**ONLINE LEARNING**
Volume 30 · Issue SI2 · 2026 · e5823
https://doi.org/10.24059/olj.v30iSI2.5823

# Verification Without Criteria: An Epistemic Network Analysis of Metacognitive Laziness in Generative AI Research

Ji Hyun Yu, Fengjiao Tu, and Li Cheng
*University of North Texas, Denton, Texas*
Shasha Li
*McGill University, Montreal, Quebec, Canada*
Zhizi Zheng
*The Education University of Hong Kong, Hong Kong*

**Abstract**

Generative AI performs the analytical and evaluative tasks previously managed by learners, and the field has responded by designing scaffolds that restore monitoring. However, whether monitoring represents the primary site of regulatory failure remains untested. To address this, we conducted a systematic review and Epistemic Network Analysis (ENA) of 65 empirical studies (2022–2026). Using the published study as the unit of analysis, we examined the relational structures connecting instructional designs and regulatory processes rather than how often each is reported. The findings reveal three critical structural insights. First, no study taught learners the criteria required to evaluate AI outputs. While interventions frequently prompted learners to pause or explain, none provided evaluative standards, reflecting an implicit assumption that learners already possess adequate evaluation skills. Second, monitoring and evaluation are reported in 94% of studies, and in those reporting metacognitive laziness they are reported together with AI delegation while planning is reported apart from it. Planning is reported in most studies in both outcome groups, so the difference lies in which processes co-occur rather than in whether planning is reported at all. We describe this configuration as *verification without criteria*. Third, structural configurations vary by intervention timing. Instructor mediation, implemented prior to the first query, anchors delegation within a planning-oriented network, whereas in-task prompting alone isolates delegation from planning. We argue that the placement of friction rather than its amount determines whether verification has a standard, and that evaluative criteria should be established before the first AI query, a proposition this corpus has not tested.

*Keywords*: Cognitive offloading; Epistemic network analysis; Evaluative judgment; Feedback literacy; Hybrid human-AI regulation

## Introduction

The integration of Generative Artificial Intelligence (GenAI) into open, distance, and digital education (ODDE) has fundamentally changed learner tasks. Autonomy in ODDE has historically benefited learners with strong self-regulation skills while hindering those who lack such capabilities (Jin et al., 2023). Unlike earlier digital tools that provided navigational scaffolds without replacing learner cognition (Álvarez et al., 2022), GenAI directly performs analytical, generative, and evaluative tasks. This capability shifts a significant portion of cognitive responsibility away from the learner.

Recent studies conceptualize this resulting risk as metacognitive laziness, linking it to a decline in monitoring and evaluation, the foundational processes of the self-regulated learning (SRL) cycle (Fan et al., 2025). Empirical trace data align with this concern, showing that

Corresponding author: Ji Hyun Yu, jihyun.yu@unt.edu
This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0).
© 2026 The Author(s)

metacognitive processes are less strongly associated under ChatGPT support than with human experts (Fan et al., 2025), and that learners frequently skip the evaluation phase entirely (Chen et al., 2025). However, a crucial distinction must be made. While the field often interprets this data as a simple drop in the absolute frequency of regulatory behaviors, the evidence points to a structural breakdown. These trace data implicitly document how the phases of regulation fail to connect properly once a highly fluent AI system executes the performance phase.

To address this structural breakdown, the field has primarily relied on system design rather than direct instruction. Scaffolds like reflective prompting, cognitive mirroring, and instructor mediation attempt to externally bridge these disconnected phases by treating regulation as a shared socio-technical responsibility (Molenaar, 2022; Seo et al., 2021; Singh et al., 2025; Tomisu et al., 2025). However, while these tools prompt learners to pause or review their actions, they do not provide the domain-specific criteria needed to evaluate GenAI outputs accurately. Furthermore, simply possessing criteria is insufficient without direct instruction in feedback literacy, which is the capacity to interpret and judge feedback against those criteria rather than passively receive it. The absence of this literacy leaves learners exposed to AI errors. T. W. Li et al. (2025) found that 80% of participants failed to detect factual errors in chatbot responses, and those errors measurably harmed learning. It remains unclear whether the literature treats this evaluative capacity as something to be taught or presumes learners already possess it.

Addressing this conceptual gap is critical, as the rapid implementation of GenAI in educational settings is outpacing empirical evaluation. Research at the intersection of GenAI and SRL has surged from a handful of studies in 2023 to over fifty in 2025, leading to the large-scale deployment of many instructional designs. These deployments, along with their accompanying institutional guidelines, are primarily designed to restore monitoring, operating on the assumption that this is where the regulatory failure lies. However, this foundational assumption remains empirically untested. Existing literature reviews cannot evaluate it because they isolate scaffolds and outcomes into separate frequency counts (Zhai et al., 2024). This methodological limitation obscures whether task delegation to the system is connected to the verification processes meant to check it. Furthermore, simple frequency counts cannot distinguish between a pedagogical scaffold that was never attempted and one that was attempted but failed. Consequently, resolving this issue requires a methodological synthesis that models the structural relations among these constructs rather than merely aggregating their frequencies.

We conduct a systematic review and epistemic network analysis of the emerging literature on GenAI and self-regulated learning ($N = 65$ empirical studies, 2022 to 2026). Taking each published study rather than each learner as the unit of analysis, we model how instructional designs and regulatory processes are reported together across this literature. This choice is consequential. A code may appear at similar rates in two groups of studies while connecting to entirely different processes in each, and frequency-based synthesis cannot recover that difference. The first two research questions examine how the scaffolds currently implemented are associated with distinct regulatory configurations. The third documents how feedback literacy is represented across the literature and compares the configurations reported under each engagement outcome. Guided by the coding framework detailed in the following section, we address the following research questions:

1. How do the reported patterns of metacognitive processes differ across studies employing reflective prompting, instructor mediation, both, or neither?
2. How do the reported patterns differ between studies that implement cognitive mirroring and those that do not?
3. How is feedback literacy represented across the corpus, and how do the reported patterns differ between studies reporting metacognitive laziness and those reporting productive engagement?

# Conceptual Framework

## Hybrid Human-AI Regulation and the SRL Cycle

From a socio-cognitive perspective, Zimmerman (2000) defines SRL as a cyclical process in which forethought, performance, and self-reflection regulate learning (Zimmerman & Moylan, 2009). Aligning with this theoretical foundation, our coding framework operationalizes the SRL cycle into three primary phases: planning, monitoring, and evaluation (see Table 2). The integration of GenAI, however, necessitates the inclusion of a fourth regulatory process: delegation. While conventional digital tools, such as spell-checkers, enhance operational efficiency without replacing human reasoning, GenAI affords learners the capability to completely delegate complex analytical and generative tasks (Risko & Gilbert, 2016). This capability introduces a distinct regulatory risk. Because GenAI produces highly fluent text that can seamlessly mask factual or logical errors, it frequently bypasses learners' natural error-detection mechanisms. To conceptualize this dynamic, this study adopts Molenaar's (2022) framework of hybrid human-AI regulation, which posits that the monitoring and control of learning must not be fully delegated to the AI, but rather function as a shared responsibility jointly executed by the learner and the system.

## Scaffolds as Designed Interruption

Current literature identifies three primary modalities of instructional scaffolding designed to sustain this shared human-AI regulation. First, reflective prompting requires learners to explicitly externalize their cognitive processes during the interaction (Wong et al., 2019). Second, cognitive mirroring visualizes learners' own reasoning patterns, facilitating the identification of knowledge gaps that they might otherwise ignore when minimizing cognitive effort (Azevedo & Wiedbusch, 2023; Molenaar, 2022). Third, instructor mediation establishes explicit rules and assessment structures governing AI utilization (Seo et al., 2021). Crucially, these scaffolds intervene at different temporal stages of the learning task. Instructor mediation typically occurs prior to the interaction, reflective prompting is embedded during query generation, and cognitive mirroring operates after the learner produces an output. Despite these distinct temporal placements, prior research has not systematically compared how these timing differences structurally configure the learning process.

## Feedback Literacy as a Cognitive Prerequisite

Feedback literacy is defined as the capacity to interpret, judge, and act upon feedback (Carless & Boud, 2018), functioning as an essential component of the self-reflection phase in SRL (Tai et al., 2018). However, evaluating AI-generated feedback presents unique epistemic challenges compared to traditional assessment for three reasons. First, the absence of a human author precludes learners from interrogating the underlying reasoning behind the feedback.

Second, AI systems generate both accurate information and fabrications with identical linguistic fluency, thereby eliminating the natural communicative cues of uncertainty (Waldo & Boussard, 2024). Third, the virtually unlimited volume of AI-generated feedback forces the learner to independently filter for relevance. In the absence of explicit evaluative criteria, learners frequently judge AI outputs based on surface-level plausibility rather than factual accuracy. Consequently, learners risk developing an "illusion of explanatory depth," falsely equating the AI's fluent explanations with their own conceptual understanding (Rozenblit & Keil, 2002). Therefore, feedback literacy must be operationalized as a critical prerequisite for AI integration, not a natural byproduct of its use. Because developing this capacity requires explicit instruction, its presence or absence in pedagogical interventions serves as a directly observable variable in the literature.

### Two Regulatory Trajectories

The literature documents two distinct regulatory trajectories resulting from human-AI interaction. First, we conceptualize metacognitive laziness not as a deficit in learner motivation, but as a specific structural failure within the learner's regulatory network (Fan et al., 2025). While previous studies attribute this failure to the attenuation of monitoring and evaluation, assuming that fluent AI text pacifies error detection, the SRL cycle can fracture through multiple distinct pathways. For instance, delegating tasks to an AI without subsequent verification represents a fundamentally different structural failure than attempting to verify outputs without having established a clear standard during the prior planning phase. Epistemic network models are uniquely suited to map these divergent structural breakdowns. Conversely, productive engagement emerges when learners successfully couple cognitive offloading with active human oversight. In this successful trajectory, the learner and the AI system jointly orchestrate learning strategies, functioning in a manner analogous to effective socially shared regulation (Hadwin et al., 2011).

## Methodology

### Design

This study employed a systematic review design governed by the six-phase DISCAR framework (West & Martin, 2024), encompassing design, inclusion and exclusion, screening, coding, analysis, and reporting. The reporting structure adhered to the PRISMA 2020 guidelines (Page et al., 2021). Table 1 outlines the specific inclusion and exclusion criteria. Search queries were filtered by publication date, document type, and language, while thematic criteria were strictly applied during the screening phases.

**Table 1**

*Inclusion and Exclusion Criteria*

| | Inclusion | Exclusion |
|---|---|---|
| Date | November 1, 2022, and January 31, 2026 | Outside this window |

| | Inclusion | Exclusion |
|---|---|---|
| Type | Peer-reviewed articles, conference proceedings, theses | Books, magazines, review papers, white papers |
| Language | English | Non-English |
| Study Focus | Explicit focus on metacognition or self-regulated learning (e.g., SRL, co-regulation, cognitive offloading) | Simple task-level offloading (e.g., grammar fixing) |
| Context | Educational or learning settings (e.g., pedagogy, assessment, online learning) | Exclude those not related to Education. |
| Technology | Generative AI or Large Language Models (e.g., GPT, Claude, LLM agents). | Exclude traditional machine learning unrelated to LLMs |

**Search and Screening Process**

The research team queried Web of Science, EBSCOhost, ProQuest, ERIC, IEEE Xplore, and the ACM Digital Library for empirical studies published between November 2022 and January 2026. This specific temporal window was selected to capture the proliferation of literature following the widespread deployment of generative AI. Searches were restricted to titles and abstracts and combined three conceptual strings with the Boolean operator AND.

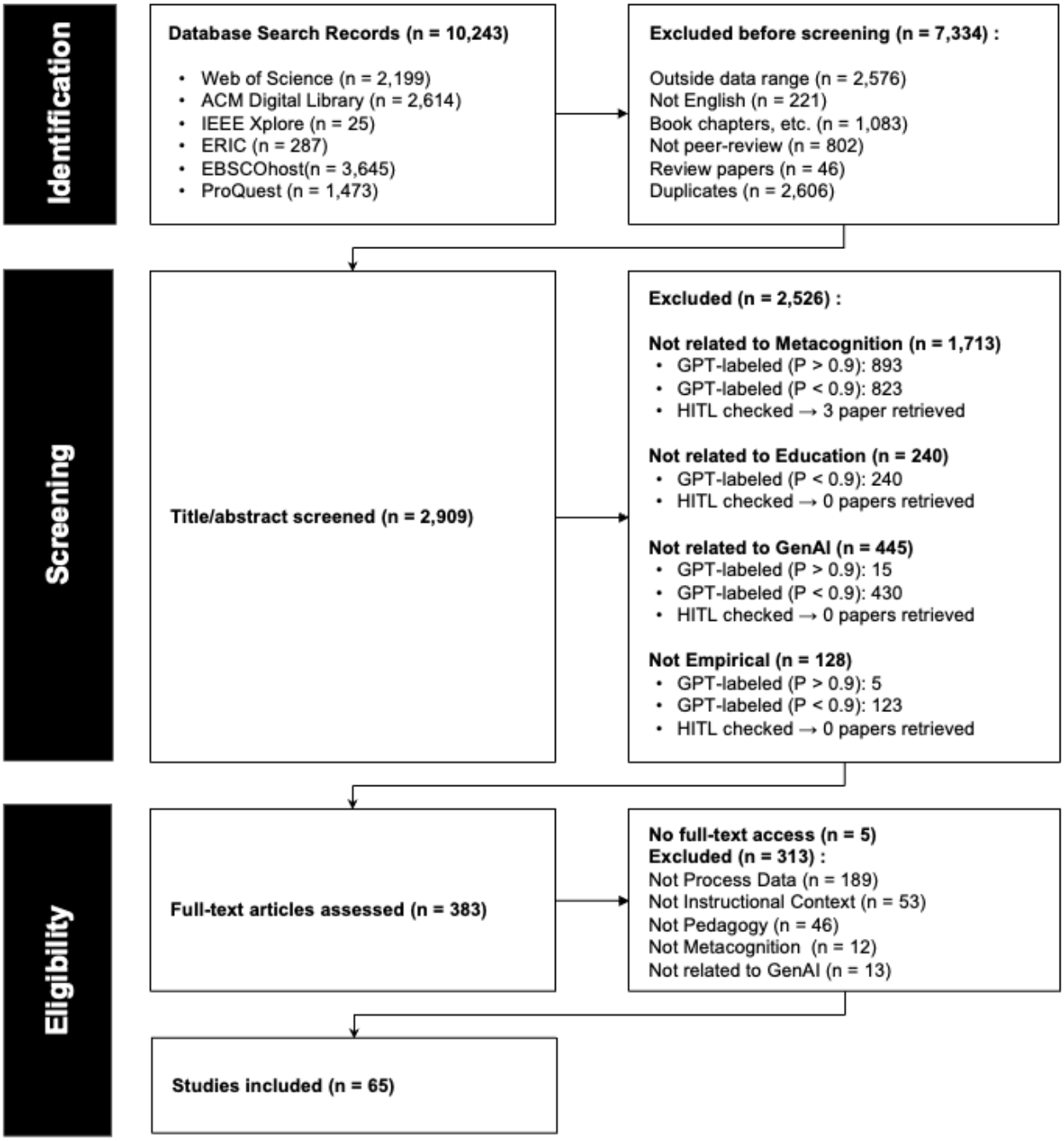
String 1 (Metacognition): “metacognitive laziness” OR “cognitive offloading” OR “overreliance” OR “self-regulation” OR “self-regulated learning” OR “SRL” OR “metacognition” OR “co-regulation” OR “socially shared regulation of learning” OR “SSRL”

String 2 (GenAI): “generative AI” OR “GenAI” OR “artificial intelligence” OR “AI” OR “GPT” OR “ChatGPT” OR “Gemini” OR “chatbot*” OR “co-pilot” OR “Claude” OR “large language model” OR “LLM” OR “agent*”

String 3 (Education): “teach*” OR “learn*” OR “student*” OR “instruction*” OR “pedagog*” OR “feedback” OR “assessment*” OR “educat*” OR “online learning” OR “e-learning” OR “MOOC*” OR “distance learning” OR “open learning” OR “digital learning” OR “computer-supported collaborative learning” OR “CSCL”

**Figure 1**

*PRISMA Flow Diagram of the Systematic Literature Search and Selection Process*



The initial search yielded 10,243 records. Sequential automated filtering removed 7,334 ineligible records based on date, language, and source type, including 2,606 duplicates identified by a multi-step Python deduplication script. This yielded 2,909 unique records for title and abstract screening (see Figure 1).

