JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 11 Issue 02
2026

Research Article
https://doi.org/10.56979/1102/2026/1664

Integrating Large Language Models into Calligraphy and Painting Education: Effects on Learning Outcomes

Qiang Li[1], and Tassanee Laknapichonchat[1*]

[1]Institute of Science, Innovation and Culture, Rajamangala University of Technology Krungthep, Bangkok, 10120, Thailand.
[*]Corresponding Author: Tassanee Laknapichonchat. Email: tassalak@gmail.com

Received: June 07, 2026 **Accepted:** August 30, 2026

Abstract: The use of LLM is rapid and we are not yet clear how much they can add to the education of art. This review explores how these systems can affect learning in calligraphy and painting, two studio arts that are deceptively embodied, based on cultural knowledge, visual judgment, and authorship. Recent peer-reviewed literature was synthesized using a structured integrative approach, concerning the use of artificial intelligence in education, the co-creation between human and artificial intelligence, creative agency, feedback, and visual-art practice. In the literature, large language models are reported to open up access to explanation, questioning, comparison and reflective feedback. These are reported or expected benefits; only wider access to explanation and feedback has consistent empirical support, and that support comes mainly from adjacent educational settings rather than from studio classes. Their benefits seem to be most likely when students make an initial intention, experiment with ideas in terms of material practice, and are confirmed by the teacher. There is less evidence of independent mastery of the skills, originality, delayed retention, or transfer once the support is removed. The main risks are: cognitive offloading, culturally generic critique, stylistic convergence, not being aware of who the author is, lack of equitable access and general confidence in fluent responses. The review suggests the use of the framework of the CRAFT-LM: cultural grounding, retained agency, adaptive feedback, formative process assessment, teacher moderation, and learning monitoring. It sets priorities for empirical testing of the discipline without offering an untested evaluation model as a result. Current evidence suggests that LLM could be used as a scaffold rather than a source for creating, assessing, or replacing studio practice, provided that they are used with caution, under the guidance of the teacher, and within explicit and culturally-aware approaches to process-based studio learning.

Keywords: Artificial Intelligence; Calligraphy Education; Creative Agency; Large Language Models; Learning Outcomes; Painting Education

1. Introduction

Large language models (LLMs) have come a long way from general-purpose conversation systems to the educational setting. They can explain concepts, compare alternatives, provide examples, offer feedback, and facilitate reflection. Some of these skills seem appealing in calligraphy and painting instruction where students may require a lot of repetition, background information, words to use when discussing art, and personal work. However, the technology makes its way into fields where learning cannot be broken down into right and wrong verbal responses. Calligraphy relies on a body-based rhythm and pressure, spacing, proportion, and culturally-founded conventions [1, 2, 3, 4]. Painting is observable, compositional, manipulative, colour-judgmental, and positionally personal [5, 6]. Supportive uses of a fluent model response can help with these processes, but can also mask a lack of observation or substitute for decision-making that the learner needs to

make. As such, reviews of AI in education say that benefits are not found in tools themselves, but rather in how they are used, and whether learners can be verified as benefiting from their use [7, 8, 9, 10, 11].

It is particularly significant in the studio learning context to differentiate between assistance and substitution. An LLM can give an explanation that can be compared to existing knowledge in subjects that have heavy amounts of text. Feedback in the visual disciplines will need to be tied to an artifact, process, and a purpose. Why does a character feel unstable? How are there no shadows or depth in a landscape? How does the change in colour affect the mood of a character? The model cannot assume the teacher's rubric, student's stage of development, cultural heritage being learned, or physical limits of brush and paper. Multimodal systems are also prone to misreading subtle formal qualities [12]. First, the question of whether or not the system can generate plausible critiques is no longer relevant. It is whether the interaction aids learners to note, test, develop and eventually perform without assistance. This concern is aligned with studies on explainable AI and the role of student agency in relation to meaningful learning, which emphasize the importance of transparency and keeping control in the learning process [13, 14].

This review separates LLM-based text systems from multimodal systems. Explanation, questioning, comparison, reflection can be facilitated with a text-only model, but it cannot examine the learner's artifact. Images, visible features, brush force, kinesthetic feedback, paper resistance, drying behavior and the temporal dynamics of a stroke can only be felt indirectly by the multimodal system if these signals are sensed by additional sensors or video [2, 3]. The claims in this review thus refer to LLM-supported instruction in the context of a larger human–tool system, rather than an LLM as a complete perceptual and/or motor tutor.

The scope for the visual arts is deliberately constrained. It includes two dimensional calligraphy and painting procedures where learning relies on visual judgement, mark-making using materials or digital media and critique. When the learning objective focuses on composition, technique, interpretation, or authorship, then traditional brush-and-ink calligraphy, ink painting, oil or water-based painting and related digital painting are included. Production of prompt-to-image, 3-D modeling, animation, and sculpture without learner controlled studio practice is not part of the review.

Cultural needs are lacking too in a lot of personalization, especially calligraphy and the painting, and are not sufficiently being met. Chinese calligraphy is a kind of writing, a kind of visual art and also a historically specific practice [1, 2, 3, 4]. Culture can refer to conventions of the script, regular script (kaishu), running script (xingshu), clerical script (lishu); stroke-order expectations; relations of horizontal (heng), vertical (shu), turning (zhe), hook (gou), left-falling (pie), and right-falling (na) strokes; balance within structure; the expectations for controlled variation; and comparison with teacher-selected exemplars from the past. For this review, cultural grounding in Chinese painting is treated as including emptiness, brush texture, ink modulation, compositional rhythm, symbolism, and historical reference; these features are used as analytic descriptors rather than as a numerical claim about model representation. A search of the literature for the reproducibility did not yield a peer-reviewed study of the representation of training-data for Chinese calligraphy versus western oil painting directly in 2019-2025. Therefore, it is not claimed to be a numerical head-to-head imbalance for that reason. Empirical evidence suggests, however, that calligraphy-analysis studies tend to rely on dynamic stroke trajectories, ordered movement, sensor data or brush–paper interaction for evaluating calligraphic features that cannot be captured by static image corpora, while general image-generation studies show that culturally underrepresented visual contexts can be flattened or stereotyped by dominant model defaults [1, 2, 3, 15]. The implication is thus not an unsupported estimate of relative corpus size, but is rather about representational inadequacy. In Chinese calligraphy, if there is not enough time to execute the strokes, the stroke order, pressure modulation, the speed of the stroke, the turning rhythm, and the lifting and dropping motion of the brush will all be rendered as if in a static contour; the stroke order, pressure modulation, the speed of the stroke, the turning rhythm, and the lifting and dropping motion of the brush are all an essential part of the visual grammar by which scripts and individual performances are judged [2, 3, 4]. Western oil paintings, on the other hand, have the potential of retaining visible color relations, composition, edge structure, and surface appearance in their static corpora, even though they lack certain embodied variables like brush loading, viscosity, drying, and the order of layers. The cross-tradition concern is thus not that Western painting is fully represented, or

that one corpus is a known multiple of the other, but rather that there is a set of culturally and materially significant variables that are coded and annotated in the one tradition which are not coded and annotated in the other.

The growing body of literature emphasizes the real opportunities. Generative systems can broaden the search for ideas, help students to express intentions, give students more than one explanation, and give quick on-the-spot prompts for comparison. These may help to make it easier for students to ask questions, especially those who don't like to ask for clarification. However, research on wider education indicates that feedback, engagement, self-directed learning and specific cognitive tasks can be beneficial, but with varying findings for different tasks and designs [16, 17, 18]. Artificial Intelligence can help create variations and externalize it in creative work. However, there is also some controlled evidence of a trade-off: too many users using the same model suggestions can lead to a reduction in the diversity of the collective output while enriching the individual output [19]. This is important for painting education because in this context stylistic convergence might manifest as the success of technique and thus the erosion of one's own voice.

The dangers go beyond incorrect feedback. Productive struggle can be reduced with immediate assistance. Productive struggle in studio learning represents the learner's attempt to look closely, to make a plan, to work through compositional or structural issues, to deal with the resistance of the material, to repeat a challenging movement, and/or to make a decision about what to alter before a response is given to the learner. A student can use a model's vocabulary without developing visual discrimination and/or a compositional suggestion without developing an understanding of its logic. In the field of education research, related patterns are mentioned in terms of cognitive offloading [20], dependency and reduced agency [21]. Aesthetic offloading in studio education can be realized by providing learners with imagery, criticism language or stylistic guidance before the learner sets an intention and tries to test it. The end result can look like a solution, but the learning outcome is limited. Educational quality, however, is not determined by ease of use and perceived usefulness [22, 23].

One of the biggest challenges is what is meant by an educational effect. Most existing literature on the Generative Artificial Intelligence (AI) is about instant performance, satisfaction or task completion. These are all important indicators but cannot be evidence of whether learners have learned the technique, transferred their learning to unfamiliar tasks, or used their sense of judgement when the support is removed. Calligraphy is not merely identification of proper "form" of a character, but also control of the brush stroke [2, 4]. Making a picture in one's mind that is formed verbally is a different experience from the ability to make a picture that is not formed verbally. Brief intervention times, a variety of tasks, and inconsistent reporting of outcomes frequently are reported in these reviews of AI in education [24, 25, 26, 27].

The problem is more acute in studio disciplines, where learning is spread out between perception, movement, language, materials, and critique [5, 6, 28]. A model may assist a student in identifying a weakness in a piece of composition but the student must be able to identify the weakness in the artwork and make the change. It may provide historical background information, but must be confirmed by the tradition itself. May offer alternatives, but the learner will need to choose which one is meaningful. Assisted output vs. the process of learning is a question about what process evidence a student observed, selected, rejected, transformed, practised and then reproduced on their own.

In this article the relationships are reviewed following a structured integrative review. This evidence base encompasses recent research in the areas of artificial intelligence in education and generative feedback, student agency, self-regulated learning, human–AI co-creation, authorship, cultural bias, and visual-art practice. Performance and image generation quality of the model is not assessed in the review. It also claims that the existing literature does not support the calculation of a pooled estimate of the causal effect on calligraphy or painting. Instead, it establishes common mechanisms, boundary conditions, and gaps of outcomes, and plans an integrated teaching and evaluation system.

They consider LLM integration as a system-design problem which makes the study suitable for the scope of the Journal of Computing & Biomedical Informatics, which publishes work on the design and evaluation of computing systems in applied settings. The elements of the related system are learner, task, artefact, model,

rubric, cultural knowledge base, and access conditions and governance rules. Innovation is not just a "chatbot" module to an existing lesson. It is a process that must be coordinated in terms of feedback, assessment, documentation, teacher role and collection of evidence. This view aligns with the current research that focuses on the interplay of technological, pedagogical, and institutional factors, and not technological determinism, in education [27, 29, 30].

The review is structured around four research questions. RQ1: Across the included 2019-2025 evidence, what is the strength and directness of support for technical skill, conceptual learning, creative development, self-regulated learning, affective experience, and durable learning in LLM-supported calligraphy and painting education? RQ2: How do text-only and multimodal LLM affordances and limitations correspond to the embodied, visual, material, and perceptual requirements of calligraphy and painting practice? RQ3: Which task-design, teacher-mediation, cultural-grounding, learner-agency, access, and governance conditions are associated with LLM use functioning as scaffolding rather than substitution? RQ4: Which methodological gaps, particularly delayed unassisted transfer, blinded artifact assessment, intervention duration, teacher mediation, and cultural validity, currently limit claims about durable and independent studio learning? These questions guide the review approach, Results, and Discussion and keep model capability, assisted performance, and evidence of learning analytically separate.

