Computers and Education: Artificial Intelligence 10 (2026) 100572

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence

Scaffolding critical thinking with generative AI: Design principles for integrating large language models in higher education

Mireia Vendrell [a,*], Samantha-Kaye Johnston [b]

[a] Multiagent Systems, Artificial Intelligence Research Institute - Spanish National Research Council (IIIA-CSIC), Barcelona, Spain
[b] Assessment and Evaluation Research Centre (AERC), University of Melbourne, Melbourne, Australia

ARTICLE INFO

Keywords:
Generative AI
Large language models
Critical thinking
Higher education
Epistemic agency
Pedagogical design
Learning technology

ABSTRACT

The rapid adoption of Large Language Models (LLMs) such as GPT-4 and DeepSeek R1 is transforming learning in higher education, yet unstructured use can weaken critical thinking by encouraging cognitive offloading, metacognitive disengagement, and reduced epistemic agency. This paper presents a conceptual and normative analysis that synthesises research from cognitive psychology, educational theory, and AI ethics to develop a design-oriented pedagogical framework for integrating LLMs in ways that strengthen, rather than displace, higher-order thinking. Grounded in design-based research principles, the framework identifies six essential processes that underpin critical engagement: conceptual interpretation, inferential reasoning, evaluative judgement, metacognitive regulation, intellectual curiosity, and epistemic integrity. These processes are translated into eight actionable design principles, including preserving cognitive friction, positioning LLMs as provisional thinking partners, embedding evaluation throughout learning, and sequencing AI-mediated with AI-free phases. Two illustrative classroom scenarios showcase practical application. The framework offers educators a theoretically grounded and practically applicable model for cultivating critical thinking and epistemic responsibility in AI-rich learning environments, contributing to emerging new systems of learning in higher education.

1. Introduction

Generative AI tools such as GPT-4 and DeepSeek R1 have moved rapidly from experimental novelties to everyday academic companions. A recent global survey reports that 86% of university students now use AI in their studies, with more than half engaging with these tools weekly, primarily to summarise documents, check grammar, paraphrase, and generate first drafts (Digital Education Council, 2024). While often perceived as convenient (though this perception is contested; see Selwyn, 2025), these tools are not designed with educational goals in mind. Large Language Models (LLMs) generate responses using probabilistic language modeling, predicting likely word sequences from training data, rather than through conceptual understanding or reasoning. This distinction has significant pedagogical implications. Without careful integration, widespread adoption risks cognitive offloading, diminished metacognitive engagement, and weakened epistemic agency, which we define as the learner's capacity to critically evaluate, justify, and take ownership of knowledge.

This paper advances the position that LLMs are not educationally neutral; their effects are contingent rather than fixed. While emerging research has documented both beneficial and harmful outcomes (e.g., Deng et al., 2024; Gerlich, 2025), their ultimate impact depends on how they are designed, including what they afford, obscure, or prioritise, and on the pedagogical context in which they are implemented. These factors jointly influence how they are used, what forms of learning they support or constrain, and whose epistemic values they reflect or exclude. In response to this complexity, we propose a normative, design-oriented pedagogical framework for the intentional integration of LLMs into higher education, with the specific goal of fostering critical thinking. Rather than banning or embracing these tools wholesale, we argue that educators must cultivate learning environments in which students engage critically with AI, using it to extend their reasoning rather than to replace it.

To understand the educational implications of LLMs, it is essential to examine how they work. Models such as GPT-4 and DeepSeek R1 are built on transformer architectures, which use self-attention mechanisms

This article is part of a special issue entitled: GenAI for Higher Edu published in Computers and Education: Artificial Intelligence.

* Corresponding author. Campus UAB Carrer de Can Planas, Zona 2, Bellaterra, 08193, Barcelona, Spain.
E-mail address: mireia.vendrell@iiia.csic.es (M. Vendrell).

https://doi.org/10.1016/j.caeai.2026.100572
Received 11 August 2025; Received in revised form 8 January 2026; Accepted 3 March 2026
Available online 7 March 2026

to weigh contextual relationships between words (see Vaswani et al., 2017). Their training occurs in two stages. First, during pre-training, the model is exposed to a massive corpora of text to optimise parameters for probabilistic word prediction. The result is a base model: a system with broad linguistic knowledge and the ability to predict likely next words in a sequence. Second, fine-tuning techniques such as instruction tuning and Reinforcement Learning from Human Feedback (RLHF) align the model's outputs with human preferences and task-specific formats (Casper et al., 2023). The result is an assistant model capable of coherent and contextually appropriate interaction. However, these systems still lack intentionality, comprehension, or reasoning. They simulate understanding but do not embody it; an epistemic distinction with critical implications for their use in educational settings.

Without deliberate integration, students may conflate linguistic fluency with epistemic validity; that is, they may mistake well-formed language for justified or reliable knowledge. When learners rely on LLMs to produce responses without engaging in the underlying reasoning, critical thinking is at risk of becoming attenuated. Ennis (1985) defines critical thinking as “reflective and reasonable thinking that is focused on deciding what to believe or do” (p. 45), a formulation that emphasises its practical orientations and its role in guiding action. The Delphi Report (Facione, 1990) elaborates this definition by distinguishing between cognitive skills, such as interpreting evidence and drawing inferences, and intellectual dispositions, including open-mindedness, curiosity, and humility. This distinction underscores that critical thinking is not only a matter of analytical ability but also of cultivating habits of mind that sustain inquiry over time. Halpern (2014) adds that critical thinking is not domain-specific but must be transferable and practiced deliberately across varied contexts, highlighting the need for intentional and repeated application if it is to become a durable competency. Importantly, these capacities do not emerge spontaneously but are cultivated through structured learning, authentic problem-solving, and dialogic engagement (Abrami et al., 2015; Dwyer, 2023; van Brussel et al., 2020). From a broader perspective, Barnett's (1997, 2015) notion of *criticality* extends this view by integrating thinking, being, and acting. In this formulation, critical thinking is not confined to the evaluation of arguments or the solving of problems in isolation. It also involves developing a reflective understanding of oneself as a learner and a citizen, recognising the values and assumptions that shape one's perspectives, and applying the awareness to engage constructively with the world. Critical thinking reaches its fullest potential when sound reasoning is linked to ethical commitment and translated into meaningful action in professional, social, and civic life. This orientation positions critical thinking as both an intellectual and moral endeavour, where the goal is not only to reason well but also to act with integrity and responsibility in the face of complex and contested issues.

Emerging research supports the concern that LLM use may disrupt this process. Stadler et al. (2024) found that students who relied on LLMs for inquiry-based tasks reported significantly lower cognitive load and produced weaker arguments compared to peers who used traditional search engines. Complementing these findings, Kosmyna et al. (2025) recently demonstrated that passive interaction with LLMs is associated with diminished attentional engagement and reduced cognitive modulation, indicating a measurable decline in sustained mental effort. Gerlich (2025) observed a negative correlation between frequent AI use and critical thinking performance, particularly among younger students, who showed higher reliance on AI-generated outputs and reduced capacity for independent evaluation. Similarly, Lee et al. (2025) found that confidence in AI tools was associated with lower levels of critical engagement, while self-confidence predicted more thoughtful reasoning. Yatani et al. (2024) noted that over-reliance on LLMs may reduce users' tendency to scrutinize content, increasing susceptibility to misinformation and hallucinated claims. Darvishi et al. (2024) concluded that repeated use of generative AI as a substitute for effortful thinking can diminish students' initiative and epistemic agency over time. Fisher et al. (2025) added that interacting with politically biased LLMs can shift users' attitudes and decisions toward the model's stance, even when the bias opposes their prior views, and that such influence occurs regardless of whether users recognise the bias. A large-scale systematic review by Zhai et al. (2024) further synthesised these concerns, showing that over-reliance on AI dialogue systems, particularly those embedded with generative models, can impair critical thinking, decision-making, and analytical reasoning across higher education contexts. These findings collectively suggest that unstructured AI use may erode not only reasoning capacity but also the inclination to reason at all.

