Scaffolding or substitution? Generative and Agentic AI in humanities learning under a Human-in-the-Loop lens*

Anna Capodieci[1,*], Alessandro Bruno[1] and Luca Quaratino[1]

[1] *IULM University, Department of Business, Law, Economics and Consumer Behaviour "Carlo A. Ricciardi", Milan, Italy*

Abstract

Humanities disciplines, which have historically been organised around the analysis of interpretation, textual judgement, and the slow process of constructing meaning, have recently been undergoing a reshaping in the way the learning process is mediated through generative and, more recently, agentic artificial intelligence. This paper presents a narrative review of generative and agentic AI in humanities education, examined through the lens of the human-in-the-loop (HITL) principle. Sources on the above-mentioned topics published between 2020 and the present were collected from several databases, such as arXiv, the ACL Anthology, Scopus, and supplementary searches conducted using Google Scholar, then classified according to their relevance to the review questions, epistemic status (empirical or conceptual/philosophical), and AI type. The review develops a descriptive taxonomy that maps documented applications along a depth-of-mediation axis. At one end are approaches that treat AI as an external object of critical analysis; at the other are those that embed AI directly within the disciplinary task. Although agentic AI is relatively well established in the broader AI-in-education literature, it remains largely unexplored in humanities-specific contexts. The review also argues that the depth-of-mediation axis and the operationalisation of the human-in-the-loop principle should not be conceived as separate dimensions. Rather, they offer complementary perspectives on the same underlying mechanism. Learning gains appear to persist when learners retain authorship over the structure and direction of the learning process. By contrast, the aforementioned gains drop when human oversight is reduced to approving AI-generated outputs at the end of the process. Finally, a hypothesis rather than a definitive conclusion is advanced: process-level authorship may represent a transversal competence linking humanities learning processes, grounded in a substantive interpretation of the human-in-the-loop principle, with forms of agency and judgment increasingly valued in labour markets shaped by AI.

Keywords

Generative AI, Agentic AI, Human-in-the-Loop, Humanities Education, Digital Humanities, Cognitive Scaffolding

1. Introduction

The rapid spread of generative and, more recently, agentic artificial intelligence is reshaping how learning is designed, mediated, and experienced across educational domains (Holmes & Tuomi, 2022; Kostopoulos et al., 2025). Within the humanities, understood here, for this review, as the disciplines centred on the interpretation of texts, artefacts, and cultural traditions (classical and modern languages, literature, history, and philosophy) — where learning has traditionally been organised around interpretation, textual judgement, and the slow construction of meaning (Chun & Elkins, 2023) — this shift raises questions that are at once pedagogical and philosophical: what happens to processes of understanding when a machine can produce plausible interpretations, translations, or narrative reconstructions on demand? Discussions of AI in education have so far tended to frame the issue as either a technical problem (how well do these tools perform) or a speculative one — for instance, the proposal to conceive human–AI interaction as a form of "System 0" thinking, in which AI functions as an extension of the human cognitive system itself — a heuristic layer of automatic, largely unconscious processing distinct from deliberate reasoning,

ReDH 2026: Rethinking the Future of Digital Humanities through Ethics and Artificial Consciousness, co-located with AIxIA 2026, October 6–9, 2026, Perugia, Italy

* Corresponding author.

anna.capodieci@studenti.iulm.it (A. Capodieci); alessandro.bruno@iulm.it (A. Bruno); luca.quaratino@iulm.it (L. Quaratino)

0009-0004-8886-9567 (A. Capodieci); 0000-0003-0707-6131 (A. Bruno); 0000-0002-8058-5357 (L. Quaratino)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-05

CEUR-WS.org/Vol-4314/ReDH9.pdf

here extended to include AI-generated content (Chiriatti et al., 2024) — concerned with what AI reveals about consciousness, authorship, or the nature of interpretation itself.

This paper takes a third, complementary route: in terms of learning processes, it asks what happens to the learner who is asked to work alongside these systems, and what is gained or lost in that encounter.

We argue that this question cannot be answered by treating "AI in the humanities classroom" as a single, undifferentiated phenomenon. Generative AI (large language models, vision-language models, diffusion models producing content on prompt) and agentic AI (systems that plan, act across multiple steps, use tools, and pursue goals with limited human supervision) differ substantially in autonomy: generative AI operating reactively to individual prompts, agentic AI orchestrating sequences of such operations toward a goal. On top of that, they also differ in the kind of output they produce, and — most importantly for our purposes — in where and how a human is expected to remain "in the loop" of the learning process. Treating them as interchangeable obscures precisely the distinctions that matter for assessing their pedagogical value.

A second, less obvious thread runs through this paper. The interpretive and hermeneutic competences that humanities education has long cultivated — decomposing a text, recognising patterns of meaning, reconstructing sense from fragmentary or ambiguous material — are structurally close to, and already aligned with, the competences increasingly demanded by an information economy reshaped by AI: the capacity to analyse, question, and act on machine-generated content rather than simply consume it. Recent policy and research literature on the future of work has begun to converge on this point (OECD, 2021; Dilinika, 2026), identifying transversal competences — critical thinking, adaptive reasoning, ethical judgement — as increasingly central to employability in AI-transformed labour markets. If this convergence holds, then humanities learning mediated by AI is not only a matter of preserving disciplinary relevance; it may also function as a training ground for fully developing exactly the set of competences that a reconfigured labour market now requires. This paper treats that possibility as a working hypothesis to be examined against the available evidence, not as a settled conclusion — and, in doing so, it also offers a way of reconnecting a learning-centred perspective with the more foundational questions about meaning, authorship, and human judgement that discussions of AI in the humanities tend to raise (Floridi, 2025; Chun & Elkins, 2023).