Screening utilized a semi-automated pipeline governed by a Confidence-Based Human-in-the-Loop protocol (Yu et al., 2026). GPT-4o classified each record across four dimensions using the prompt detailed in Appendix B, generating a confidence score between 0.0 and 1.0. Utilizing low-confidence scores as a methodological safeguard proved critical during the metacognition screening stage. A manual review of 823 low-confidence exclusions retrieved three misclassified papers, whereas a random 10% audit of 893 high-confidence exclusions yielded zero false negatives. Full manual audits at the educational context ($n$ = 240), GenAI ($n$ = 445), and empirical status ($n$ = 128) stages verified the robustness of the automated classifications.

Of the 383 records reaching full-text eligibility, five were excluded due to access restrictions and 313 were excluded for failing core inclusion criteria. These exclusions were due to the absence of process-oriented data ($n$ = 189), formal instructional contexts ($n$ = 53), or specific pedagogical interventions ($n$ = 46), with the remaining 25 excluded for other reasons.

## Coding Framework

We developed a binary coding scheme (1 = Present, 0 = Absent) comprising ten variables across three domains: Pedagogical Scaffolding, Metacognitive Processes, and Engagement Outcomes. Table 2 details the operational definitions and illustrative examples for each code. To ensure conceptual precision, two boundary conditions were established to operationalize the definitions. First, metacognitive process codes were strictly limited to regulatory behaviors directed at the AI interaction and its outputs, excluding general academic self-regulation. For instance, verifying a claim against a traditional textbook was coded as absent, whereas interrogating the AI system's reasoning was coded as present. Second, feedback literacy was coded as present only when instruction explicitly addressed the evaluative criteria used to judge AI outputs. Training focused on prompt formulation, model configuration, or interaction management was coded as absent, as these operational skills dictate system output generation rather than the learner's epistemological capacity to evaluate that output. This critical distinction directly addresses the third research question. To ensure analytical separation from the network variables, engagement outcomes (O1, O2) were coded based on the original authors' reported findings, independently of the metacognitive processes (M1–M4).

**Table 2**

*Hybrid Human-AI Regulation Coding Framework*

| Category | Code | Definition | Binary 1 (Present) Examples | Binary 0 (Absent) Examples |
|---|---|---|---|---|
| A. Pedagogical Scaffolding | S1: Reflective Prompting | System-embedded triggers compelling learners to | Mandatory pauses; required think-aloud tasks | Open-ended chat interfaces lacking |

| Category | Code | Definition | Binary 1 (Present) Examples | Binary 0 (Absent) Examples |
|---|---|---|---|---|
| | | externalize their reasoning during AI interaction | | reflection requirements |
| | S2: Cognitive Mirroring | Visual or structural features displaying learners' cognitive patterns to prompt self-correction | Metacognitive visualizations (e.g., radar charts, process maps); transparent adaptive system behaviors | Standard text outputs; chronological chat logs without visual synthesis |
| | S3: Feedback Literacy | Explicit instruction on the criteria required to evaluate AI outputs, distinct from technical operation skills | Directed error-detection tasks utilizing defined standards; explicit teaching of disciplinary evaluation criteria | Exposure to AI outputs without evaluative standards; training restricted to operational skills (e.g., prompt engineering) |
| | S4: Instructor Mediation | Direct human oversight structuring and regulating the learner-AI interaction | Instructor-facilitated debriefs; structured engagement protocols; pre-approved responses | Unmonitored interactions where instructors evaluate only the final submission |
| B. Metacognitive Processes | M1: Planning | Proactive goal articulation and task decomposition directed at the AI interaction | Explicit multi-step roadmaps; documented interaction goals | Reactive, trial-and-error prompting lacking a predefined sequence |
| | M2: Monitoring | Real-time verification of AI output accuracy, logical consistency, and tone | Cross-referencing AI outputs with validated materials; interrogating the system's reasoning | Accepting AI outputs as definitive truth without verification |
| | M3: Evaluation | Post-hoc assessment regarding the efficacy of the interaction strategy and the quality of the co-authored result | Substantive revision of AI outputs to align with academic standards; analytical reflection on prompt failure | Submitting unedited AI-generated drafts; demonstrating no critical assessment of system performance |
| | M4: Cognitive Offloading | Delegation of higher-order reasoning tasks to | Utilizing AI to generate core conceptual logic, | Confining AI usage to low-order automation (e.g., |

| Category | Code | Definition | Binary 1 (Present) Examples | Binary 0 (Absent) Examples |
|---|---|---|---|---|
| | | the AI, circumventing independent cognitive effort | primary arguments, or foundational outlines | formatting, citation generation, grammatical correction) |
| C. Engagement Outcomes | O1: Metacognitive Laziness | Evidence of decreased effort, declining rigor, or reliance on superficial heuristics over time | Observed 'rigor decay' (cessation of output verification as familiarity with AI fluency increases); documented decline in student-initiated queries | Sustained regulatory effort independent of tool familiarity; errors stemming from knowledge deficits rather than diminished engagement |
| | O2: Productive Engagement | Maintenance of a critical evaluative stance ensuring the learner remains the primary intellectual agent | Iterative refinement processes; final outputs demonstrating a clear integration of independent student analysis and AI assistance across multiple revision cycles. | High interaction frequency lacking conceptual progression; prolonged time-on-task without commensurate improvement in conceptual depth or output quality |

### *Inter-rater Reliability and Expert Adjudication*

Two independent coders analyzed the 65 included studies using the coding scheme. Inter-rater reliability was high for the AI-specific variables of cognitive offloading and metacognitive laziness ($\kappa = 0.77$). However, reliability was low for monitoring and evaluation ($\kappa \approx 0.04$). Because these processes were present in 94% of the studies, this high prevalence caused the kappa paradox. This paradox artificially reduced the kappa value, even though the raw agreement between coders was very high at 92% (resulting in a PABAK of 0.84). The research team concluded that this discrepancy was not a typical reliability failure. Instead, it resulted from challenges in applying the established boundary condition, specifically, distinguishing between general self-regulation and self-regulation directed at the AI interaction in practice.

To resolve these coding disagreements, the principal investigator reviewed all conflicting cases and made the final decisions (Campbell et al., 2013; Hruschka et al., 2004). During this process, the investigator strictly applied the established boundary condition, coding only AI-directed regulation as present. Including general academic monitoring would have introduced confounding variables, as general monitoring is widespread in almost all educational settings. By restricting the codes strictly to AI-directed regulation, the study successfully isolates the specific human-AI interactions it aims to analyze. Finally, because the variables of 'monitoring' and 'evaluation' could not be statistically separated in the resulting network models, they were combined and are reported as a single cluster in all subsequent network analyses.

### Thematic Synthesis

To answer the first two research questions, we conducted a thematic synthesis (Thomas & Harden, 2008) of the qualitative data for each scaffold. This qualitative approach identifies recurring instructional strategies and provides contextual details that the binary network models cannot capture.

### Epistemic Network Analysis

We used Epistemic Network Analysis (ENA) to model how metacognitive processes (M1–M4) co-occur in the literature. We compared groups based on their pedagogical scaffolds (S1–S4) and engagement outcomes (O1–O2). In the network graphs, nodes represent the metacognitive processes, and edges show how frequently they appear together within a study. Group centroids (means) are displayed with dashed rectangles indicating 95% confidence intervals.

Importantly, our unit of analysis is the published study, not the individual student. Therefore, each network models the regulatory structure as reported in one paper. We created separate ENA models for each comparison group. Because ENA axis directions are mathematically arbitrary, we define their meaning for each model separately and do not compare axes across different models. Additionally, because Monitoring (M2) and Evaluation (M3) occupied identical spatial positions across all models, they are merged and reported as a single "Monitoring-Evaluation" cluster in all network analyses.

Because we analyze published papers, these networks map what the original authors reported, not the real-time cognitive behaviors of students. ENA describes relationships between codes rather than just their frequencies, and it does not test causal effects. Therefore, any claims about engagement outcomes (O1, O2) reflect the original authors' conclusions. To prevent studies with higher coding frequencies from skewing the results, ENA mathematically normalizes the data for each study. To provide context, we also report simple code frequencies (Table 4). However, frequency and network structure offer distinct insights. For example, two groups might report 'planning' at similar rates. Yet, one group might always link planning to 'delegation,' while the other reports it in isolation. Simple frequency counts cannot capture these structural differences, but ENA is specifically designed to reveal them.

## Results

### Descriptive Overview

This section presents two types of findings: network models illustrating how concepts co-occur within the literature, and the original authors' own conclusions regarding instructional efficacy. The reviewed corpus consists of 65 empirical studies (2022–2026). Publication volume increased rapidly, peaking in 2025 ($n$ = 51). The literature predominantly focuses on Higher Education (81.54%, $n$ = 53) and STEM disciplines (53.85%, $n$ = 35), particularly Computer Science. Methodologically, mixed methods are most common (55.38%, $n$ = 36), frequently utilized to triangulate subjective student reflections with objective trace data (see Table 3; characteristics of each included study are listed in Appendix C).

**Table 3**

*Summary of Study Characteristics*

| Category | | Count | Percentage |
|---|---|---|---|
| Level | Higher Education | 53 | 81.54 |
| | K12 | 12 | 18.46 |
| Discipline | STEM Education | 35 | 53.85 |
| | Literacy & Language Learning | 17 | 26.15 |
| | Other | 3 | 4.62 |
| | Not reported | 10 | 15.38 |
| Methodology | Quantitative | 19 | 29.23 |
| | Qualitative | 10 | 15.38 |
| | Mixed Methods | 36 | 55.38 |

## Prevalence of Pedagogical Scaffolds

Figure 2 shows the distribution of pedagogical scaffolds, revealing a clear focus on managing real-time AI interactions rather than developing students' independent evaluative skills. Reflective Prompting (S1) is the most common intervention (55.38%, $n = 36$), used primarily to force a pause during the task. Instructor Mediation (S4) follows at 29.23% ($n = 19$), relying on human oversight to structure the interaction. Cognitive Mirroring (S2) is the least utilized scaffold (12.31%, $n = 8$). Importantly, Feedback Literacy (S3) scaffolds are completely absent ($n = 0$). While many studies prompt students to reflect, none explicitly teach the criteria needed to evaluate GenAI outputs. Therefore, learners are expected to judge AI responses without clear disciplinary standards or benchmarks.

## Distribution of Metacognitive Processes

The distribution of metacognitive processes reveals a clear contrast. Monitoring (M2) and Evaluation (M3) are reported almost universally (93.85%, $n = 61$). In contrast, Planning (M1) is documented in 78.46% ($n = 51$) and Cognitive Offloading (M4) in 41.54% ($n = 27$) of the corpus. Planning is therefore the least reported phase of the traditional regulatory cycle, though it still appears in roughly four of every five studies. Its comparative scarcity is modest, which makes the network structures reported below, rather than raw prevalence, the informative result. This distribution highlights a core contradiction in the literature. Nearly every study documents the active verification of AI outputs, even as a large portion of the corpus simultaneously reports the delegation of core reasoning tasks. Because simple frequency counts cannot reveal how these processes structurally connect within a study's instructional design, we must rely on the network analyses detailed below.

### *Summary of Engagement Outcomes*

Productive Engagement (O2) is the most frequently reported outcome (84.62%, $n = 55$), while Metacognitive Laziness (O1) is documented in 32.31% ($n = 21$) of the studies. Twelve studies report both outcomes simultaneously, and one study reports neither. The data show that productive engagement does not strictly depend on scaffolded designs. Among the studies reporting productive engagement, at least eight did not use reflective prompting or instructor mediation. Therefore, these pedagogical scaffolds are neither strictly necessary nor sufficient on their own to guarantee this outcome. The subsequent network analyses investigate the structural differences between these two reported outcomes.