This trajectory, however, is not inevitable. Under thoughtfully scaffolded conditions, generative AI has demonstrated the potential to enhance cognitive development. Studies report improvements in academic performance (Deng et al., 2024), creativity and problem-solving (Pardos & Bhandari, 2024), language development (Karataş et al., 2024), and student motivation (Heung & Chiu, 2025). Critically, Kosmyna et al. (2025) found that when students first engaged with a task independently before consulting an LLM, their outputs were significantly stronger. This underscores the importance of sequencing and instructional framing. These results parallel earlier technology integrations, such as calculators and search engines, which yield learning gains when embedded in pedagogical designs that preserve key cognitive processes. The question, then, is not whether LLMs support or hinder learning, but under what conditions they do so.

Answering that question requires more than reactive policy or short-term restrictions. Generative AI is not a passing disruption but a structural shift in academic and professional practice. Limiting access may delay misuse, but does little to cultivate the skills and dispositions students need to think and act critically in AI-rich environments. LLMs challenge educators to rethink how critical thinking, intellectual autonomy, and epistemic responsibility are taught. The ability to collaborate with AI systems without surrendering cognitive agency is fast becoming a core competency in higher education (Lokesh et al., 2024).

This paper, therefore, does not advocate uncritical adoption of LLMs, nor their exclusion. Instead, it develops a theoretically grounded, pedagogically actionable framework for scaffolding critical thinking in AI-enhanced learning environments. To guide the development of the proposed pedagogical model, this paper adopts a design-oriented conceptual synthesis methodology grounded in the logics of Design-Based Research (DBR) and instructional-design theory; that is, a normative, theory-informed approach that prescribes how instruction should be structured to promote specific educational goals (Reigeluth, 2013). While it does not report an empirical study, it draws from the epistemological commitments of DBR, particularly its emphasis on theory-informed design, practical relevance, and iterative refinement (McKenney & Reeves, 2018). Rather than using formal conjecture mapping (see Sandoval, 2014), the framework advances a set of design principles explicitly aligned with specific cognitive, metacognitive, and epistemic processes that support critical thinking in AI-mediated environments. It also draws on considerations from AI ethics and on traditions of critical pedagogy, particularly the Freirean view of education as a dialogic, emancipatory practice aimed at fostering reflection, action, and social transformation (Freire, 1970, 1972). This positions the framework not only as a heuristic guide for instructional design but also as a normative intervention aimed at preserving epistemic agency and intellectual autonomy in higher education.

Guided by this position, the aim of the study is to develop a conceptually grounded, design-oriented pedagogical framework for integrating LLMs into higher education in ways that preserve and extend critical thinking.

To this end, the paper addresses the following research questions:

(Sub-RQ1) *How does unstructured student interaction with LLMs affect students' engagement and the development of critical thinking?*

(Sub-RQ2) *What aspects of students' thinking and learning should be preserved or enhanced when integrating LLMs into educational practice?*

(Sub-RQ3) *What pedagogical design principles can guide the use of LLMs to scaffold critical thinking effectively?*

2. Unstructured use of LLMs and its impact on the development of critical thinking

(Sub-RQ1) *How does unstructured student interaction with LLMs affect students' engagement and the development of critical thinking?*

To identify the functions that should be preserved when integrating LLMs into education (explored in Section 3), we must first examine how unstructured use affects students' thinking and engagement with knowledge. This section addresses three interconnected risks: cognitive offloading, metacognitive disengagement, and epistemic narrowing. *Cognitive* refers to effortful processes of interpreting, reasoning, and evaluating information, *metacognitive* refers to the monitoring and regulation of those processes, and *epistemic* refers to the norms and responsibilities governing how knowledge claims are evaluated, justified, and ethically engaged with. Each presents a distinct challenge to the cultivation of critical thinking in higher education.

2.1. Cognitive offloading: how LLMs diminish effortful thought

LLM tools offer linguistic fluency and rapid responses, but their convenience introduces a fundamental pedagogical tension: they may displace the cognitive effort essential to deep learning. When students rely on LLMs to summarise texts, generate ideas, or explain complex concepts, they often bypass the productive struggle that underpins critical thinking. For example, when prompted to explain philosophical theories or scientific principles, LLMs produce fluent summaries that mask ambiguity, historical context, or conceptual nuance. The result is a shift from process to product; a form of what Morozov (2013) critiques as *technological solutionism*, where education becomes a matter of efficiency rather than intellectual engagement.

This concern spans disciplines. In STEM fields, students may use LLMs to generate code or solve equations without engaging with underlying principles. Neurocognitive studies (e.g., Kosmyna et al., 2025) show that such passive interaction reduces sustained attention and mental effort. These effects are analogous to the decline in spatial memory observed with habitual GPS use (Dahmani & Bohbot, 2020).

Over-reliance on LLMs can erode the capacity to construct arguments, interrogate assumptions, and navigate conceptual uncertainty (Darvishi, 2024; Lee et al., 2025; Stadler et al., 2024), key skills for sustained inquiry and independent thought. These cognitive and epistemic dispositions are often bypassed when students default to the fluency and immediacy of AI-generated responses. The risk is especially acute among younger learners who may lack the epistemic norms and metacognitive strategies needed to critically assess AI content. For these students, LLMs can function as *epistemic surrogates*, appearing authoritative while offering conclusions without justification (Gerlich, 2025).

Cognitive Load Theory helps explain this dynamic. While LLMs reduce extraneous load by handling surface-level tasks, they may also suppress germane load, which is the effort involved in constructing meaning and engaging in deep reasoning (Jose et al., 2025). Stadler et al. (2024) found that students using LLMs for inquiry-based tasks reported lower cognitive effort and produced weaker arguments, suggesting a trade-off between fluency and depth. Yet it is precisely this cognitive friction, involving engagement with ambiguity and complexity, that fosters deeper learning (D'Mello & Graesser, 2014; Kapur, 2008).

This trade-off underscores a broader concern: generating coherent text is not equivalent to critical thought. LLMs streamline superficial effort but may inhibit the kind of friction necessary for epistemic growth. As Facione (1990) emphasises, critical thinking includes not only cognitive skills but also dispositions such as curiosity, humility, and intellectual perseverance. These traits are cultivated through sustained engagement with complexity. Yet these very conditions are often circumvented when students rely on AI outputs designed to maximize fluency and minimize friction. Without pedagogical scaffolding, LLMs risk encouraging a style of engagement that prioritises immediacy over inquiry, gradually eroding the habits of mind that critical thinking and meaningful learning demand.

2.2. Metacognitive disengagement: how LLMs undermine self-regulation

Beyond reducing cognitive effort, unstructured AI use weakens students' capacity to monitor and regulate their thinking. This metacognitive disengagement is particularly concerning in settings where students accept fluent outputs without reflection or verification.

Because these systems generate linguistic fluent and context-conditioned responses, students may mistake surface-level fluency for conceptual accuracy. This reflects the fluency heuristic, a cognitive bias in which ease of processing is misinterpreted as truth (Oppenheimer, 2008). Without instructional scaffolding, students may accept AI-generated outputs at face value, bypassing the reasoning and verification that critical thinking requires. This weakens epistemic vigilance, a key component of reflective judgement.

LLMs also obscure their own reasoning. Their explanations are generated using the same probabilistic methods as their answers, offering no genuine logic or traceable evidence. Repeated exposure to such black-box outputs can, therefore, normalise a cognitive style that values fluency over justification and closure over critical inquiry.

Fan et al. (2024) describe a resulting pattern of *metacognitive laziness*, where students not only generate content with LLMs but also defer judgement about its quality. This often manifests in classroom contexts as passive acceptance of feedback or resistance to revision. From the standpoint of critical thinking, it diminishes reflective doubt and iterative reasoning. From a design ethics perspective, it shows how systems built for convenience discourage sustained mental effort. Ultimately, these tendencies can lead to epistemic automation. LLMs become default authorities, even when their claims are opaque or unjustified. The deeper risk, therefore, is not just poor reasoning but the loss of the disposition to think critically at all.

2.3. Epistemic narrowing: how LLMs constrain intellectual diversity

In addition to diminishing cognitive and metacognitive engagement, unstructured LLM use contributes to epistemic narrowing, a process that constrains the conditions for critical thinking by reducing students' exposure to diverse, ambiguous, or contested forms of knowledge. These systems are typically optimised for convergence, producing singular, self-contained answers that prioritise fluency and coherence over exploration and uncertainty. As a result, learners are less likely to face conceptual ambiguity, provisional reasoning, and dialogic exploration, all of which are essential for inquiry and intellectual growth (Dwyer, 2023).