2. Review approach

2.1. Design and scope

The selection of a structured integrative review was opted for due to the broadness of the literature reviewed both conceptually and methodologically. Direct studies of LLM in calligraphy and painting are relatively under-studied to date. The most striking parallels that can be identified are in generative artificial intelligence (AI) in education, feedback research in education, research on human-AI creative activity, and research on the nature of artificial-intelligence art. Sources include information on various facets of the problem. To carry out a quantitative meta-analysis, the participants, interventions, outcome measures and effect statistics should be the same. Those conditions don't exist now. When working towards an integrative design, it is important to make careful comparisons but not to interchange the unlike evidence.

Eligibilities: 2019-2025. The work in this area has also been important to include, as it links to early studies on embodied calligraphy and represents the first steps towards the literature in the educational-AI transformer era, yet maintaining a close connection to the synthesis of the present day generative and multimodal systems. Not all engineering papers were counted simply for announcing model benchmarks, but were included when they explained a capability or limitation for the studio learning. Hence, the review is more focused on education mechanisms and results than on the architecture's performance.

2.2. Evidence identification and search record

On 27 August 2026, the auditable search was executed in the SciSpace scholarly index for peer-reviewed journal and conference papers published from 2019 through 2025. The date is the retrieval date, not the publication-date criterion: records published within 2019-2025 were accessed retrospectively through the search conducted on that day. Eligibility was judged by the date of first online publication. Jiang et al. [31] was published online on 30 December 2025 and later assigned to a 2026 issue, so it is counted within the 2019-2025 window and in the 2025 publication count. Four preset full-question searches were used: (i) Education in Chinese calligraphy and brush-stroke learning, and the motor control of handwriting and computer-vision analysis; (ii) Embodied learning, material practice, and dialogic feedback in painting or studio art; (iii) Automation bias, fluency effects, overreliance, and trust in AI-generated feedback; (iv) Multimodal generative AI, computer vision, image analysis, and visual-art education. The first 10 records returned for each query under SciSpace's default relevance ordering were taken forward to screening, giving 40 records before duplicate handling. The Search Log and Screening Log in Supplementary File S1 record the retrieved records, screening decisions, and exclusion reasons.

One duplicate was identified among the 40 retrieved records, leaving 39 unique records for eligibility screening. Seventeen records met the date, publication-type, provenance, and topical criteria and form the auditable analytic core; 22 records were excluded. Table 1 sets out the four search questions as entered, the core concepts each question covered, the number of records carried forward at each stage, and the number of records excluded for each reason; the record flow follows the reporting structure of PRISMA 2020 [32]. The 17 analytic-core studies are the only records used for study counts, publication trends, and coded evidence-frequency statements. Sources used only for theory, definition, policy, or interpretation are treated as contextual references and are not additional included studies. The Extraction Matrix in Supplementary File S1 records the study-level design, population or material, technology, outcome, evidence role, blinded-rating information, finding used in the synthesis, and directness for each of the 17 records.

Table 1. Search questions and record flow (SciSpace scholarly index, searched 27 August 2026; limits: 2019-2025, peer-reviewed journal or conference).

ID	Search question (as entered)	Core concepts	Records
S1	What peer-reviewed studies from 2019 to 2025 examine Chinese calligraphy education, brush-stroke learning, handwriting motor control, or computer-vision analysis of Chinese calligraphy that are relevant to embodied learning and visual-form assessment?	Calligraphy education; brush-stroke learning; handwriting motor control; computer vision	10
S2	What peer-reviewed studies from 2019 to 2025 examine embodied learning, motor skill, material practice, or dialogic feedback in painting or studio art education?	Embodied learning; motor skill; material practice; dialogic feedback; painting and studio art	10
S3	What peer-reviewed studies from 2019 to 2025 explain automation bias, fluency effects, overtrust, or misplaced trust in AI-generated feedback among learners, especially novices with low prior knowledge?	Automation bias; fluency effects; overtrust; AI-generated feedback; novice learners	10
S4	What peer-reviewed studies from 2019 to 2025 examine multimodal generative AI, computer vision, or image-analysis systems in art education, painting, or visual-arts learning, including composition and color analysis?	Multimodal generative AI; computer vision; image analysis; art and painting education	10
	Records retrieved (first 10 per question under default relevance ordering)		**40**
	Duplicate records removed		**1**
	Unique records screened against the eligibility criteria in Section 2.3		**39**
	Excluded: outside the 2019-2025 window		**8**
	Excluded: preprint or arXiv record without verified peer-reviewed publication		**4**
	Excluded: book, book chapter, or unclear record type		**3**
	Excluded: repository-only item		**3**
	Excluded: insufficient methodological detail in the retrieved record		**2**
	Excluded: unclear publication provenance		**1**
	Excluded: mini-review of lower methodological weight		**1**
	Records included in the analytic core		**17**

Table 1 makes the search reproducible at the level of each question. The four questions were chosen to cover the main parts of the review problem: the embodied character of calligraphy, material and dialogic learning in painting, the reliability of AI-generated feedback, and the capabilities of multimodal systems. The lower rows show how the 40 retrieved records were reduced to the 17-study core and why each of the 22

exclusions was made, counted by the primary reason recorded in the Screening_Log of Supplementary File S1. Every count reported in the Results can therefore be traced back to a defined stage of selection.

SciSpace's default relevance ordering determined which 10 records per query entered screening. The authors did not define or re-rank this ordering, and the platform's underlying weighting criteria were not available in the search record. Relevance rank therefore functioned as a retrieval boundary, not as the eligibility rule: the retrieved records were screened against the criteria in Section 2.3. Because records below the top-10 cut-off were not screened, platform ranking could have affected which studies were available for inclusion; this limitation is stated in Section 4.6. Two additional cross-check searches for cultural-representation bias in generative-image systems and for foundational educational-AI theory were documented in Supplementary File S1 but were used only for interpretation. The pre-existing theory/policy bibliography was likewise treated as contextual material and did not contribute to the 17-study analytic core or to study-frequency statements.

Supplementary File S1: 342-revised_Supplementary_S1.xlsx contains five worksheets: Search_Log, Screening_Log, Extraction_Matrix, Publication_Trend, and Provenance_Notes. Together they provide the reproducibility materials requested for source identification, screening, study-level extraction, publication-trend reporting, and provenance documentation.

2.3. Eligibility and evidence roles

Evidence was considered for the analytic core when it met all of the following inclusion criteria: publication between 2019 and 2025 inclusive; peer-reviewed journal or conference status; identifiable publication provenance; relevance to at least one of four domains (educational uses of LLMs or generative AI; feedback, self-regulation, engagement, and student agency; human-AI co-creation, authorship, and creativity; or visual-art analysis or production with clear studio-learning implications); and sufficient extractable information for at least one coded outcome, mechanism, evidence role, or governance construct. Eligible designs included empirical studies, evidence syntheses, and peer-reviewed theoretical papers with a clear educational or governance contribution. Records were excluded when they fell outside the date window, were duplicate returns, were dissertations, repository-only items, book chapters outside the journal/conference criterion, preprint-only records, commercial or promotional sources, had unclear publication origin, or were purely technical studies without educational or studio-learning relevance. Two further records were excluded because the retrieved record gave too little methodological detail for coding, and one mini-review was excluded for its lower methodological weight, as its topic was already covered by the retained studies. Opinion pieces and commentaries could be retained for conceptual, ethical, cultural, or policy interpretation, but they were not used for causal or study-frequency claims.

Directness was evaluated based on the question that the review was based on. A study in a classroom provided direct evidence of educational interaction and adjacent evidence for calligraphy and painting. A study of artificial-intelligence art provided direct evidence regarding authorship or audience judgment, and indirect evidence regarding student learning. A self-determination study might help to explain motivation and agency, but it would not be a technical skill. This multi-layered interpretation is vital in emerging fields because of the danger of assuming that technological capability, user acceptance and learning are all the same thing. The directness code is reported at study level in Supplementary File S1 and is used below to separate direct or adjacent evidence from author interpretation. Table 2 summarizes the evidence domains, the contribution each makes to the review, and the limits on how it can be used.

Table 2. Review domains, evidence roles, and limits.

Evidence domain	Primary contribution	Typical evidence	Limit for this review
Artificial intelligence in education	Learning mechanisms, feedback, engagement, agency, and implementation	Experimental studies, surveys, and reviews	Most studies concern higher education and text-rich tasks rather than studio disciplines

Human–AI co-creation and AI art	Authorship, creative contribution, aesthetic judgment, and disclosure	Creative experiments, audience studies, and conceptual analyses	Product judgments do not directly measure student learning
Visual-art technologies	Capabilities of image analysis and generation	Technical reviews and system studies	Model performance cannot establish educational effectiveness
Learning theory and governance	Autonomy, competence, self-regulation, literacy, integrity, access, and policy	Theory reviews, policy frameworks, and empirical studies	Mechanisms require discipline-specific testing in calligraphy and painting

Table 2 shows that no single evidence domain answers the review question on its own. Research on artificial intelligence in education explains learning mechanisms but rarely involves studio work, while studies of AI art and visual-art technologies describe capability or audience judgement without measuring student learning. For this reason, evidence from each domain was weighted by its directness rather than pooled.

2.4. Outcome and mechanism coding

In the synthesis, the learning outcomes were clustered into six learning outcome families. Technical skill was considered an outcome family and was conceptualised as ensuring that the actions carried out within the discipline were executed in a controlled manner. The conceptual learning included terminology, historical understanding and explanation of the visual decisions. Creative development included the areas of ideational breadth, purposeful transformation, personal style and originality. Self-regulated learning included planning, monitoring, and using feedback, as well as independent completion. Affective experience included engagement, confidence, cognitive load, and perceived agency. Delayed retention and transfer to a new task was included in durable learning with no assistance from the model.

Mechanisms were coded as based on the actions actually performed by the learner and the teacher. The use of models was categorized as: explanation; dialogic questioning; comparison; critique preparation; practice planning; idea generation; reflection; assessment support. There was a three-level procedural code to differentiate between assistance and substitution: 0 = advisory support, the system suggests an idea while the learner assesses and modifies the suggestion; 1 = mixed assistance, the system significantly narrows down the learner's choice but the learner still tests and modifies it; and 2 = substitution, the system presents the idea, judgment, or product without any material change by the learner. This is not a psychometric scale validated for use. Evidence directness was also rated as D2 (direct evidence of the learner's outcome in a relevant studio or educational experience); D1 (direct evidence of the learner's outcome in an adjacent studio or educational experience); D0 (conceptual/technical evidence only). Design Type and Evidence Type, Primary Construct and Primary Outcome, Duration/Temporal Depth, and Teacher Mediation were recorded as separate types of data. These fields constitute the coding rubric and explain why technological capability was not considered as an educational effectiveness [13, 14, 20]. Table 3 lists the six outcome families with the indicators for calligraphy and painting and the kind of evidence that would count as strong.

Table 3. Outcome families and observable evidence for studio learning.

Outcome family	Calligraphy indicators	Painting indicators	Strong evidence
Technical skill	Stroke execution, spacing, structure, rhythm, and brush control	Composition, material technique, color or ink control, and spatial organization	Blinded ratings with discipline-specific rubrics
Conceptual learning	Explanation of style, structure, terminology,	Explanation of composition, medium,	Oral or written explanation plus transfer questions

	and exemplar differences	visual reference, and intended meaning	
Creative development	Purposeful variation within a tradition and personal interpretation	Ideational breadth, transformation, expressive coherence, and personal voice	Process portfolio, rejected alternatives, and independent artifact
Self-regulated learning	Practice planning, error monitoring, and independent repetition	Goal setting, critique use, and independent completion	Interaction records, reflection, and an unassisted task
Affective experience	Engagement, confidence, cognitive load, and perceived agency	Engagement, confidence, cognitive load, and perceived ownership	Validated scales at planned time points
Durable learning	Delayed independent writing and transfer to unfamiliar characters	Delayed independent work and transfer to an unfamiliar subject or medium	Delayed assessment completed without model access

Table 3 sets out what would count as strong evidence for each outcome family. In every row, strong evidence requires either independent performance or assessment that does not depend on the model being present. This standard is applied to the included studies in Section 3 and explains why most current findings are graded as adjacent rather than direct.

