-language publications and primarily address English-medium university tasks involving evidence evaluation, argument, and revision. Some cited studies took place in other instructional settings or with younger learners. They supply theoretical resources and comparisons, not direct validation in our target setting. Non-English scholarship and local pedagogical traditions are not systematically represented.

The unit of analysis is a learner–task episode within a sequence of university work. Humanities and social-science assignments fit the model most directly because students must make the reasons for a judgement explicit. STEM tasks may also fit when they involve explaining a model, examining causal assumptions, or defending a solution strategy. Numerical accuracy, routine symbolic manipulation, and code completion would require additional indicators; they cannot be treated as equivalent to argumentation.

A critique, design decision, or evaluation of constraints may offer an application in creative writing, art, or design. The claim would concern that judgement, not aesthetic achievement, originality, embodied making, or writer's voice. These remain possible applications rather than demonstrated transfers.

The clearest setting for this model is a course in which students must eventually perform a specified reasoning function without AI performing it for them. Reference materials and accessibility resources can remain. We follow the individual's developing capability, while recognising the contribution of teachers, peers, and institutions. Memorisation-only exercises and purely procedural drills are excluded. Multilingual teaching and other disciplinary standards raise further questions about whether the same account holds.

3. Conceptual Distinctions and Theoretical Foundations

3.1. Performance, independent thinking, and cognitive responsibility

Independence has to be specified relative to a task; it is not a general label for a student. We must identify the reasoning function and resources available. AI-assisted performance describes quality, accuracy, or efficiency with documented help. Independent higher-order thinking instead concerns the learner's capacity to generate, evaluate, and revise reasoning without AI providing

the target inference or directing essential monitoring.

Typing an argument personally leaves open who noticed that a claim needed checking; AI may have supplied that judgement. By contrast, a data table or accessible reading tool can leave evaluation entirely to the learner. Bearman et al. (2024) treat evaluative judgement as deciding the quality of one's own and others' work. For our purposes, the decisive question is whether the student can exercise it beyond the supported episode. The origin of the keystrokes cannot answer that question.

Cognitive responsibility identifies who proposes a reason, checks evidence, resolves a contradiction, or initiates revision. It describes contributions to a task, not a moral quality or a trait defined by test scores. Those contributions can be examined and compared with independent performance. Learner-generated reasoning need not be academically original: reconstructing a familiar relation can be meaningful, whereas reproducing AI wording alone does not demonstrate that work.

Table 1. Core constructs and function-relative evidence boundaries

Panel A. Working definitions		
Construct	What to observe	What is not enough
Assisted performance	Work quality under recorded AI-support conditions	Treating the product score as mastery
Independent higher-order thinking	Target reasoning and essential monitoring without AI supplying either	Avoiding tools or recalling the wording
Substantive learner reconstruction	A learner-performed inference about a claim or its warrant	Paraphrase, polish, or unsupported agreement
Metacognitive calibration	Task-linked understanding judgements against demonstrated capability	Confidence or self-reported reflection alone
Learner-attributable evidence	An observation supporting a bounded claim about a learner function	A claim of entirely human authorship
Function-specific support	Content assistance and regulatory prompts recorded separately	Combining prompts or time into one dependence score
Panel B. Boundary cases for a causal-argument task		
Activity	What may be attributed	What it does not show alone
Using a reference book or AI to locate passages	The later evaluation, if the resource does not supply the target inference	Independent searching when AI did the search
AI ranks sources or explains their causal relevance	The learner's additional, demonstrably reconstructed reasoning	Independent source judgement from repeating the synthesis
AI names a confounder or drafts the reflection	A later learner explanation under documented conditions	Spontaneous detection or mastery from the supplied reflection

Note. Attribution depends on the learning target and how the evidence was produced. Access to resources does not rule out independence, just as permission to use AI does not establish it. Section 4.2 sets out the reconstruction codes.

3.2. Scaffolding and the adjustment of assistance

Scaffolding explains how students can work beyond their current unaided capability (Wood et al., 1976). Contingency, fading, and the transfer of responsibility are central to this account (van de Pol et al., 2010). Our extension concerns the decision to change help: what evidence shows that the learner is ready to assume the particular function being supported?

Help has to leave something for the student to learn, yet a student may need to see a worked step before being able to practise it. Koedinger and Aleven (2007) describe this as the assistance dilemma. Prior knowledge changes that decision: Tetzlaff et al.'s (2025) expertise-reversal meta-analysis supports adaptive assistance, with less conclusive findings in humanities and language learning. Our framework therefore permits initial modelling, continued guidance, and renewed help with unfamiliar tasks. It also permits ongoing shared control, as in Molenaar (2022); retaining a useful tool is not itself evidence that learning failed.

3.3. Generative processing and its limits

Generative learning offers a rationale for asking students to work out relations among ideas (Fiorella, 2023), while ICAP connects forms of engagement with learning opportunities (Chi & Wylie, 2014). Neither makes a long interaction log proof of substantive thought. Thurn et al. (2023) challenge assumptions about how observable behaviour maps onto cognitive activity. We therefore examine what reasoning an exchange contains, rather than treat dialogue or an interactive interface as evidence that reconstruction occurred.

Reconstruction might involve explaining why a conclusion follows, testing a premise, or changing an argument after a counterexample. A worked explanation can prepare students for this activity; reading is not inherently unproductive. Equally, a request to write an explanation can elicit little more than compliance. The distinction concerns the reasoning performed, not the number of required interface actions.