This narrowing of thought is compounded by hallucination, the tendency of LLMs to generate fluent but inaccurate or fabricated information. Students lacking domain expertise may accept these outputs uncritically. As McClure et al. (2024) warns, rhetorical fluency can mask factual weakness. Seamless interfaces further obscure uncertainty, encouraging passive trust rather than active questioning.

Even when accurate, LLM outputs typically reinforce dominant knowledge structures. Trained on large-scale internet corpora, they reproduce mainstream viewpoints while marginalising counter-hegemonic or underrepresented perspectives. Much of the training data originates from Euro-American sources, embedding dominant cultural, linguistic, and epistemic norms, and thereby creating an uneven landscape of knowledge. For example, students from underrepresented backgrounds, this limits the visibility of their ways of knowing and narrows the horizon of whose knowledge is valued. For students situated within mainstream viewpoints, the risk is enclosure within familiar worldviews, where alignment with dominant narratives goes unchallenged, reducing opportunities for critical questioning, cultural awareness, or engagement with alternative perspectives. This runs counter to the aims of critical thinking, which include the ability to ask: *Whose knowledge is represented? What assumptions are embedded in the framing? What perspectives are excluded?* (Kudina et al., 2025). In this way, LLMs influence not only access to information but also the boundaries of what is considered legitimate knowledge (i.e., what is considered worth knowing, questioning, or reimagining). This concern is compounded by the fact that, to date, many of the most widely used models in education, including versions of ChatGPT, are not open source; their foundational architectures, training data, and design choices remain largely inaccessible. As a result, these systems not only mediate access to knowledge but also obscure how content is selected, shaped or silenced.

As Ozalp et al. (2022) argue, this opacity reflects a broader strategy through which dominant technology platforms consolidate control over knowledge infrastructures, especially in regulated domains like education. Komljenovic et al. (2023) further argue that venture capital-funded edtech companies are not neutral service providers, but political and economic actors whose priorities increasingly shape public education policy. This convergence of platform dominance and investor influence risks subordinating educational values to commercial logics, with significant consequences for critical thinking. When access to knowledge is governed by proprietary systems aligned with market interests, opportunities for open inquiry, pluralism, and democratic deliberation are systematically constrained.

These issues, therefore, are not only cognitive but also ethical and educational. Critical thinking entails more than analytical skill; it involves intellectual honesty, reflective judgment, and a willingness to question the foundations of knowledge itself (Paul & Elder, 2007). When students default to AI-generated answers, they risk bypassing this responsibility, opting for efficiency over reflection and surface fluency over meaningful understanding. This shift encourages a pattern of uncritical acceptance, where the ease of automated outputs displaces the struggle and curiosity that authentic learning requires. More deeply, it raises urgent questions about the values we encode into our learning environments: *Are we fostering independent thinkers equipped to question dominant and commercially driven narratives, or are we conditioning passive users to accept inherited assumptions without critique?* Importantly, the way we integrate LLMs into education will shape not only how students learn, but who they become as knowers, questioners, and participants in a shared intellectual and civic life.

3. Essential cognitive and metacognitive processes

(Sub-RQ2) *What aspects of students' thinking and learning should be preserved or enhanced when integrating LLMs into educational practice?*

If Section 2 examined how unstructured use of LLMs can attenuate critical thinking, this section identifies the specific intellectual processes that must be deliberately supported in AI-enhanced learning environments. Drawing from Bloom's revised taxonomy (Anderson & Krathwohl, 2001), Halpern's psychological model of critical thinking (2014), and the APA Delphi Report (Facione, 1990), we articulate six essential and interrelated processes: conceptual interpretation, inferential reasoning, evaluative judgement, metacognitive regulation, intellectual curiosity, and epistemic integrity (see Table 1).

Each process underpins higher-order thinking and critical engagement and is increasingly relevant as learners navigate the epistemic affordances and limitations of AI systems. These capacities are not only cognitive but also dispositional, as critical thinking involves habits of inquiry, intellectual humility, and epistemic responsibility. Aligning with UNESCO's (2023) guidance on digital learning and human agency, these processes also speak to broader aims of education, including personal autonomy, social participation, and ethical knowledge production.

Table 1
Framework integration: from risks and theory to pedagogical design.

Risks of unstructured LLM use	Intellectual processes at risk	Educational & theoretical foundations	Ethical/ Normative foundations
Cognitive offloading: Offloading effort; skipping deep engagement	(1) Conceptual interpretation	Bloom: Understand, Analyse Facione: Interpretation Halpern: Verbal reasoning	Human agency & Thinking as part of being
Cognitive offloading: Accepting fluent but unexamined responses	(2) Inferential reasoning	Bloom: Analyse, Evaluate, Create Facione: Inference Halpern: Hypothesis testing	Epistemic agency & Interrogation of knowledge
Metacognitive disengagement: Uncritical acceptance of AI content	(3) Evaluative judgement	Bloom: Evaluate Facione: Evaluation, Explanation Halpern: Decision-making	Epistemic responsibility & Reflexivity
Metacognitive disengagement: Diminished reflective self-monitoring	(4) Metacognitive regulation	Bloom: Metacognitive layer Facione: Self-regulation Halpern: Monitoring	Self-directed learning & Critical self-understanding
Epistemic narrowing: Over-reliance on convergent outputs → curiosity erosion	(5) Intellectual curiosity	Facione: Inquisitiveness Halpern: Flexible thinking	Equitable exploration & Questioning norms
Epistemic narrowing: Reproduction of bias and reduction of perspective diversity	(6) Epistemic integrity	Facione: Truth-seeking, Fair-mindedness Halpern: Ethical reasoning	Inclusive participation & Social critique and justice

Beyond cognitive performance, these processes contribute to what Davies and Barnett (2015) describe as *criticality*: a broader educational aim encompassing thinking, being, and acting in the world. From this perspective, critical thinking includes the capacity for self-understanding, social critique, and principled action. Cultivating these processes is therefore not only a cognitive task, but also an ethical and political imperative, especially in the context of AI systems that may entrench dominant norms and obscure epistemic pluralism.

3.1. Conceptual interpretation

Conceptual interpretation is the ability to actively construct meaning by selecting, organising, and integrating information into coherent mental models. It involves clarifying concepts, discerning relationships, and translating complex ideas into one's own understanding. This process aligns with Bloom's “Understand” and “Analyse” levels (Anderson & Krathwohl, 2001) and is foundational in Facione's (1990) taxonomy, where interpretation entails decoding significance and articulating meaning with precision. Unlike passive comprehension, conceptual interpretation requires the learner to distinguish core ideas from peripheral details, reconcile competing views, and connect new input to prior knowledge. As Dwyer (2017) emphasises, it is through this constructive engagement that understanding becomes robust, transferrable, and epistemically grounded. In AI-mediated environments, where outputs may offer polished explanations without revealing conceptual nuance or complexity, fostering interpretation ensures that learners remain actively engaged in making meaning rather than deferring to surface-level coherence.

3.2. Inferential reasoning

Inferential reasoning refers to the disciplined process of generating warranted conclusions from evidence. It involves identifying assumptions, discerning logical relationships, and predicting implications. Bloom's taxonomy situates it across the "Analyse", "Evaluate", and "Create" levels (Anderson & Krathwohl, 2001), while Facione (1990) defines it as securing relevant elements to draw reasoned conclusions. For Halpern (2014), inference is essential to adaptive decision-making, particularly when reasoning under uncertainty. Effective inferential reasoning goes beyond recognising plausible answers; it entails constructing arguments, identifying causal links, and testing hypotheses across contexts. As Ennis (1985) notes, inference is the cognitive bridge between information and action; a skill that must be practised through intentional, evidence-based tasks rather than substituted by automated outputs. In AI-mediated learning, where models can produce convincing yet unfounded responses, inferential reasoning is critical for helping students assess the strength of connections, evaluate plausibility, and resist uncritical acceptance of generated conclusions.