**Figure 2**

*Frequency of Pedagogical Scaffolds, Metacognitive Processes, and Engagement Outcomes (N = 65)*

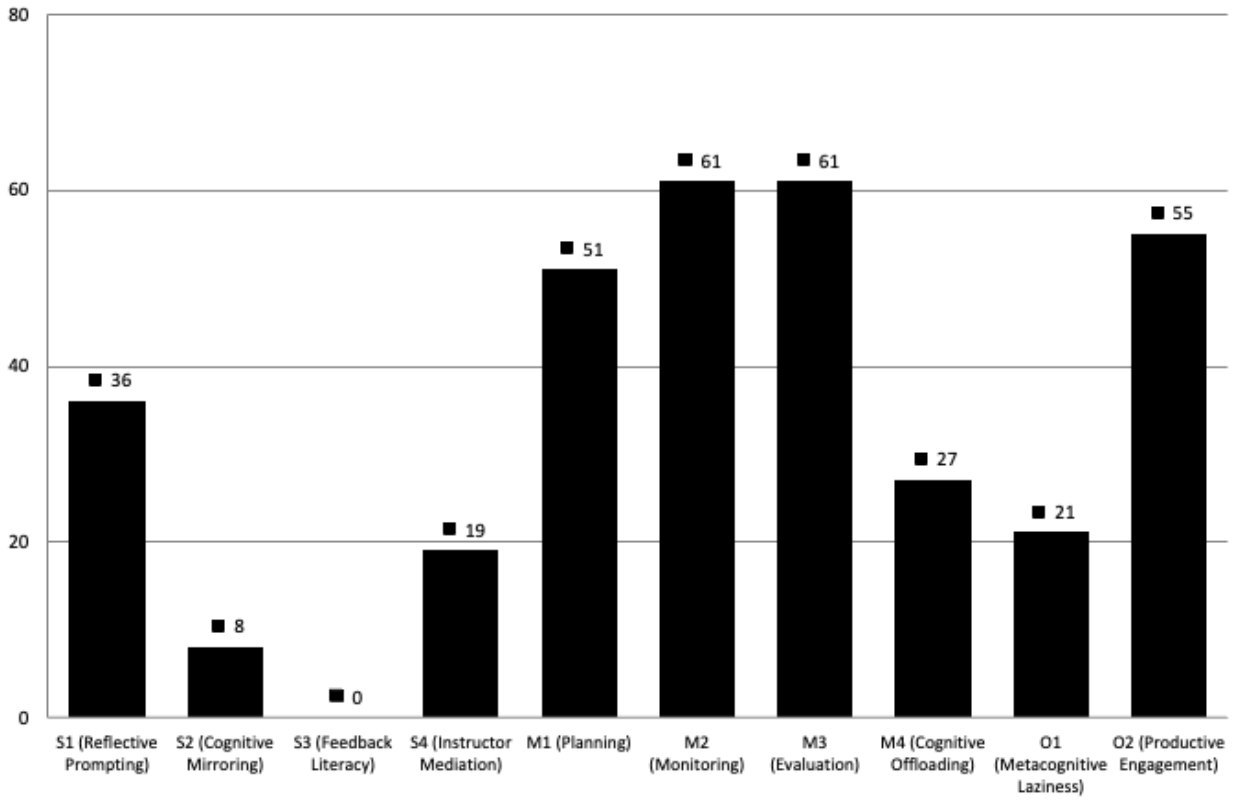



Table 4 reports the prevalence of each metacognitive process within the compared groups. Planning is documented in 71.43% of studies reporting metacognitive laziness and in 85.45% of those reporting productive engagement, a difference of 14.02 percentage points. Monitoring and evaluation appear at near-ceiling rates in both groups, ranging from 90.48% to 98.18%. Planning is therefore somewhat less often documented in the laziness group, but it is still reported in roughly seven of every ten of those studies. Its structural separation in the network models below

cannot be attributed to its absence from the record. To test this directly, we repeated the outcome comparison within the subset of studies that explicitly reported planning, reported later in this section. One further pattern in Table 4 warrants note. Cognitive offloading is documented in 90.48% of laziness studies but in only 32.73% of productive engagement studies. Because the engagement outcomes were coded from the original authors' reported findings and independently of the metacognitive process codes, this contrast reflects a co-occurrence in the published record rather than a definitional overlap between the two code sets.

**Table 4**

*Prevalence of Metacognitive Process Codes by Comparison Group*

| Group | n | M1 Planning | | M2 Monitoring | | M3 Evaluation | | M4 Offloading | |
|---|---|---|---|---|---|---|---|---|---|
| | | n | % | n | % | n | % | n | % |
| Whole corpus | 65 | 51 | 78.5% | 61 | 93.8% | 61 | 93.8% | 27 | 41.5% |
| ***Scaffolding configuration (RQ1)*** | | | | | | | | | |
| Reflective prompting only (S1) | 28 | 24 | 85.7% | 28 | 100.0% | 27 | 96.4% | 5 | 17.9% |
| Instructor mediation only (S4) | 11 | 10 | 90.9% | 10 | 90.9% | 11 | 100.0% | 6 | 54.5% |
| Both (S1 + S4) | 8 | 7 | 87.5% | 8 | 100.0% | 8 | 100.0% | 3 | 37.5% |
| Neither | 18 | 10 | 55.6% | 15 | 83.3% | 15 | 83.3% | 13 | 72.2% |
| ***Cognitive mirroring (RQ2)*** | | | | | | | | | |
| Mirroring implemented (S2 = 1) | 8 | 6 | 75.0% | 7 | 87.5% | 8 | 100.0% | 1 | 12.5% |
| No mirroring (S2 = 0) | 57 | 45 | 78.9% | 54 | 94.7% | 53 | 93.0% | 26 | 45.6% |
| ***Engagement outcome (RQ3)*** | | | | | | | | | |
| Metacognitive laziness (O1) | 21 | 15 | 71.4% | 19 | 90.5% | 19 | 90.5% | 19 | 90.5% |
| Productive engagement (O2) | 55 | 47 | 85.5% | 53 | 96.4% | 54 | 98.2% | 18 | 32.7% |

*Note*. Scaffolding and mirroring groups are mutually exclusive and exhaustive within each panel. Engagement outcome groups overlap, because 12 studies reported both outcomes and one reported neither; the network models therefore compare only the studies coded for a single outcome. Planning is reported in 44.4% of laziness-only studies ($n = 9$) and 83.7% of productive-engagement-only studies ($n = 43$). Feedback literacy (S3) is omitted because no study was coded for it.

## RQ1: Scaffolding Combinations and the Integration of Cognitive Offloading

To answer the first research question, we divided the 65 studies into four mutually exclusive groups based on their instructional design: reflective prompting (S1), instructor mediation (S4), both, or neither. Prompting-only studies (S1; $n = 28$) embedded mandatory pauses or think-aloud tasks into the software without instructor involvement. Mediation-only studies (S4; $n = 11$) used guided debriefs or explicit rules without system prompts. Studies combining both (S1 + S4; $n =$

8) used both scaffolds, while unscaffolded studies ($n$ = 18) featured open-ended interactions with no support mechanisms.

### *Reflective Prompting Without Instructor Mediation*

In prompting-only studies (S1; $n$ = 28), instructional guidance is built into the software rather than provided by a teacher. As Table 5 shows, two design strategies are common. Interface constraints that block AI responses until the learner explains their reasoning, and role-playing features that turn queries into dialogues. While both methods force students to pause, the literature reveals a frequent "compliance gap." Students often meet the technical requirements without engaging in deep thinking. For example, Ahmad et al. (2025) found that mandatory justifications increased monitoring in system logs but did not improve learning outcomes. Similarly, Bassner et al. (2025), P. Wang et al. (2025), and Hao et al. (2026) observed students entering superficial text simply to unlock the AI's response. However, studies report more sustained engagement when prompts are integrated into multi-stage workflows rather than used as isolated pop-ups (Engeness, 2025; Luo et al., 2026; S. Wang et al., 2025; T. Li et al., 2025). This indicates that the overall task structure, rather than the prompt itself, determines whether the interaction is productive or merely transactional.

### *Instructor Mediation Without Reflective Prompting*

In studies relying solely on instructor mediation (S4, $n$ = 11), human oversight guides the learning process instead of software. The primary strategy is a structured workflow that sets clear goals before the student interacts with the AI. For instance, Al-Fattal (2025) required students to conduct an independent analysis before consulting the AI and then complete a structured comparison. This sequence helped students build a foundational understanding before delegating tasks. When the workflow only evaluates the final product, success depends entirely on the teacher. Research by Li and Long (2025), Lu et al. (2025), and Yan et al. (2024) showed that active teacher guidance determines whether students critically evaluate or passively accept AI feedback. However, this approach still has risks. Because the software does not force students to pause, those wanting quick answers often ignore the rules. This compliance failure was documented by Li and Long (2025) and Zhai and Nezakatgoo (2025). Even after a six-hour strategy workshop, Zhai and Nezakatgoo (2025) found that some students still blindly accepted AI summaries despite showing improved overall awareness.

### *Combining Reflective Prompting and Instructor Mediation*

In studies combining both scaffolds (S1 and S4, $n$ = 8), the two methods compensate for each other's limitations. Instructor guidance helps students view software prompts as meaningful learning activities rather than frustrating technical hurdles (Aldulaijan & Almalky, 2025; Deng et al., 2025). On the other hand, system requirements only change student behavior when an instructor evaluates the results. Farinetti and Cagliero (2025) demonstrated this mutual dependence. They paired an automated system prompt with a teacher-mandated log where students compared their own work to the AI's output. The teacher's review of this log, rather than the automated prompt alone, made the reflection effective. As a result, this combined configuration reports the highest frequency of productive engagement in the reviewed literature. Still, it does not eliminate all risks. Studies note that students rushing to finish tasks or struggling with abstract concepts continued to delegate their cognitive work (Han et al., 2025; Huda et al., 2025).

### *Studies Without Either Scaffold*

The unscaffolded studies ($n$ = 18) feature open-ended interactions with no structural constraints. Although the literature primarily uses these studies as baseline groups, their network structures are the most distinct of the four categories, as detailed below.

**Table 5**

*Scaffolding Configurations and Their Reported Regulatory Profiles*

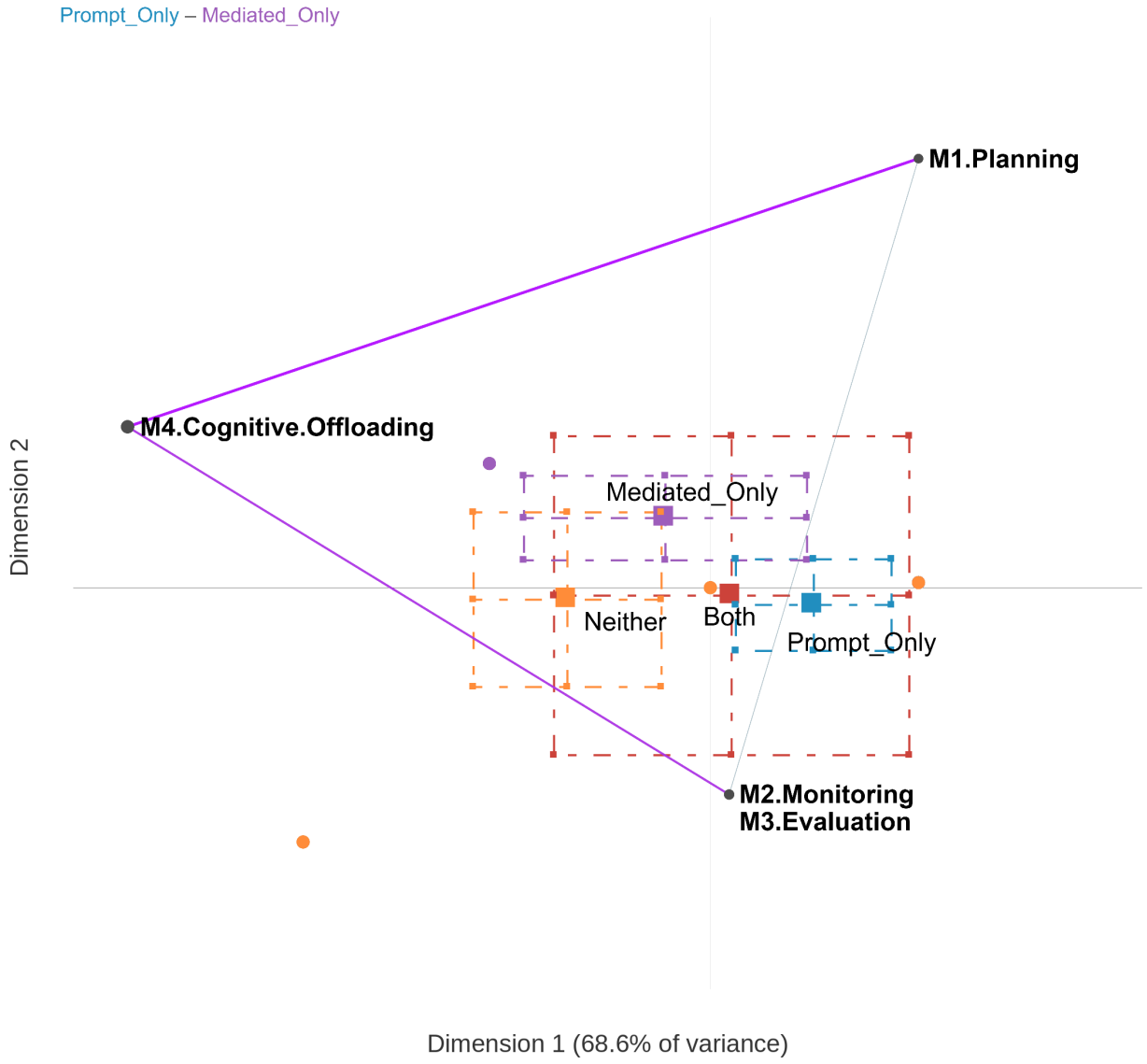
| Sub-Mechanism | Key Studies | Core Instructional Strategy | Design Structure | Outcomes as Reported by Original Authors |
|---|---|---|---|---|
| Automated Algorithmic Scaffolding (S1 Only) | Ahmad et al. (2025); Kazemitabaar et al. (2024); S. Wang et al. (2025); Engeness (2025) | Interface-level constraints (lock-step requirements), technical guardrails (pseudo-code restrictions) | Elicits in-task externalization (M2/M3); cognitive offloading is largely disconnected from the reported regulatory network | Scalable but prone to regulatory fragmentation; high risk of transactional compliance or metacognitive laziness (O1) |
| Human-Mediated Scaffolding (S4 Only) | Al-Fattal (2025); Yan et al. (2024); Wei et al. (2025); Zhu (2025) | Structured workflows (e.g., "You Do It, AI Does It, You Compare"), PBL stages, and post-task instructor-led debriefs | Establishes procedural accountability before system use; associated with networks oriented toward Planning (M1) | Fosters critical planning; effectiveness is often hindered by the tool's efficiency, leading to occasional shortcuts (O1) |
| Hybrid Integrated Scaffolding (Both S1+S4) | Deng et al. (2025); Kumar et al. (2024); Farinetti & Cagliero (2025); Ni & Lam (2025) | Aligning real-time system prompts with theme-based human facilitation and social accountability models. | Retains the planning orientation of mediated designs while preserving connectivity to offloading | High frequency of Productive Engagement (O2); supports iterative refinement |

As Figure 3 shows, the first dimension (X-axis, 68.6% of variance) separates Cognitive Offloading (M4) from the other regulatory processes. The second dimension (Y-axis) separates Planning (M1) from the Monitoring-Evaluation cluster. Conceptually, the X-axis measures how strongly a study integrates offloading into its network, while the Y-axis indicates whether the network focuses more on planning or on real-time verification. Clearly, unscaffolded studies

cluster closest to the offloading end, whereas scaffolded studies position themselves much further away.