3.4. Calibration, regulation, and target-relative offloading

The question is not only whether a prompt improves a response, but whether the student eventually starts the check. Guan et al. (2025) review chatbot support for goals, monitoring, strategies, and reflection, including nongenerative systems. Monitoring accuracy can improve, although the intervention matters (Janssen & Lazonder, 2024). In Xu et al.'s (2025) four-week university quasi-experiment, metacognitive prompts improved aspects of task strategy and self-evaluation without a significant achievement difference between groups. These studies motivate, rather than settle, the question of self-initiated checking.

Recognising a gap is not the same as knowing how to resolve it. A learner may doubt an explanation yet lack a reliable source to consult. Regulation should make an appropriate response possible: accepting a well-supported recommendation, questioning a weak one, or seeking help. Tankelevitch et al. (2024) discuss the planning, checking, and revision involved in GenAI use. Cues that initiate these activities differ from explanations or solution steps that supply content, and a student may need quite different amounts of each.

Cognitive offloading changes a task's demands through external resources or actions (Risko & Gilbert, 2016). We propose that its educational value depends on the delegated activity and the use made of the capacity released. Language polishing may free attention for evidence evaluation in one course and replace the intended writing practice in another. Even peripheral offloading offers little benefit if the learner disengages. The relevant relation is therefore among delegation, the learning goal, and subsequent activity.

3.5. Evidence attribution as an instructional inference

Learner-attributable evidence supports a claim about a student's performance of a specified function. It does not estimate how much text was human-written. Following Kane (2013), the issue is what interpretation and decision an observation can justify. AI may locate passages while the student evaluates their causal relevance. If AI supplies the causal analysis too, repeating it gives different evidence. A later task may establish that the relation has been learned, but the assisted response alone cannot settle the question.

To interpret an episode, record the target function and available resources alongside what AI supplied and what the learner did. A mixed contribution leaves understanding uncertain; it does not settle it negatively. Before adjusting help, consider language, interface difficulty, anxiety, and familiarity with the task. Section 6.3 sets out ways to obtain further evidence without turning a learning error into a misconduct judgement.

4. A Process Framework for Developing Independent Higher-Order Thinking

4.1. Establishing learning goals and cognitive responsibilities

The instructor first needs to decide what students should learn to do. A strong assignment and the capability to reason behind it are related, but they are not the same aim. Specifying the target function, prerequisite knowledge, and permitted resources makes the division of work intelligible. Without that basis, calling a student dependent says little: an AI action may be incidental help in one assignment and a substitute for its central learning activity in another.

In the learning-support example, students must judge a causal claim and reconsider a recommendation when its assumptions change. AI could summarise the supplied documents or suggest alternatives. With suitable guidance, students would identify the claim and explain what the evidence warrants. Making this responsibility explicit does not require every learner to solve the entire problem alone before receiving help.

A prerequisite check might show that one student still needs an example of association versus causation, while another is ready to attempt an ambiguous case. Either starting point can fit the framework. Productive failure involves designed problem solving followed by instruction, not leaving difficulty unresolved (Sinha & Kapur, 2021). We therefore prescribe no fixed student-first, AI-second order. The sequence should help students acquire reasoning they can later use independently.

Support can change again when the task changes. A student who evaluates one kind of study confidently may need modelling for another. Readiness concerns a function in the current task, not a permanent label of independent thinker.

4.2. Reconstructing and evaluating AI-supported reasoning

Substantive reconstruction establishes, tests, or modifies a relation that matters to the conclusion. A justified agreement may qualify; an unsupported disagreement does not. Rewording an explanation is also insufficient. Reading a good explanation may support learning, but reading alone provides no observable reconstruction to code. Box 1 shows how these distinctions can be applied.

Box 1. Proposed coding rules and worked examples

Use a claim–reason episode as the coding unit. It includes the relevant AI output, the student's response or revision, and any associated consultation of sources. An episode may span several turns; a sentence is not necessarily a complete unit. First identify the inferential move. Then record its quality, provenance, and degree of prompting separately. The examples below concern a hypothetical AI recommendation to expand a support programme because its participants scored

higher than non-participants.

A — Assumption analysis. “The comparison assumes the groups were similar before enrolment. But students chose whether to join.” This counts because the response connects an unstated premise to the comparison being made. “AI made assumptions” does not identify the relevant relation.

E — Evidence–conclusion evaluation. “We have a score difference, but no baseline measure. The attainment gap could have been there already.” The student explains why the evidence leaves a question unresolved. Copying the scores, however accurately, does not perform that evaluation.

C — Counterexample or alternative explanation. “More motivated students might both join the programme and study longer. That would produce the same pattern even without a programme effect.” The response offers a plausible rival mechanism. Naming “motivation” without explaining its role is only a partial attempt.

V — Evidence-based revision. “I would keep the association claim, but drop the claim that the programme has been shown to cause the improvement. We need a stronger comparison.” Code revision when the change is linked to an identified evidential limitation. Replacing “proves” with “suggests” without giving a reason remains surface editing.

Exclusions and separate ratings. First ask whether the response makes an inferential move. An incorrect move can still be a reconstruction attempt, so correctness is judged afterward. Paraphrase, polishing, and agreement or rejection without a reason do not qualify. For a qualifying move, retain the three quality ratings: invalid or unsupported, partially justified, or adequately justified.