3.3. Evaluative judgement

Evaluative judgement is the ability to assess the credibility, coherence, and evidentiary basis of information and arguments. It corresponds to the "Evaluate" level in Bloom's taxonomy and is central in both Facione's (1990) and Halpern's (2014) models of critical thinking. It entails scrutinising claims, comparing competing interpretations, and determining whether conclusions follow logically from premises. Importantly, it also involves judging the quality and relevance of sources, a process that Paul and Elder (2019) link to intellectual fairness and humility. In AI-mediated environments, evaluative judgement becomes essential for interrogating the reliability of outputs that may appear fluent but lack transparency. As UNESCO (2025) warns, cultivating evaluative capacity is vital to resisting epistemic passivity and sustaining critical engagement in digital learning ecosystems.

Evaluative judgment thus functions as the core process by which learners assess the quality, credibility, and warrant of specific claims in AI-mediated environments. Epistemic vigilance operates alongside the process by sensitising learners to when evaluation is necessary, particularly in relation to issues of source reliability, bias, and omission. Epistemic agency then shapes how learners respond to these judgments, shaping whether and how evaluations are acted upon their ongoing engagement with knowledge, reasoning practices, and epistemic tools.

3.4. Metacognitive regulation

Metacognitive regulation involves the monitoring, evaluation, and strategic control of one's cognitive processes during learning and reasoning. It represents the metacognitive dimension of Bloom's revised taxonomy and is a key element in both Facione's (1990) emphasis on self-regulation and Halpern's (2014) focus on adaptive cognition. Effective regulation allows learners to assess task demands, recognise cognitive biases, and modify their approach when errors or gaps are detected. This process transforms critical thinking from an episodic act into a sustained, self-directed practice. As Dwyer (2017) argues, metacognitive oversight is indispensable in complex, ill-structured tasks where solutions are not predefined. In AI-rich environments, it empowers students to maintain epistemic agency by continuously interrogating their own reasoning, rather than defaulting to automated outputs.

3.5. Intellectual curiosity

Intellectual curiosity is the motivational disposition to explore ideas, ask questions, and pursue knowledge beyond instrumental goals. Although not explicitly listed in Bloom's taxonomy, it is a core disposition in both Facione's (1990) concept of inquisitiveness and Halpern's (2014) discussion of flexible, open-minded thinking. Curiosity fuels cognitive persistence and epistemic openness, prompting learners to seek novelty, tolerate ambiguity, and explore alternative perspectives. Ennis (1985) positions it as the drive that sustains inquiry beyond immediate answers. In digitally mediated learning, where LLMs often offer polished but singular responses, curiosity must be cultivated through tasks that encourage divergence, dialogic inquiry, and iterative exploration. Ultimately, curiosity is central not only to academic growth but also to democratic participation and lifelong learning.

3.6. Epistemic integrity

Epistemic integrity is the ethical orientation to seek truth, evaluate knowledge claims fairly, and engage with complexity conscientiously. It combines intellectual honesty with critical reflexivity and aligns closely with Facione's (1990) traits of truth-seeking and fair-mindedness, as well as Halpern's (2014) framing of ethical reasoning. While epistemic agency involves constructing and defending knowledge claims, integrity governs the values that shape those claims, such as humility, justice, and respect for difference. This orientation is especially critical in contexts where LLMs reproduce dominant discourses, conceal provenance, or obscure bias. As Davies and Barnett (2015) argue, safeguarding epistemic integrity requires that learners not only analyse content, but interrogate systems of knowledge production, asking whose voices are amplified or excluded. In this sense, epistemic integrity is both a personal commitment and a democratic imperative.

Together, these six processes provide a foundation for rethinking how critical thinking can be cultivated in AI-mediated learning environments. They clarify not only the cognitive skills and dispositions at risk when LLMs are used uncritically, but also the developmental capacities that must be intentionally scaffolded through pedagogy. Rather than framing AI tools as replacements for reasoning, this framework positions them as prompts for deeper engagement; tools that must be situated within learning designs that preserve interpretation, foster inference, demand evaluation, support self-regulation, stimulate curiosity, and uphold ethical inquiry. In the next section, we translate these processes into pedagogical principles that can guide educators in designing learning environments that integrate LLMs without compromising the intellectual and moral aims of higher education.

4. A pedagogical model for AI-enhanced critical thinking

> (Sub-RQ3) *What pedagogical design principles can guide the use of LLMs to scaffold critical thinking effectively?*

Building on the six intellectual processes outlined in Section 3, this section proposes a pedagogical model for integrating LLMs into higher education in ways that strengthen, rather than diminish, cognitive effort, metacognitive regulation, and epistemic integrity. Rather than advocating blanket policies for or against LLM use, this model emphasises *deliberate orchestration*: the intentional sequencing of pre-AI, during-AI, and post-AI learning activities designed to scaffold critical engagement and preserve core intellectual processes.

Central to this model is a rejection of AI as a pedagogical default. LLMs are positioned not as authoritative sources, but as *conditional tools*, activated at the right moment for the right learning purpose. The goal is to preserve the learner's role as an agentive, reflective thinker; someone who uses AI to deepen inquiry rather than shortcut it.

4.1. Core design principles

The following eight principles offer actionable guidelines for integrating LLMs into learning environments without undermining the development of critical thinking. Each principle is directly aligned with one or more of the six foundational processes defined in Section 3 and is supported by evidence from educational research on AI, cognition, and instructional design (See Table 2).

P1. Preserve cognitive friction. Critical thinking develops through effortful engagement with ambiguity, contradiction, and complexity (Jaramillo Gómez et al., 2025). As outlined in Section 2.1, LLMs often reduce cognitive load but also risk bypassing the very friction that stimulates deep understanding. Therefore, instructional designs should intentionally preserve productive struggle by requiring students to formulate hypotheses, analyse problems, or generate arguments independently before consulting AI tools.

Friction should also be sustained during and after interaction with LLMs. AI can be used to generate friction by introducing counterarguments, flawed logic, or alternative interpretations that prompt critical response. After engagement, learners should be encouraged to evaluate AI-generated content, compare it against their own reasoning, and revise their conclusions where necessary. This ongoing challenge maintains cognitive effort and supports epistemic agency.

This dialogic tension, akin to Socratic questioning (see Paul & Elder, 2007), fosters deeper inquiry and metacognitive reflection. Friction, in this sense, is not a barrier but a condition for cultivating epistemic virtues (Bowell & Kingsbury, 2015) and supporting what Barnett (2015) calls *critical being*; that is, the capacity to think, reflect, and act with integrity in uncertain contexts.

P2. Scaffold LLMs as thinking partners. To support critical thinking without displacing it, LLMs should be positioned as provisional collaborators, not authoritative sources (UNESCO, 2023). Their role is to assist students in breaking down tasks, exploring alternatives, and clarifying reasoning, while ensuring that learners retain cognitive ownership.

Scaffolding should guide students through structured inquiry, prompting them to actively question, compare, and revise ideas rather than accept AI outputs at face value. Within this framework, LLMs can model argument structures, suggest counterpoints, or simulate dialogue that stimulates reflection. Aligned with the 3H model: helpful, harmless, and honest (Askell et al., 2021), LLMs function best when used to extend thinking, not shortcut it.

Following UNESCO's AI literacy progression (2024), students should first understand how LLMs generate content, then critically engage with their outputs, and finally apply them in creating original responses. Used in this way, LLMs can prompt deeper reasoning and foster metacognitive awareness.

Table 2
Design principles and the cognitive or epistemic processes they support.

Design principle	Main supported processes	Justification
P1. Preserve cognitive friction	(1) Conceptual interpretation, (2) Inferential reasoning	Reinforces deep engagement before using AI, supporting reasoning and conceptual construction.
P2. Scaffold LLMs as partners	(2) Inferential reasoning, (5) Intellectual curiosity, (6) Epistemic integrity	Positions AI as a dialogic tool that fosters exploration and reflective thinking.
P3. Embed evaluation as standard practice	(2) Inferential reasoning, (3) Evaluative judgement	Encourages justification and structured comparison of claims.
P4. Activate metacognitive regulation	(4) Metacognitive regulation	Supports monitoring and self-adjustment through explicit strategies and cognitive tools.
P5. Encourage humility and curiosity	(5) Intellectual curiosity, (6) Epistemic integrity	Fosters open-ended inquiry and critical questioning of knowledge.
P6. Foster epistemic integrity	(6) Epistemic integrity	Reinforces ethical reasoning and critical awareness in AI interaction.
P7. Align assessment with thinking	(3) Evaluative judgement	Rewards quality of reasoning and supports critical engagement with AI tools.
P8. Balance AI-mediated and AI-free phases	(1) Conceptual interpretation, (4) Metacognitive regulation	Develops independent reasoning before using AI, strengthening agency and self-regulation.