2.5. Appraisal and synthesis

The quality of the evidence was judged using the following criteria: design strength, outcome clarity, directness and temporal depth. Results of controlled (and quasi-experimental) research were more highly weighted when claims were made about effects. The assessment of the durable learning effect was shifted to a greater emphasis on the longitudinal and delayed evaluations. Surveys and acceptance studies were not used to cause change, but were used to inform implementation. Studies of audience response, or co-creative by the professionals, informed authorship and creativity, but were not included in the classroom tests. Convergence, limitations and gaps in measurement were identified through the review process [11, 25, 33, 34].

Analysis was done in four stages. Each of the sources retained was coded for year, evidence domain, design type, AI modality, participant context, learning outcome, measurement type, AI duration, teacher mediation, cultural relevance, evidence directness, scaffolding/substitution code, principal finding, and limitation. Second, the findings were grouped by outcome family and related to the four research questions. Third, proposed benefits were accompanied by counter-risks and alternative explanations. Fourth, implications for the educational or studio environment were translated from the evidence when they were supported by at least adjacent educational or studio evidence. A staged logic that connects model performance with educational effectiveness: Technical capability is considered the first link, the interaction between teacher and students is the second, the role of the students is the third, and the independent performance or transfer of the students is the fourth. In the remainder of the links, a capability claim is not sufficient to be interpreted as an educational-effect claim without evidence. The following study-level fields include the duration/temporal depth, teacher mediation, cultural relevance, directness, scaffolding/substitution status and appraisal limitation: Supplementary File S1: Extraction_Matrix (L-Q). If the retrieved report did not contain a field, it is not inferred as not stated.

Due to the different subject matter, student groups, model types, comparison conditions, length and outcome definitions of the retained studies, a pooled effect size was not calculated. Weight of evidence was therefore reported qualitatively as being stronger, moderate, mixed, limited, or weak based on directness of the evidence, strength of the design, the measurement of the outcome, and the age of the evidence. Acceptance or conceptual papers carried the least weight for effect claims, whereas controlled or quasi-experimental studies with independent performance measures carried greater weight. Table 3 contains references to blinded

artifact scoring, but it is a recommended evidence standard rather than a fact that is already evidenced in the literature in the LLM–calligraphy or LLM–painting discipline. Author-developed teaching or framework propositions are identified as such and are not assigned empirical effect strength. Figure 1 summarises this evidence architecture: retrieval and screening of the 17-study core, coding of outcomes and mechanisms, separate treatment of contextual references, and the translation of the synthesis into the CRAFT-LM framework.

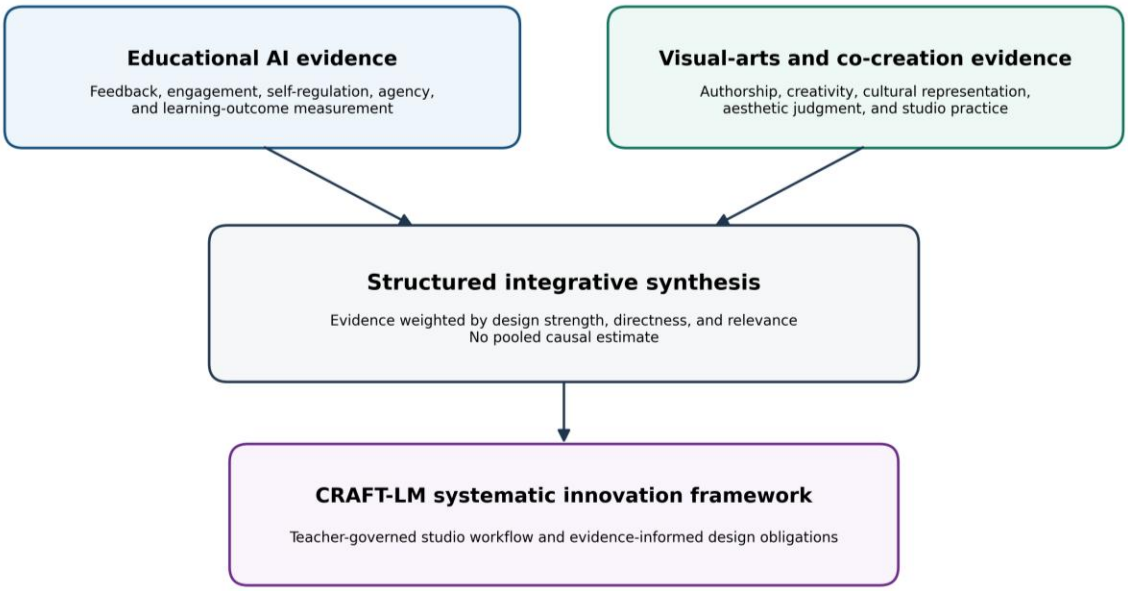


Figure 1. Evidence architecture for the structured integrative review and development of the CRAFT-LM framework.

3. Results

3.1. Learning outcomes and evidence strength (RQ1)

The 17-study auditable core also shows a clear recent concentration of publications: 2019 = 1, 2020 = 0, 2021 = 0, 2022 = 2, 2023 = 3, 2024 = 4, and 2025 = 7. The pattern indicates rapid growth in the evidence base after 2022, while also showing that the field remains recent rather than longitudinally mature. The same counts are provided in Supplementary File S1, Publication_Trend.

Throughout the Results, the included-study count remains n = 17. Citations represented in the S1 Extraction_Matrix are the analytic-core studies and are the only sources used for numerical study counts and coded evidence-frequency statements. Citations that appear in the Results but are not part of that 17-study matrix are contextual or theory/policy sources used to interpret mechanisms, boundary conditions, or implications; they are not additional included studies and do not change n = 17. Their provenance role is recorded in S1, Provenance_Notes.

The general educational research literature contains mixed support for LLM-supported learning. The advantages of the system have been reported in providing timely explanation, targeted feedback, or structured interaction [16, 17, 35]. Furthermore, reviews also suggest that benefits are more likely to be realized when LLM use is guided by the task and the teacher rather than limited to requesting answers [24, 34]. But engagement and satisfaction are subjectively reported and exposure durations are brief and conditions are variable. The best defence is that the model can enhance access to feedback and opportunities for reflection, and that the quality of learning is related to the amount of reasoning that stays in the learners' hands.

The second condition added by the visual-arts layer is that output quality and artistic learning are not the same. While AI systems can produce a finished product and a believable critique, the sophistication of the appearance isn't sufficient to indicate student intent. Philosophical and empirical approaches to AI art break down the difference between formal novelty and authorship, meaning and perceived authenticity [36, 37, 38]. A characteristic of human–AI co-creation research is the setting up of AI outputs, choosing them, editing them,

adapting them, and discarding them as results [39, 40]. These actions may be educationally useful but must be made apparent in the process.

The most consistent benefits reported in the literature are for proximal learning processes. Pupils have immediate explanations, no social cost of asking the same question again, have the ability to compare options and are prompted to reflect. These affordances could lead to more and earlier formative interaction. They do not ensure they provide accurate feedback nor do they ensure that students act upon this feedback productively. The research on generative feedback and tutoring demonstrates that results are tied to the design of the task, the learner's knowledge, and the verification [16, 17, 35]. There is, therefore, more evidence supporting a mechanism claim than a universal effect claim. A recent art-education synthesis similarly reports promising but early-stage effects across learning, creativity, engagement, and cultural understanding, with outcomes dependent on structured pedagogy and instructor design [31].

It's also more straightforward to show immediate performance than durable learning. If a model is present, a student can achieve the same objective in a better way, or create a more aesthetically pleasing artifact. Such improvement may be a result of actual learning or of temporary support or direct substitution. There are tasks that can only be transferred later, which requires a delayed transfer task. There is a recurrent pattern of short exposure time, self-reported outcomes, and inconsistent conditions of comparison when evaluating student engagement and educational AI, as noted below. The following list summarizes recurrent concerns when reviewing student engagement and educational AI: short exposure time, self-reported outcomes, and inconsistent conditions of comparisons. These weaknesses are particularly significant in studio education where control and visual judgment are acquired through practice. Overall, support for proximal educational processes is moderate and mainly adjacent (D1), whereas support for durable, independent studio learning is weak; the latter is not established by a direct LLM-calligraphy or LLM-painting delayed-transfer study.

Table 4 places the benefits most often claimed for LLM use beside the evidence currently available for each of them.

Table 4. Claimed benefits of LLM use and their current empirical status.

Claimed benefit	What the evidence shows	Directness	Status
Faster access to explanation and feedback	Reported in tutoring and feedback studies [16, 17, 35]	D1	Supported for proximal processes only
Higher engagement and confidence	Largely self-reported [22, 34]	D1	Supported as perception, not as learning
Broader idea generation	Individual output enriched while group diversity falls [19, 41]	D0-D1	Supported, with a convergence trade-off
Improved technical skill in stroke and composition	No LLM-calligraphy or LLM-painting trial; the closest calligraphy study tested a computer-vision system rather than an LLM [42]	D2 for a non-LLM system; none for LLMs	Not established
Retention and independent performance	Trials in other subjects show that assisted gains can disappear once support is removed [43, 44]	D1	Not established; counter-evidence exists
Cultural understanding	Early-stage effects reported in an art-education synthesis [31]	D2 for art education generally	Plausible; untested in calligraphy and painting

Table 4 separates what LLMs are expected to offer from what has been shown. Only proximal processes, such as quicker access to explanation and feedback, have consistent empirical support, and even this comes from adjacent educational settings rather than studio classes. Gains in technical skill, retention and

independent performance remain unestablished for calligraphy and painting, and evidence from other subjects indicates that performance achieved with assistance should not be read as learning.

3.2. Conceptual learning and feedback access (RQ1)

Conversational explanation is well suited for large language models. They can remove the meaning of words, contrast different versions of it, translate technical jargon, and dissect a broad question into a series of more specific ones. In calligraphy, it can help identify structural balance or simple symmetry or relationships between pressure, speed, and spacing. In painting, it can assist them to express their intention when they are making a painting, the relationship between colours, the significance of a visual reference etc. These uses can help to remove participation barriers and to open up critique language.

It is theoretically better to use dialogic interaction when it involves the learner making an inspection and a prediction, as well as comparison and justification, rather than simply accepting a declarative judgment. It is not the reviewer's opinion that dialogue is always better for manual painting skills. It is procedural value: If a question is set like 'Which edge has the focal contrast and what evidence in the painting would lead one to that judgement?' then the learner retains the process of making the judgement and could accept the verdict without making inspection. Explainable-AI research also focuses on the aspects of usability, contestability, and learner choice [14]. In a studio environment, the language transfer is pedagogically valuable only if it focuses on the artefact and provides an opportunity for the learner to make something himself. The Painting Dialogues method's surrounding evidence is consistent with this procedural account: shared painting, joint analysis, and reflection were a learning process used to support expression, sensorial engagement, and reflective communication [45]. This is not a direct comparison of which is better, LLM dialogue or declarative instruction when teaching manual painting; it's evidence for dialogic reflection in manual painting. Accordingly, the evidence for conceptual explanation and feedback access is moderate and mainly D1; the studio questioning example above is an interpretive teaching application, not a direct comparative result.

3.3. Embodied and visual mechanisms (RQ2)

Technical mastery evidence is less than conceptual support. The calligraphy needs to coordinate the movement, timing, pressure and sensory input while painting, etc. These actions can be linguistically described but must be learnt through the coupling with perception and by repeated motor calibration. Embodied cognition research shows that action and manipulation of the body can be a part of the learning process, rather than merely a means of delivering it [28]. Even with multimodal systems, the analysis of composition, colour, contour and selected formal features is available, but there is pedagogically validated and style-sensitive diagnosis of embodied studio technique that is limited. Technical recognition/technical description, cultural interpretation/educational judgment should thus not be understood as one and the same, but rather as distinct functions [12, 46]. The visual and kinesthetic art-learning observations further suggest that bodily engagement and material interaction are not an alternative way of delivery, but rather become a part of learning [5], while gestural-painting analysis provides neighbouring observations that the information conveyed by movement and material traces cannot be fully verbalised [6].