Record its provenance as student-added, reconstructed after a prompt, reproduced from AI, or uncertain. Several different moves may occur in one episode; restating the same reason does not add depth. This keeps the occurrence of reconstruction separate from its correctness and source.

A response may repeat a relation that AI supplied several turns earlier. Coders therefore need the surrounding exchange, not just the student's last message. Anchor examples and piloted episode boundaries should guide coding; a prespecified subset should be double-coded. Report initial agreement and the resolution of disagreements, then compare the process codes with independent tasks. Where a log is missing, retain the uncertain category rather than infer that no thinking occurred. These are coding proposals, not validated rules.

Fault-finding is a poor task for a novice who lacks the knowledge or sources needed to judge a claim. It can reward confident guessing. Reconstruction needs checked material and corrective feedback. Steiss et al. (2024), comparing feedback in secondary-school writing, caution against assuming equivalent quality from human and GenAI sources. Following successive judgements reveals how the student's reasons change; an end-of-task reflection cannot recreate an undocumented sequence.

4.3. Adjusting content and regulatory support

Content and regulatory assistance need not change together. Content help supplies a relation, example, explanation, or solution step. Regulatory help directs attention or initiates planning, checking, or revision without necessarily performing the analysis. Either may be reduced as the learner develops, and either may be restored for an unfamiliar problem.

The distinction is easiest to see in a causal-analysis task. “Prior motivation can cause both enrolment and higher scores, so the group difference need not be a programme effect” gives the learner the analysis. It is content support. “Check whether another variable could explain the comparison; give your reason” prompts a check but leaves the variable and its relevance for the student to work out. That is regulatory support.

A question can do both: “Could prior motivation explain enrolment and scores?” names a

candidate explanation while inviting evaluation. It should be coded as mixed support. The grammatical form of a prompt is not its cognitive function. Even a purely regulatory cue supplies help when noticing the need to check is itself the learning target.

Record the function of support, the target relation supplied, the specificity of the cue, and who initiated the action. A mixed message can be segmented or given both labels; record duration separately. P3 can then be tested without assuming that content and regulation act independently of each other.

Consider two students in a hypothetical tutorial. A reminder to check the comparison enables one to explain the assumption immediately; supplying another piece of the answer would miss the difficulty. The other notices the problem but needs an example to work through its causal implications. Both might receive one message. Yet a brief cue supplies essential monitoring for the first, while a longer explanation provides content for the second. Message counts alone would obscure this difference.

To decide whether causal-evaluation help can be reduced, ask the student to evaluate a new claim without an AI-supplied answer. To reduce monitoring prompts, observe whether checking begins without a reminder. Neither a single successful assisted report nor a failed probe should determine the decision.

A complete interaction log is not always needed for the teaching decision. An oral explanation, or a response to a changed assumption, may give a teacher useful evidence at lower stakes. Such checks should preserve accessibility support and remain proportionate to the decision. When responses disagree, Section 6.3 offers an interpretation process; averaging them into one score would hide the difficulty.

A diagnostic task may itself produce learning. P4 concerns a separate benefit: using its result to choose the next support. The proposed factorial study gives every group the same diagnostic opportunities so that these two effects can be distinguished.

4.4. Independent reconstruction, transfer, and renewed support

A later attempt can ask students to use the reasoning under a changed constraint, in a new setting, or after a delay. Simply removing help does not guarantee a learning benefit. Gonçalves et al. (2025) identify feedback as a relevant condition in comparisons of retrieval and elaborative study. Cao and Carvalho (2026) examine how initial instruction and task variability shape generalisation in non-GenAI experiments. This evidence supports investigating independent reconstruction under specified conditions, rather than treating every unaided exercise as training in general critical thinking.

Transfer claims must specify what changes. Barnett and Ceci (2002) distinguish changes in content and context that make one application more distant than another. Familiar-task success, delayed retention, near transfer, and more distant transfer may diverge. Remembering a rule does not ensure that a student recognises when to use it.

A second non-randomised support initiative offers a near-transfer task in the programme example. A change in examination practice could be more distant, with different evidence and constraints. Specify the shared structure and required knowledge. Failure could call for further teaching, different examples, renewed prompts, or a narrower transfer claim.

Independent performance feeds back into the next round of support. Help can increase when complexity increases. The aim is to develop reasoning functions that students can exercise and deliberately augment, not to separate students permanently from AI.

Table 2. Process framework and decision points

Process	What to observe	Support decision
---------	-----------------	------------------

Process	What to observe	Support decision
Specify goals and responsibilities	Target function and prerequisite knowledge	Allocate a feasible initial responsibility
Reconstruct and evaluate	Assumptions, warrants, counterexamples, revisions	Check whether the learner performs the reasoning
Adapt from attributable evidence	Diagnostic and process evidence for each function	Retain, reduce, or reinstate the relevant support
Reconstruct independently and transfer	Reasoning under specified new conditions	Establish its limits or renew assistance

Table 2 follows both the production of an assisted artifact and the development of capability across tasks. Their relationship is what the framework seeks to explain. A strong product cannot be assumed to indicate either substantial learning or its absence.

5. Theoretical Propositions and Boundary Conditions

All six propositions require a judgement about what improvement would be worth the effort. Before collecting data, researchers should specify a meaningful change in reasoning, support decisions, or teacher time. Confidence intervals and, where appropriate, equivalence tests can distinguish negligible effects from estimates too imprecise to resolve the question (Lakens et al., 2018). A universal standardised cutoff cannot make that educational judgement.