P3. Embed evaluation as standard practice. Critical thinking depends not only on generating ideas but on evaluating them rigorously. To cultivate evaluative judgement, learners must be consistently required to assess the credibility, coherence, and evidentiary support of both human-derived and AI-generated claims (Facione, 1990; Ennis, 1985).

Rather than treating evaluation as a final step, instructional design should embed it throughout the learning process. This includes integrating checkpoints where students cross-reference AI outputs with diverse sources, apply structured criteria (e.g., relevance, bias, fallacies), and articulate reasons for accepting or rejecting claims.

Such habits cultivate what Bielik and Krell (2025) describe as *epistemic vigilance*: the capacity to critically assess both the credibility of information sources and the validity of their claims, using structured reasoning and scientific heuristics. They also reflect what Bailin and Battersby (2015) describe as the dialectical nature of critical thinking: an iterative process of weighing alternatives, responding to objections, and refining judgements in light of new evidence. Evaluation, in this sense, becomes a norm, not an exception, within AI-supported learning.

P4. Activate metacognitive self-regulation. In AI-supported learning, students must remain in control of their thinking processes rather than outsourcing them to the system. To sustain metacognitive regulation, tasks should incorporate tools such as planning templates, reflective journals, and AI prompt logs that make thinking explicit and subject to review (Teng & Yue, 2023).

Explicit instruction and repeated practice in applying these strategies, especially in authentic, complex tasks, reinforce students' ability to monitor, evaluate, and direct their own thinking (Manalo et al., 2015). This helps preserve agency and fosters deliberate, critical engagement with both their ideas and AI-generated content.

P5. Encourage intellectual humility and curiosity. Developing critical thinkers demands openness to complexity and a willingness to question one's own assumptions (Halpern, 1998; Paul & Elder, 2007). Tasks should prompt students to examine the limitations of AI output by generating alternative perspectives, simulating counterfactuals, or identifying omissions. This practice cultivates both intellectual humility and epistemic curiosity, dispositions essential for navigating uncertainty and resisting overconfidence in algorithmic authority.

By positioning the learner in active epistemic dialogue with AI, educators help students internalise habits of questioning, thereby deepening their understanding and strengthening critical judgement.

P6. Foster epistemic integrity. Tasks should be designed to reinforce the ethical and epistemic responsibilities of learners. This includes justifying claims, considering multiple perspectives, and reasoning under uncertainty; core aspects of what Paul and Elder (2007) call fair-minded critical thinking. Barnett (2015) emphasises that critical being requires coherence between thought and action, linking formal critique with ethical engagement in the world.

In AI-supported contexts, maintaining epistemic integrity involves resisting the temptation to accept plausible output uncritically and instead cultivating the disposition to question, verify, and take ownership of one's thinking.

P7. Align assessment with intended cognition. Assessment should prioritise the quality of reasoning over the surface fluency of AI-assisted outputs. Rubrics must explicitly reward higher-order skills such as analysis, evaluation, and reflection, core components of critical thinking across disciplines (Ennis, 1991; Halpern, 2014). To support deep learning, it is essential that instructional goals are closely aligned with assessment criteria. This alignment not only promotes meaningful cognitive engagement but also facilitates the transfer of critical thinking skills beyond isolated tasks. By assessing the reasoning behind students' judgements, including how they interpret, question, and integrate

AI-generated content, educators reinforce epistemic responsibility and reduce incentives for superficial engagement.

P8. Balance AI-mediated and AI-free task phases. Effective critical thinking pedagogy requires a deliberate integration of AI-supported and AI-free phases. Tasks involving hypothesis generation, argument construction, or self-regulated planning are best performed without AI to preserve cognitive autonomy and avoid overreliance. These "AI-free zones" cultivate original reasoning and metacognitive control before learners engage with AI. Subsequent AI use should serve as a reflective extension, supporting critique, revision, or comparison, not as a substitute for thinking.

This sequencing fosters independent judgement and reinforces the learner's role as an active epistemic agent. Table 2 presents each design principle alongside the specific cognitive and epistemic processes it is intended to support, with pedagogical justifications for each alignment. From a learning science perspective, these principles draw on well-established mechanisms such as self-regulated learning cycles of planning, monitoring, and evaluation (Panadero, 2017; Zimmerman, 2000), as well as feedback literacy, which emphasises learners' capacity to judge the quality of feedback and decide how to act on it (Carless & Boud, 2018; Hopfenbeck, 2020; Nicol & Macfarlane-Dick, 2007). This structure highlights the framework's dual commitment to instructional intentionality and the cultivation of critical thinking capacities in AI-mediated learning environments.

4.2. Practical scenarios and applications

To operationalise the design principles outlined in Section 4.1, this section presents two scaffolded activities that integrate generative AI in ways that preserve students' cognitive ownership, promote epistemic integrity, and foster critical engagement with both knowledge and technology. These scenarios are not only technical exercises; they are designed to cultivate thinking that is self-aware, world-aware, and action-oriented, as well as a capacity to interrogate and shape technology's role in knowledge production. Importantly, the scenarios are offered not as fixed templates or "best practices", but as situated examples intended to provoke adaptation, critique, and redesign in response to local pedagogical and institutional conditions.

Each scenario activates the six essential intellectual processes outlined in Section 3 (i.e., conceptual interpretation, inferential reasoning, evaluative judgement, metacognitive regulation, intellectual curiosity, and epistemic integrity) while reflecting all eight pedagogical design principles articulated in Section 4.1. Taken together, they demonstrate how structured engagements with LLMs can move beyond efficiency-oriented use toward the cultivation of thoughtful and ethically grounded inquiry.

4.2.1. Scenario A: designing and deconstructing AI prompts

Students engage in a two-part activity focused on the intentional crafting and critical evaluation of prompts used to query LLMs (Table 3). The activity treats prompt design not as a technical skill to be optimised, but as an epistemic practice through which assumptions, values, and boundaries of inquiry are articulated and contested.

- In Part I, students are given a conceptual topic relevant to their course (e.g., "democracy in the digital age," "climate justice," or "data privacy") and are asked to design 2–3 prompts that aim to draw out depth, nuance, or multiple perspectives from an LLM. This phase foregrounds students' own interpretative agency and resists premature reliance on automated outputs.
- In Part II, they analyse the responses generated by their prompts for epistemic breadth, assumptions, and omissions. They then revise at least one prompt in response to this critique and reflect on how the framing of a question shapes the construction of knowledge.

Table 3
Scenario A. Processes activated and design principles enacted.

Processes activated	Design principles enacted	Illustrative operational moves
Conceptual interpretation is required to distill the essence of the topic into an effective, purposeful prompt.	P1. Preserve friction	1-Students draft prompts individually without AI access and briefly justify why each prompt is expected to elicit depth or complexity.
Inferential reasoning is developed as students anticipate how LLMs may interpret or misinterpret their queries.	P2. Scaffold LLMs as thinking partners	2-Before submitting prompts, students predict possible AI responses and identify potential assumptions or simplifications the model might make.
Evaluative judgement emerges in assessing the completeness and validity of AI-generated responses.	P3. Embed evaluation as standard practice	3-Students analyse AI outputs using guiding questions such as: *Which perspectives are prioritised? Which are absent? What claims lack justification?*
Metacognitive regulation is engaged through iterative revision and strategic questioning.	P4. Activate metacognitive regulation	4-Students revise at least one prompt and provide a short reflection explaining what epistemic limitations the revision was intended to address.
Intellectual curiosity is fostered by exploring alternative framings and possible answers.	P5. Foster curiosity and humility	5-Students generate alternative prompt framings that deliberately surface uncertainty, disagreement, or marginal perspectives.
Epistemic integrity is supported as students examine biases in both prompts and outputs.	P6. Reinforce epistemic integrity	6-Learners identify ethical or political implications of prompt framing, including the risks of leading, exclusionary, or overly convergent questions.
	P7. Align assessment with cognition	7-Assessment emphasises quality of prompt reasoning, depth of critique, and justification of revisions rather than the fluency of AI outputs.
	P8. Balance AI-free and AI-mediated phases	8-The activity follows an AI-free design phase, an AI-mediated exploration phase, and an AI-free synthesis phase.