Defensible calligraphy sequence can be proposed from the synthesis but it is an author-developed design proposition instead of an empirically proven protocol. A learner could start by comparing two models of a heng-to-zhe transition produced by different teachers and make a prediction about where the pressure and direction of the turn should be placed, practise some controlled repetitions and then compare the result with the other two models and discuss the resulting corner shape and rhythm. The model can explain the terminology or elicit what has changed in the movement, the modeler (learner) executes the movement and the teacher checks to see if the correction is correct for the script and stage of learning. Later unassisted writing is more predictive of learning than is recognition accuracy or verbal recall.

In the field of painting, for example, model support might include composition, medium, edge control, tonal organization or color relationships, but the system doesn't involve the physical material. The feedback should thus be considered as a prediction that can be tested. The learner will be able to make two quick sketches and then judge if there is a more successful relationship between positive and negative space in the model and

which of the two sketches is more successful in establishing the desired relationship. If the change in glaze is suggested, it is the learner's responsibility to change the glaze, observing, accepting, rejecting, and/or adjusting the material change to create a cooler glaze which will provide less visual weight. By doing this it maintains the learner as the person who creates art from ink, oil, watercolor or pixels. Evidence for embodied technical mastery is limited: the supporting work is primarily D0-D1, and the calligraphy and painting teaching sequences described here should be read as author-developed applications rather than tested LLM protocols.

3.4. Creativity, authorship, and aesthetic diversity (RQ1/RQ3)

Thirdly, it's creativity. Generative systems can expand the spectrum of ideas by providing alternatives from which the learner can compare or manipulate. The advantage is maximized in the case of provisional production. Doshi and Hauser [19] found that generative support can enhance the creativity of individual users while decreasing the diversity of output for users. In art education, this is the esoterica of the same “style” of art but a more elaborate art class. Furthermore, in design applications of artificial intelligence, dependence and evolution are important issues brought up by Wu et al. [47]. Accordingly, there is a need to praise and appreciate differences, localisation, and material change, and efficiency in the pedagogical response.

AI-Art research also reveals that creativity is spread among framing, promoting the ideas, picking, editing, discarding and reshaping the outputs. The student may however, overestimate the contribution of the model to the text; when formal novelty emerges suddenly after a prompt, and the system's easy explanation creates the impression that the result is the student's own creation. This attribution can muddy the learner's own framing and selection decisions, or, at the other end of the spectrum, mask the extent to which the learner's own framing and selection came from the model [36, 37, 38, 39, 40]. In this case, process documentation is essential, if writing is part of the learning outcome.

Aesthetic convergence is particularly important in the case of painting. Doshi and Hauser [19] found that generative support can boost creativity scores at an individual level while simultaneously lowering the diversity within a group. Convergence in the classroom might be measured as a visual-embedding similarity pairwise, palette/compositional feature similarity and blinded qualitative coding of recurring stylistic motifs across the group. A convincing analysis should seek for less "between-student dispersion" following the help and not assess convergence through a single polished product. This tendency can be counteracted by assignments that demand local references, a transformation of material, the rejection of alternatives and an explanation of how the end product diverges from the suggestions made in the model. Evidence for creative-development benefits is mixed. The diversity and authorship findings are drawn largely from co-creation or AI-art studies and therefore provide adjacent rather than direct evidence of student learning in calligraphy or painting.

3.5. Cultural, trust, and governance conditions (RQ3)

Representation of culture entered into the system at three levels. During training, model-level data and defaults will affect the types of examples and visual notations generated. At the level of interaction, the prompts, the sources of retrieval, the teacher-approved exemplars and the image inputs decide the aspects of the model that are supposed to be noticed in the interaction. The utilization of the output is institutionally decided based on the curriculum, assessment, procurement, disclosure, access and privacy rules. Calligraphic feedback might be an over-emphasis on symmetry or standardization or a lack of recognition for style-specific variation; in painting it might be a regurgitation of commercialized motifs or culturally thin descriptions. Thus, cultural grounding is a necessity and not an optional enrichment to the system [15, 48].

Cultural authenticity raises a further problem that accuracy checks alone do not solve. A response can describe a script or a painting convention correctly and still treat a living practice as a set of visual features detached from its lineage of teachers, exemplars and permitted variation. Recent work on an AI-powered robotic system for calligraphy creation [49] shows how far the technical reproduction of calligraphic form has advanced, yet reproduction of form does not settle whether a work is authentic within a tradition; that judgement remains with practitioners. Generative image models also tend to fall back on dominant visual stereotypes when a cultural context is underrepresented [15]. The authenticity question for calligraphy and

painting education is therefore whether model output supports learning within a tradition or replaces the standards of that tradition with generic ones.

Trust is not a given. Automation bias, authority bias, anchoring on the first possible explanation, confirmation bias, and a fluency heuristic (where ease of reading is interpreted as being correct) can arise from fluency responses. These risks are heightened for students whose prior knowledge is lower since they lack the schemas for drawing comparisons to an historical attribution, stylistic rule, or technical claim. The results of the Educational over-reliance reviews show that the use of AI recommendations without critical thinking results in a decrease in analytical thinking and decision making [50]. Thus, for a student of calligraphy, AI literacy involves the awareness of the existence of a cited masterpiece or historical reference, the authenticity of an exemplar, uncertainty, the comparison of claims with teacher-approved sources and the awareness of lack of confidence in the use of language without evidence [51]. Studies also show that the trust of the learner is affected by perceived technological performance, user experience and individual differences; therefore, relying on an AI-generated response should not be treated as a substitute for accuracy, and it is not guaranteed that the level of trust of the response is equivalent to the level of its accuracy [52]. The literature on cognitive ease also shows that fluent explanations may lead to the illusion of competence before real competence [53].

The reliability of AI-generated feedback has to be judged separately from its fluency. In a randomised study of writing, learners supported by ChatGPT produced higher-scoring essays without corresponding gains in knowledge or transfer, a pattern the authors described as metacognitive laziness [44]. Arts teachers themselves name misinformation and copying among their main concerns about the classroom use of ChatGPT [54]. For studio subjects, feedback should therefore be treated as reliable only when it has been checked against approved exemplars and confirmed through the learner's own making.

There's also the impact on the reception. The awareness of assistive Artificial Intelligence has been demonstrated to impact perceptions of authenticity and creativity through audience perception studies [38, 55]. This does not mean the work performed by the “assisted worker” is not valuable work. It connects to the concept of making contributions “visible.” Students should document the system they used, the reason for engaging in the interaction, the outputs they considered, and changes that they made. This sort of documentation can help to ensure integrity is maintained and teachers have more evidence of learning.

Authorship and stylistic convergence are closely linked. When many learners draw on the same model, the risk is not only that individual work becomes derivative but that a whole class drifts toward similar solutions. Anderson et al. [56] found that ideas produced with LLM assistance were more alike across users than ideas produced without it, and Doshi and Hauser [19] reported reduced collective diversity even where individual creativity ratings rose. A large-scale study of artists who adopted text-to-image tools found higher productivity alongside a decline in the average novelty of their work over time [41]. In a painting class, such convergence becomes visible only across the group, which is why authorship records and group-level comparison belong together.

A real challenge is disclosure fatigue, that is, multiple AI tools in the same task. If you record all of the prompts, you may end up with attribution confusion and documentation becoming paperwork. One activity log that is structured and includes the following is a more useful classroom rule: the tool used, the stage of the task, the purpose of use, suggestion accepted/rejected, and material/concept transformation of the learner. This will allow the author's name to be maintained intact whilst not counting everything that happens at the lowest level as a separate authorship event. Evidence for trust and over-reliance mechanisms is moderate, while cultural and governance conclusions rely more heavily on contextual or adjacent evidence; the classroom rules proposed here are interpretive implications rather than tested causal effects.

3.6. Calligraphy-specific integration (RQ2)

There are four possible uses of the calligraphy synthesis. The first step an LLM can take is to define terms and talk about teacher-endorsed exemplars. Second, it can pose questions about the structure, spacing, rhythm and style. Third, it can be used to transform the bounded observations observed by a task-validated component of computer vision into instructions that can be read by the learner. All are related components: video-based stroke-trajectory tracking, stroke segmentation, contour/stroke comparison, writing-action analysis using a

sensor, CNN/LSTM or vision-transformer classifiers specialized for specific formal features [2, 3, 42]. These modules are only able to certify a very limited technical skill such as trajectory or stroke recognition, but not artistic skill, brush vitality, fidelity to a culture or learning itself. Finally, the LLM may help students to organically plan the practice and record refinements. Technical recognition and the artistic evaluation should be separated.

The model steps onto the stage for a realistic calligraphy exercise to start. Students' understanding of script, exemplar, structure and quality is desired. An independent baseline sample is used to determine performance at the present time. The model can then ask questions and provide terms or lead in the comparison of approved references. The student carries out tests with help and practice. The movement quality, the degree of image fidelity and the energy for expression are no longer measured by the degree of image similarity, but with teacher feedback. The final transfer task will be given using unfamiliar characters, and will not have model support.

The design also ensures that cultural learning is protected. Calligraphy lessons are not limited to a neat appearance of this lettering. It contains line, variation, disciplined copying and the 'growing up' of personal interpretation. A model can give or form the context, but it is the teacher's choice which is valid and in which way it is linked with the style being practised. The goal is to break away from the idea of creating a standardized form and instead concentrate on the skill of observation, execution, comparison and reasons for the learner. [4]. Calligraphy-specific evidence remains limited for LLM-supported learning: embodied and computer-vision studies provide adjacent support, while the instructional sequence in this subsection is author-developed.

3.7. Painting-specific integration (RQ2)

The broader ideational role is identified by the painting-specific synthesis. With LLM, students are able to formulate a theme, think about alternative compositional choices, visualize a comparison in historic context, generate questions to critique compositions, and think about revision. Feedback may be given in terms of colour relations or spatial organization in multimodal systems as well, but the feedback has to be treated as a hypothesis. In spite of all this, there will ultimately be a formal decision, which must be based on the intent, medium and context and which will still have to be made by humans. Research on the effects of AI on art has uncovered that individuals' perception and appreciation of AI-generated artworks are changing as they are made aware of its AI origins, and that their perception of created art could also shift when they use AI to help create artworks [55, 57]. These findings are additional support for explicit author and contribution comments.

The best thing to do is extend the question, not to provide the answer. A model can make divergent compositional suggestions, it can suggest changes to a palette and it can ask questions for discussion. The student then draws out ideas, experiments materials and chooses a direction that has significance. There is no time to spare the details and go straight to the pretty picture. Suggestions must be made and the learner should be able to explain what was used, why it wasn't used and what was changed. This is a generative process like studio-based art education, where successive prompts and transformations are used rather than a prompt-to-product process [58]. In art education, using AI-generated images as tools to aid learning, for example, has shown that students' immersion, imagination, and expression can be improved when the technology is used in the classroom within a pedagogical framework [59]; while in the analysis of image-based AIGC in art education, it has been found that technical proficiency does not always correspond to educational efficacy [60].

This can be used for traditional and contemporary painting. Culture and sensitivity towards materials are characteristics of ink painting. The limitations can be of medium, location, social context or experience, etc. The framework thus does not stipulate an aesthetic. It establishes a relationship: The output of the model is provisional, the disciplinary judgement is still localised and the learner still responsible for the output. Evidence for LLM-supported painting learning is mixed to limited: art-education and AIGC studies support plausible mechanisms, but durable, independent painting learning has not been directly established.