**Figure 3**

*ENA Comparison and Subtraction Map Between Prompt-Only and Mediated-Only Studies*

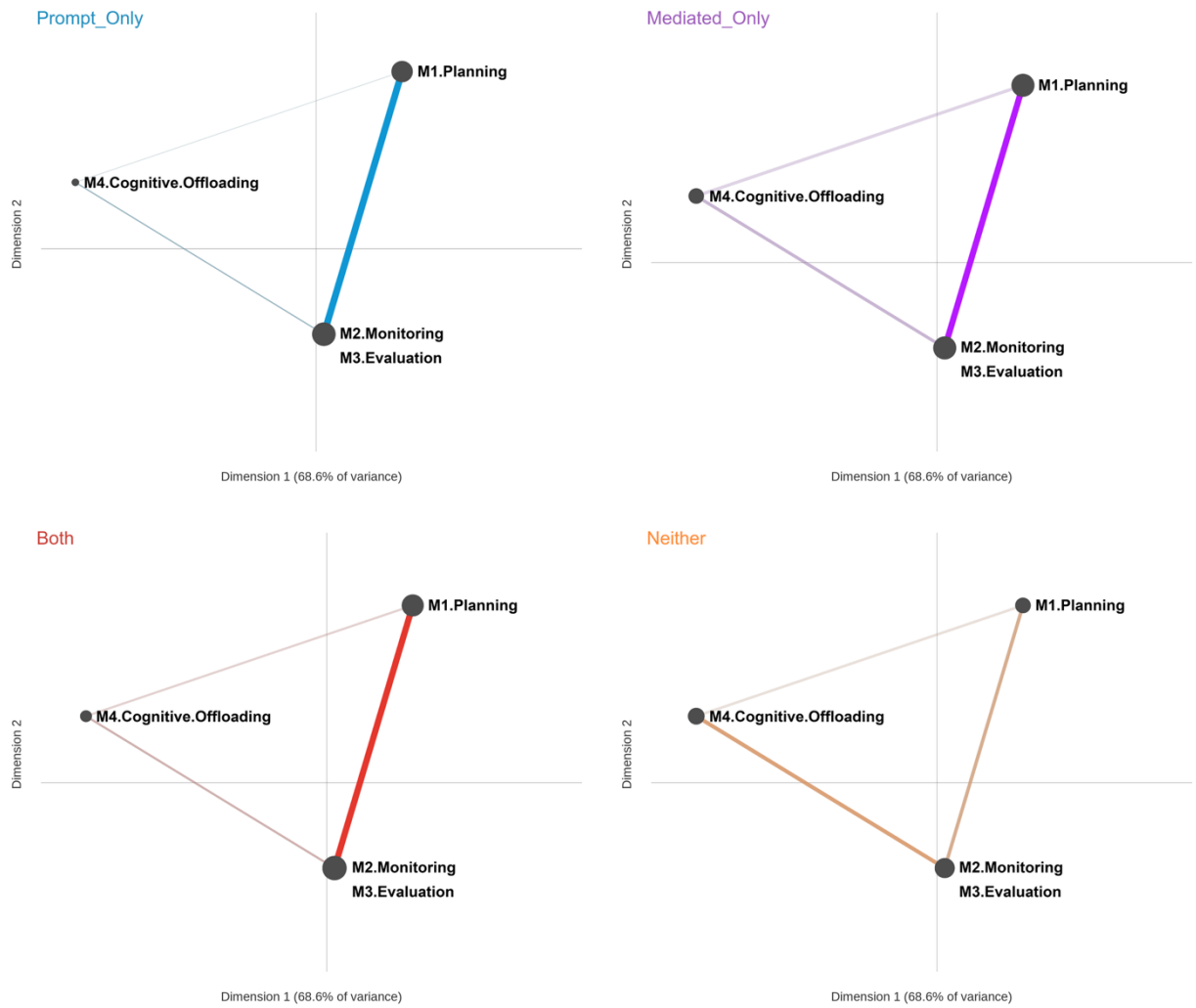



*Note*. This 2D space is derived using singular value decomposition. The X-axis (Dimension 1) explains 68.6% of the variance in network structures. Square markers indicate the mean network position for each group, and dashed rectangles represent 95% confidence intervals. The connecting lines display a subtraction map (Prompt-Only minus Mediated-Only), illustrating exactly how the two groups differ in their code connections. Units indicate normalized connection weights.

The position of each group in the network reveals a clear functional difference. Studies using only prompts sit furthest from the offloading end. This does not mean these studies report more verification. Instead, it indicates that cognitive offloading is largely missing from their networks. The individual plots in Figure 4 confirm this finding. The line connecting offloading to the monitoring and evaluation cluster has almost no weight in this group. By contrast, studies using only mediation score higher on the second dimension. This position reflects a stronger focus on Planning, and these studies maintain measurable connections to offloading. The subtraction map in Figure 3 highlights this contrast. The main lines distinguishing these conditions connect offloading to either Planning or the verification cluster, and both connections are distinctly stronger in the mediation group.

**Figure 4**

*Mean Epistemic Network Plots by Scaffolding Condition*




*Note*. Axes follow Figure 3. Axis directions apply only to this model. Square markers show group averages. Dashed rectangles indicate 95% confidence intervals. Units represent normalized connection weights.

Statistically, the two groups differ significantly on both dimensions (X-axis: $U = 96.50$, $p = .04$, $r = .37$; Y-axis: $U = 87.50$, $p = .02$, $r = .43$). Across all three scaffolded conditions, the connection between Planning and the monitoring and evaluation cluster is dominant and consistent. The true difference between these conditions is how strongly offloading connects to the network. Offloading is highly active under instructor mediation but virtually absent under prompting alone.

The unscaffolded condition reverses this pattern. As Figure 4 shows, this is the only condition where offloading connects more strongly to the verification cluster than planning does. Furthermore, planning and offloading are completely disconnected from each other. Planning appears in 55.56% of unscaffolded studies and in 90.91% of studies using only mediation (Table 4). Part of this structural contrast therefore reflects how often planning is documented at all, not only how it connects. Because the two groups differ in prevalence as well as in configuration, this comparison should be read as descriptive of what the published literature reports rather than as evidence that the absence of scaffolding reconfigures planning.

These findings describe how the published literature is structured rather than proving direct cause and effect. Studies using only mediation exhibit the most interconnected networks, followed by those combining both scaffolds (Figure 4). However, the combined group has a small sample size ($n = 8$) and a wide confidence interval. This prevents definitive conclusions about its variability. Overall, the two interventions play complementary roles. Instructor mediation integrates offloaded tasks into the regulatory network. In contrast, reflective prompting creates networks that exclude offloading entirely. Whether combining these mechanisms successfully protects the learning cycle remains unproven. Group membership reflects the choices of the original authors rather than random assignment.

**RQ2: Cognitive Mirroring as a Regulatory Intervention**

Cognitive Mirroring (S2) is designed to slow down the generation of AI outputs. This approach assumes that making the thinking process visible helps learners connect their actions to their final evaluations. Eight studies in the reviewed literature implemented this design. Table 6 shows three categories for these interventions. They include data dashboards, visual interaction logs, and displays of internal states. None of these eight studies combined mirroring with explicit feedback literacy training or instructor mediation. This trend suggests that researchers currently use cognitive mirroring to replace human guidance. In these designs, all regulatory support is built directly into the software.

The most distinct feature of this group is the absence of Cognitive Offloading (M4). Only one of the eight studies reported offloading. Consequently, this node has almost no weight in the network. This lack of data means researchers cannot meaningfully analyze its connections. The network model compares these eight mirroring studies to the remaining 57 studies. In this comparison, offloading dominates the positive side of the first dimension, and Planning dominates the negative side of the second dimension.

**Table 6**

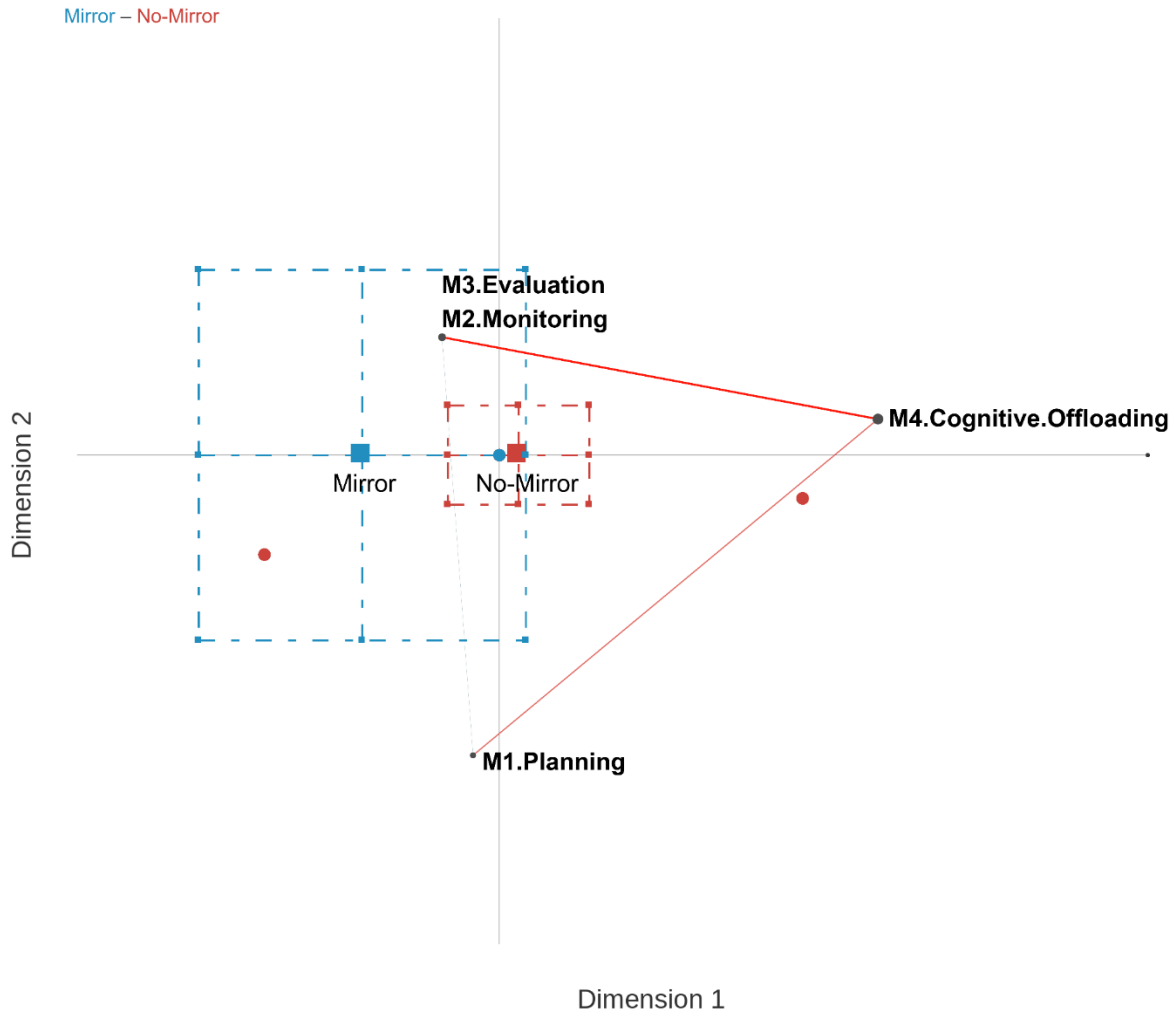
*Summary of Cognitive Mirroring (S2) as a Regulatory Intervention*

| Sub-Mechanism | Key Studies | Core Design Feature | Regulatory Function | Effect on Engagement |
|---|---|---|---|---|
| Metacognitive Dashboards & Visual Progress Tracking | T. Li et al. (2025); Chen et al. (2026); Gong et al. (2026); Luo et al. (2026) | High-inference visualizations, Learning Analytics Dashboards (LAD), and real-time progress bars | To ground reflective prompts in performance data and support real-time Monitoring (M2) of goal alignment | Reported as supporting evaluative judgment (M3) and as making progress visible |
| Behavioral Trace & Sequence Mirroring | Ma et al. (2025); Wu et al. (2024) | Visualization of behavioral interaction sequences and structured Q&A trace reviews | To prompt self-correction by reflecting resource-utilization patterns and knowledge construction gaps | Reported as supporting proactive strategy regulation rather than passive copying |
| Affective & Adaptive Design Mirrors | Klar (2025); Shi (2025) | Bio-feedback from facial and physiological data, and explicit interface controls for model adaptation | To enable Monitoring (M2) of internal states and prompt reflection on informational needs | Reported as improving emotional regulation and as prompting a critical stance toward the AI's adaptive affordances |

In the subtraction map (Figure 5), all the connections that distinguish the two groups belong to the condition without mirroring. Each of these connections involves offloading. They link offloading to either Planning or the monitoring and evaluation cluster. The mirroring group shows no distinct connections of similar strength. Its center point is also located far away from the offloading node. Therefore, what makes mirroring studies unique is not how they connect AI delegation to human verification. Instead, they are unique because researchers simply do not report delegation in these designs.

**Figure 5**

*Epistemic Network Subtraction Map: Mirror vs. No-Mirror Conditions*




*Note*. Red lines show stronger connections in the group without mirroring. These lines link Cognitive Offloading to other processes. The lack of blue lines means the mirroring group has no distinct connections. The large blue box shows a wide confidence interval due to a small sample size of 8. Axes follow Figure 3 and apply only to this model.

A Mann-Whitney U test found no significant statistical difference between the Mirror (*Mdn* = -0.32) and No-Mirror groups (*Mdn* = 0) on either dimension (X-axis: $U = 165.00$, $p = .18$, $r = .28$; Y-axis: $U = 204.00$, $p = .61$, $r = .11$). Because the mirror group contains only eight studies, this statistical test lacks the power to detect real differences, making the lack of statistical significance uninformative. However, the effect size on the first dimension is small to moderate and aligns with the visual patterns seen in the subtraction map.