A statistically detectable effect may still be too small to justify changing teaching. A wide interval, on the other hand, may leave a worthwhile effect unresolved. Throughout this section, we separate those practical judgements from evidence that a mechanism operates. P3 and P5 concern differentiated effects or interactions, so neither requires a positive average gain. Content, feedback, prior knowledge, and learning time must be comparable, or their differences examined as competing explanations.

5.1. P1: Substantive learner reconstruction

P1. Given comparable content quality and learning opportunities, AI-supported activities that require students to generate, evaluate, and revise reasons will lead to stronger independent reasoning than activities in which AI supplies the reasoning and students mainly receive or restate it.

Being assigned to reconstruct an explanation does not mean that the learner worked through its reasoning. P1 needs evidence of that work. Fiorella (2023) supplies the generative-learning rationale, whereas Thurn et al. (2023) caution against reading cognition directly from observable engagement. The analysis should identify the inferential moves and their quality; word counts or the order of turns do not establish either.

Without prerequisite knowledge and feedback, generating an explanation may reinforce a misconception. A worked explanation could be more useful at that point. P1 therefore depends on giving students a feasible opportunity to assume the reasoning as learning proceeds.

P1 would be challenged if well-supported students demonstrably reconstructed the reasoning, yet a precise estimate ruled out a meaningful independent-learning advantage. A gain explained entirely by extra study time would favour that simpler account. The common practical-value criterion above still applies even when a processing effect is statistically detectable.

5.2. P2: Calibration linked to regulatory action

P2. AI-supported learning will foster greater regulatory independence when better

calibration of understanding leads students to initiate appropriate checking, help seeking, or revision than when monitoring remains dependent on prompts or does not guide action.

The important step in P2 comes after the judgement of understanding. Does a student who recognises a gap seek help, and can someone with adequate grounds for acceptance move on? Ngai and Gilbert (2026) link predictions and feedback to improved calibration and choices of external reminders in memory tasks. Their study provides a precedent for judgement guiding action. P2 asks whether this connection can support regulatory independence in GenAI coursework.

Elicit a judgement about a particular task, then observe action without a reminder. This sequence would be lost in a general reflection score. Analyses should separate calibration from achievement and establish that students had workable checking resources. Recognising a gap is of limited use when checking is unavailable or prohibitively costly.

Better calibration without appropriate self-initiated action, despite a feasible opportunity to act, would weaken P2. If checking ceases when reminders stop, prompt compliance is the more plausible account. Changes in confidence or checking frequency alone do not establish improved regulatory decisions.

5.3. P3: Differentiated content and regulatory support

P3. Withdrawing content assistance and withdrawing regulatory prompts will have distinguishable consequences. Successful reasoning without answer support does not, by itself, show that a student can initiate monitoring without reminders.

Jin et al. (2023) used design storyboards and interviews to explore how students envisaged AI support for cognitive and metacognitive regulation. P3 takes up a different question: what follows when each kind of help is removed? Solving a problem but overlooking an assumption differs from noticing the assumption without being able to explain it. A single dependence score would make those support needs look alike.

A useful test of P3 would manipulate content help and regulatory cues separately. It would measure spontaneous checking as well as reasoning accuracy, allowing for prior knowledge and task novelty. One possible result is that accuracy survives removal of answer help while checking ceases without reminders. That pattern would distinguish the functions, not establish a general judgement of independent thinking. No single order of support withdrawal is assumed.

Separate support functions would be less useful as an explanation if repeated, well-measured manipulations produced indistinguishable outcomes across learner profiles with adequate precision. Another result is possible: effects might differ detectably yet not enough to change a support decision. That finding would limit the distinction's educational value, even though the functions remain conceptually different.

5.4. P4: Learner-attributable evidence for adaptation

P4. When AI contributes materially to the reasoning used to infer mastery, a support policy that also uses learner-attributable diagnostic evidence will better match assistance to need and improve subsequent independent learning compared with a policy based on assisted artifact quality alone. The added evidence must be reliable and relevant to the decision.

The proposed mechanism in P4 is better matching of support to need. Diagnostic evidence should help retain assistance for functions not yet demonstrated and reduce it where the learner is ready. Its expected value rises when the assisted product reveals less about independent capability. The size of this information gap must be measured; AI involvement alone cannot establish it.

Compare two uses of AI. In one, it formats citations while the student builds the argument. In the other, it supplies the causal chain, counterarguments, and evaluation. We expect the first product to be more informative about the learner's reasoning. Assistance with core reasoning may

therefore make readiness judgements less dependable than assistance with presentation.

The contrast does not guarantee a loss of predictive validity. A skilled learner may learn from extensive assistance; a rubric dominated by presentation may miss reasoning even when no AI contributed to it. We use the comparison to identify where diagnostic information might be useful. It is not a prescription for extra testing in every case.

To test this prediction, pair assisted artifacts with several independent tasks addressing the same function. First fix the readiness thresholds; then use held-out students or tasks to test whether diagnostics improve prediction beyond artifact scores and prior knowledge. Some evaluation tasks must never enter the support policy. Otherwise, evidence used to choose help would also be used to declare the choice successful.