3.8. Teacher mediation as the stable enabling condition (RQ3)

In both disciplines, teacher mediation was the most consistent enabling condition. The teacher sets the learning target, selects legitimate exemplars, determines what cannot be outsourced, checks cultural and technical claims, and decides what feedback suits the learner's stage. Curriculum values that cannot be delegated to a general model include the historical lineage being taught, the relationship between disciplined copying and personal interpretation, the educational meaning assigned to brush vitality or restraint, culturally appropriate symbolism, the standard of evidence for an art-historical claim, and the course's expectations for authorship. The LLM can increase the frequency of low-stakes interaction, but it does not determine these values. Policy frameworks likewise require literacy, assessment, integrity, and governance to be coordinated [29, 61, 62].

Teacher mediation isn't just about fixing factual mistakes. It arranges the sequence of help, safeguards the productive struggle, chooses culturally relevant materials and determines when a demonstration or more language is needed. It also establishes a common ground for disclosure and privacy. In their absence, this system might reward speed and polish, rather than the capacities that the course is designed to develop. Policy and integrity frameworks also illustrate the need for the design of technology, assessment and governance to be co-developed [29, 61, 63]. Teacher mediation is supported as a moderate, cross-source enabling condition, mainly from adjacent educational and policy evidence; it has not been isolated as a causal effect in a direct LLM-calligraphy or LLM-painting trial. Table 5 brings these findings together by learning outcome.

Table 5. Integrative evidence synthesis by learning outcome.

Outcome	Most plausible benefit	Main counter-risk	Strength of current support	Implication for teaching
Conceptual learning	Rapid explanation, comparison, vocabulary, and contextual prompts	Plausible but inaccurate or culturally thin explanation	Moderate, mainly from adjacent educational evidence	Use approved references and require evidence-based explanation
Technical skill	Practice planning and targeted attention to formal features	Verbal understanding mistaken for embodied mastery	Limited for calligraphy and painting	Pair every explanation with making and independent performance
Creative development	More alternatives, reframing, and externalization of ideas	Aesthetic convergence, imitation, and weakened authorship	Mixed	Reward transformation, divergence, and process documentation
Self-regulated learning	Goal setting, monitoring, comparison, and reflection	Dependence on immediate guidance and reduced productive struggle	Mixed	Delay assistance and require learner decisions
Affective experience	Lower threshold for questions and increased confidence	Novelty effects, frustration, or misplaced confidence	Moderate but often self-reported	Measure cognitive load and agency, not satisfaction alone

Durable learning	Potential support through repeated guided practice	Assisted performance disappears when support is removed	Weak	Use delayed, unassisted transfer tasks

Table 5 shows that the strength of support falls as outcomes move from proximal to durable. Conceptual learning and affective experience have moderate support, whereas technical skill and durable learning have limited or weak support. Each plausible benefit is paired with a specific counter-risk, which is why the teaching implications in the final column concentrate on making independent work and learner decisions visible.

3.9. Practical skill, originality, retention, and independent learning (RQ1)

Four outcomes matter most for studio education, and the evidence for each is weaker than the evidence for proximal support. For practical skill, no included study tested whether LLM-supported instruction improves brush control, stroke execution or material technique; verbal guidance can describe a correction, but motor calibration still depends on repeated physical attempts [2, 28]. For originality, generative support raised the rated quality of individual work while narrowing collective diversity [19], and artists who adopted text-to-image tools became more productive while the average novelty of their content declined [41].

For retention, a field experiment in high-school mathematics found that unrestricted GPT-4 access improved performance during practice but lowered later examination performance once access was removed, whereas a tutor version with built-in learning safeguards largely avoided this loss [43]. For independent learning, ChatGPT support in a randomised writing study improved the quality of the submitted product without improving knowledge gain or transfer [44]. In calligraphy, the computer-vision teaching system studied by Zhang and Xu [42] reported retention outcomes, but it was not an LLM condition and assessor blinding was not stated. None of the LLM studies concerns calligraphy or painting, yet together they show why assisted output should not be taken as evidence of skill, originality or durable learning in studio subjects. Evidence on these four outcomes is therefore weak for LLM-supported calligraphy and painting, and the most informative findings come from other subjects.

3.10. Evidence gaps: durability, directness, and transfer (RQ4)

The clearest evidence gap is the absence of direct, discipline-specific tests that separate assisted output from independent learning. In the 17-study auditable core, none reported a blinded rating of a delayed, independent calligraphy or painting artifact after an LLM-supported intervention. The calligraphy computer-vision experiment by Zhang and Xu [42] reported performance and retention outcomes, but the retrieved record did not state that artifact assessors were blinded and the intervention was not an LLM condition. The art-education studies and reviews likewise did not provide a directly comparable blinded LLM–calligraphy or LLM–painting endpoint. Thus, the blinded-rating result of the review is not a positive effect estimate; it is a documented zero for the specific design feature needed to support durable studio-learning claims. This coding is shown study by study in Supplementary File S1.

The same gap limits conclusions about dialogic feedback, autonomy, and aesthetic diversity. Existing studies justify these mechanisms as plausible and educationally relevant, but the causal chain from model interaction to later independent studio performance has not been established. The Results therefore end with the boundary of the evidence; the CRAFT-LM framework is presented in the Discussion as an author-proposed, evidence-informed model rather than as an empirical result. The evidence strength for durable, direct LLM-studio learning is therefore weak; the documented absence of delayed blinded endpoints defines an evidence gap rather than a negative effect estimate.

4. Discussion

The main conclusion is conditional. LLM integration may be useful in calligraphy and painting education when it extends feedback and reflection without replacing the learner's intention, judgment, and making. This claim is narrower than the general claim that generative artificial intelligence supports learning, because it keeps physical studio practice, perceptual discrimination, cultural awareness, and authorship at the center of

studio training. Evidence is strongest for proximal benefits such as explanation, engagement, support for reworking drafts, and access to examples. Evidence is weakest for delayed transfer, independent performance, and long-term artistic development.

This understanding is in line with the concept of self-determination theory. Autonomy, competence, and support [64, 65, 66] will make students more likely to persist in their learning. Competence can be facilitated through an LLM that provides easygoing explanations and non-threatening feedback. May encourage independence as students test their own ideas. Indirectly, it can devalue both needs when it directs the work, gives the idea and/or creates dependence on the moment's approval. The design goal is not to provide the maximum assistance possible. It is supportive in that it leaves the learner responsible for the artistic problem.

The second explanation is from self-regulated learning. Effective learners plan, monitor, evaluate and adjust. LLMs can help to make these processes more visible by helping students establish goals, make comparisons, explain revisions, look for evidence, etc. They can also assist the student in the processes. Chiu [20] proposes that generative support might be classified according to its relation with self-regulation. This approach is continued here with regard to education in the studio. In a strong interaction, the learner is prompted to make an observation, justify it, and make a decision. A weak interaction has a complete judgement which the learner can only mimic.

The results are also part of the knowledge about the role of multimodality. A text-only LLM is not able to analyze a painting or handwritten character. Inspection doesn't make a disciplinary valid, you can inspect an image using a multimodal model. There are four distinct aspects of technical recognition, formal description, cultural interpretation, and educational judgment. A number of coordinated components are thus required for establishing a reliable system: image preprocessing, feature extraction, an approved knowledge base, a rubric, a conversational model and human verification. It is not enough to have one of these skills to be effective as an educator, as Rombach et al. [46] demonstrate in their research, and Cetinic and She [12] show how flexible and useful AI analysis of art is.

4.1. Interpreting learning in embodied and visual disciplines

The results have shown a significant restriction on verbal intelligence. Modelling can facilitate thinking about an artefact, but studio learning does not dispense with perceptual discrimination and action. Feedback in calligraphy should lead the learner back to the brush, stroke and structure. In painting, it should bring the learner back to the surface, material, composition and intended meaning of the painting. Dialogue's teaching worth comes when it alters the learner's noticing and doing. That's why process and transfer are better than just a finished product.

Preserving embodied correction is the most important design principle in calligraphy because procedural skill develops through repeated perception–action coupling, proprioceptive and tactile feedback, and motor error correction. A student may understand verbally that a stroke is too heavy yet remain unable to regulate pressure. The next step must therefore return to the brush: a concise explanation should be followed by a defined action such as repeating a stroke under a pressure constraint, comparing two exemplars, or correcting structural imbalance. The physical attempt supplies sensory evidence that language alone cannot provide [2, 28]. Improvement should be judged in later independent work.

The biggest threat when painting is aesthetic outsourcing. A model can easily suggest themes, compositions, palettes or stylistic references. This can be useful to a novice to get away from an empty page but can be limiting when working around model defaults. A more powerful alternative brings generative workflows back to physical making. In this way, the best educational schemes cycle between language, generated possibilities, sketches, materials, and critique. The student is not just to select, but to transform.

The concept of 'ideation' versus 'authorship' is at the core of painting. Generative suggestions may be needed when the learner is stuck on the first correct answer, or they may also be too soon. Timing, then, is important for support. The need for a preliminary sketch, written intention or material experiment before model interaction safeguards a record of the learner's own thinking. Later conversations subsequently challenge or extend that thinking, not obliterating its source.

The qualitative weight labels in Table 5 are anchored to the sources cited in the corresponding Results subsections and to the study-level coding in Supplementary File S1. Conceptual learning and feedback draw primarily on educational studies and reviews; technical-skill claims rely on adjacent embodied, computer-vision, and visual-art evidence, with no direct LLM–calligraphy or LLM–painting trial; creative-development claims draw mainly on co-creation and AI-art literature; and self-regulation and affective outcomes are supported mainly by adjacent educational evidence. Durable learning remains the weakest category because no direct LLM–studio study in the retained analytic core combined delayed transfer with blinded artifact scoring. Contextual references inform interpretation but are not counted as additional included studies or study-frequency evidence.

4.2. Proposed CRAFT-LM framework

CRAFT-LM was derived from the synthesis rather than designed in advance. Each recurring condition identified in Section 3, whether it supported learning or created risk, was examined to see whether a design requirement could respond to it. Six such requirements emerged, and Table 6 links each one to the findings, sources and strength of evidence behind it. Because several components rest on adjacent rather than direct evidence, Table 6 also names an indicator through which each component can be tested.

Table 6. Derivation of the CRAFT-LM components from the review findings.

Component	Finding it responds to	Section	Key sources	Evidence strength	Testable indicator
Cultural grounding	Generic or flattened cultural feedback; embodied features missing from static corpora	3.5, 3.6	[2, 4, 15]	Contextual and adjacent	Expert rating of feedback against approved exemplars
Retained agency	Loss of learner agency; stylistic convergence	3.4, 3.5	[13, 19]	Moderate (adjacent)	Share of suggestions rejected or transformed; between-student dispersion
Adaptive feedback	Feedback access is the most consistent proximal benefit	3.1, 3.2	[14, 16, 17]	Moderate (D1)	Teacher-rated specificity and accuracy of feedback
Formative process assessment	Polished products hide the respective contributions of learner and model	3.4, 4.4	[39, 61]	Contextual	Completeness of the process portfolio
Teacher moderation	Most consistent enabling condition across sources	3.8	[29, 61]	Moderate (cross-source)	Rate of corrected model errors; implementation fidelity
Learning monitoring	No delayed, blinded, unassisted endpoints in the analytic core	3.9, 3.10	[42, 43]	Weak (documented gap)	Delayed unassisted transfer score

Table 6 shows that each letter of the acronym corresponds to a specific condition reported in the Results rather than to a general teaching principle. It also makes the uneven evidence base visible: adaptive feedback, retained agency and teacher moderation draw on moderate adjacent evidence, whereas learning monitoring

answers a documented gap. The indicators in the final column are the measures through which the framework can be tested in the comparison design described in Section 4.4.