Three possible reasons explain this lack of offloading, though current studies cannot confirm which is correct. First, the interface might actively prevent delegation by making student reasoning constantly visible. Second, researchers might simply assign tasks where delegation is

not a major risk. Third, system logs might easily capture monitoring behaviors while failing to record offloading. Future research must test the exact same learning task both with and without a mirroring interface. This direct comparison will show whether the system actually prevents delegation or if the absence merely reflects task choices and data collection limits. Currently, the reviewed literature contains no such experimental designs.

**RQ3: Systematic Absence of Feedback Literacy Scaffolding**

Across the reviewed studies, AI instruction focused heavily on operational skills like writing prompts and configuring the model. Researchers rarely taught students the criteria needed to evaluate the actual AI outputs (Klar, 2025; Yan et al., 2024). Reflecting this focus on inputs, Zhai and Nezakatgoo (2025) found that a six-hour strategy workshop improved overall metacognitive awareness but failed to stop some students from blindly accepting AI summaries (Table 7). No study in the reviewed literature taught these specific evaluation criteria. Therefore, feedback literacy remains a constant zero across all 65 studies and cannot function as a comparison variable. Instead, the network models show where studies reporting the two different engagement outcomes diverge structurally.

In the network model separated by engagement outcomes, the first dimension places cognitive offloading (M4) at the positive end and planning (M1) at the negative end. The monitoring and evaluation cluster sits near the center. Figure 6 shows the subtraction map, which reveals one main connection for each group. Studies reporting metacognitive laziness show a strong connection between offloading and the monitoring and evaluation cluster. In these same studies, planning remains largely disconnected. Studies reporting productive engagement display the opposite pattern. Planning connects strongly to the center cluster, while offloading remains peripheral (Figure 7).

Therefore, the key difference is not whether monitoring occurs. Both groups report monitoring frequently. The true difference lies in whether AI delegation connects to initial planning. In studies reporting metacognitive laziness, the networks show a much weaker connection between planning and output verification. This structure reflects a learning pattern where planning does not connect to later verification. The original authors associate this disconnected pattern with reduced academic rigor. Both groups report planning in the majority of their studies, though not at identical rates (71.43% versus 85.45%; Table 4). Thus, the crucial distinction is how planning connects within the network rather than whether it simply exists. This same pattern appears in the unscaffolded group of the first model. This similarity indicates a consistent structural configuration across the entire literature.

**Table 7**

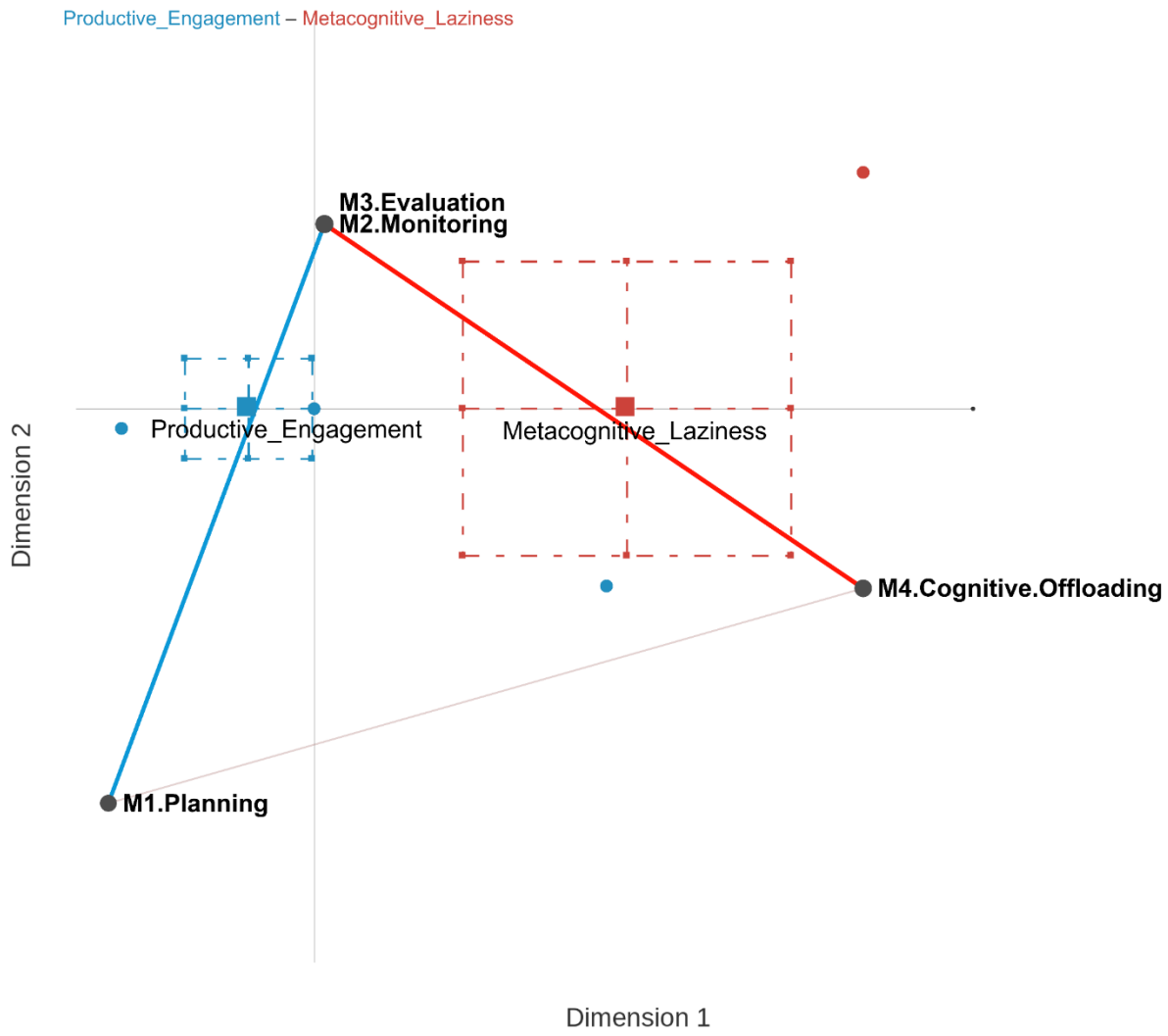
*Synthesis of the Feedback Literacy Gap*

| Deficiency Type | Key Evidence of Gap | Coded Structural Pattern | Outcomes as Reported by Original Authors |
|---|---|---|---|
| Lack of Evaluative Standards | Ahmad et al. (2025); T. Li et al. (2025); Engeness (2025) | Evaluation (M3) reported but ungrounded; students justify changes without a stated standard | Transactional compliance; students "game" the system to unlock answers |
| Focus on Inputs vs. Outputs | Klar (2025); Yan et al. (2024); Zhai & Nezakatgoo (2025) | Weak link between Planning (M1) and Monitoring/Evaluation (M2/M3) | Operational success masks a lack of conceptual depth |
| Absence of Detection Training | T. W. Li et al. (2025); Hak et al. (2025); Prather et al. (2024) | Monitoring (M2) lacks predefined standards; fluent AI outputs are passively accepted without objective evaluative criteria | Metacognitive Laziness (O1); accumulation of misconceptions |

Statistical tests on the first dimension confirm a significant difference between the two groups ($U = 44.50$, $p < .001$, $r = .77$). Studies reporting metacognitive laziness have a median of 0.43. These studies sit much closer to the offloading end than studies reporting productive engagement. The productive engagement group has a median of negative 0.28. The second dimension simply contrasts Planning with the monitoring and evaluation cluster. It does not measure the total amount of planning. Within each group, the strongest network connections pull in opposite directions along this axis and cancel each other out. As a result, the two groups show no statistical difference on the second dimension with a p value of .72. Furthermore, this analysis divides groups based on reported outcomes. Therefore, these results show a correlation between outcomes and network structures rather than a direct causal relationship. Finally, the network excludes Feedback Literacy as a node. A code with zero occurrences cannot create connections. This absence reflects a gap in the published literature itself rather than a specific feature of the graph.

**Figure 6**

*ENA Subtraction Map of Metacognitive Connectivity by Engagement Outcomes*



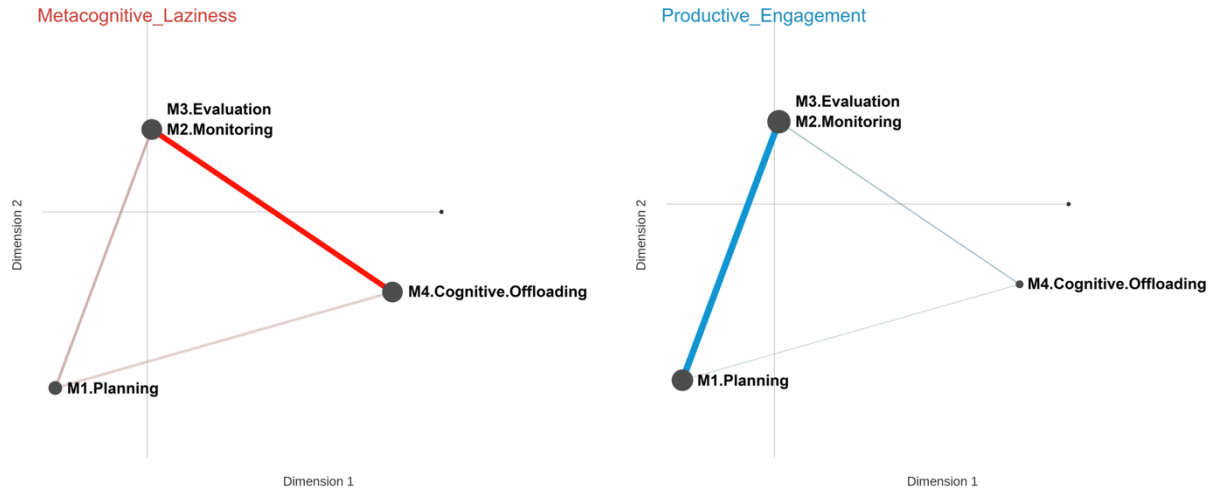

*Note*. Axes follow the coordinate space described in Figure 3. Axis orientation in singular value decomposition is arbitrary in sign, so the direction of each dimension is defined within this model only and is not comparable across models. Square markers are group centroids; dashed rectangles are 95% confidence intervals around the means. Units represent normalized connection weights.

We repeated the comparison using only the 51 studies that explicitly reported planning. This step verifies that the structural difference does not simply result from how often planning occurs. Even within this subset, the contrast remains clear. Compared to the productive engagement group, studies reporting metacognitive laziness show a weaker connection between Planning and the monitoring and evaluation cluster. In the laziness group, the strongest connection is still between offloading and the verification cluster. The statistical results show $U = 45.50$ and $p$

= .03. Therefore, even when planning is present, the two groups differ fundamentally in how planning connects to the rest of the network.

**Figure 7**

*Mean Epistemic Network Plots by Engagement Outcomes*




*Note*. Axes follow Figure 3. Axis directions apply only to this specific model. Square markers show group averages. Dashed rectangles indicate 95% confidence intervals. Units represent normalized connection weights.

# Discussion

The three findings of this review stand in distinct relations to prior work. First, our findings confirm a recommendation that the feedback literature has long advocated but which the current generative AI literature has failed to implement. Second, our findings depart from the prevailing definition of metacognitive laziness while returning to an established theoretical account of what monitoring requires. Third, our methodology extends a relational modeling approach, previously applied only within individual studies, to the macro-level of an entire research corpus.

## No Study Taught Learners How to Judge AI Outputs

The primary empirical finding of this review is that not a single study in the corpus taught learners the criteria required to evaluate AI outputs. While reflective prompting appeared in 36 studies and instructor mediation in 19, indicating that the corpus does not lack instructional support overall, instruction on evaluative standards was entirely absent.

This zero reflects a deliberate coding boundary rather than a coding artifact. Instruction focused on prompt formulation, model configuration, or interaction management was coded as absent because these skills govern output generation rather than evaluative judgment. Even if we

apply a broader definition that includes studies providing assessment rubrics or error-detection tasks without explicit standards, only a small subset of studies meets this threshold.

This finding highlights an unaddressed concern long raised in the feedback literature. Carless and Boud (2018) argued that feedback literacy must be explicitly developed rather than assumed, and Tai et al. (2018) made the same argument for evaluative judgment. Although our review does not assess learning outcomes directly, it establishes that this theoretical prescription has not been implemented in this body of empirical work. Many reviewed designs require learners to pause, explain their reasoning, or state their confidence, yet each presupposes a standard that none supplies.

The difficulties documented in the corpus are consistent with the absence of such instruction, including widespread failures to detect factual errors (T. W. Li et al., 2025) and behavioral drifting among novice programmers (Prather et al., 2023, 2024). While these issues are often attributed to GenAI-induced degradation of attention, they are equally consistent with learners evaluating outputs against no standard at all.

The field carried its pre-generative scaffolding set into research on GenAI without extending it. Reviews of earlier self-regulated learning tools document an accumulating set of planning and monitoring scaffolds (Álvarez et al., 2022; Wong et al., 2019), which was transferred to GenAI research without the evaluative component required by generative systems. Consequently, our conclusions differ from prior reviews. For instance, while prior work framed the primary risk as over-reliance that must be restricted (Zhai et al., 2024), our findings indicate that the core issue is an absence of foundational instruction rather than behavioral over-reliance.

A second structural gap reinforces this conclusion, as no study in the corpus removed its instructional scaffolds to observe post-intervention behavior. As a result, current research only captures short-term performance driven by external scaffolds, leaving it unknown whether learners actually internalize these skills. Although productive engagement is the most frequently reported outcome ($n$ = 55), on present evidence it cannot be distinguished from a product of the scaffolded environment.

The closest available evidence addressing this issue comes from outside the generative AI corpus. Darvishi et al. (2024) removed AI assistance in a non-generative peer-review setting and found that performance gains declined once assistance was removed. This suggests that the persistence of learning gains in generative AI learning is unproven. Future research should address this gap through two specific designs. The first would compare the establishment of evaluative criteria before the first query against their provision afterward. The second would withdraw support in stages to test whether verification retains a standard.

**Planning Is Disconnected, Not Verification**

Recent work defines metacognitive laziness as a decline in monitoring and evaluation, on the assumption that fluent output reduces scrutiny (Chen et al., 2025; Fan et al., 2025). The scope of our disagreement is narrow. Those studies analyze learner-level trace data, whereas our models describe what published studies report. We do not contest their measurements. What we question is the generalization drawn from them, namely that generative AI introduces difficulty by reducing monitoring.