Through this cycle, the LLM becomes an object of inquiry rather than a resource of authority, making visible how technological systems participate in shaping what counts as knowledge.

This scenario cultivates a meta-awareness of inquiry itself. By foregrounding the role of question design in shaping the contours of knowledge, students learn that critical thinking is not just about evaluating answers, but it is equally about formulating the right questions. In this way, prompt engineering becomes a pedagogical vehicle for developing intentionality, epistemic responsibility, and reflective doubt in AI-mediated learning.

4.2.2. Scenario B: constructing, critiquing, and reframing AI-mediated arguments

Students complete a multi-phase writing activity that integrates original argument construction, critical evaluation of LLM-generated counterarguments, and epistemic reframing from marginalised perspectives (Table 4). The activity is designed to surface the situated and debated nature of knowledge production, specially when AI systems are introduced as participants in argumentative processes.

- In Part I, students independently formulate a thesis on a complex issue (e.g., "Should predictive policing be banned?", "Does the use of AI tools in decision-making processes enhance or undermine

Table 4
Scenario B. Processes activated and design principles enacted.

Processes activated	Design principles enacted	Illustrative operational moves
Conceptual interpretation: Required in formulating and reframing the argument.	P1. Preserve friction	Students submit an initial thesis and supporting reasoning before any AI interaction.
Inferential reasoning: Engaged in building, testing, and revising claims.	P2. Scaffold LLMs as thinking partners	Students prompt the LLM to generate counterarguments and identify the strongest assumption underlying each response.
Evaluative judgement: Central to critiquing AI-generated reasoning.	P3. Embed evaluation as standard practice	Learners annotate AI outputs, marking logical strengths, weaknesses, unsupported claims, and potential bias.
Metacognitive regulation: Practiced through planning, revision, and reflection.	P4. Activate metacognitive regulation	Students write a brief reflection explaining how and why their argument changed after engaging with AI-generated counterarguments.
Intellectual curiosity: Stimulated by engaging with contrasting and unfamiliar perspectives.	P5. Foster curiosity and humility	Students reframe the original question from a marginalised or non-dominant perspective and compare resulting AI outputs.
Epistemic integrity: Reinforced through attribution, transparency, and ethical awareness.	P6. Reinforce epistemic integrity	Learners reflect on whose knowledge is amplified or marginalised by different framings and justify their final position.
	P7. Align assessment with cognition	Assessment focuses on quality of reasoning, depth of critique, and epistemic awareness rather than persuasive polish.
	P8. Balance AI-free and AI-mediated phases	The sequence alternates independent reasoning, AI engagement, and AI-free revision to preserve cognitive ownership.

democratic values?"), giving initial reasoning and evidence. This phase established cognitive ownership and positions students as primary authors of meaning.

- In Part II, they prompt an LLM to generate counterarguments and critique these outputs using structured criteria focused on logic, bias, and credibility. Rather than treating AI responses as neutral or comprehensive, students are encouraged to examine how framing, training data, and optimisation goals shape the arguments produced.
- Finally, in Part III, students reframe the original question from a marginalised perspective (e.g., feminist, Indigenous, postcolonial), generate a new AI response, and compare how epistemic framing alters argumentative possibilities. This phase makes explicit the political and cultural dimensions of knowledge construction and highlights whose voices are amplified or obscured through AI mediation.

This scenario engages students in a full cycle of critical thinking, from independently argument construction to interrogation of AI-generated reasoning and epistemic reframing. By alternating between AI-free and AI-mediated phases, it preserves cognitive ownership while leveraging LLMs as tools for epistemic challenge and expansion. Ultimately, this activity helps learners develop a more self-aware, ethically grounded, and socially attuned understanding of how knowledge is constructed, contested, and reframed in the age of generative AI.

5. Discussion and conclusion

This paper advanced a normative, design-oriented pedagogical framework for integrating LLMs into higher education in ways that preserve and extend critical thinking. We argued that LLMs are neither educationally neutral nor inherently beneficial or harmful. Their impact depends on how they are designed: what they afford, obscure, or prioritise, and on the pedagogical conditions under which they are implemented. Our central claim is that the value of generative AI in education is not technical but fundamentally pedagogical, shaped by how learners are guided to think with, about, and beyond these tools.

To achieve this, we drew on design-based research, critical pedagogy, and AI ethics to identify six essential intellectual processes underpinning critical thinking: conceptual interpretation, inferential reasoning, evaluative judgement, metacognitive regulation, intellectual curiosity, and epistemic integrity. These were translated into eight pedagogical design principles and illustrated through scaffolded classroom scenarios. Together, these components offer a practical model for cultivating cognitive effort, epistemic agency, and reflective inquiry in AI-mediated learning. Importantly, this framework is not intended as a definitive or universal solution, but as provisional pedagogical orientation that must remain open to critique, adaptation, and contextual interpretation.

Beyond its conceptual contribution, the framework introduces an educational innovation in how generative AI is pedagogically positioned in higher education. Rather than treating LLMs as instructional tools, tutors, or efficiency supports, the framework reframes them as objects of inquiry that learners must interrogate, evaluate, and work against. This shifts the role of pedagogical scaffolding in AI mediated learning from reducing difficulty or cognitive load toward deliberately preserving productive cognitive friction through structured sequencing of AI free and AI mediated phases, explicit evaluation of AI outputs, and support for metacognitive regulation.

From a pedagogical perspective, the framework extends established educational theories rather than introducing a new instructional paradigm. It operationalises constructivist and socio-cognitive views of learning by foregrounding active meaning making, inferential reasoning, and evaluation (e.g., Bruner, 1996; Piaget, 1970; Vygotsky, 1978), while drawing on research on metacognition and self regulated learning to emphasise monitoring, reflection, and strategic control of thinking. Likewise, the framework aligns with traditions of critical pedagogy (e.g., Barnet, 2015; Freire, 1970) by treating knowledge as provisional, situated, and open to interrogation, particularly in relation to the perceived authority of AI generated content. In this way, the framework translates well established educational theories into concrete pedagogical design principles for AI mediated learning environments.

Taken together, these design commitments reframe not only how AI is used in learning, but also the role of the educator in AI-mediated environments. This reorientation demands a shift in the educator's role from facilitator to cognitive orchestrator, as outlined by UNESCO (2023). Educators must design scaffolded and dialogic learning environments that sustain cognitive friction, preserve interpretive autonomy, and prompt students to engage with bias, ambiguity, and conflicting perspectives. Joksimovic et al. (2023) highlight that human-AI collaboration must evolve through iterative and co-constructed interaction, rather than automation of cognitive tasks. This calls for AI-free zones of engagement, process-oriented assessment rubrics, and embedded cycles of reflection.

At the curricular level, AI literacy must include technical competence while also developing students' capacity to question, interpret, and evaluate AI-generated outputs. This includes the capacity to tolerate uncertainty, interrogate automated outputs, and weigh competing interpretations. Marmolejo-Ramos et al. (2025) show that individuals with higher levels of statistical literacy are less likely to accept algorithmic outputs uncritically, especially in high-stakes contexts. Their findings suggest that careful scrutiny and informed scepticism are markers of critical competence. Embedding such forms of statistical and critical literacy in curricula is therefore not only desirable but essential for protecting learner autonomy in increasingly datafied learning environments.

In teacher education, the implications are no less significant. Educators must be prepared to critically engage with AI tools and to guide students in doing the same. This includes understanding how LLMs work, designing learning experiences that preserve student agency, and fostering ethical sensitivity to issues such as data provenance, algorithmic bias, and platform dependency. Marrone et al. (2025) note that educational leaders are increasingly concerned about the pedagogical risks of uncritical AI adoption. Their work underscores the need for robust professional learning frameworks that prepare teachers to engage with AI confidently and ethically.