Figure 2 illustrates an evidence-informed, author-proposed CRAFT-LM model developed through interpretive synthesis of the review findings. It has not yet been empirically validated. The system is culturally grounded in local traditions and approved exemplars. Retained agency means that students make choices and intentions which have consequences. Adaptive feedback offers explanations and questions related to the learner's current work. Formative process assessment appreciates drafts, trials, rejections and reflection. Teacher moderation verifies disciplinary and cultural quality. Learning monitoring assesses what is left after help is taken away. The framework is thus a proposal for design testing, not a result of the literature review.

The acronym describes a set of design obligations rather than six independent features. Cultural grounding shapes the knowledge and examples from which feedback is drawn. Retained agency protects the learner's intention and consequential choices. Adaptive feedback matches the learner's current work while disclosing uncertainty. Formative process assessment makes trials, decisions, and rejected alternatives visible. Teacher moderation verifies disciplinary and cultural quality. Learning monitoring tests what remains after assistance is removed. The center of the framework is learner-led practice, not model output.

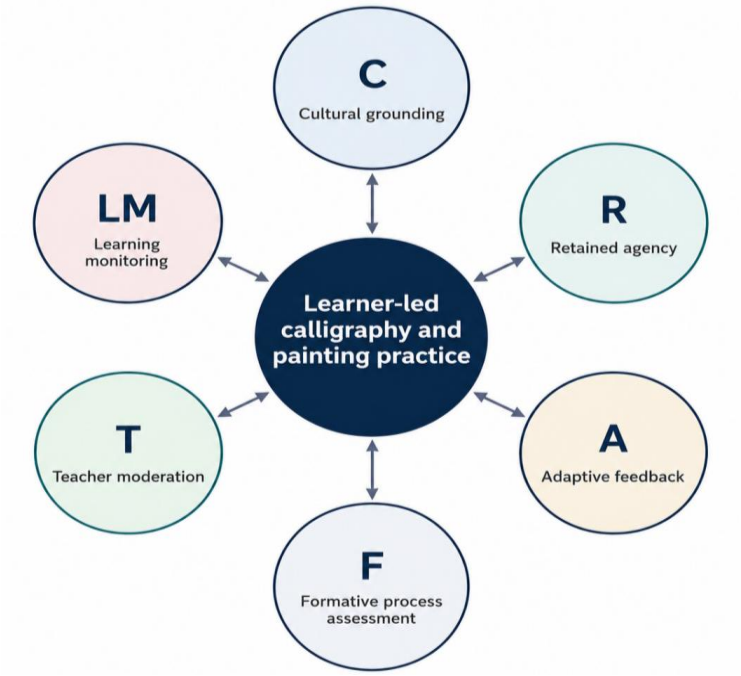


Figure 2. Proposed CRAFT-LM framework for learner-led, culturally grounded, and teacher-moderated integration.

Figure 3 translates the framework into a studio learning cycle. The sequence begins with intention and cultural reference, followed by an independent baseline task. LLM guidance enters only after the learner has produced evidence of current understanding. The student then tests feedback through physical or digital making. Teacher critique verifies the guidance and helps distinguish productive experimentation from misunderstanding. Refinement and reflection lead to an independent transfer task. This sequence protects the role of practice and makes the contribution of the model observable.

As an illustrative classroom translation rather than a research-evaluation protocol, the six-stage loop can fit a 90-minute studio class as follows: 0-10 minutes for intention and cultural reference, 10-20 minutes for an independent baseline attempt, 20-35 minutes for model-guided questions or comparison, 35-60 minutes for material making and physical testing, 60-75 minutes for teacher critique and verification, and 75-90 minutes for refinement, reflection, and a brief statement of what should transfer to the next unassisted task. The timing is adaptable to medium and learner level; its purpose is to preserve a return from language to making within the same class period.

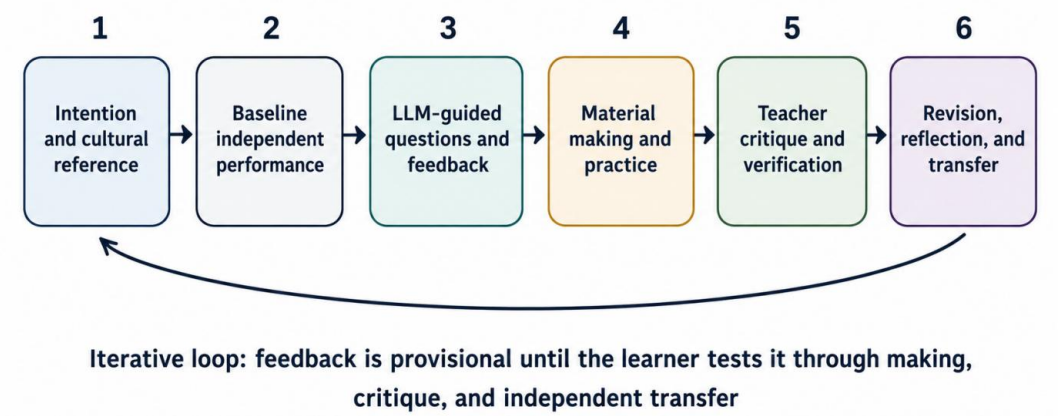


Figure 3. Studio learning cycle that places model support between independent baseline performance and teacher-verified refinement.

The outcome pathway is shown in Figure 4. LLM support first affects proximal processes such as attention, monitoring, comparison, and refinement. These processes may influence disciplinary learning, including stroke control, composition, technique, and visual reasoning. Durable outcomes require transfer, retention, and agency after support is removed. The pathway is moderated by task design, teacher mediation, artificial-intelligence literacy, cultural fit, and access. This model explains why the same technology can support one class and hinder another.

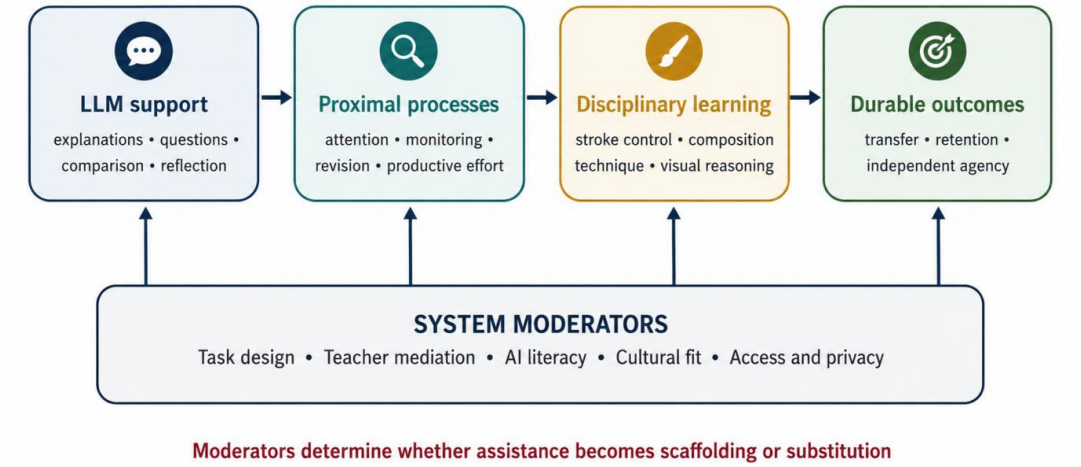


Figure 4. Pathway from model support to proximal processes, disciplinary learning, and durable outcomes under system-level moderators.

Together, the figures show why implementation cannot be reduced to access. The model affects learning through attention, monitoring, comparison, and productive effort. Those processes may contribute to disciplinary learning, but only when task design, teacher mediation, literacy, cultural fit, access, and privacy are adequate. The system may otherwise shift from scaffolding to substitution. This conditional pathway explains why similar tools can produce different outcomes across classrooms. Table 7 sets out the enabling and risk conditions for each element of the CRAFT-LM system.

Table 7. Enabling and risk conditions within the CRAFT-LM system.

System element	Enabling condition	Risk condition	Expected educational consequence
Task design	Model enters after intention and baseline work	Prompt-to-product task with no independent attempt	Reflection and learning versus substitution

Feedback	Specific questions, alternatives, evidence, and uncertainty	Authoritative scoring or generic praise	Improved noticing versus false confidence
Student agency	Selection, rejection, transformation, and explanation are required	First plausible suggestion is accepted	Ownership versus dependency
Teacher role	Verification, demonstration, critique, and cultural framing	Tool access without disciplinary mediation	Valid scaffolding versus misinformation
Assessment	Process-weighted rubric plus delayed independent task	Surface polish is rewarded alone	Learning evidence versus assisted performance
Culture	Local exemplars, terminology, bias checks, and plural traditions	Model defaults are treated as neutral standards	Cultural fit versus flattening
Access and governance	Equivalent access, disclosure, privacy rules, and an alternative route	Unequal subscriptions and unclear data use	Fair participation versus hidden advantage

Table 7 shows that each element of the system can either support learning or lead to substitution, depending on how it is set up. The same model can therefore have opposite effects in two classrooms. The table also gives each CRAFT-LM component an observable condition that can be checked when the framework is put into practice.

4.3. Theoretical contribution

This study is a continuation of the self-determination theory (SDT) and contains information about how autonomy and competence may be expressed in decision making within a studio. Autonomy is not the right to use a tool, but instead the constant responsibility of the learner, with regard to his/her intention, selection, transformation and judgment. Competence is more the control in independent work rather than the interfacing fluency. Other related factors also have an impact on how feedback is interpreted, which affects teacher feedback, critique, and cultural membership [64, 65, 67, 68]. The perception and behavior of learners in an applied studio setting can be used for assessing the autonomous nature of the retained. The validated autonomy/agency scale can be used as a measure of the degree of perceived choice, and the percentage of decisions made by the learner, percentage of model suggestions that the learner refuses to use as is or significantly changes, and quality of justification for model suggestions the learner retains can be measured by the performance on a subsequent unassisted task. These indicators are reported separately and CRAFT-LM does not validate a new composite autonomy score.

The contribution of self-regulated learning to learning is also clearly defined. LLMs can collaborate with planning and monitoring by having a conversation. That visibility can be beneficial to the beginner, but can also be given to the model for him/her to control. Delayed assistance, explanation of choices, and independent transfer are then deemed to be design controls by CRAFT-LM. These controls make this agency a proof. They also explain why use of a product is not a good measure of the quality of its product in education.

Creativity theory is also connected with assessment. One of the key elements of human-AI collaboration is the evaluation of their work that is based on novelty or attractiveness of the final result. When it comes to education, a different question is: What skill did the student learn? Creative learning is thinking, comparing, discarding and altering possibilities and keeping a personal and culturally grounded perspective. This is more easily observed in a process portfolio or an unassisted task, rather than in one artefact or creativity score.

4.4. Methodological contribution

The main methodological contribution is the differentiation between assisted performance and lasting learning. Several studies have taken the stance that improvement is the basis for learning that takes place through the use of a tool. If the system offers explanations, structure or content in the assessment, then this inference is doubtful. The improved design includes an independent baseline, documented support phase (teacher verification of both support phase and practice), immediate post test, and delayed transfer task (no model). The last two stages offer information about the capability that could be extended beyond the interaction or not.

In addition to discipline-specific rubrics, blinding (double-blind scoring) is also required for artifact assessment. There should be a contrast in the execution of the strokes, structure, spacing, rhythm, and fidelity to the style in calligraphy rubrics. The following good practice advice needs to be incorporated in painting rubrics: composition, material technique, tonal/color control, visual reasoning, and expressive coherence. The rater should NOT know about the condition of the learner. Inter-rater reliability should be reported and disagreements should be dealt with by using a stated rule. These practices safeguard the study from learning mistakes. Recent synthesis of the arts assessments literature also advocates for more transparent and participatory, and context-specific, strategies than opaque technical scoring, aligning with the use of explicit rubrics and adjudication rules in studio subjects [69].