Against this backdrop, the paper conducts a narrative review of generative and agentic AI approaches used in humanities learning, combining a descriptive mapping of the field with a critical synthesis of when and why human oversight over these approaches remains substantive rather than merely nominal. The manuscript tackles three questions: which approaches are currently employed, and with what distribution and maturity (RQ1); how these approaches position AI relative to deep cognitive processes — interpretation, critical judgement, sensemaking — and in what forms the human-in-the-loop principle is operationalised within that positioning (RQ2); and what evidence exists on their effects, positive and negative, on learners' humanistic and transversal competences (RQ3). Section 3 addresses RQ1 through a descriptive taxonomy of the approaches documented in the corpus; Section 4 takes up RQ2 and RQ3 jointly, reading the depth-of-mediation axis established in Section 3 for its implication about cognitive scaffolding or substitution, distilling from that reading an account of why human engagement holds or fails; Section 5 concludes with a synthesis, the review's limitations, and directions for future research (Figure 1).

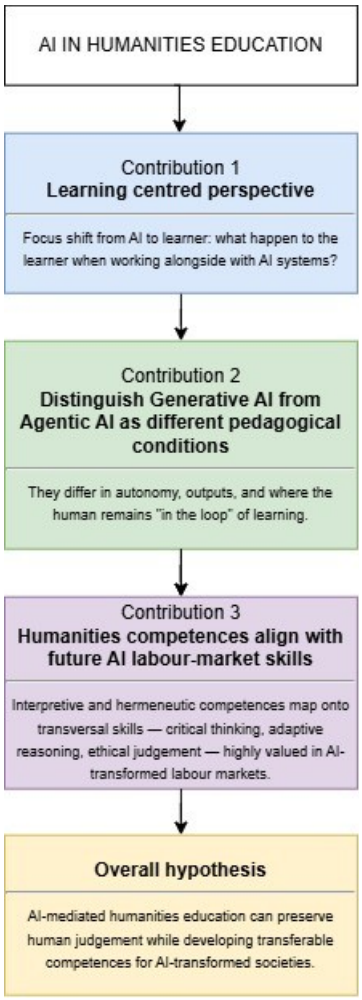


Figure 1: Contributions of the review and overall working hypothesis.

2. Methodology

This review focuses on humanities and classics education — the disciplines centred on textual and cultural interpretation — as mediated by AI. Given the exploratory stage of this research strand and the breadth of the topic across disciplinary and methodological traditions, a narrative approach rather than a systematic review was adopted. Sources were identified through Google Scholar, arXiv, ACL Anthology, and Scopus, using queries intersecting artificial intelligence/generative AI /large language models/agentic AI with learning/education, and humanities/classics/hermeneutics. The temporal window was set from 2020 to the present, with a practical emphasis on the period following the public release of ChatGPT in late 2022 for generative AI, and on the most recent eighteen months for agentic AI. Each candidate source was classified as core, background, or marginal relative to the review's questions, and, separately, as empirical (reporting observed or experimentally induced effects) or conceptual/philosophical (proposing frameworks, definitions, or theoretical arguments) — a distinction maintained throughout the review rather than applied only as a preliminary filter, since the two kinds of source play different argumentative roles in the analysis that follows. Searches conducted between June and July 2026 across these databases identified approximately 42 candidate sources judged relevant to the review's questions; nearly all were retained as core or background material, reflecting the exploratory, non-systematic nature of the search rather than a formal multi-stage screening process, with sources judged unrelated to generative or agentic AI in educational contexts excluded at this stage. The corpus assembled through this process spans AI-in-education literature more broadly, humanities-specific pedagogical studies, and conceptual work on authorship, human-in-the-loop design, and

transversal competences. Two limitations follow directly from this design. First, the corpus is not exhaustive: a systematic search across databases not covered here might surface additional humanities-specific cases, particularly for agentic AI, where the literature is thin enough that a small number of additional sources could alter the picture. Second, no formal screening protocol — dual independent coding, a PRISMA flow diagram — was applied; the authors carried out source selection and classification.

3. Taxonomic-descriptive analysis of AI approaches in humanities learning

3.1. Generative AI

Generative AI — large language models and, more marginally, tools performing narrower forms of natural-language analysis — is by far the more mature and widely documented of the two paradigms examined here, with most humanities-specific applications concentrated after the public release of ChatGPT in late 2022. Rather than listing isolated use cases, it is more informative to read the documented applications along a single axis: the depth of mediation between the learner and the primary text — that is, whether the AI output is positioned as an external object to be examined and questioned, or as a component woven into the interpretive task itself.

At the more peripheral end of this axis lies an activity in which students apply IBM Watson's Tone Analyzer to a poem by Catullus, translated into English for compatibility with the tool (Iannella, 2018). The activity's value lies specifically in what it asks students not to do: rather than checking whether the machine's output matches a correct or expected reading, students are asked to observe how the tool assigns emotional and linguistic tones to specific lines, compare that output against their own reading of Catullus's oxymoronic odi et amo, and reflect on where the two interpretations converge or diverge — and why. The exercise treats disagreement between human and machine reading as the productive centre of the task, not as a failure to be corrected. Predating the current generative-AI wave by several years and relying on a different underlying technology — tone and sentiment analysis rather than a language model — this case anticipates, in earlier technical form, a stance that resurfaces throughout the more recent literature (Peddar, 2025; Caserta, 2025): AI output as a prompt for critical reasoning about interpretation itself, rather than as an authority to be accepted or a target to be reproduced correctly.

Closer to the centre of the axis, generative AI has been evaluated directly on interpretive tasks over primary sources: ChatGPT's capacity to analyse and interpret poems and song lyrics ranging from ancient Greek works to modern compositions was tested by Virvou et al. (2023), who found notably strong performance on content interpretation and analogy identification, but weaker results on translation, historical contextualisation, and factual accuracy — with errors at times produced without any signal of uncertainty. A related quasi-experimental design tested generative feedback combined with interactive video assessment on philological content specifically (Pellas, 2025).