Measure the two consequential errors: withdrawing help too soon and retaining it unnecessarily. Calibration and prediction error can complement these measures, but correlation alone does not show that diagnosis improved a decision at an acceptable cost.

For P4, more informative diagnosis is only the first step. It must change the matching of help to need, and that change must support later independent learning under a reasonable policy. Either link could fail. Learning gains with no change in support decisions would suggest practice, attention, or another explanation. The timing and thresholds for diagnosis need local validation against both learning value and the burden imposed.

5.5. P5: Target-relative offloading and reinvestment

P5. The effect of cognitive offloading on independent learning will depend on whether the delegated activity is a learning target and whether students use the resources released for goal-relevant processing.

Formatting assistance may release time for evaluating an argument. Delegating evaluation can instead remove the target practice, unless the student subsequently reconstructs and tests it. Building on Risko and Gilbert (2016), P5 therefore predicts an interaction with the learning goal. It does not assign a fixed educational effect to a tool feature.

Classifying delegation requires an already specified learning goal. Researchers can then examine time allocation and reasoning to see whether released resources were reinvested, and assess independent outcomes. Reduced effort alone cannot establish that sequence. Failure of goal–delegation alignment to moderate learning in suitable tests would challenge P5. Benefits without a plausible redistribution of activity would favour access, reduced anxiety, or better information instead. The resources released need not be limited to effort.

5.6. P6: Reconstruction under changing conditions

P6. AI-supported instruction followed by independent reconstruction across relevant task variations will produce more robust transfer to specified new tasks than repeated success on similar assisted tasks, provided that students have the knowledge and feedback needed to make sense of the variations.

P6 concerns reconstructing reasoning and recognising where it applies, a narrower claim than improving critical thinking generally. Cao and Carvalho (2026) provide a conditional rationale for reconstruction and variation; Barnett and Ceci (2002) clarify what makes an application new. Retention, near transfer, and more distant transfer therefore remain separate outcomes.

Tests must address differences in time, feedback, practice, and task variation before attributing a transfer gain to the proposed mechanism. Precise evidence of no gain after successful implementation, or improvement limited to familiar wording, would challenge P6. Gains confined to closely related tasks would support only the corresponding narrower claim.

Variation can also overwhelm novices who cannot yet see the shared structure. Tasks should change in ways students can interpret with support. The proposal is neither steadily increasing

difficulty nor permanent withdrawal of help.

6. An Empirical Research Agenda

6.1. A focused factorial test of reconstruction and adaptation

The proposed 2×2 experiment separates two decisions. For the first, students either construct and revise reasons with bounded AI support or study a matched explanation through guided reading and comprehension activities. This contrasts responsibility for building the inferential relations, not thought with its absence. For the second, assistance is adjusted from artifact quality alone or from artifact quality together with learner-attributable diagnostics. P1 concerns the first contrast and P4 the second. The study would test these claims rather than the framework in its entirety.

All four groups receive comparable diagnostic tasks, timing, and feedback. Only the use of those results to guide help differs. Prespecify the content and regulatory actions available, their maximum intensity, and the decision rules; record what was actually delivered. Equal diagnostic opportunities make it possible to distinguish practice benefits from the value of adaptation.

Pilot a common reading period, a fixed window for explanation work, and matched feedback time. Early finishers could use prespecified ancillary reading or a filler activity until the next checkpoint. This material should provide neither target-reasoning practice nor new target concepts or corrective feedback. Record active reasoning, source reading, waiting, and filler time separately.

Equal clock time does not guarantee equal learning opportunities. Filler may alter engagement or substantive exposure. Where feasible, a replication should compare the time-balanced design with an active-reading alternative; otherwise, these differences remain possible explanations for the result.

Both groups need the same checked sources and comparably accurate information. Assignment alone cannot ensure the intended contrast: students may reconstruct an explanation they receive or copy one they are asked to reconstruct. Pilot coding and output constraints should check the distinction. Analysis should then follow the reasoning actually performed, rather than infer it from the condition label.

Measure effort and perceived load at matched points. Because they may mediate the intervention, automatically adjusting for post-treatment load could remove part of its total effect. Baseline knowledge, reading proficiency, interface familiarity, and pre-existing demands are separate concerns. Any secondary process analysis should state its causal assumptions.

The assisted artifact and later learning should have separate scores. A new, structurally related task completed independently after a delay could supply the primary outcome. Sample planning needs a prespecified meaningful effect, with provision for the factorial interaction, clustering, repeated measures, attrition, and measurement error.

P4 requires observations between diagnosis and the final outcome. The question is whether evidence changed support choices and improved their match to need, judged using tasks withheld from the policy. Sensitivity analyses should examine diagnostic reliability, the help actually delivered, and AI use outside the study. A difference in final scores alone would leave the evidence-to-decision-to-learning explanation untested.

6.2. Repeated-task investigation of regulation

A second study could follow checking and help seeking across successive assignments. Within each learner-task episode, align interactions, draft revisions, independent explanations, and responses to altered evidence. Brief stimulated-recall interviews may clarify why a suggestion

was accepted or rejected, but recollection should be scrutinised alongside the log rather than treated as a definitive account.

For repeated tasks, Box 1's record of inferential moves, quality, and provenance should also show who initiated each check. Establish coding reliability before linking these observations to independent outcomes. Multilevel analysis accommodates repeated episodes, but it cannot turn temporal order into a causal explanation. Where educationally appropriate, randomly withdrawing a particular prompt would provide a more direct test of the regulatory prediction.