These principles could guide future assessment. If grade will be based solely on the last polish, then students will easily be able to conceal collaboration and willingly accept coherence that emerges from the models. A process weighted rubric should separate out the intent, the presence of exploration, disciplinary skill, value of student change, cultural awareness, sharing, contemplation, and self-propulsion. The final product is still significant but a part of the learning history. In addition, this approach can help solve the problem of academic dishonesty, since it makes the authorship more transparent than detection only [61, 63].

A fair test of teacher-guided LLM use requires comparison with conventional instruction under matched conditions. A three-arm quasi-experimental or cluster-randomised design would serve this purpose: (a) traditional teacher-led instruction, (b) teacher-guided LLM use following the CRAFT-LM sequence, and (c) unguided LLM access. Instructional time, exemplars and tasks should be identical across the arms. Each learner would complete an independent baseline, an immediate post-test and a delayed transfer task four to six weeks later, with no model access during either test. Artifacts should be scored by at least two raters blind to condition, using the indicators in Table 3, and inter-rater reliability should be reported as an intraclass correlation coefficient. Process measures should include the proportion of model suggestions that learners rejected or substantially changed, and stylistic convergence should be estimated as between-student dispersion within each arm. Because learners are nested within classes, mixed-effects models with class as a random effect are appropriate, and an implementation-fidelity checklist should confirm that teachers in arm (b) followed the intended sequence. Only a difference between arms (a) and (b) on the delayed, unassisted task would support the claim that teacher-guided LLM use adds value beyond traditional teaching.

4.5. Practical and policy implications

The practical and policy recommendations below are tied to the findings reported in Section 3. Teacher moderation and source checking follow from the trust, cultural-grounding, and teacher-mediation findings in Sections 3.5 and 3.8; process documentation and authorship rules follow from Sections 3.4 and 3.5; independent baseline and transfer assessment follow from Sections 3.1 and 3.10; and access, privacy, and alternative-pathway recommendations follow from the governance conditions in Section 3.5. Where these recommendations extend beyond direct studio evidence, they are presented as evidence-informed design or policy implications rather than as demonstrated effects.

Teacher moderation should be planned as a part of the system and not as an informal safeguard. It is essential that teachers can have control over approved exemplars, prompt templates, prohibited uses and escalation rules, and be able to review feedback histories and fix model errors. AI literacy for disciplines involves verifying historical attributions, examining the provenance of exemplar images, being able to detect that something is evidence while the other is an assertion, identifying uncertainty, understanding privacy of

images uploaded, and knowing when a model is demonstrating a stylistic preference as a rule. The general prompt writing workshops are insufficient [51].

The CRAFT-LM framework is a combination of both technical, pedagogical and governance elements and is represented by a single operating model that is offered here for empirical testing rather than as a proven solution. The conversational interface is often optimized instead of curriculum and assessment which is why many implementations fail. CRAFT-LM takes off from cultural foundation and learner's intention. It then restricts support via feedback from teacher and process assessment. Learning monitoring is intended to close the loop with tests for transfer and independence. The level of implementation of the framework could be assignment, course or program level.

The model can be placed in the process by instruction, at the assignment level. Clarification of terminology and planning of an assignment may be required as part of a calligraphy assignment but may not be automatically scored. For a painting task, it may be acceptable to prepare ideas and critiques, but it must be accompanied by transformation of material and a written description of rejected ideas. At level of the course, assisted tasks can be followed by unassisted tasks to enable students to learn strategies as well as their own performing skills. Program-level tracking of access, privacy, cultural fit and equity can be provided by institutions.

Digital divide is an immediate risk to equitable scaffolding for AI. The quality of a student's assistance is influenced by the device, bandwidth, language support, accessibility requirements, previous technical knowledge and experience, and the quality of the multimodal function offered by paid subscriptions. Whilst the studio task itself may not be delivered in English, English-language proficiency is likely to make a difference in generating more useful outputs for students with greater English ability. Institutions must offer equivalent access to a tool, and ensure a credible non-LLM alternative exists if one is needed. Terms used in Chinese, art-historical mentions, privacy conditions, accessibility, and consistency of performance based on the subscription levels should be considered by procurement [22, 23].

Both authorship and privacy are of equal value. Personal artwork, unpublished portfolios or images with personal information should not be uploaded unless institutional privacy and consent requirements allow for such use. There should be explicit regulations for the retention of data, training of models and permitted uploads in institutions. The statements of learner's primary changes, the purpose of use and the system should be noted and released. In research about how AI-generated art is attributed [38], it is found that people's perception of authenticity is influenced by the attribution. So it is useful to have openness in the documentation of processes to guarantee integrity and artistic communications.

Another issue raised by the evidence is the fundamental concept of creativity. There are many ways to measure creativity, such as novelty, usefulness, personal voice, divergent ideation, cultural significance, etc. The number of ideas can be increased with a system, but not ownership. It can be used to improve the coherence of the formal structure, though it might reduce diversity. The outcome is dependent on the ability of the learners to compare, reject and transform suggestions. The process evidence, reflection and independent transfer should also be measured in art education, along with the product rating, in order to measure creativity.

The institutional decision making process of procurement and governance should be based on the educational need. When choosing a platform, you should weigh not just the quality of the text or imagery that they are able to produce, but also other important. The following questions must be addressed: Do teachers have control of the reference materials? Is there a way to check the interaction record? Are student’s works saved for training? Is Chinese terminology passed on correctly? Are there any inequalities in access? There should be a viable alternative (but not a model) pathway for individuals who are not required to participate in learning.

4.6. Limitations

Several limitations remain. This is an integrative review, not a systematic review. The auditable search used one cross-publisher scholarly index and screened only the first 10 SciSpace-ranked records for each of four queries. Because the authors did not control SciSpace's default relevance ordering and the underlying weighting criteria were not available in the search record, potentially eligible studies below the retrieval cut-

off may have been missed. The 17-study core is therefore a transparent but bounded evidence set rather than an exhaustive census of the literature. Theoretical, policy, and other contextual references are used for interpretation only and do not contribute to study-frequency claims. Each of the 17 analytic-core studies corresponds to an included record documented in Supplementary File S1 and to its respective citation in the manuscript reference list. Contextual, theoretical, and policy references are not counted among the 17 included studies, and their roles are documented separately in the Provenance_Notes worksheet.

There is little direct evidence of intervention in calligraphy and painting. The discussion thus draws on triangulation, achieved in the following way: educational research, creativity studies, visual-art scholarship, and learning theory. This can be helpful for system design, but it is not helpful for determining an average treatment effect. There are also different definitions for outcomes. Although they address different questions, the concepts of engagement, satisfaction, performance, creativity, and learning are sometimes commingled. The review discusses this issue at a conceptual level but future empirical studies should be based on consistent terms and a specific measurement program.

This is because technology advances faster than educational evaluation. Platform capabilities, data policies and model behavior may change throughout a course or across study sites. Cultural performance can also vary in terms of language, reference material, and prompt formulation. The CRAFT-LM framework should then be considered a design proposition to be tested. It is dependent on implementation fidelity, the ability to adapt it to the disciplines, and evidence from delayed independent performance.

Lastly, the review is not based on the assumption that all calligraphy and painting traditions have a common standard. Cultural grounding has to be determined locally and on an expert basis. Reducing generic feedback does not eliminate debate in a tradition, but it can be addressed through teacher-approved references. Cross-cultural studies in the future should include the conditions under which the same system facilitates interpretation, conditions under which it imposes external conventions, and how the learners negotiate the differences.

5. Conclusions

These are useful applications of LLM that can be incorporated into the teaching of calligraphy and painting, but there is no clear evidence that LLM learning is superior. The functions with empirical support are access to explanation, reflective questioning, feedback on drafts and systematic comparison, and this support comes mainly from adjacent educational settings; gains in technical skill, originality, retention and independent performance remain potential rather than demonstrated. These functions can be useful for reinforcing learning, while students still have the freedom to make decisions about what they learn and how; feedback is assessed through action, and students can also have a voice in assessing the quality of the learning based on disciplinary and cultural standards.

Substitution, aesthetic convergence, culturally generic feedback, opaque authorship and unequal access are the principal risks. A great deal more emphasis is placed on immediate performance, as opposed to retention, transfer or independent creative development. This can, therefore, amplify the educational value within a system that can produce a polished response. CRAFT-LM is proposed as a way to address this issue through cultural grounding, retained agency, adaptive feedback, formative process assessment, teacher moderation and learning monitoring.

It's important to start a high-quality implementation with the learning problem, not the tool. Students should set themselves a goal before they are given assistance and do their baseline work. Model suggestions should be tried, self-evaluated and teacher-evaluated by the learner. Process, disciplinary technique, cultural understanding, authorship and independent transfer should be valued during assessment. Feedback should be planned with access, privacy and disclosure.

CRAFT-LM should be read as a set of untested design requirements derived from the reviewed literature, not as a validated model. Its value depends on empirical work that does not yet exist, and five directions are the most pressing. The first is controlled comparison of teacher-guided LLM use, unguided use and traditional instruction, assessed on delayed and unassisted tasks. The second is blinded artifact scoring with discipline-

specific rubrics for calligraphy and painting, with inter-rater reliability reported. The third is measurement of stylistic convergence across groups of learners rather than judgement of single products. The fourth is testing whether LLM feedback on calligraphic and painterly conventions is accurate and culturally valid when checked against expert teachers. The fifth is longer interventions that follow retention and independent creative development across a full course. Until such evidence is available, claims about learning effects should remain cautious and tied to the instructional design in which the model is used.

Declarations

Funding: This research received no external funding.

Conflict of Interest: The authors declare no conflict of interest.

Ethics approval: Not applicable. This review reports no new research involving human participants or identifiable personal data.

Data availability: No new participant data were generated. The complete reproducibility materials for the structured literature search are supplied as Supplementary File S1, containing the full search questions, screening decisions, exclusion reasons, and study-level extraction matrix for the auditable core evidence set.

Author Contributions: Both authors contributed to the conception and development of the review, interpretation of the literature, preparation of the manuscript, and critical review of the final version. Both authors read and approved the final manuscript.