At the more integrated end, generative AI has been embedded as a working component of an extended disciplinary task rather than as an object to interrogate on its own. Peddar (2025) documents its structured integration into a secondary-school Latin classroom, built around guided student reflection on the tool's uses and misuses rather than unmediated adoption. A further step in integration is visible in a didactic sequence where generative tools support the construction and testing of counterfactual historical scenarios anchored in Herodotus, Thucydides, and Livy (Caserta, 2025), with the activity explicitly framed as strengthening — rather than replacing — source-based historical judgement. A broader mapping of generative AI applied to ancient-language and history teaching, including its documented limits around lexical ambiguity, poetic metre, and textual criticism, is offered by Cuffari (2025).

Read along this axis, the corpus assembled here shows a trajectory from AI-as-object-of-scrutiny toward AI-as-embedded-component of the disciplinary task. Whether this progressive

integration deepens or displaces the interpretive work it accompanies is a question the cases themselves do not settle — one that the descriptive picture assembled so far can only pose, not answer (Figure 2).

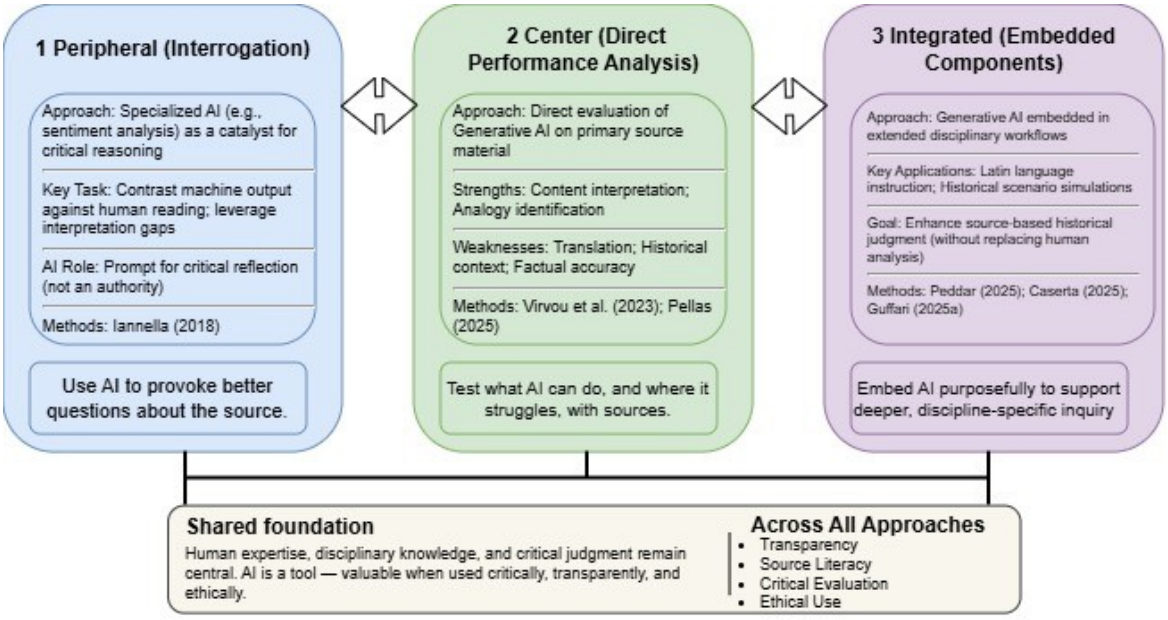


Figure 2: Generative AI in the corpus: from external object of scrutiny to embedded task component.

3.2. Agentic AI

Agentic AI is approached in this review to denote systems characterised by autonomy, multi-step planning, tool use, and goal pursuit with limited continuous human supervision — typically built on generative models as their underlying engine, so that the two are not alternative categories but different levels of system organisation— a definition adopted explicitly because the term circulates in the literature with considerably looser meanings, at times applied to any system exhibiting adaptive or interactive behaviour. This distinction can be operationalised through a set of verifiable criteria — autonomy, planning and reasoning, persistent memory, goal-directed multi-turn interaction, tool use, and safeguards — requiring a system to satisfy a majority of them before qualifying as agentic (proactive) rather than merely reactive (Kostopoulos et al., 2025); this threshold is adopted here to keep the label from collapsing into a synonym for "AI in general."

Agentic AI in education more broadly is an active and rapidly growing field. Recent work organises documented applications around a small set of recurring paradigms — reflection, planning, tool use, and multi-agent collaboration (Kamalov et al., 2025) — spanning multi-agent tutoring systems, automated essay scoring, and simulated classroom environments; Kostopoulos et al. (2025) offer a comparable synthesis across the wider AIED (Artificial Intelligence in Education) literature. Within the humanities specifically, however, this activity does not appear to extend: the corpus examined here yields a single documented case — an AI system autonomously producing an extended essay on the Aeneid with minimal step-by-step human direction, presented by Cuffari (2025) as an illustration of what agentic behaviour looks like in a classical-literature classroom rather than as an evaluated pedagogical intervention. Here the student's relation to the process is essentially one of reception: the multi-step reasoning happens inside the system, and what reaches the learner is the finished product.

A structurally different configuration is conceivable, though it does not appear in the literature examined here and is proposed on conceptual grounds rather than drawn from an observed case: one in which the learner constructs or configures the agentic workflow itself — for instance, sequencing and supervising the discrete steps of a translation task (source analysis, lexical

disambiguation, syntactic parsing, semantic interpretation, terminology verification, target-text drafting, and revision) rather than delegating the whole sequence to the system and evaluating only its output. On this reading, agentic AI would function less as an author to be checked and more as a process made newly visible and manipulable — one the learner assembles rather than merely receives. Whether such a configuration exists in practice, and with what effect, is a question this review can only raise, not answer, in the absence of documented cases.