A task asking students to find an error has already prompted monitoring. To test P3, leave room for students to notice that a check is needed. There must also be room for justified acceptance: careful reasoning need not involve an elaborate verification routine for every well-supported claim.

6.3. Measurement and interpretation

Table 3. Evidence layers and complementary interpretations

Evidence layer	Possible indicator	Limit of interpretation
Assisted artifact	Argument quality with recorded support	Shows the collaborative result, not mastery alone
Reconstruction	Inferential move, quality, prompting, and provenance	Behavioural evidence, not direct access to thought
Calibration and action	Task-linked judgement followed by checking or help seeking	Lower confidence need not mean better calibration
Adaptation decision	Specific support retained, faded, or reinstated	Separate probe practice from use of the result
Immediate independent capability	New reasoning without target-solving assistance	A task sample, not a global trait
Delayed performance and transfer	Retention and specified new-task outcomes	Keep transfer distances separate

A counterexample in a log may support the same interpretation as a reasoned revision and success on a new independent case. They corroborate different aspects of the same learner, task family, and function; they are not interchangeable observations. Give each weight according to relevance, provenance, reliability, and coverage. Collaborative work addresses collaborative competence, whereas an independent probe more directly samples the target function. Any combined score would require validation.

A mismatch between observations needs interpretation. After a strong assisted report and a weak independent response, retain provisional help and obtain another targeted observation. Weak writing alongside sound oral reasoning calls for checking language or response-mode demands. Good logs followed by poor delayed transfer may instead suggest task-bound learning or uncertain attribution. Repeated independent success should not be discounted merely because logs are missing.

Prespecify how conflicts will be handled: check provenance and scoring, examine accessibility and task fit, and collect an alternative low-stakes observation where feasible. Keep an uncertain category when the conflict remains unresolved. Report these profiles rather than conceal them in a total score.

The rubric should reward warranted conclusions, evidence–conclusion links, consideration of alternatives, and responses to changed assumptions. Language quality belongs in the score when

it is a learning target, not as a substitute for reasoning. Raters need anchor responses, blinding where feasible, and reported agreement. A combined score also needs a rationale and evidence for its intended interpretation.

Yan et al. (2024) identify transparency, reproducibility, and privacy as concerns in educational LLM research. For this study, an interpretable record would include the model and access date, prompts, support policy, permitted tools, and relevant configuration. Altered or missing logs should be disclosed. Unrelated personal material need not be collected.

Consent and data governance should fit the study. De-identified examples, a codebook, and policy specifications may allow scrutiny without publishing complete conversations.

7. Discussion and Implications

7.1. Theoretical contribution and dialogue with GenAI frameworks

The contribution centres on a support decision. A student may evaluate a weakness once it is identified but still need help to notice it. Evidence about those two functions should lead to different choices. Table 4 situates this emphasis alongside four GenAI approaches. Additional diagnosis is expected to be useful where it resolves a consequential uncertainty in the assisted artifact.

Table 4. Relationship to published GenAI higher-education approaches

Published approach	Contribution retained	Proposed extension
Evaluative judgement (Bearman et al., 2024)	Judging AI outputs and processes; using AI to assess student judgements	Attribute evidence of a judgement before adjusting support for that function
Critical-thinking design principles (Vendrell & Johnston, 2026)	Conditions for critical engagement and supported or unsupported activity	Specify which function changes, the evidence for the change, and its predicted effects
Refined AI Assessment Scale (Perkins et al., 2025)	Outcome-sensitive redesign and communication of AI arrangements	Adapt to individual readiness within the agreed AI arrangement
Assessment twins (Roe et al., 2026)	Complementary evidence for shared outcomes and validity-focused cross-checks	Use the evidence to guide successive support decisions; separate policy gains from probe-practice gains

Note. These are conceptual accounts and pedagogical or assessment-design frameworks, not four validated causal models. Molenaar's (2022) broader account of hybrid learning is an antecedent rather than a GenAI-specific higher-education comparator.

Bearman et al. (2024) make defensible quality judgements an educational aim. Our question arises when such a judgement appears in assisted work: did the student also recognise that it was needed? Evaluating an identified limitation provides different evidence from noticing it unaided. Changes to content help and regulatory prompts may therefore require different grounds.

Vendrell and Johnston (2026) place AI-supported and AI-free activity within cognitive and epistemic pedagogy. P3 examines what a teacher changes between those activities. Once a student can explain a supplied counterargument, the teacher may withdraw it while retaining a cue to look for one, withdraw both, or keep both. Transferring responsibility is the shared purpose. We specify the evidence needed to decide which of these functions the learner is ready to take on.

An assignment-wide AI rule cannot by itself establish each student's readiness. The refined AI Assessment Scale links redesign and communication to learning outcomes (Perkins et al., 2025). Within that common arrangement, one student may construct an argument unaided while another still needs a cue to begin checking. Our model concerns this variation in the help students need, rather than a different rule about permitted AI use.

Assessment twins provide the closest comparison. Roe et al. (2026) pair components addressing shared outcomes so that complementary evidence can be cross-checked. We build on that logic rather than claim paired assessment as new. The extension is to use the evidence when choosing subsequent help. P4 isolates this decision by holding diagnostic practice constant and varying whether its results guide assistance. Better matching and later learning must then be explained beyond the benefit of completing another assessment.