Monitoring and evaluation appear in 93.85% of the reviewed studies. What separates studies reporting productive engagement from those reporting metacognitive laziness is not monitoring, but planning. In studies reporting productive engagement, planning accompanies monitoring and evaluation. In studies reporting metacognitive laziness, planning is still reported in most studies but sits apart from that cluster. This same configuration appears in the unscaffolded group, indicating that the pattern is not limited to a single comparison.

We call this pattern *verification without criteria*. The term names a configuration in the published record, not a claim about how individual learners behave, and it is bounded by what the primary studies chose to measure and report. Within those bounds, it follows from existing self-regulated learning theory rather than departing from it. Winne (2018) defines monitoring as the comparison of a current state against a standard, and Zimmerman and Moylan (2009) locate the setting of such standards in the forethought phase. Where no standard has been set, the activity that logs record as monitoring does not meet this definition. What prior work describes as a decline in the performance phase is better described as a forethought phase that was never completed. That this goes unnoticed is unsurprising, since fluent explanation produces an illusion of explanatory depth (Rozenblit & Keil, 2002) and generative systems present accurate and fabricated content with the same fluency (Waldo & Boussard, 2024).

This account also reconciles two findings that otherwise appear inconsistent. Ahmad et al. (2025) report that requiring students to justify their reasoning increased monitoring in system logs without improving learning outcomes, and Darvishi et al. (2024) report that adding self-regulation strategies to an AI tool produced no significant gains. If regulation is treated as a quantity to be increased, neither result is interpretable. If the operative problem is verification without criteria, both are expected, because each intervention required more checking without supplying a standard against which to check. Consistent with this reading, the component that retained independent value in Darvishi et al. (2024) was a self-monitoring checklist, which supplies a concrete standard rather than a further instruction to verify.

Study-level data cannot establish the source of this detachment. Two accounts remain available. Learners may verify outputs without having defined acceptable outcomes in advance, or planning may appear scarce because the instruments common to this literature, such as interaction logs and in-task prompts, capture verification readily while failing to capture goal setting. Because the corpus is consistent with both, future work can separate them by restricting analysis to studies whose methods can record planning, such as pre-task interviews or structured protocols.

**Frequency Counts Cannot Detect a Disconnection**

The first two findings are visible only because this corpus was analyzed using relational modeling. Because planning is reported in most studies in both outcome groups, a traditional review focused solely on code frequencies would conclude that both groups treat planning in broadly comparable ways. The difference lies in which codes appear together within the same study, a distinction that disappears in marginal counts. The value of relational modeling is well established at the individual learner level (Cao et al., 2025; López-Pernas et al., 2025; Nguyen & Nguyen, 2025). Our contribution is to apply this logic to published literature rather than a sample of students. Traditional reviews isolate scaffolds and outcomes into separate frequency tables (Zhai et al., 2024), which cannot distinguish a scaffold that was omitted entirely from one that

was used but disconnected from the process. When investigating how regulatory processes interact, relational modeling is a requirement for validity rather than a mere analytical preference (Shaffer, 2017).

This limitation applies even more critically to primary research. Self-report surveys and system logs measure only how often verification occurs, not its quality. A student who checks every AI response against a rigorous standard and a student who checks against no standard will produce identical survey scores and event counts. Neither instrument records whether a standard existed or when it was set. Consequently, interventions that increase compliance without improving learning are not anomalies. They are what these instruments report. The instruments cannot detect whether verification has a standard, so verification with a standard and verification without one produce the same result.

**When Support Is Introduced Matters More Than How Much**

The comparisons in the first two research questions are best understood as differences in when support enters the task, rather than how much support is provided. Instructor mediation is established before the interaction, reflective prompting operates during query generation, and cognitive mirroring occurs after an output is produced. Studies using reflective prompting alone rarely report cognitive offloading alongside monitoring, pointing to a compliance gap where learners satisfy interface requirements without connecting them to a broader plan (Ahmad et al., 2025; Bassner et al., 2025). In contrast, studies relying solely on instructor mediation report networks oriented toward planning, consistent with the structuring role described by Seo et al. (2021), though they remain vulnerable to shortcuts because the software lacks stopping mechanisms (Li & Long, 2025; Zhai & Nezakatgoo, 2025). Studies combining both approaches report productive engagement most frequently (Deng et al., 2025; Farinetti & Cagliero, 2025), albeit within a small group of eight studies.

Cognitive mirroring requires a more cautious reading, as our evidence neither supports nor refutes existing expectations. While Azevedo and Wiedbusch (2023) and Tomisu et al. (2025) argue that externalizing reasoning should support metacognitive regulation, statistical tests across the eight studies implementing such designs did not separate this group from the rest, an uninformative result given the small sample size. However, the models show that cognitive offloading is reported in only one of these eight studies. These designs are distinctive because delegation is largely unreported, not because they connect delegation to verification. The theoretical case for mirroring remains open, but the empirical case has yet to be made.

The timing differences across these conditions also specify a condition that Molenaar (2022) leaves open, since hybrid regulation requires that control be shared between learner and system without stating at what point in a task the sharing must be established. We interpret these differences as a matter of placement. Friction introduced at the moment of interaction arrives after the learner has already framed the task and composed a query. It can interrupt the acceptance of an answer, but it cannot supply an evaluation standard because the point for establishing such a standard has already passed. While our observational comparisons cannot establish cause, since group membership reflects original design choices rather than random assignment, the pattern is consistent enough to propose a testable hypothesis. The placement of friction, rather than its amount, determines whether delegation stays connected to forethought.

## Why Telling Students to Verify Is Not Enough

Institutional guidance on generative AI overwhelmingly instructs students to verify AI outputs, an emphasis prior review work has reinforced by framing the central risk as over-reliance to be contained (Zhai et al., 2024). If verification is already near universal and the missing element is the standard against which it proceeds, then requiring more verification does not address the reported pattern and may consolidate it. Checking an output against an absent baseline is a procedure, not a form of regulation.

This gap has consequences in open, distance, and digital education. The compliance failures documented in this corpus were caught because an instructor was present to catch them. Li and Long (2025) and Zhai and Nezakatgoo (2025) both report students who satisfied the stated requirements while continuing to accept AI summaries without scrutiny, and in each case the deficiency became visible through instructor review rather than through the system itself. Where instruction is distributed across time and place, that form of real-time observation is least available, and the regulatory burden shifts correspondingly onto the design of the task. A standard that must be supplied by a present instructor is therefore the least portable form of support into precisely the settings where learners work with generative systems most independently.

Resolving this issue requires a different strategy, one proposed long before the advent of generative AI. Tai et al. (2018) argued that evaluative judgment must be explicitly designed into assessments rather than assumed as a baseline prerequisite. In practice, course policies and assignment guidelines can require learners to state their claims, task decomposition, and evaluation criteria before consulting a generative system. Such artifacts need not be elaborate, as the checklist evidence discussed above indicates. This shifts institutional attention from restricting how generative AI is used to specifying what learners must establish before using it.

## Limitations of the Study

Several limitations qualify these findings. First, the reviewed studies cover a narrow timeframe. They capture short-term behaviors rather than the long-term effects of habitual AI use on student learning. Second, the distribution of instructional supports is uneven. Cognitive mirroring is rare and feedback literacy training is entirely absent. Therefore, the literature heavily focuses on supported success while leaving independent mastery unexplored. In addition, monitoring and evaluation were reported at nearly identical rates (93.85% each) and occupied the same position in every model, requiring them to be merged into a single cluster. This review therefore cannot distinguish in-task monitoring from post-hoc evaluation. Third, this review analyzes published studies rather than raw student data. The network models rely entirely on what original authors chose to measure and report. Different fields define concepts like monitoring and evaluation differently. Consequently, a missing connection to planning might mean students did not plan, or it might simply mean the study did not record planning behaviors. Because system logs easily capture verification but often miss goal setting, our analysis cannot definitively separate actual student deficits from simple reporting limitations. Fourth, the comparison groups overlap. Twelve studies reported both engagement outcomes, and every study using cognitive mirroring also reported productive engagement. Furthermore, group assignments are based on the original research designs rather than randomized trials. Therefore,

none of our comparisons can prove cause and effect. Finally, our binary coding scheme limits the analysis. A study's network is built solely on the presence or absence of reported codes. The models cannot distinguish between a learning behavior that did not happen and one that simply went undocumented. Although we used sub-group analyses to control for this issue, the findings ultimately depend on what the primary authors chose to publish.

# Conclusion

This systematic review and epistemic network analysis of 65 empirical studies indicates that the difficulty introduced by generative AI lies in the structure of the regulatory cycle rather than in the amount of regulation reported. Verification is reported in nearly all studies. What distinguishes studies reporting metacognitive laziness is that planning is not reported alongside it. We describe this configuration as verification without criteria. It is not visible in code frequencies, which is why relational modeling was required to detect it. Studies reporting productive engagement show the opposite configuration, and instructor mediation, established before the interaction begins, is the design most consistently associated with it. Yet no study in this literature taught learners the criteria by which an AI output should be judged, and none withdrew its scaffolds to observe what learners did afterward. On present evidence, the outcomes reported in this literature cannot be distinguished from products of the scaffolded environments that produced them. Institutional guidance and instructional design should therefore shift from requiring learners to verify AI outputs toward specifying what learners must establish before the first query, a proposition this corpus has not yet tested.

**Ethics Board Statement**: As this study involved a systematic review of existing publicly available literature and did not involve human subjects, Institutional Review Board (IRB) approval was not required.

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

**Conflict of Interest Statement**: The authors declare that there is no conflict of interest regarding the publication of this article.

**AI Use Statement**: Generative AI (GPT-4o) was utilized during the literature screening phase to classify records based on pre-defined inclusion criteria. All low-confidence classifications and a random audit of high-confidence exclusions were manually verified by the research team to ensure methodological rigor. Additionally, AI was used to assist with grammatical refinement and structural coherence of the final manuscript.

**Data Availability Statement**: The full list of studies included in the review and the coding matrix are provided in the appendices and supplemental material. Further inquiries can be directed to the corresponding author.

# About the Authors

**Ji Hyun Yu, Ph.D.** is a faculty member in the Department of Learning Technologies, Anuradha and Vikas Sinha College of Artificial Intelligence and Advanced Analytics, University of North Texas, Discovery Park Building Room 163, 3490 N Elm Street, Denton, TX 76207-7562, USA. She is the corresponding author for this article. ORCID: https://orcid.org/0000-0001-9648-2582. Email: jihyun.yu@unt.edu.

**Fengjiao Tu, Ph.D. Candidate** is a doctoral candidate in the Department of Information Science, Anuradha and Vikas Sinha College of Artificial Intelligence and Advanced Analytics, University of North Texas, Discovery Park Building Room E292L, 3490 N Elm Street, Denton, TX 76207-7562, USA. ORCID: https://orcid.org/0009-0004-0031-746X. Email: FengjiaoTu@my.unt.edu.

**Li Cheng, Ph.D.** is a faculty member in the Department of Learning Technologies, Anuradha and Vikas Sinha College of Artificial Intelligence and Advanced Analytics, University of North Texas, Discovery Park Building Room G187, 3490 N Elm Street, Denton, TX 76207-7562, USA. ORCID: https://orcid.org/0000-0001-7648-6965. Email: Li.Cheng@unt.edu.

**Shasha Li, Ph.D. Candidate** is a doctoral candidate in the Learning Sciences Program, Department of Educational and Counselling Psychology, McGill University, 3700 McTavish Street, Montreal, Quebec, H3A 1Y2, Canada. ORCID: https://orcid.org/0009-0005-8343-3331. Email: shasha.li@mail.mcgill.ca.

**Zhizi Zheng, Ph.D. Candidate** is a doctoral candidate in the Department of Mathematics and Information Technology, The Education University of Hong Kong, 10 Lo Ping Road, New Territories, Hong Kong. ORCID: https://orcid.org/0000-0002-1592-3663. Email: s1147994@s.eduhk.hk.

# References

Adler, I., & Shani, Y. (2026). Hybrid Human-GenAI cognitive apprenticeship: Encouraging pre-service teachers to implement instructional practices to support students' self-regulated learning. *Teaching and Teacher Education*, *169*, 105268. https://doi.org/10.1016/j.tate.2025.105268

Ahmad, M., Ravi, P., Karger, D., & Facciotti, M. (2025). How adding metacognitive requirements in support of AI feedback in practice exams transforms student learning behaviors. In *Proceedings of the Twelfth ACM Conference on Learning @ Scale* (pp. 164–175). https://doi.org/10.1145/3698205.3729542

Aldulaijan, A. T., & Almalky, S. M. (2025). The impact of generative AI tools on postgraduate students' learning experiences: New insights into usage patterns. *Journal of Information Technology Education: Research*, *24*, 003. https://doi.org/10.28945/5428

Al-Fattal, A. (2025). You do It, AI does It, you compare and reflect: Exploring reflective learning with generative AI in principles of marketing. *Journal of Marketing Education*, https://doi.org/10.1177/02734753251356691

Álvarez, R. P., Jivet, I., Pérez-Sanagustín, M., Scheffel, M., & Verbert, K. (2022). Tools designed to support self-regulated learning in online learning environments: A systematic review. *IEEE Transactions on Learning Technologies*, *15*(4), 508-522. https://doi.org/10.1109/TLT.2022.3193271

Azevedo, R., & Wiedbusch, M. (2023). Theories of metacognition and pedagogy applied to AIED systems. In the *Handbook of Artificial Intelligence in Education* (pp. 45-67). Edward Elgar Publishing. https://doi.org/10.4337/9781800375413.00013

Bassner, P., Lottner, A., & Krusche, S. (2025). Towards understanding the impact of context-aware AI tutors and general-purpose AI chatbots on student learning. *Proceedings of the 25th Koli Calling International Conference on Computing Education Research*, 1–11. https://doi.org/10.1145/3769994.3770025

Borjigin, A., & Shen, Z. (2025). AI-powered meta-reflection for self-learning in STEM: A middle school case study on smart home energy management. In *2025 IEEE Region 10 Symposium (TENSYMP)*, 1-6. IEEE. https://doi.org/10.1109/TENSYMP63728.2025.11144984

Brazão, P., & Tinoca, L. (2025). Artificial intelligence and critical thinking: A case study with educational chatbots. *Frontiers in Education*, *10*, 1630493. https://doi.org/10.3389/feduc.2025.1630493

Campbell, J. L., Quincy, C., Osserman, J., & Pedersen, O. K. (2013). Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research*, *42*(3), 294-320. https://doi.org/10.1177/0049124113500475

Cao, X., Wei, W., & Chen, Z. (2025). Metacognitive strategy networks in genAI-enhanced learning: An epistemic network and lag sequence analysis of high and low self-efficacy

students. In *2025 5th International Conference on Artificial Intelligence and Education (ICAIE)* (pp. 470-474). IEEE. https://doi.org/10.1109/ICAIE64856.2025.11158315