Beyond this general landscape of applications, one conceptual framework speaks directly to the same dimension this section has been tracing — where, exactly, agency sits between learner and system. Agentic systems have been proposed as prospective "socio-cognitive teammates," collaborators embedded in the learning process rather than passive tools, with the locus of agency shared and negotiated between learner and system (Yan, 2025). It is the only framework identified in this review that addresses this dimension explicitly, which makes it a useful reference point — though of a different nature from the configuration outlined above: Yan's framing concerns how autonomy is distributed between learner and system step by step, while the question raised here concerns who authored the sequence of steps itself, regardless of how autonomously the AI then executes them.

This asymmetry between a mature general field and an almost unexplored humanities-specific application is itself worth stating as a finding of the mapping exercise. Whether this absence reflects a genuine limit of agentic AI in supporting humanistic reasoning, or simply a domain not yet explored, is a question the mapping itself cannot settle.

3.3. Comparative dimensions of generative and agentic AI

Table 1 summarises the dimensions along which the two paradigms mapped in Sections 3.1 and 3.2 differ, restricted to what is documented in the literature examined.

Table 1
Comparative dimensions of generative and agentic AI in humanities learning

Dimension	Generative AI	Agentic AI
Degree of autonomy	Single-turn, reactive; no multi-step planning — each output is generated in response to a discrete prompt	Multi-step, goal-directed; the system plans and executes a sequence of actions with minimal ongoing direction
Type of output	A discrete artefact produced per request (an interpretation, a translation, a tone analysis, a narrative reconstruction)	An extended, composite artefact (e.g., a complete essay) resulting from an internal, unobserved sequence of steps
Form of interaction with the learner	Prompt–response; the learner initiates and the system replies once per cycle, with re-prompting possible but each cycle independent	Delegated task; the learner specifies a goal and the system carries out the entire sequence autonomously
Point of human insertion in the loop	After generation — the learner (often teacher-guided) examines, questions, or integrates the output post hoc	After completion — the learner reviews the finished product; intermediate steps are not directly accessible

Read together, Sections 3.1–3.3 map where documented AI systems currently sit along a single axis — depth of mediation with the primary text and degree of autonomy over the process — without yet asking what this positioning means for the cognitive work of interpretation itself, or for the human oversight meant to accompany it.

4. Critical analysis

The previous section traced a single axis (Figure 3) along which the applications examined here differ: the depth of mediation between the learner and the object of learning — whether an AI output, or an AI-driven process, is positioned as an external object to be examined and questioned, or as a component woven into the task itself. What that section could describe but not settle is what this depth of mediation does to the learner engaged in it: whether being woven into a task,

rather than observing an output from outside it, changes the quality of cognitive engagement, and whether it changes how meaningfully a human remains in the loop of the process. These are the two questions this section now takes up.

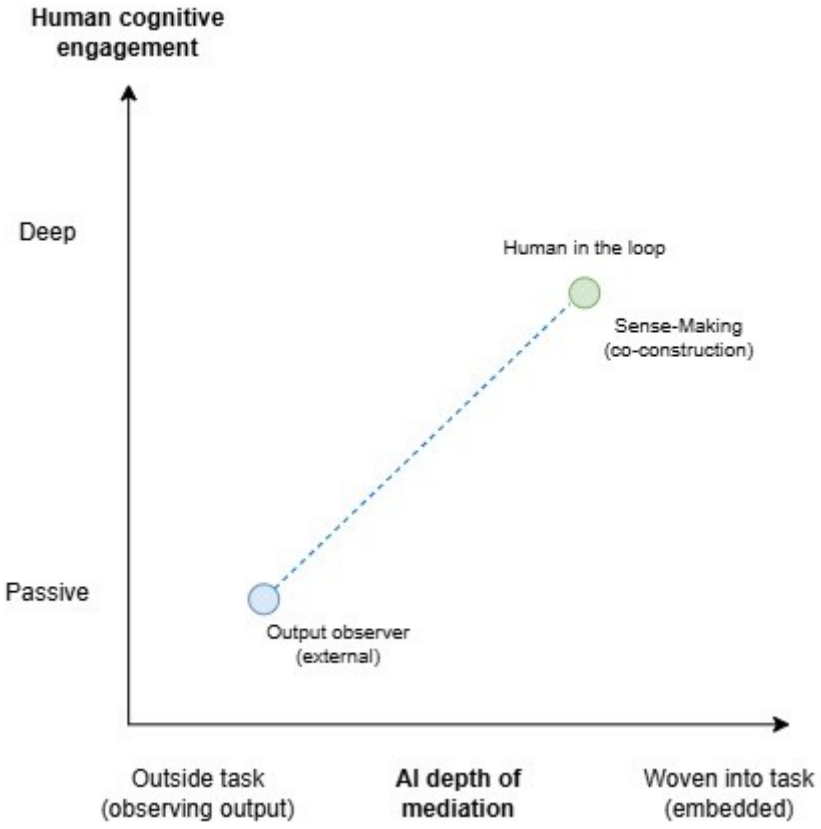


Figure 3: The depth-of-mediation axis

4.1. When AI scaffolds: cognitive affordances and active human-in-the-loop

The claim this subsection defends is that depth of mediation and operationalisation of the human-in-the-loop principle are not two variables to be assessed side by side, but the same mechanism read from two angles: when an AI output or process is woven into a task rather than held apart from it, the human is not merely checking a result afterwards but authoring the process the result belongs to — and it is this authorship, more than the mere presence of a check, that keeps a cognitive gain intact once the system is removed. The clearest demonstration that sequencing, not merely the presence of oversight, determines cognitive outcome comes from an experimental study outside the humanities. Wong and Qiu (2026) randomly assigned university students to a creative task performed independently, with unrestricted access to ChatGPT, or under a "think first, ChatGPT later" protocol requiring an independently generated idea before any engagement with the system. Unrestricted access typically meant instructing the system to generate a solution outright; the regulated group used it to develop ideas already their own. The distinction registered only once the tool was withdrawn: the unrestricted group's creativity reverted to baseline, while the regulated group's gains persisted into an unaided task. This is direct evidence that when human engagement occurs, the AI's contribution is what determines durability, not merely whether a check occurs at all.