The claim lies in this combination of function-specific attribution and a testable support policy. It must add value beyond prior knowledge, support intensity, and existing assessment designs. The coding rules and boundary cases make that added value open to examination.

7.2. Instructional application and assessment boundaries

A course need not begin by auditing every AI exchange. It could identify one reasoning function, state which resources students may use, and specify what evidence would warrant changing help. Collaborative competence and independent capability may both matter. The assessment must keep their conditions visible, so that success with assistance does not silently become evidence of success without it.

Illustrative sociology assignment. The proposed lesson asks whether a university learning-support programme improved attainment. Students receive checked readings and a small dataset comparing participants with non-participants. In this research-methods task, they must judge the causal claim and consider alternative explanations. The lesson illustrates a design decision; it is not an intervention reported in this article.

AI may locate passages, assemble provisional summaries, and format references, with students checking summaries against the originals. The boundary follows the work: a summary that supplies the causal warrant has already performed part of the target reasoning.

Students receive the same AI-generated analysis and write a reconstruction memo of about 200 words. They identify an assumption, examine an evidence–conclusion link, and justify retaining or changing the recommendation. Sources and accessibility aids remain available, but AI does not draft this analysis. Box 1 guides formative feedback; the word count is a guide to length, not evidence of reasoning.

For the next diagnostic, present another programme and ask students to evaluate its recommendation without telling them to find a confounder. Supplied materials remain available, but task-solving AI does not. Record the first response before offering a targeted cue, then score prompted work separately. This leaves room to observe whether checking begins independently.

A student who explains the problem after a cue may next work with lighter regulatory prompts, adjusted as further evidence appears. Someone who notices the problem but cannot explain its causal implications may instead need a worked example. Reduce help for a function only after repeated adequate evidence; conflicting results call for further low-stakes observations. Before treating this design as a lesson protocol, pilot its timing and thresholds locally.

Academic integrity and instructional inference. The model assumes that GenAI is permitted under stated teaching conditions. Collaborative assignments and independent probes need separately announced resources, purposes, and criteria. Disclosure addresses compliance with those rules; attribution addresses what performance demonstrates. An assisted–independent gap alone establishes neither misconduct nor a basis for replacing institutional integrity procedures

with developmental diagnosis.

In a closed-book examination with a zero-AI rule, adaptive AI assistance would breach the stated conditions. Only the distinctions about evidence could inform interpretation or subsequent teaching; the full support cycle could not operate. Institutions remain responsible for choosing outcomes, assessment modes, and approved accommodations.

7.3. Limitations, equity, and rival interpretations

The framework offers questions for validation, not established pathways or diagnostic accuracy. Whether the same codes work across disciplines remains unresolved. The individual learner is our focus, leaving peers, motivation, affect, disciplinary communities, and institutional incentives in need of fuller treatment. A focused English-language synthesis can frame those issues, but findings from non-GenAI settings cannot directly establish the mechanisms proposed here.

Application in other languages requires more than translated prompts. Bilingual responses, local argumentation practices, and standards of evidence may change how codes and diagnostic decisions work. Comparability should be established before results are pooled across languages.

Task boundaries matter too. STEM calculation, creative production, and art-making require additional theories and measures. Memorization-only training and pure procedural drills are outside this model, despite possible overlaps in underlying learning processes.

A student may reason well yet struggle with the interface or fail to retrieve a transcript. Common orientation, comparable access, separate measures of tool familiarity and subject knowledge, and alternative response modes are possible safeguards. Researchers should test whether interface demands change what the indicators measure; statistical adjustment alone is insufficient. Weak digital traces cannot simply become weak-thinking scores.

Independence concerns who supplies the inference, not whether the learner has access to the task. Translation may preserve access, whereas a tool that changes the argument's logic also changes the evidence. Oral, written, and supported-input responses should sample the same function. Uncertain comparability must be recorded rather than assumed away. Fluency, typing speed, and complete logs remain different from demonstrated reasoning.

Diagnosis can itself get in the way of learning: it takes time, may interrupt work, and may increase anxiety. In some settings, informed teacher judgement may serve better than a rigid rule. Validation must therefore consider burden alongside thresholds and measurement quality. It must also allow rival explanations to succeed. Better information, additional practice, motivation, or reduced anxiety could account for gains otherwise attributed to reconstruction or adaptation.

8. Conclusion

The next support decision depends on more than the quality of an assisted argument. A learner may be ready to explain or evaluate a reason while still needing a reminder to start. We connect evidence about these functions to separate choices about content help and regulatory prompts. Six propositions, coding examples, and a focused experiment make the account testable. The immediate application is university argumentation; extending it requires evidence. Even a demonstrated mechanism must offer a worthwhile benefit before implementation is justified. The purpose of help at each stage is to support what the learner is ready to undertake next.