Institutional adoption of the proposed framework will depend not only on its pedagogical soundness but also on systemic capacity for implementation. A recent meta-review by Bond et al. (2024) reinforces this point, identifying persistent barriers to AI adoption in higher education. These include ethical concerns, fragmented curricula, inadequate infrastructure, and limited faculty training. Without addressing these systemic constraints, even well-designed pedagogical models risk limited impact. Jin et al. (2025) arrive at similar conclusions in their global analysis of institutional AI policies. While many universities are actively developing ethical guidelines, designing authentic assessments, and offering training for staff and students, they also found that comprehensive policy frameworks remain uneven, particularly with regard to data privacy, equity, and continuous evaluation. Their findings highlight the importance of well-defined roles, transparent communication, and adaptive governance to ensure responsible AI integration. These insights reinforce the need to align institutional capacity with principled design.

To operationalise this vision, institutions should build robust educator support systems. These must include interdisciplinary design teams, alignment with digital and civic literacies, and thoughtful integration of global frameworks such as those provided by UNESCO's (2023; 2024). Importantly, such frameworks should not be treated as prescriptions to be implemented, but as starting points for collective reflection and adaptation.

6. Implications for pedagogical data analytics

Although this study is conceptual, it has direct implications for pedagogical data analytics in AI-mediated learning environments. The framework positions analytics as interpretative supports for pedagogical judgment, rather than as direct measures of learning, because the intellectual processes it targets are not directly observable and must be inferred from patterns of learner interaction across time and tasks. Within this orientation, the six processes identified in the framework can inform process-oriented analytics that foreground how students engage with learning activities, rather than privileging final products alone. Relevant analytic traces include patterns of AI prompt formulation and revision, sequences of draft modification, engagement with counterarguments or alternative perspectives, and transitions between AI-free and AI-mediated phases of work. When analysed longitudinally and in relation to task design, these traces provide context-dependent indicators of cognitive effort and self-regulation that require pedagogical interpretation rather than automated classification.

Used in this way, pedagogical analytics can support instructional decision making by helping educators identify patterns such as early dependence on AI generated outputs, limited revision following feedback, or uncritical acceptance of responses. These patterns may indicate cognitive offloading or weak self-regulatory engagement. Conversely, sustained cycles of evaluation, revision, and comparison across AI-mediated and AI-free phases can function as conditional indicators of critical engagement and reflective judgment. Importantly, such indicators are not substitutes for professional judgment but resources for informing when to introduce additional cognitive friction, adjust task sequencing, or redesign learning activities.

7. Limitations and future research

This study is conceptual in nature and does not present empirical validation of the proposed framework. While the design principles are grounded in established theory and supported by recent research, their effectiveness and applicability may vary across disciplines, institutional contexts, and learner populations. Higher education systems differ in technological infrastructure, faculty expertise, and policy environments, which may influence how the framework can be implemented and sustained. Consequently, the principles should be adapted with sensitivity to local needs, cultural considerations, and resource constraints.

Future research should empirically investigate the effectiveness of the proposed design principles, using methodologies such as classroom-based interventions, design-based research, or comparative studies to examine how different instructional sequences and patterns of AI use influence student reasoning, metacognitive regulation, and epistemic dispositions. Further investigation is also needed into how institutional structures, such as curriculum committees, faculty development programs, and digital policy teams, can support sustainable and ethically grounded AI integration across diverse educational settings.

In conclusion, the transformation of education through generative AI will not be driven by automation alone. It will require intentional pedagogical design. Educators must help students to engage critically with AI, interrogate its assumptions, and transcend its limitations, cultivating not only intellectual skill but also ethical responsibility and integrity in their engagement with knowledge.

CRediT authorship contribution statement

Mireia Vendrell: Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Samantha-Kaye Johnston:** Writing – review & editing, Writing – original draft, Supervision, Conceptualization.

Statements on ethics and open data

This study is conceptual and theoretical in nature and did not involve human participants, human data, or human subjects research. Accordingly, ethical approval from an institutional review board or ethics committee was not required, and informed consent was not applicable. No datasets were generated or analysed during the current study. All sources drawn upon are publicly available and are listed in the reference section.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2015). Strategies for teaching students to think critically: A meta-analysis. *Review of Educational Research, 85*(2), 275–314. https://doi.org/10.3102/003465431455106

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing : A revision of Bloom's taxonomy of educational objectives*. Longman.

Askell, A., Bai, Y., Chen, A., Drain, D., & Ganguli, D. (2021). A general language assistant as a laboratory for alignment. *arXiv preprint arXiv.*. https://doi.org/10.48550/arXiv.2112.00861

Bailin, S., & Battersby, M. (2015). Teaching critical thinking as inquiry. In M. Davies, & R. Barnett (Eds.), *The palgrave handbook of critical thinking in higher education* (pp. 123–138). Palgrave Macmillan.

Barnett, R. (2015). A curriculum for critical being. In M. Davies, & R. Barnett (Eds.), *The palgrave handbook of critical thinking in higher education* (pp. 63–76). Palgrave Macmillan.

Bielik, T., & Krell, M. (2025). Developing and evaluating the extended epistemic vigilance framework. *Journal of Research in Science Teaching, 62*(3), 869–895. https://doi.org/10.1002/tea.21983

Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., … Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education, 21*(1), 4. https://doi.org/10.1186/s41239-023-00436-z

Bowell, T., & Kingsbury, J. (2015). Virtue and inquiry: Bridging the transfer gap. In M. Davies, & R. Barnett (Eds.), *The palgrave handbook of critical thinking in higher education* (pp. 233–245). Palgrave Macmillan.

Bruner, J. (1996). *The culture of education*. Cambridge, MA: Harvard University Press.

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

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., … Hadfield-Menell, D. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*. https://doi.org/10.48550/arXiv.2307.15217

D'Mello, S., & Graesser, A. (2014). Confusion and its dynamics during device comprehension with breakdown scenarios. *Acta Psychologica, 151*, 106–116. https://doi.org/10.1016/j.actpsy.2014.06.005

Dahmani, L., & Bohbot, V. D. (2020). Habitual use of GPS negatively impacts spatial memory during self-guided navigation. *Scientific Reports, 10*(1), 6310. https://doi.org/10.1038/s41598-020-62877-0

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

Davies, M., & Barnett, R. (2015). *The palgrave handbook of critical thinking in higher education*. Palgrave Macmillan. https://doi.org/10.1057/9781137378057_13

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2024). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education*. , Article 105224. https://doi.org/10.1016/j.compedu.2024.105224

Digital Education Council. (2024). Digital education council global AI student survey 2024. *Digital Education Council*. https://www.digitaleducationcouncil.com/post/digital-education-council-global-ai-student-survey-2024.

Dwyer, C. P. (2017). *Critical thinking*. Cambridge University Press.

Dwyer, C. P. (2023). An evaluative review of barriers to critical thinking in educational and real-world settings. *Journal of Intelligence, 11*(6), 105. https://doi.org/10.3390/jintelligence11060105

Ennis, R. H. (1985). A logical basis for measuring critical thinking skills. *Educational Leadership, 43*(2), 44–48. https://pdfs.semanticscholar.org/80a7/c7d4a98987590751df4b1bd9adf747fd7aaa.pdf.

Facione, P. A. (1990). *Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction. Research Findings and Recommendations*.

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., … Gašević, D. (2024). 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

Fisher, J., Feng, S., Aron, R., Richardson, T., Choi, Y., Fisher, D. W., … Reinecke, K. (2025). g. *Biased ai can influence political decision-makin*. arXiv preprint arXiv:2410.06415.

Freire, P. (1970). *Pedagogy of the oppressed (M. B. Ramos, Trans.). Herder and Herder* (Original work published 1968).

Freire, P. (1972). Education: Domestication or liberation? *Prospects, 2*(2), 173–181.

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies, 15*(1), 6. https://doi.org/10.3390/soc15010006

Halpern, D. F. (2014). *Thought and knowledge: An introduction to critical thinking* (5th ed.). Psychology Press.