Within the humanities, Caserta's (2025) sequence on counterfactual reasoning in Herodotus offers a design, not yet tested, in which the same principle is built directly into pedagogical structure: students move through guided variation of a single detail within a structured prompt, development of counterfactuals already latent in the source text, and finally independent composition of original scenarios with no prompt structure left to lean on. Cognitive scaffolding

and human oversight are not separate features of this design, since at each stage what the student is asked to author expands as what the prompt supplies contracts. A comparator with actual classroom evidence is offered by Voicu (2026), whose corpus of AI-mediated interpretation episodes identifies a trajectory of AI-critical interpretation — students triangulating, questioning, and contextualising AI output rather than adapting it — as the pattern in which scaffolded, non-delegating use of AI is associated with genuine interpretive autonomy, though also the least frequent of the three observed trajectories (26.2% of episodes); this descriptive finding motivates, but does not itself test, a developmental pedagogy proposed to cultivate that trajectory deliberately, in a design not unlike Caserta's own. It is worth being explicit about what this evidence base does and does not support: only Wong and Qiu's result demonstrates the mechanism causally; the humanities evidence available at this stage is either untested by design (Caserta) or associational rather than causal (Voicu).

Whether the disciplinary competence these cases cultivate also constitutes a transversal, labour-market-relevant one remains, at this stage, a hypothesis rather than a finding: bibliometric evidence associates process-oriented competences with resilience to AI-driven automation (Dilinika, 2026), and a structurally similar argument has been made for another discipline undergoing the same reconfiguration, statistics, whose practitioners are argued to remain valuable where their expertise concerns designing analytical processes rather than executing them (Liu et al., 2024) — a parallel that supports the hypothesis without, on the evidence assembled here, settling it.

4.2. When AI substitutes: cognitive risks and absent human-in-the-loop

Depth of mediation cuts both ways. The same feature that made the scaffolding cases durable — the learner's authorship over what is checked, and when — is precisely what is missing when an AI output is held apart from the learner's own work rather than folded into it, and its absence takes two overlapping forms: an error in the system's output that nothing invites the learner to catch, and a decline, over time, in the learner's own capacity to catch such errors independently.

The first is documented with unusual clarity within the humanities. Ross (2023) reports asking ChatGPT to translate Catullus 16, an explicit poem the system's content policy would not reproduce; rather than declining, the system substituted a different Catullan poem while presenting it as Carmen 16, with no disclaimer of any kind. As the author notes, a student taking this output at face value would develop a specific and traceable misconception about the text, not a vague one — the failure is not that the system is occasionally wrong, but that nothing in the output signals that this is the moment verification is required. This is what Cuffari (2025) terms hallucination, but the case matters here for a more specific reason: it identifies exactly the point in the interaction — the point of human insertion mapped in Table 1 as occurring after generation — at which the human-in-the-loop principle fails, because the check it presupposes is not one an unprompted learner has reason to perform.

Decline in independent capacity is documented outside the humanities, but by a design that isolates the mechanism with matched experimental conditions. Kosmyna et al. (2025) assigned participants to three groups — writing essays using an LLM, a search engine, or no external tool — and found that neural connectivity and reported ownership of one's own writing scaled with the degree of external reliance: strongest unaided, intermediate for search-engine use, weakest where an LLM supplied the content. Participants who wrote unaided before being reassigned to LLM use in a later session showed higher neural engagement than those who had used the LLM from the outset, independently replicating, by different means, the sequencing effect already demonstrated experimentally for the opposite case (Wong & Qiu, 2026).

Within the humanities, the same decline appears as the more frequent of two classroom trajectories: AI-dependent interpretation, in which output is adopted with minimal elaboration or verification — what Suriano et al. (2025) would classify as dependence on quick solutions — accounts for 35.7% of the episodes Voicu (2026) observes, against 26.2% for the critical pattern

discussed in the preceding subsection. A further piece of the same picture, this time about self-assessment rather than performance, comes from Kelly et al. (2023): confidence in using generative AI did not reliably track actual experience with some of the least experienced users reporting the highest confidence (Kelly et al., 2023) — a miscalibration that plausibly explains why the kind of verification required by the case discussed above (Ross, 2023) often does not occur even when nothing prevents it. Ramos-Benitez et al. (2026) describe comparable findings across the wider literature as "troubling associations between frequent AI use and diminished critical reasoning," explicitly qualifying them as emerging rather than settled.

None of this is a new anxiety in substance: the concern that a tool for producing answers might erode the capacity to reason toward them has a documented precedent in Plato's account, in the Phaedrus, of writing itself as a threat to memory (Cuffari, 2025). What is specific to the cases assembled here is not the anxiety but its location: in each of them, the point at which a human was meant to intervene in the loop is exactly the point that failed to register as one requiring intervention at all.