References

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, 128(4), 612–637. <https://doi.org/10.1037/0033-2909.128.4.612>

- 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), Article e2422633122. <https://doi.org/10.1073/pnas.2422633122>
- Bearman, M., Tai, J., Dawson, P., Boud, D., & Ajjawi, R. (2024). Developing evaluative judgement for a time of generative artificial intelligence. *Assessment & Evaluation in Higher Education*, 49(6), 893–905. <https://doi.org/10.1080/02602938.2024.2335321>
- Cao, M., & Carvalho, P. F. (2026). Striking the balance: How variability shapes retrieval practice and worked examples for transfer learning. *Educational Psychology Review*, 38, Article 71. <https://doi.org/10.1007/s10648-026-10169-w>
- Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4), 219–243. <https://doi.org/10.1080/00461520.2014.965823>
- 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. *British Journal of Educational Technology*, 56(2), 489–530. <https://doi.org/10.1111/bjet.13544>
- Fiorella, L. (2023). Making sense of generative learning. *Educational Psychology Review*, 35, Article 50. <https://doi.org/10.1007/s10648-023-09769-7>
- Gonçalves, A. O., Muniz, B. F. B., & Jaeger, A. (2025). Retrieval practice versus elaborative encoding: A systematic and meta-analytic review. *Educational Psychology Review*, 37, Article 100. <https://doi.org/10.1007/s10648-025-10076-6>
- Guan, R., Raković, M., Chen, G., & Gašević, D. (2025). How educational chatbots support self-regulated learning? A systematic review of the literature. *Education and Information Technologies*, 30(4), 4493–4518. <https://doi.org/10.1007/s10639-024-12881-y>
- Jaakkola, E. (2020). Designing conceptual articles: Four approaches. *AMS Review*, 10(1–2), 18–26. <https://doi.org/10.1007/s13162-020-00161-0>
- Janssen, N., & Lazonder, A. W. (2024). Meta-analysis of interventions for monitoring accuracy in problem solving. *Educational Psychology Review*, 36, Article 96. <https://doi.org/10.1007/s10648-024-09936-4>
- Jin, S.-H., Im, K., Yoo, M., Roll, I., & Seo, K. (2023). Supporting students' self-regulated learning in online learning using artificial intelligence applications. *International Journal of Educational Technology in Higher Education*, 20, Article 37. <https://doi.org/10.1186/s41239-023-00406-5>
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>
- 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. *Scientific Reports*, 15, Article 17458. <https://doi.org/10.1038/s41598-025-97652-6>
- Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with Cognitive Tutors. *Educational Psychology Review*, 19(3), 239–264. <https://doi.org/10.1007/s10648-007-9049-0>
- Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2), 259–269. <https://doi.org/10.1177/2515245918770963>
- Molenaar, I. (2022). Towards hybrid human–AI learning technologies. *European Journal of Education*, 57(4), 632–645. <https://doi.org/10.1111/ejed.12527>
- Ngai, C., & Gilbert, S. J. (2026). Metacognitive training facilitates optimal cognitive offloading.

Cognitive Research: Principles and Implications, 11, Article 21.
<https://doi.org/10.1186/s41235-026-00714-0>

Perkins, M., Roe, J., & Furze, L. (2025). Reimagining the Artificial Intelligence Assessment Scale (AIAS): A refined framework for educational assessment. *Journal of University Teaching and Learning Practice*, 22(7). <https://doi.org/10.53761/rrm4y757>

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676–688. <https://doi.org/10.1016/j.tics.2016.07.002>

Roe, J., Perkins, M., & Giray, L. (2026). Assessment twins: An approach for strengthening assessment validity in the age of generative AI. *Journal of Applied Learning & Teaching*, 9(2). <https://doi.org/10.37074/jalt.2026.9.2.3>

Sinha, T., & Kapur, M. (2021). When problem solving followed by instruction works: Evidence for productive failure. *Review of Educational Research*, 91(5), 761–798. <https://doi.org/10.3102/00346543211019105>

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, 10(2), 176–199. <https://doi.org/10.1177/1745691615569000>

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, 91, Article 101894. <https://doi.org/10.1016/j.learninstruc.2024.101894>

Tankelevitch, L., Kewenig, V., Simkute, A., Scott, A. E., Sarkar, A., Sellen, A., & Rintel, S. (2024). The metacognitive demands and opportunities of generative AI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Article 680, pp. 1–24). Association for Computing Machinery. <https://doi.org/10.1145/3613904.3642902>

Tetzlaff, L., Simonsmeier, B., Peters, T., & Brod, G. (2025). A cornerstone of adaptivity – A meta-analysis of the expertise reversal effect. *Learning and Instruction*, 98, Article 102142. <https://doi.org/10.1016/j.learninstruc.2025.102142>

Thurn, C. M., Edelsbrunner, P. A., Berkowitz, M., Deiglmayr, A., & Schalk, L. (2023). Questioning central assumptions of the ICAP framework. *npj Science of Learning*, 8, Article 49. <https://doi.org/10.1038/s41539-023-00197-4>

van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher–student interaction: A decade of research. *Educational Psychology Review*, 22(3), 271–296. <https://doi.org/10.1007/s10648-010-9127-6>

Vendrell, M., & Johnston, S.-K. (2026). Scaffolding critical thinking with generative AI: Design principles for integrating large language models in higher education. *Computers and Education: Artificial Intelligence*, 10, Article 100572. <https://doi.org/10.1016/j.caeai.2026.100572>

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, 17(2), 89–100. <https://doi.org/10.1111/j.1469-7610.1976.tb00381.x>

Xu, X., Qiao, L., Cheng, N., Liu, H., & Zhao, W. (2025). Enhancing self-regulated learning and learning experience in generative AI environments: The critical role of metacognitive support. *British Journal of Educational Technology*, 56(5), 1842–1863. <https://doi.org/10.1111/bjet.13599>

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1), 90–112. <https://doi.org/10.1111/bjet.13370>