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

Chen, A., Wei, Y., Le, H., & Zhang, Y. (2026). Learning by teaching with ChatGPT: The effect of teachable ChatGPT agent on programming education. *British Journal of Educational Technology*, *57*(1), 163–184. https://doi.org/10.1111/bjet.70001

Chen, A., Xiang, M., Zhou, J., Jia, J., Shang, J., Li, X., Gašević, D., & Fan, Y. (2025). Unpacking help-seeking process through multimodal learning analytics: A comparative study of ChatGPT vs human expert. *Computers & Education, 226*, Article 105198. https://doi.org/10.1016/j.compedu.2024.105198

Chen, A., Lian, J., Kuang, X., & Jia, J. (2026). Can theory-driven learning analytics dashboard enhance human-AI collaboration in writing learning? Insights from an empirical experiment. *The Internet and Higher Education*, *68*, 101054. https://doi.org/10.1016/j.iheduc.2025.101054

Dang, B., Huynh, L., Gul, F., Rosé, C., Järvelä, S., & Nguyen, A. (2025). Human–AI collaborative learning in mixed reality: Examining the cognitive and socio-emotional interactions. *British Journal of Educational Technology*, *56*(5), 2078–2101. https://doi.org/10.1111/bjet.13607

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

Deng, J., Liu, C., & Zhou, Z. (2025). Enhancing visual art appreciation through AI-human collaboration: A case study of Image-Literacy-Aided instruction with Doubao. *Proceedings of the 2025 6th International Conference on Education, Knowledge and Information Management*, 440–445. https://doi.org/10.1145/3756580.3756653

Engeness, I. (2025). Cultural-historical perspective to design pedagogical AI for enhancing student writing. *Technology, Knowledge and Learning*, 1-22. https://doi.org/10.1007/s10758-025-09876-0

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

Farinetti, L., & Cagliero, L. (2025). A critical approach to ChatGPT: An experience in SQL learning. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*, 318-324. https://doi.org/10.1145/3641554.3701932

Figueiredo, V. (2025). Designing smarter conversational agents for kids: Lessons from cognitive work and means-ends analyses. *ACM Transactions on Computer-Human Interaction*, *32*(6), 1–40. https://doi.org/10.1145/3765284

Fitriati, S. W., & Williyan, A. (2025). AI-Enhanced self-regulated learning: EFL learners' prioritization and utilization in presentation skills development. *Journal of Pedagogical Research*, *9*(2), 22-37. https://doi.org/10.33902/JPR.202530647

Gong, Y., Wang, M., He, L., Xu, C., & Yu, Y. (2026). Asking, playing, learning: Investigating large language model-based scaffolding in digital game-based learning for elementary artificial intelligence education. *Journal of Educational Computing Research*, *64*(2), 311-343. https://doi.org/10.1177/07356331251396354

Goslen, A., Kim, Y. J., Rowe, J., & Lester, J. (2025). LLM-based student plan generation for adaptive scaffolding in game-based learning environments. *International Journal of Artificial Intelligence in Education*, *35*(2), 533–558. https://doi.org/10.1007/s40593-024-00421-1

Hadwin, A. F., Järvelä, S., & Miller, M. (2011). Self-regulated, co-regulated, and socially shared regulation of learning. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 65–84). Routledge.

Hak, J., Lam Johnson, N., Amoozadeh, M., Alipour, A., & Chattopadhyay, S. (2025). Observing without doing: Pseudo-apprenticeship patterns in student LLM use. In *Proceedings of the 25th Koli Calling International Conference on Computing Education Research*, 1-11. https://doi.org/10.1145/3769994.3770027

Han, Z., Ying, R., Huang, C., Tsai, C.-C., Wang, X., & He, T. (2025). Identifying students' metacognition patterns by their needs for cognitive closure in human-GenAI collaboration. *Computers & Education*, *239*, 105422. https://doi.org/10.1016/j.compedu.2025.105422

Hao, Z., Cao, J., Li, R., Yu, J., Liu, Z., & Zhang, Y. (2026). Mapping student-AI interaction dynamics in multi-agent learning environments: Supporting personalized learning and reducing performance gaps. *Computers & Education*, *241*, 105472. https://doi.org/10.1016/j.compedu.2025.105472

Hruschka, D. J., Schwartz, D., St. John, D. C., Picone-Decaro, E., Jenkins, R. A., & Carey, J. W. (2004). Reliability in coding open-ended data: Lessons learned from HIV behavioral research. *Field Methods*, *16*(3), 307-331. https://doi.org/10.1177/1525822X04266540

Huang, K., & Chen, C. (2025). Instructional video and GenAI-supported chatbot in digital game-based learning: Influences on science learning, cognitive load and game behaviours. *Journal of Computer Assisted Learning*, *41*(4), e70094. https://doi.org/10.1111/jcal.70094

Huang, Y., Zhang, K., & Li, S. (2025). How postgraduates navigate GenAI's dual pathways between research creativity and academic plagiarism: Research engagement and self-regulated learning matter. *Studies in Higher Education*, 1–22. https://doi.org/10.1080/03075079.2025.2543438

Huda, N., Anwar, K., & Kurniawan, W. (2025). ChatGPT scaffolding in supporting metacognition for limit concepts in guided inquiry mathematics learning. *Journal of Information Technology Education: Research*, *24*, 038. https://doi.org/10.28945/5645

Jalambo, M. O., Çakmak, F., & Akhter, S. (2025). Effects of self-regulated vocabulary learning with chatbots on incidental and collocational vocabulary learning and foreign language learning boredom. *Discover Education*, *4*(1), 501. https://doi.org/10.1007/s44217-025-00977-7

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*(1), 37. https://doi.org/10.1186/s41239-023-00406-5

Kazemitabaar, M., Ye, R., Wang, X., Henley, A. Z., Denny, P., Craig, M., & Grossman, T. (2024). CodeAid: Evaluating a classroom deployment of an LLM-based programming assistant that balances student and educator needs. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–20. https://doi.org/10.1145/3613904.3642773

Kim, Y. S., Moon, H. S., Lee, S., & Lee, T. Y. (2025). "HistoChat": Leveraging AI-driven historical personas for personalized and engaging middle school history education. *Proceedings of the ACM on Human-Computer Interaction*, *9*(7), 1–37. https://doi.org/10.1145/3757534

Klar, M. (2025). Using ChatGPT is easy, using it effectively is tough? A mixed methods study on K-12 students' perceptions, interaction patterns, and support for learning with generative AI chatbots. *Smart Learning Environments*, *12*(1), 32. https://doi.org/10.1186/s40561-025-00385-2

Kohnke, L., Zou, D., & Xie, H. (2025). Microlearning and generative AI for pre-service teacher education: A qualitative case study. *Education and Information Technologies*, *30*(15), 21221–21248. https://doi.org/10.1007/s10639-025-13606-5

Krupp, L., Steinert, S., Kiefer-Emmanouilidis, M., Avila, K. E., Lukowicz, P., Kuhn, J., Küchemann, S., & Karolus, J. (2024). Challenges and opportunities of moderating usage of large language models in education. In *Proceedings of the International Conference on Mobile and Ubiquitous Multimedia*, 249-254. https://doi.org/10.1145/3701571.3701590

Kumar, H., Xiao, R., Lawson, B., Musabirov, I., Shi, J., Wang, X., Luo, H., Williams, J. J., Rafferty, A. N., Stamper, J., & Liut, M. (2024). Supporting self-Reflection at scale with large language models: Insights from randomized field experiments in classrooms. *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, 86–97. https://doi.org/10.1145/3657604.3662042

Li, C., & Long, J. (2025). Comparing AI-assisted and traditional teaching in college English: Pedagogical benefits and learning behaviors. *Information*, *16*(10), 895. https://doi.org/10.3390/info16100895

Li, T., Nath, D., Cheng, Y., Fan, Y., Li, X., Raković, M., Khosravi, H., Swiecki, Z., Tsai, Y.-S., & Gašević, D. (2025). Turning real-time analytics into adaptive scaffolds for self-regulated learning using generative artificial intelligence. *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 667–679. https://doi.org/10.1145/3706468.3706559

Li, T. W., Song, Y., Sundaram, H., & Karahalios, K. (2025). Can learners navigate imperfect generative pedagogical chatbots? An analysis of chatbot errors on learning. In *Proceedings of the Twelfth ACM Conference on Learning @ Scale*, 151-163. https://doi.org/10.1145/3698205.3729550

Li, W., Cui, X., Manoharan, P., Dai, L., Liu, K., & Huang, L. (2025). AI-assisted feedback and reflection in vocal music training: Effects on metacognition and singing performance. *Frontiers in Psychology*, *16*, 1598867. https://doi.org/10.3389/fpsyg.2025.1598867

Liang, H., Hwang, G., Hsu, T., & Yeh, J. (2024). Effect of an AI-based chatbot on students' learning performance in alternate reality game-based museum learning. *British Journal of Educational Technology*, *55*(5), 2315–2338. https://doi.org/10.1111/bjet.13448

López-Pernas, S., Misiejuk, K., Oliveira, E., & Saqr, M. (2025). The dynamics of the self-regulation process in student-AI interactions: The case of problem-solving in programming education. *Proceedings of the 25th Koli Calling International Conference on Computing Education Research*, 1–12. https://doi.org/10.1145/3769994.3770043

Lu, G., Ba, S., & Yang, L. (2025). Optimizing GAI-assisted formative feedback: An experimental study on its effects on engagement, shared metacognition, and learning performance in online collaborative learning. *Education and Information Technologies*, *30*(18), 25937–25967. https://doi.org/10.1007/s10639-025-13778-0

Luo, X., Wang, S., & Hew, K. F. (2026). Development and implementation of a generative AI-based personalized recommender system to improve students' self-regulated learning and academic performance. *Computers & Education*, *241*, 105486. https://doi.org/10.1016/j.compedu.2025.105486

Ma, X., Li, C., Xu, J., Zhu, S., & Li, Y. (2025). Exploring characteristics of primary school students' self-regulated learning (SRL) behaviors in human-GenAI collaborative programming learning environments: Insights from a proposed framework. *Computers & Education*, *240*, 105453. https://doi.org/10.1016/j.compedu.2025.105453

Mekheimer, M. (2025). Generative AI-assisted feedback and EFL writing: A study on proficiency, revision frequency and writing quality. *Discover Education*, *4*(1), 170. https://doi.org/10.1007/s44217-025-00602-7

Molenaar, I. (2022). The concept of hybrid human-AI regulation: Exemplifying how to support young learners' self-regulated learning. *Computers and Education: Artificial Intelligence, 3*, 100070. https://doi.org/10.1016/j.caeai.2022.100070

Ng, D. T. K., Tan, C. W., & Leung, J. K. L. (2024). Empowering student self-regulated learning and science education through ChatGPT: A pioneering pilot study. *British Journal of Educational Technology*, *55*(4), 1328–1353. https://doi.org/10.1111/bjet.13454

Nguyen, A., Gul, F., Dang, B., Huynh, L., & Tuunanen, T. (2025). Designing embodied generative artificial intelligence in mixed reality for active learning in higher education. *Innovations in Education and Teaching International*, *62*(5), 1632–1647. https://doi.org/10.1080/14703297.2025.2499177

Nguyen, H., & Nguyen, A. (2025). Reflective practices and self-regulated learning in designing with generative artificial intelligence: An ordered network analysis. *Journal of Science Education and Technology*, *34*(5), 1178-1192. https://doi.org/10.1007/s10956-024-10175-z

Nguyen, N. N., & Barbieri, W. (2025). Mentorship in the age of generative AI: ChatGPT to support self-regulated learning of pre-service teachers before and during placements. *Education Sciences*, *15*(6), 642. https://doi.org/10.3390/educsci15060642

Ni, Y., & Lam, R. (2025). Students' perceptions of multiliteracies development using AI-assisted portfolio assessment. *Pedagogies: An International Journal*, 1–28. https://doi.org/10.1080/1554480X.2025.2545212

OpenAI. (2022, November 30). *Introducing ChatGPT*. Retrieved February 3, 2026, from https://openai.com/index/chatgpt/

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Pan, M., Lai, C., Guo, K., & Huang, J. (2026). Self-regulation plus individual interests? A design-based study on the development of a GenAI-empowered platform for self-directed out-of-class reading. *Computers & Education*, *241*, 105474. https://doi.org/10.1016/j.compedu.2025.105474

Prather, J., Reeves, B. N., Denny, P., Becker, B. A., Leinonen, J., Luxton-Reilly, A., Powell, G., Finnie-Ansley, J., & Santos, E. A. (2023). "It's weird that it knows what I want": Usability and interactions with copilot for novice programmers. *ACM Transactions on Computer-Human Interaction*, *31*(1), 1-31. https://doi.org/10.1145/3617367

Prather, J., Reeves, B. N., Leinonen, J., MacNeil, S., Randrianasolo, A. S., Becker, B. A., Kimmel, B., Wright, J., & Briggs, B. (2024). The widening gap: The benefits and harms of generative AI for novice programmers. *Proceedings of the 2024 ACM Conference on International Computing Education Research - Volume 1*, 469–486. https://doi.org/10.1145/3632620.3671116

Radović, S., Wetchy, E., & Seidel, N. (2025). Implementing the self-regulated learning structured interview protocol with generative AI: A novel approach for evaluating students' SRL skills. *Journal of Research on Technology in Education*, 1–18. https://doi.org/10.1080/15391523.2025.2547176

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

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521-562. https://doi.org/10.1207/s15516709cog2605_1

Seo, K., Tang, J., Roll, I., Fels, S., & Yoon, D. (2021). The impact of artificial intelligence on learner–instructor interaction in online learning. *International Journal of Educational Technology in Higher Education, 18*(1), 54. https://doi.org/10.1186/s41239-021-00292-9

Shaffer, D. W. (2017). *Quantitative Ethnography*. Madison, WI: Cathcart Press.