4.3. Methodological reflections on the corpus

Before drawing the two preceding subsections together, it is worth pausing on what kind of evidence has been assembled across this review as a whole, and not only on what it shows. Self-report is by far the most common method in the corpus examined here, whether in the form of single-session surveys (Kelly et al., 2023; Suriano et al., 2025) or structured classroom observation coded after the fact (Voicu, 2026); a smaller number of cases rest on qualitative testing by a single author (Ross, 2023) or on a pedagogical design proposed but not yet tested (Caserta, 2025). Only one study departs from this pattern by manipulating conditions experimentally and tracking a subsequent unaided task (Wong & Qiu, 2026), and only one, across the entire body of literature this review draws on, employs a neurophysiological measure of cognitive engagement rather than inferring it from behaviour or self-report (Kosmyna et al., 2025). This asymmetry is worth stating plainly. Kosmyna et al.'s EEG connectivity technique is well established in neuroscience more broadly, but its specific interpretation here — as evidence of "cognitive debt" — rests on a single, not yet replicated study. Elsewhere in the corpus, the reverse gap appears: theories drawn from cognitive neuroscience are invoked to interpret findings (Suriano et al., 2025, via dual-process and cognitive load theory) without any neurophysiological measurement to support them. Across this literature, then, the vocabulary of cognitive neuroscience and its empirical methods rarely appear together.

4.4. Synthesis: the relationship between cognitive positioning and HITL operationalisation

The two preceding discussions are not independent inventories of positive and negative cases. Where cognitive scaffolding held — the gradual withdrawal of structure, from a fully structured prompt to none, in Caserta's (2025) sequence, the durable gains tied to sequencing in Wong and Qiu (2026), the rarer but identifiable trajectory of triangulation and revision in Voicu's (2026) classroom corpus — the human-in-the-loop principle was, in each case, substantively operationalised: the human's engagement did work, not merely register a check. Where cognitive substitution took hold instead — the unflagged error a student had no cue to question (Ross, 2023), the graded decline in neural engagement tracking depth of reliance (Kosmyna et al., 2025), the more frequent pattern of direct uptake without verification (Voicu, 2026), the confidence untethered from experience (Kelly et al., 2023) — the same principle was present only in name, a checkpoint the interaction gave no reason to use. This is not two phenomena running in parallel; it is one relation observed from either side.

This relationship is not to be accounted for by the presence of AI, nor by its degree of autonomy. It is to be focused on whether the human retains authorship over the structure and sequencing of the process that generates the output, rather than only over its final content.

Rodrigues (2026) highlighted this very point: after analysing AI-assisted authorship through virtue epistemology, the authorship survives delegation to a generative system through specific curatorial virtues. Rodrigues names an architectural virtue, which is exercised precisely when the human holds responsibility for the way a process is structured. The latter goes beyond the act of approving what AI produces. Floridi (2025) identifies three learning roles within a system of rules: the player, the observer, and the designer. The player uses their skills framed within the rules as given; the observer reflects on those rules and can critique them without changing them; and the designer, who reconceives the rules themselves, creating new structures rather than only using or questioning existing ones.

While the first two roles stick within an architecture built by someone else, the third role accounts for authoring the architecture itself. By looking back at what Rodrigues's architectural virtue identifies, independently of the underlying system as generative or agentic, the shift from player or observer to designer comes into play. It is not the withholding of delegation; it is the holding of architectural account for a process while the same is delegated. The latter is meant as opposed to something elicited only after delegation has ended.

What none of the cases discussed above reaches is this same shift at a larger scale: authorship not over a single interpretive judgement, nor over a sequence of prompts within a fixed activity, nor over the structure of one artefact, but over the architecture of a multi-phase task as a whole — the progression from guided variation to independent composition in Caserta's sequence was itself designed by someone else, however much authorial autonomy it grants the student within it. Independently of whether such a task is realised through generative prompting, an agentic workflow, or some combination of the two, this larger-scale authorship remains, on the evidence assembled here, unreached.

This shift in scale is not to be confused with a shift in kind already addressed elsewhere: a four-level continuum of AI autonomy during interaction (Yan, 2025) describes how much initiative a system exercises moment-to-moment, a question orthogonal to the one traced here, which concerns who authored the sequence of steps regardless of how autonomously it is then executed.

If the relationship described in this section holds, the competence at stake across both subsections is narrower than "using AI well": it is the capacity to author, verify, and revise the structure of a process rather than only its content — a capacity bibliometric and policy literature already associate with resilience in an AI-reconfigured labour market (Dilinika, 2026; Liu et al., 2024; OECD, 2021). Whether this capacity, demonstrated here only at the scale of a single artefact, extends to the architecture of a process as a whole remains the open question this review has been able to pose, but not to settle.

5. Conclusions and future perspectives

This review has traced how generative and agentic AI are currently mediating humanities learning and the consequences for the cognitive processes that learning is meant to cultivate. On the descriptive side (RQ1), the corpus shows a marked asymmetry: generative AI is comparatively mature and well-documented across a range of humanities applications, while agentic AI, aside from general education literature, remains almost entirely unexplored in humanities-specific settings. On the critical side (RQ2), the evidence assembled does not support treating depth of mediation and human-in-the-loop operationalisation as separate axes: they appear closely linked, and this association is explained by authorship over the structure of a process, not merely oversight of its output. This association is observed within the cases assembled here and should not yet be read as a general mechanism. Read against this, RQ3 receives a qualified answer: where such authorship is present, cognitive gains persist once AI support is withdrawn; where it is absent, an unflagged error or a quietly eroding capacity for independent verification takes its place — and only one of these patterns, on the evidence available, extends beyond the scale of a single artefact to the architecture of a task as a whole, and even there it remains unreached. This last point carries a broader implication that this review has treated throughout as hypothesis rather than finding: the