Heung, Y. M. E., & Chiu, T. K. (2025). How ChatGPT impacts student engagement from a systematic review and meta-analysis study. *Computers and Education: Artificial Intelligence, 8*, Article 100361. https://doi.org/10.1016/j.caeai.2025.100361

Hopfenbeck, T. N. (2020). The need for actionable feedback in assessment literacy. *Assessment in Education: Principles, Policy & Practice, 27*(3), 249–251. https://doi.org/10.1080/0969594X.2020.1771665

Jaramillo Gómez, D. L., Álvarez Maestre, A. J., Parada Trujillo, A. E., Pérez Fuentes, C. A., Bedoya Ortiz, D. H., & Sanabria Alarcón, R. K. (2025). Determining factors for the development of critical thinking in higher education. *Journal of Intelligence, 13*(6), 59. https://doi.org/10.3390/jintelligence13060059

Jin, Y., Yan, L., Echeverria, V., Gašević, D., & Martinez-Maldonado, R. (2025). Generative AI in higher education: A global perspective of institutional adoption policies and guidelines. *Computers and Education: Artificial Intelligence, 8*, Article 100348. https://doi.org/10.1016/j.caeai.2024.100348

Joksimovic, S., Ifenthaler, D., Marrone, R., De Laat, M., & Siemens, G. (2023). Opportunities of artificial intelligence for supporting complex problem-solving: Findings from a scoping review. *Computers and Education: Artificial Intelligence, 4*, 1–12. https://doi.org/10.1016/j.caeai.2023.100138

Jose, B., Cherian, J., Verghis, A. M., Varghise, S. M., S, M., & Joseph, S. (2025). The cognitive paradox of AI in education: Between enhancement and erosion. *Frontiers in Psychology, 16*, Article 1550621. https://doi.org/10.3389/fpsyg.2025.1550621

Kapur, M. (2008). Productive failure. *Cognition and Instruction, 26*(3), 379–424. https://doi.org/10.1080/07370000802212669

Karataş, F., Abedi, F. Y., Ozek Gunyel, F., Karadeniz, D., & Kuzgun, Y. (2024). Incorporating AI in foreign language education: An investigation into ChatGPT's effect on foreign language learners. *Education and Information Technologies*, 1–24. https://doi.org/10.1007/s10639-024-12574-6

Komljenovic, J., Williamson, B., Eynon, R., & Davies, H. C. (2023). When public policy 'fails' and venture capital 'saves' education: Edtech investors as economic and political actors. *Globalisation, Societies and Education*, 1–16. https://doi.org/10.1080/14767724.2023.2272134

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X. H., Beresnitzky, A. V., … Maes, P. (2025). Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. arXiv preprint arXiv:2506.08872 https://doi.org/10.48550/arXiv.2506.08872.

Kudina, O., Ballsun-Stanton, B., & Alfano, M. (2025). The use of large language models as scaffolds for proleptic reasoning. *Asian Journal of Philosophy, 4*(1), 1–18. https://doi.org/10.1007/s44204-025-00247-1

Lee, H. P. H., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. https://doi.org/10.1145/3706598.3713778.

Lokesh, G. R., Harish, K. S., Sangu, V. S., Prabakar, S., Kumar, V. S., & Vallabhaneni, M. (2024). AI and the future of work: Preparing the workforce for technological shifts and skill evolution. *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS), 1*, 1–6. IEEE.

Manalo, E., Kusumi, T., Koyasu, M., Michita, Y., & Tanaka, Y. (2015). Do students from different cultures think differently about critical and other thinking skills? In E. M. Davies, & R. Barnett (Eds.), *The palgrave handbook of critical thinking in higher education* (pp. 299–316). https://doi.org/10.1057/9781137378057_19

Marmolejo-Ramos, F., Marrone, R., Korolkiewicz, M., Gabriel, F., Siemens, G., Joksimovic, S., … Tejada, J. (2025). Factors influencing trust in algorithmic decision-making: An indirect scenario-based experiment. *Frontiers in Artificial Intelligence, 7*, Article 1465605. https://doi.org/10.3389/frai.2024.1465605

Marrone, R., Fowler, S., Bathakur, A., Dawson, S., Siemens, G., & Singh, C. (2025). Perceptions and perspectives of Australian school leaders on the integration of artificial intelligence in schools. *School Leadership & Management, 45*(1), 30–52. https://doi.org/10.1080/13632434.2024.2425019

McClure, J., Zheng, J., Bickel, F., Jiang, S., Rosé, C. P., Chao, J., & Winling, L. (2024). Modeling with primary sources: An approach to teach data bias for artificial intelligence and machine learning education. In *Proceedings of the 18th International Conference of the Learning Sciences-ICLS 2024* (pp. 514–521). International Society of the Learning Sciences.

McKenney, S., & Reeves, T. (2018). *Conducting educational design research*. Routledge.

Morozov, E. (2013). *To save everything, click here: The folly of technological solutionism*. Public Affairs.

Nicol, D., & Macfarlane-Dick, D. (2007). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199–218. https://doi.org/10.1080/03075070600572090

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences, 12*(6), 237–241.

Ozalp, H., Ozcan, P., Dinckol, D., Zachariadis, M., & Gawer, A. (2022). "Digital colonization" of highly regulated industries: an analysis of big tech platforms' entry into health care and education. *California Management Review, 64*(4), 78–107. https://doi.org/10.1177/00081256221094307

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

Pardos, Z. A., & Bhandari, S. (2024). ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *PLoS One, 19*(5), Article e0304013. https://doi.org/10.1371/journal.pone.0304013

Paul, R., & Elder, L. (2007). *The thinker's guide to the art of Socratic questioning*. Foundation for Critical Thinking. www.criticalthinking.org.

Paul, R., & Elder, L. (2019). *A guide for educators to critical thinking competency standards: Standards, principles, performance indicators, and outcomes with a critical thinking master rubric*. The Foundation for Critical Thinking. www.criticalthinking.org.

Piaget, J. (1970). *Science of education and the psychology of the child*. Trans. D. Coltman. Orion.

Reigeluth, C. M. (2013). *Instructional-design theories and models: A new paradigm of instructional theory* (Vol. II). Routledge.

Sandoval, W. (2014). Conjecture mapping: An approach to systematic educational design research. *The Journal of the Learning Sciences, 23*(1), 18–36. https://doi.org/10.1080/10508406.2013.778204

Selwyn, N., Ljungqvist, M., & Sonesson, A. (2025). *When the prompting stops: Exploring teachers' work around the educational frailties of generative AI tools* (pp. 1–14). Learning, Media and Technology. https://doi.org/10.1080/17439884.2025.2537959

Stadler, M., Bannert, M., & Sailer, M. (2024). Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior, 160*, Article 108386. https://doi.org/10.1016/j.chb.2024.108386

Teng, M. F., & Yue, M. (2023). Metacognitive writing strategies, critical thinking skills, and academic writing performance: A structural equation modeling approach. *Metacognition and Learning, 18*(1), 237–260. https://doi.org/10.1007/s11409-022-09328-5

UNESCO. (2023). *Guidance for generative AI in education and research*. United Nations Educational, Scientific and Cultural Organization. https://doi.org/10.54675/VKKE7525

UNESCO. (2024). *AI competency framework for students*. United Nations Educational, Scientific and Cultural Organization. https://doi.org/10.54675/JKJB9835

UNESCO. (2025). *Red teaming artificial intelligence for social good: The playbook*. United Nations Educational, Scientific and Cultural Organization. https://doi.org/10.54675/QXHZ9733

Van Brussel, S., Timmermans, M., Verkoeijen, P., & Paas, F. (2020). 'Consider the opposite'–effects of elaborative feedback and correct answer feedback on reducing confirmation bias–A pre-registered study. *Contemporary Educational Psychology, 60*, Article 101844. https://doi.org/10.1016/j.cedpsych.2020.101844

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (Vol. 86). Harvard university press.

Yatani, K., Sramek, Z., & Yang, C. L. (2024). AI as extraherics: Fostering higher-order thinking skills in Human-AI interaction. https://doi.org/10.48550/arXiv.2409.09218.

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

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13–39). Academic Press.