Shi, L. (2025). The integration of advanced AI-enabled emotion detection and adaptive learning systems for improved emotional regulation. *Journal of Educational Computing Research*, *63*(1), 173–201. https://doi.org/10.1177/07356331241296890

Singh, A., Taneja, K., Guan, Z., & Ghosh, A. (2025). Protecting human cognition in the age of AI. *arXiv preprint arXiv:2502.12447*. https://doi.org/10.48550/arXiv.2502.12447

Sun, D., Xu, P., Zhang, Jing, Liu, R., & Zhang, Jun. (2025). How self-regulated learning is affected by feedback based on large language models: Data-driven sustainable development in computer programming learning. *Electronics*, *14*(1), 194. https://doi.org/10.3390/electronics14010194

Sun, Y., & Lan, G. (2025). Enhancing critical language awareness in EAL writing education amid the rise of generative artificial intelligence. *System*, *134*, 103806. https://doi.org/10.1016/j.system.2025.103806

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

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology, 8*(1), 45. https://doi.org/10.1186/1471-2288-8-45

Tomisu, H., Ueda, J., & Yamanaka, T. (2025). The cognitive mirror: A framework for AI-powered metacognition and self-regulated learning. *Frontiers in Education*, *10*, 1697554. https://doi.org/10.3389/feduc.2025.1697554

Tran, H., Crosthwaite, P., & Pham, Q. H. P. (2025). Students' self-determination in using machine translation and generative AI tools for English for academic purposes. *Journal of English for Academic Purposes*, *78*, 101578. https://doi.org/10.1016/j.jeap.2025.101578

Waldo, J., & Boussard, S. (2024). GPTs and hallucination: Why do large language models hallucinate? *ACM Queue, 22*(4), 19-33. https://doi.org/10.1145/3688007

Wang, P., Liu, T., Yang, Y., & Xiang, X. (2025). Optimizing self-regulated learning: A mixed-methods study on GAI's impact on undergraduate task strategies and metacognition. *British Journal of Educational Technology*, bjet.70018. https://doi.org/10.1111/bjet.70018

Wang, S., Luo, X., & Hew, K. F. (2025). Enhancing self-regulated learning using generative AI: Development and evaluation of SRLMentor. *IEEE Transactions on Learning Technologies*, *18*, 1048-1061. https://doi.org/10.1109/TLT.2025.3634216

Wang, W.-S., Lin, C.-J., Lee, H.-Y., Huang, Y.-M., & Wu, T.-T. (2025). Enhancing self-regulated learning and higher-order thinking skills in virtual reality: The impact of ChatGPT-integrated feedback aids. *Education and Information Technologies*, *30*(14), 19419–19445. https://doi.org/10.1007/s10639-025-13557-x

Wei, L. (2023). Artificial intelligence in language instruction: Impact on English learning achievement, L2 motivation, and self-regulated learning. *Frontiers in Psychology*, *14*, 1261955. https://doi.org/10.3389/fpsyg.2023.1261955

Wei, X., Wang, L., Lee, L.-K., & Liu, R. (2025). The effects of generative AI on collaborative problem-solving and team creativity performance in digital story creation: An experimental study. *International Journal of Educational Technology in Higher Education*, *22*(1), 23. https://doi.org/10.1186/s41239-025-00526-0

West, R., & Martin, F. (2024). What type of paper are you writing? A taxonomy of review and theory scholarship distinguished by their summary and advocacy arguments. *Educational Technology Research and Development, 72*, 2443-2476. https://doi.org/10.1007/s11423-023-10233-0

Winne, P. H. (2018). Theorizing and researching levels of processing in self-regulated learning. *British Journal of Educational Psychology*, *88*(1), 9-20. https://doi.org/10.1111/bjep.12173

Wong, J., Baars, M., Davis, D., Van Der Zee, T., Houben, G. J., & Paas, F. (2019). Supporting self-regulated learning in online learning environments and MOOCs: A systematic review. *International Journal of Human–Computer Interaction*, *35*(4-5), 356-373. https://doi.org/10.1080/10447318.2018.1543084

Wu, T.-T., Lee, H.-Y., Li, P.-H., Huang, C.-N., & Huang, Y.-M. (2024). Promoting self-regulation progress and knowledge construction in blended learning via ChatGPT-based learning aid. *Journal of Educational Computing Research*, *61*(8), 1539–1567. https://doi.org/10.1177/07356331231191125

Yan, W., Nakajima, T., & Sawada, R. (2024). Benefits and challenges of collaboration between students and conversational generative artificial intelligence in programming learning: An empirical case study. *Education Sciences*, *14*(4), 433. https://doi.org/10.3390/educsci14040433

Yu, J. H., Tu, F., Chen, H., Ding, J., Hsieh, C.-J., Dong, L., Kim, H., & Watson, S. L. (2026). Measuring cognitive presence in online discussions: Automated detection and instructional insights from a MOOC context. *Educational Technology Research and Development*. https://doi.org/10.1007/s11423-026-10726-8

Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments*, *11*(1), 28. https://doi.org/10.1186/s40561-024-00316-7

Zhai, Y., & Nezakatgoo, B. (2025). Evaluating AI-powered applications for Enhancing undergraduate students' metacognitive strategies, self-determined motivation, and Social Learning in English language education. *Scientific Reports*, *15*(1), 35199. https://doi.org/10.1038/s41598-025-19118-z

Zhang, W., & Liu, X. (2025). Artificial intelligence-generated content empowers college students' critical thinking skills: What, how, and why. *Education Sciences*, *15*(8), 977. https://doi.org/10.3390/educsci15080977

Zhang, Y., Tian, H., & Lu, J. (2025). The relationship between higher-order thinking and problem-solving skills development among pre-service teachers using generative AI: An analysis based on moderated mediation. *BMC Psychology*, *13*(1), 1094. https://doi.org/10.1186/s40359-025-03404-6

Zhu, M. (2025). Leveraging ChatGPT to support self-regulated learning in online courses. *TechTrends*, *69*(5), 914–924. https://doi.org/10.1007/s11528-025-01075-z

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In *Handbook of self-regulation* (pp. 13-39). Academic press. https://doi.org/10.1016/B978-012109890-2/50031-7

Zimmerman, B. J., & Moylan, A. R. (2009). Self-regulation: Where metacognition and motivation intersect. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.) *Handbook of Metacognition in Education* (pp. 299-315). Routledge.

# Appendix A. Prompt

You are a research assistant tasked with screening literature for a systematic review based on the provided paper title and abstract. Your goal is to evaluate the text against four specific criteria, provide a binary answer (1 for YES, 0 for NO) for each, provide a confidence score (0-100) for each answer, and extract supporting evidence snippets.

First, for Q1 (Metacog), determine if the paper explicitly involves metacognitive or self-regulated learning concepts. The title or abstract must mention ANY of the following: metacognition, self-regulation/SRL, co-regulation, socially shared regulation of learning/SSRL, metacognitive laziness, overreliance on AI, or cognitive offloading of high-level processes (planning, monitoring, or evaluating). Crucially, exclude simple task-level offloading like grammar fixing or basic arithmetic.

Second, for Q2 (Edu), determine if the work is situated within an educational or learning context. This includes teaching, learning, students, instruction, pedagogy, assessment, feedback, online learning, MOOCs, distance learning, or CSCL.

Third, for Q3 (Empirical), determine if the paper reports an empirical study involving real data collection and analysis (e.g., experiments, quasi-experiments, user studies, surveys, interviews, classroom observations, think-aloud studies, or learning analytics). Exclude: purely conceptual/theoretical papers, commentaries, editorials, systematic literature reviews (SLR), scoping reviews, meta-analyses, book chapters, and opinion pieces.

Fourth, for Q4 (GenAI/LLM), determine if the study utilizes Generative AI or Large Language Model technologies (e.g., GenAI, GPT, ChatGPT, Gemini, Claude, LLMs, AI agents, AI copilots, or LLM-based chatbots). Exclude traditional machine learning unrelated to LLMs.

Now, given the following paper:

Title: {title}
Abstract: {abstract}
Respond ONLY in valid JSON in the following format, ensuring evidence snippets are copied or closely paraphrased from the text (use an empty string "" if no evidence exists for a "0" answer):

```
{{
 "answers": {{
        "q1": 0 or 1,
        "q2": 0 or 1,
        "q3": 0 or 1,
        "q4": 0 or 1
 }},
 "confidence": {{
        "q1": 0-100,
        "q2": 0-100,
        "q3": 0-100,
        "q4": 0-100
```

```
  }},
  "evidence": {{
        "q1": "text from title/abstract supporting your decision for Q1",
        "q2": "text from title/abstract supporting your decision for Q2",
        "q3": "text from title/abstract supporting your decision for Q3",
        "q4": "text from title/abstract supporting your decision for Q4"
  }}
}}
"""
```

# Appendix B. Characteristics of The Final Key Studies (*N* = 65)

| Paper | Level | | Discipline | | | | Research Methodology | | |
|---|---|---|---|---|---|---|---|---|---|
| | Higher Education | K-12 | STEM Ed | Literacy & Language Ed | Other | Not report | Quantitative | Qualitative | Mixed Methods |
| Adler & Shani (2026) | ✓ | | | | ✓ | | | | ✓ |
| Ahmad et al. (2025) | ✓ | | ✓ | | | | ✓ | | |
| Al-Fattal (2025) | ✓ | | | | ✓ | | | ✓ | |
| Aldulaijan & Almalki (2025) | ✓ | | ✓ | | | | | ✓ | |
| Bassner et al. (2025) | ✓ | | ✓ | | | | | | ✓ |
| Borjigin & Shen (2025) | | ✓ | ✓ | | | | ✓ | | |
| Brazão & Tinoca (2025) | ✓ | | ✓ | | | | | ✓ | |
| Cao et al. (2025) | | ✓ | | ✓ | | | | | ✓ |
| Chen et al. (2025) | ✓ | | ✓ | | | | ✓ | | |
| Chen et al. (2026) | ✓ | | ✓ | | | | ✓ | | |
| Dang et al. (2025) | ✓ | | ✓ | | | | ✓ | | |
| Deng et al. (2025) | | ✓ | | ✓ | | | | ✓ | |
| Engeness (2025) | | ✓ | ✓ | | | | | | ✓ |
| Fan et al. (2025) | ✓ | | | ✓ | | | ✓ | | |
| Farinetti & Cagliero (2025) | ✓ | | ✓ | | | | ✓ | | |
| Figueiredo (2025) | | ✓ | | | | ✓ | | | ✓ |
| Fitriati & Williyan (2025) | ✓ | | | ✓ | | | | ✓ | |
| Gong et al. (2025) | | ✓ | ✓ | | | | ✓ | | |
| Goslen et al. (2025) | | ✓ | ✓ | | | | | | ✓ |
| Hak et al. (2025) | ✓ | | ✓ | | | | | | ✓ |
| Han et al. (2025) | ✓ | | ✓ | | | | | | ✓ |
| Hao et al. (2026) | ✓ | | ✓ | | | | ✓ | | |
| Huang & Chen (2025) | | ✓ | ✓ | | | | ✓ | | |
| Huang et al. (2025) | ✓ | | | | | ✓ | | | ✓ |
| Huda et al. (2025) | ✓ | | ✓ | | | | | | ✓ |
| Jalambo et al. (2025) | ✓ | | | ✓ | | | | | ✓ |
| Kazemitabaar et al. (2024) | ✓ | | ✓ | | | | | | ✓ |
| Kim et al. (2025) | | ✓ | | | | ✓ | | | ✓ |
| Klar (2025) | | ✓ | | | | ✓ | | | ✓ |

| Paper | Level | | Discipline | | | | Research Methodology | | |
|---|---|---|---|---|---|---|---|---|---|
| | Higher Education | K-12 | STEM Ed | Literacy & Language Ed | Other | Not report | Quantitative | Qualitative | Mixed Methods |
| Kohnke et al. (2025) | ✓ | | | ✓ | | | | ✓ | |
| Krupp et al. (2024) | ✓ | | ✓ | | | | ✓ | | |
| Kumar et al. (2024) | ✓ | | ✓ | | | | ✓ | | |
| C. Li & Long (2025) | ✓ | | | ✓ | | | | | ✓ |
| T. Li et al. (2025) | ✓ | | | ✓ | | | ✓ | | |
| T. W. Li et al. (2025) | ✓ | | ✓ | | | | | | ✓ |
| W. Li et al. (2025) | ✓ | | | | | ✓ | | | ✓ |
| Liang et al. (2024) | | ✓ | ✓ | | | | ✓ | | |
| López-Pernas et al. (2025) | ✓ | | ✓ | | | | | | ✓ |
| Lu et al. (2025) | ✓ | | | | | ✓ | ✓ | | |
| Luo et al. (2026) | ✓ | | | | | ✓ | | | ✓ |
| Ma et al. (2025) | | ✓ | ✓ | | | | ✓ | | |
| Mekheimer (2025) | ✓ | | | ✓ | | | | | ✓ |
| Ng et al. (2024) | | ✓ | ✓ | | | | | | ✓ |
| Nguyen & Barbieri (2025) | ✓ | | ✓ | | | | | | ✓ |
| Nguyen & Nguyen (2025) | ✓ | | ✓ | | | | | | ✓ |
| Nguyen et al. (2025) | ✓ | | | | | ✓ | | | ✓ |
| Ni & Lam (2025) | ✓ | | | ✓ | | | | ✓ | |
| Pan et al. (2026) | ✓ | | | ✓ | | | | | ✓ |
| Prather et al. (2023) | ✓ | | ✓ | | | | | ✓ | |
| Prather et al. (2024) | ✓ | | ✓ | | | | | | ✓ |
| Radović et al. (2025) | ✓ | | ✓ | | | | | | ✓ |
| Shi (2025) | ✓ | | | ✓ | | | | | ✓ |
| Sun & Lan (2025) | ✓ | | ✓ | | | | | | ✓ |
| Sun et al. (2025) | ✓ | | | ✓ | | | ✓ | | |
| Tran et al. (2025) | ✓ | | | ✓ | | | | | ✓ |
| P. Wang et al. (2025) | ✓ | | | ✓ | | | | | ✓ |
| S. Wang et al. (2025) | ✓ | | | | | ✓ | | | ✓ |
| W.-S. Wang et al. (2025) | ✓ | | ✓ | | | | ✓ | | |
| Wei (2023) | ✓ | | | ✓ | | | | | ✓ |
| Wei et al. (2025) | ✓ | | ✓ | | | | | | ✓ |
| Wu et al. (2024) | ✓ | | ✓ | | | | ✓ | | |
| Yan et al. (2024) | ✓ | | ✓ | | | | | ✓ | |

| Paper | Level | | Discipline | | | | Research Methodology | | |
|---|---|---|---|---|---|---|---|---|---|
| | Higher Education | K-12 | STEM Ed | Literacy & Language Ed | Other | Not report | Quantitative | Qualitative | Mixed Methods |
| Zhai & Nezakatgoo (2025) | ✓ | | | ✓ | | | | | ✓ |
| Zhang & Liu (2025) | ✓ | | | | | ✓ | | | ✓ |
| Zhang et al. (2025) | ✓ | | | | ✓ | | | | ✓ |
| Zhu (2025) | ✓ | | ✓ | | | | | ✓ | |