competence that distinguishes scaffolding from substitution — authoring, verifying, and revising the structure of a process rather than only consuming its output — is transversal rather than humanities-specific, aligning structurally with the competences that policy and bibliometric literature increasingly associate with resilience in an AI-reconfigured labour market. If this alignment holds, humanities education mediated by AI under a substantive human-in-the-loop principle is not only a matter of preserving disciplinary relevance against a technology reshaping how texts are read and written; it may also be one of the more direct pedagogical routes toward exactly the competences that employability now requires — a way of showing that the interpretive judgement humanities has always cultivated has a stake beyond the discipline itself, not a replacement for that stake. Two limitations qualify these conclusions and set the direction for what follows. First, this is a narrative rather than a systematic review, and the humanities-specific agentic AI literature is so thin that a small number of additional sources could shift the picture presented here. Second, the transversal-competence hypothesis advanced above rests on parallel literatures — labour-market and bibliometric — rather than on direct empirical evidence within humanities education itself; testing whether the competences this review has traced as durable within a single artefact actually transfer, and whether they extend to the architecture of a process as a whole, is the natural next step this review has been able to pose but not to settle. Third, the account offered here treats depth of mediation as a single descriptive axis; a more fully operationalised framework — separating the structure of mediation from the depth of cognitive engagement it affords, and treating agentic AI's autonomy along multiple, potentially co-occurring dimensions (goals, constraints, tool access, and degree of intermediate-step autonomy) rather than as a single categorical property — is left to future work.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude (Anthropic) in order to: Formatting assistance. The author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

Berger, J., Burton, J. W., Hertwig, R., Kosche, T., Kurvers, R. H. J. M., Kurzenberger, B., Lazik, C., Onnasch, L., Rieger, T., Thoma, A. I., Wulff, D. U., & Herzog, S. M. (2026). Fostering human learning is crucial for boosting human–AI synergy. https://doi.org/10.48550/arXiv.2512.13253

Caserta, C. (2025). Oltre il fatto: storia controfattuale e Intelligenza Artificiale nella didattica dell'antico. In G. Mosconi (Ed.), Antico & IA. Intelligenze artificiali e didattica del mondo antico (pp. 153–194). Edizioni Università di Cassino.

Chiriatti, M., Ganapini, M., Panai, E., Ubiali, M., & Riva, G. (2024). The case for human–AI interaction as system 0 thinking. Nature Human Behaviour, 8(10), 1829–1830. https://doi.org/10.1038/s41562-024-01995-5

Chun, J., & Elkins, K. (2023). The crisis of artificial intelligence: A new digital humanities curriculum for human-centred AI. International Journal of Humanities and Arts Computing, 17(2), 147–167. https://doi.org/10.3366/ijhac.2023.0310

Cianfrocca, G. (2025). Intelligenza Artificiale e traduzione contrastiva. Il caso Sulpicia. In G. Mosconi (Ed.), Antico & IA. Intelligenze artificiali e didattica del mondo antico (pp. 85–128). Edizioni Università di Cassino.

Cuffari, S. V. (2025). Prometeo o automa? L'Intelligenza Artificiale come risorsa critica per la didattica della storia e delle letterature antiche. In G. Mosconi (Ed.), Antico & IA. Intelligenze artificiali e didattica del mondo antico (pp. 19–48). Edizioni Università di Cassino.

Darvin, R. (2025). The need for critical digital literacies in generative AI-mediated L2 writing. Journal of Second Language Writing, 67, 101186. https://doi.org/10.1016/j.jslw.2025.101186

De Nonno, M. (2021). Latino per la scuola, latino per la società. In Centro Studi "La permanenza del Classico" (Ed.), Lucrezio, Seneca e noi. Studi per Ivano Dionigi. Pàtron Editore.

Dilinika, J. M. S. (2026). Artificial intelligence and transversal skills for future workforce: A bibliometric analysis of research in higher education (2020–2025). Education for Information. Advance online publication. https://doi.org/10.1177/01678329261468180

DivyaVenkatesh, J., Jaiswal, A., & Nanda, G. (2024). Comparing human text classification performance and explainability with large language and machine learning models using eye-tracking. Scientific Reports, 14, 14295. https://doi.org/10.1038/s41598-024-65080-7

Essien, A., Bukoye, O. T., O'Dea, X., & Kremantzis, M. (2024). The influence of AI text generators on critical thinking skills in UK business schools. Studies in Higher Education, 49(5), 865–882. https://doi.org/10.1080/03075079.2024.2316881

Floridi, L. (2025). On the Future of Education in a Digital Society. Available at SSRN: https://ssrn.com/abstract=5654551 or http://dx.doi.org/10.2139/ssrn.5654551

Gonsalves, C. (2024). Generative AI's impact on critical thinking: Revisiting Bloom's Taxonomy. Journal of Marketing Education, 48(1), 4–19. https://doi.org/10.1177/02734753241305980

Gu, P., Duan, F., Li, W., Xu, B., Cai, Y., Yao, T., Zhuo, C., Liu, T., & Ge, B. (2025). Bridging technology and humanities: Evaluating the impact of large language models on social sciences research with DeepSeek-R1. https://doi.org/10.48550/arXiv.2503.16304

Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. European Journal of Education, 57(4), 542–570. https://doi.org/10.1111/ejed.12533

Iannella, A. (2018). Digital authentic learning: introdurre alla tone analysis per favorire un pensiero critico digitale. Thamyris, n.s. 9, 319–336.

Kamalov, F., Calonge, D. S., Smail, L., Azizov, D., Thadani, D. R., Kwong, T., & Atif, A. (2025). Evolution of AI in education: Agentic workflows. https://doi.org/10.48550/arXiv.2504.20082

Kelly, A., Sullivan, M., & Strampel, K. (2023). Generative artificial intelligence: University student awareness, experience, and confidence in use across disciplines. Journal of University Teaching and Learning Practice, 20(6), 1–16. https://doi.org/10.53761/1.20.6.12

Kolhatin, A. O. (2025). From automation to augmentation: a human-centered framework for generative AI in adaptive educational content creation. In Proceedings of AREdu 2025: 8th International Workshop on Augmented Reality in Education. CEUR Workshop Proceedings.

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. MIT Media Lab.

Kostopoulos, G., Gkamas, V., Rigou, M., & Kotsiantis, S. (2025). Agentic AI in education: State of the art and future directions. IEEE Access, 13, 177467–177491. https://doi.org/10.1109/ACCESS.2025.3620473

Lazaros, K., Vrahatis, A. G., & Kotsiantis, S. (2026). Human-in-the-loop artificial intelligence: A systematic review of concepts, methods, and applications. Entropy, 28(4), 377. https://doi.org/10.3390/e28040377

Lim, K. Y., & Darvin, R. (2026). Critical digital literacies, generative AI, and the negotiation of agency in human-AI interactions. System, 136, 103904. https://doi.org/10.1016/j.system.2025.103904

Liu, J., Chen, K., & Lyu, W. (2024). Embracing artificial intelligence in the labour market: the case of statistics. Humanities and Social Sciences Communications, 11, 1112. https://doi.org/10.1057/s41599-024-03557-6

Mosconi, G. (2025). Intelligenze Artificiali, didattica dell'antico e traduzione dalle lingue classiche. Perché essere sia Apocalittici che Integrati. In G. Mosconi (Ed.), Antico & IA. Intelligenze artificiali e didattica del mondo antico (pp. 49–84). Edizioni Università di Cassino.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: a state of the art. Artificial Intelligence Review, 56, 3005–3054. https://doi.org/10.1007/s10462-022-10246-w

Nasr, N. R., Tu, C.-H., Werner, J., Bauer, T., Yen, C.-J., & Sujo-Montes, L. (2025). Exploring the impact of generative AI ChatGPT on critical thinking in higher education: Passive AI-directed use or human–AI supported collaboration? Education Sciences, 15(9), 1198. https://doi.org/10.3390/educsci15091198

OECD. (2021). AI and the future of skills, volume 1: Capabilities and assessments (Educational Research and Innovation). OECD Publishing. https://doi.org/10.1787/5ee71f34-en

Peddar, D. (2025). Utilising Generative AI in the Classics classroom. The Journal of Classics Teaching, 26, 8–13. https://doi.org/10.1017/S2058631024001363

Pellas, N. (2025). Effects of generative AI feedback and interactive video assessment on student learning achievement in philological content creation courses. Journal of Educational Computing Research, 64(1), 16–58.

Pérez-Paredes, P., Curry, N., & Ordoñana-Guillamón, C. (2025). Critical AI literacy for applied linguistics and language education students. Journal of China Computer-Assisted Language Learning. Advance online publication. https://doi.org/10.1515/jccall-2025-0005

Quattrociocchi, W., Capraro, V., & Perc, M. (2025). Epistemological fault lines between human and artificial intelligence. https://doi.org/10.48550/arXiv.2512.19466

Ramos-Benitez, M. J., García-Osorio, M. E., & Delgado, Y. (2026). Critical alliance of AI in education: A pedagogical framework for safeguarding cognitive skills. International Medical Education, 5(1), 22. https://doi.org/10.3390/ime5010022

Rodrigues, T. V. (2026). The apt curation model: An epistemic virtue theory of AI-assisted authorship. Philosophy & Technology, 39(39). https://doi.org/10.1007/s13347-026-01038-z

Romero, M. (2025). From consumption to co-creation: A systematic review of six levels of AI-enhanced creative engagement in education. Multimodal Technologies and Interaction, 9(10), 110. https://doi.org/10.3390/mti9100110

Ross, E. A. S. (2023). A new frontier: AI and ancient language pedagogy. The Journal of Classics Teaching, 24, 143–161. https://doi.org/10.1017/S2058631023000430

Schwartz, D. G., Teeni, D., & Yahav, I. (2023). Reciprocal Human Machine Learning (RHML): Human-AI collaboration based on theories of dyadic learning. In Proceedings of the AAAI Summer Symposium Series (SuSS-23). AAAI.

Suriano, R., Plebe, A., Acciai, A., & Fabio, R. A. (2025). Student interaction with ChatGPT can promote complex critical thinking skills. Learning and Instruction, 95, 102011. https://doi.org/10.1016/j.learninstruc.2024.102011

Virvou, M., Tsihrintzis, G. A., Sotiropoulos, D. N., Chrysafiadi, K., Sakkopoulos, E., & Tsichrintzi, E.-A. (2023). ChatGPT in artificial intelligence-empowered e-learning for cultural heritage: The case of lyrics and poems. In 2023 14th International Conference on Information, Intelligence, Systems & Applications (IISA). IEEE. doi: 10.1109/IISA59645.2023.10345878.

Voicu, C.-G. (2026). AI-mediated learning and the restructuring of interpretive cognition: A developmental-critical model for social sciences and humanities education. Journal of Digital Pedagogy, 5(1), 37–51. https://doi.org/10.61071/JDP.2665

Wong, S. S. H., & Qiu, S. X. (2026). Think first, ChatGPT later: Guiding human–AI collaboration for learning gains in independent human creativity. Educational Psychology Review, 38(45). https://doi.org/10.1007/s10648-026-10118-7

Yan, L. (2025). From passive tool to socio-cognitive teammate: A conceptual framework for agentic AI in human-AI collaborative learning. https://doi.org/10.48550/arXiv.2508.14825

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenič, A., Spector, M., Liu, J.-B., Yuan, J., & Li, Y. (2021). A review of artificial intelligence (AI) in education from 2010 to 2020. Complexity, 2021, Article 8812542. https://doi.org/10.1155/2021/8812542