References

1. Chen, W., He, Y., Chen, C., Zhu, M., Bi, S., Liu, J., Xia, M., Lin, Q., Wang, Y., & Wang, W. (2019). Long-term Chinese calligraphic handwriting training has a positive effect on brain network efficiency. PLOS ONE, 14(1), e0210962. https://doi.org/10.1371/journal.pone.0210962
2. Fang, C., Nakagaki, K., Labrune, J. B., & Ishii, H. (2022). Embodied calligraphy learning through direct kinesthetic feedback. In Proceedings of the Symposium on Computational Fabrication. ACM. https://doi.org/10.1145/3559400.3565595
3. Lin, J., Ji, Y., & Cai, Y. (2023). A digital Chinese brush system design based on embodied interaction theory. In Proceedings of the Eleventh International Symposium of Chinese CHI. ACM. https://doi.org/10.1145/3629606.3629648
4. Wei, M., Talib, O. B., Chua, Y. P., Gui, X., Pan, X., & Song, Q. (2025). The power of brushstrokes: How Chinese calligraphy transforms learning – A systematic review. Multidisciplinary Reviews, 8(11), e2025356. https://doi.org/10.31893/multirev.2025356
5. Kangas, J., Sundstedt, M., Kaihovirta, H., & Harju-Luukkainen, H. (2024). Visual and embodied learning in young children's art education. Nordisk Barnehageforskning, 21. https://doi.org/10.23865/nbf.v21.368
6. Torres, S., & Pinheiro, D. (2025). Gestural painting and viewer embodiment: Insights from neuroaesthetics. In Proceedings of the 12th SWS International Scientific Conference on Arts and Humanities (ISCAH 2025). https://doi.org/10.35603/sws.iscah.2025/s04.44
7. Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. IEEE Access, 8, 75264–75278. https://doi.org/10.1109/ACCESS.2020.2988510
8. Grassini, S. (2023). Shaping the future of education: Exploring the potential and consequences of AI and ChatGPT in educational settings. Education Sciences, 13(7), 692. https://doi.org/10.3390/educsci13070692
9. Hwang, G.-J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of artificial intelligence in education. Computers and Education: Artificial Intelligence, 1, 100001. https://doi.org/10.1016/j.caeai.2020.100001
10. Ouyang, F., & Jiao, P. (2021). Artificial intelligence in education: The three paradigms. Computers and Education: Artificial Intelligence, 2, 100020. https://doi.org/10.1016/j.caeai.2021.100020
11. Zhai, X., Chu, X., Chai, C. S., Jong, M. S.-Y., Istenic, A., Spector, M., Liu, J.-B., Yuan, J., & Li, Y. (2021). A review of artificial intelligence in education from 2010 to 2020. Complexity, 2021, 8812542. https://doi.org/10.1155/2021/8812542
12. Cetinic, E., & She, J. (2022). Understanding and creating art with AI: Review and outlook. ACM Computing Surveys, 55(3), Article 60. https://doi.org/10.1145/3475799
13. Darvishi, A., Khosravi, H., Sadiq, S., Gašević, D., & Siemens, G. (2024). Impact of AI assistance on student agency. Computers & Education, 210, 104967. https://doi.org/10.1016/j.compedu.2023.104967
14. Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. Computers and Education: Artificial Intelligence, 3, 100074. https://doi.org/10.1016/j.caeai.2022.100074
15. Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., & Caliskan, A. (2023). Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 1493–1504). ACM. https://doi.org/10.1145/3593013.3594095
16. Ding, L., Li, T., Jiang, S. Y., & Gapud, A. (2023). Students' perceptions of using ChatGPT in a physics class as a virtual tutor. International Journal of Educational Technology in Higher Education, 20, Article 63. https://doi.org/10.1186/s41239-023-00434-1
17. Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. International Journal of Educational Technology in Higher Education, 20, Article 55. https://doi.org/10.1186/s41239-023-00425-2
18. Lin, X. (2024). Exploring the role of ChatGPT as a facilitator for motivating self-directed learning among adult learners. Adult Learning, 35(3), 156–166. https://doi.org/10.1177/10451595231184928
19. Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. Science Advances, 10(28), eadn5290. https://doi.org/10.1126/sciadv.adn5290

20. Chiu, T. K. F. (2024a). A classification tool to foster self-regulated learning with generative artificial intelligence by applying self-determination theory: A case of ChatGPT. Educational Technology Research and Development, 72, 2401–2416. https://doi.org/10.1007/s11423-024-10366-w
21. Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274
22. Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. International Journal of Educational Technology in Higher Education, 20, Article 43. https://doi.org/10.1186/s41239-023-00411-8
23. Delcker, J., Heil, J., Ifenthaler, D., Seufert, S., & Spirgi, L. (2024). First-year students' AI-competence as a predictor for intended and de facto use of AI-tools for supporting learning processes in higher education. International Journal of Educational Technology in Higher Education, 21, Article 18. https://doi.org/10.1186/s41239-024-00452-7
24. Chiu, T. K. F., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. Computers and Education: Artificial Intelligence, 4, 100118. https://doi.org/10.1016/j.caeai.2022.100118
25. Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. International Journal of Educational Technology in Higher Education, 20, Article 22. https://doi.org/10.1186/s41239-023-00392-8
26. Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. Education Sciences, 13(4), 410. https://doi.org/10.3390/educsci13040410
27. Ouyang, F., Zheng, L., & Jiao, P. (2022). Artificial intelligence in online higher education: A systematic review of empirical research from 2011 to 2020. Education and Information Technologies, 27, 7893–7925. https://doi.org/10.1007/s10639-022-10925-9
28. Castro-Alonso, J. C., Ayres, P., Zhang, S.-G., de Koning, B. B., & Paas, F. (2024). Research avenues supporting embodied cognition in learning and instruction. Educational Psychology Review, 36. https://doi.org/10.1007/s10648-024-09847-4
29. Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. International Journal of Educational Technology in Higher Education, 20, Article 38. https://doi.org/10.1186/s41239-023-00408-3
30. Chiu, T. K. F. (2024b). Future research recommendations for transforming higher education with generative AI. Computers and Education: Artificial Intelligence, 6, 100197. https://doi.org/10.1016/j.caeai.2023.100197
31. Jiang, Y., Fan, Y., & Liu, Z. (2026). Generative AI in art education: A systematic review of research trends, tool applications, and outcomes (2019-2025). Education Sciences, 16(1), 47. https://doi.org/10.3390/educsci16010047
32. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71
33. [33] Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2024). A SWOT analysis of ChatGPT: Implications for educational practice and research. Innovations in Education and Teaching International, 61(3), 460–474. https://doi.org/10.1080/14703297.2023.2195846
34. Lo, C. K., Hew, K. F., & Jong, M. S.-Y. (2024). The influence of ChatGPT on student engagement: A systematic review and future research agenda. Computers & Education, 219, 105100. https://doi.org/10.1016/j.compedu.2024.105100
35. Essel, H. B., Vlachopoulos, D., Essuman, A. B., & Amankwa, J. O. (2024). ChatGPT effects on cognitive skills of undergraduate students: Receiving instant responses from AI-based conversational large language models. Computers and Education: Artificial Intelligence, 6, 100198. https://doi.org/10.1016/j.caeai.2023.100198
36. Browne, K. (2022). Who (or what) is an AI artist? Leonardo, 55(2), 130–134. https://doi.org/10.1162/leon_a_02092
37. Chatterjee, A. (2022). Art in an age of artificial intelligence. Frontiers in Psychology, 13, 1024449. https://doi.org/10.3389/fpsyg.2022.1024449
38. Horton, C. B., White, M. W., & Iyengar, S. S. (2023). Bias against AI art can enhance perceptions of human creativity. Scientific Reports, 13, 19001. https://doi.org/10.1038/s41598-023-45202-3

39. Messer, U. (2024). Co-creating art with generative artificial intelligence: Implications for artworks and artists. Computers in Human Behavior: Artificial Humans, 2(1), 100056. https://doi.org/10.1016/j.chbah.2024.100056
40. O'Toole, K., & Horvát, E.-Á. (2024). Extending human creativity with AI. Journal of Creativity, 34(2), 100080. https://doi.org/10.1016/j.yjoc.2024.100080
41. Zhou, E., & Lee, D. (2024). Generative artificial intelligence, human creativity, and art. PNAS Nexus, 3(3), pgae052. https://doi.org/10.1093/pnasnexus/pgae052
42. Zhang, Y., & Xu, L. (2025). Application of computer vision-based dynamic stroke analysis in intelligent calligraphy teaching systems. Journal of Computational Methods in Sciences and Engineering. https://doi.org/10.1177/14727978251361547
43. Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. Proceedings of the National Academy of Sciences, 122(26), e2422633122. https://doi.org/10.1073/pnas.2422633122
44. Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. British Journal of Educational Technology, 56(2), 489–530. https://doi.org/10.1111/bjet.13544
45. Soto, A., & Karczmarzyk, M. A. (2022). Expresión, comunicación y autoconocimiento mediante el método Painting Dialogues. Human Review, 11. https://doi.org/10.37467/revhuman.v11.4280
46. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10684–10695). https://doi.org/10.1109/CVPR52688.2022.01042
47. Wu, J., Cai, Y., Sun, T.-T., Ma, K., & Lu, C. (2025). Integrating AIGC with design: Dependence, application, and evolution—a systematic literature review. Journal of Engineering Design, 36(5–6), 758–796. https://doi.org/10.1080/09544828.2024.2362587
48. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). ACM. https://doi.org/10.1145/3442188.3445922
49. Hu, S., Wang, N., Zhao, K., Jing, Y., Zhang, Y., Wang, Z., Guo, B., & Yu, Z. (2025). Masterpiece creation: An AI-powered robotic calligraphy creation system. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 9(3), 1–34. https://doi.org/10.1145/3749957
50. 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, Article 28. https://doi.org/10.1186/s40561-024-00316-7
51. Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. Computers and Education: Artificial Intelligence, 2, 100041. https://doi.org/10.1016/j.caeai.2021.100041
52. Cui, Y., Zeng, M., Du, X., & He, W. (2025). What shapes learners' trust in AI? A meta-analytic review of its antecedents and consequences. IEEE Access, 13. https://doi.org/10.1109/ACCESS.2025.3611367
53. Abubakar, M., Rashid, S., & Sehar, N. (2025). Fluency without learning: How AI-induced cognitive ease disrupts the mechanisms of durable education. Journal of Applied Linguistics and TESOL, 8(3), 1–16. https://doi.org/10.63878/jalt1560
54. Kladaki, M., Kostas, A., & Alexopoulos, P. (2025). Exploring teachers' beliefs about ChatGPT in arts education. Education Sciences, 15(7), 795. https://doi.org/10.3390/educsci15070795
55. Latikka, R., Bergdahl, J., Savela, N., & Oksanen, A. (2023). AI as an artist? A two-wave survey study on attitudes toward using artificial intelligence in art. Poetics, 99, 101839. https://doi.org/10.1016/j.poetic.2023.101839
56. Anderson, B. R., Shah, J. H., & Kreminski, M. (2024). Homogenization effects of large language models on human creative ideation. In Proceedings of the 16th Conference on Creativity & Cognition (pp. 413–425). ACM. https://doi.org/10.1145/3635636.3656204
57. McLoughlin, J. (2025). The work of art in the age of artificial intelligibility. AI & Society, 40(2), 371–383. https://doi.org/10.1007/s00146-023-01845-4
58. Cotroneo, P., & Hutson, J. (2023). Generative AI tools in art education: Exploring prompt engineering and iterative processes for enhanced creativity. Metaverse, 4(1). https://doi.org/10.54517/m.v4i1.2164

59. Chang, Y. P. S., Yoon, I.-H., Park, J. W., Je-gal, Y.-J., & Lee, J. (2023). Understanding AI image generator and exploring educational possibilities in art education. Art Education Review, 88. https://doi.org/10.25297/aer.2023.88.277
60. Xu, J. (2024). Exploration of the applications of image-based AIGC in art education. Advances in Social Development and Education Research, 3(1). https://doi.org/10.61935/asder.3.1.2024.p60
61. Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International, 61(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148
62. Chiu, T. K. F. (2024c). The impact of generative AI on practices, policies and research direction in education: A case of ChatGPT and Midjourney. Interactive Learning Environments, 32(10), 6187–6203. https://doi.org/10.1080/10494820.2023.2253861
63. Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. Smart Learning Environments, 10, Article 15. https://doi.org/10.1186/s40561-023-00237-x
64. Guay, F. (2022). Applying self-determination theory to education: Regulation types, psychological needs, and autonomy-supporting behaviors. Canadian Journal of School Psychology, 37(1), 75–92. https://doi.org/10.1177/08295735211055355
65. Ryan, R. M., & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. Contemporary Educational Psychology, 61, 101860. https://doi.org/10.1016/j.cedpsych.2020.101860
66. Vansteenkiste, M., Ryan, R. M., & Soenens, B. (2020). Basic psychological need theory: Advancements, critical themes, and future directions. Motivation and Emotion, 44, 1–31. https://doi.org/10.1007/s11031-019-09818-1
67. Chiu, T. K. F. (2021). Digital support for student engagement in blended learning based on self-determination theory. Computers in Human Behavior, 124, 106909. https://doi.org/10.1016/j.chb.2021.106909
68. Chiu, T. K. F. (2022). Applying the self-determination theory to explain student engagement in online learning during the COVID-19 pandemic. Journal of Research on Technology in Education, 54(S1), S14–S30. https://doi.org/10.1080/15391523.2021.1891998
69. Rojas Lievano, M. S. (2025). Prácticas, tensiones e innovaciones en la evaluación de la educación artística: Una revisión sistemática (2019-2024). Ibero Ciencias - Revista Científica y Académica, 4(3), 4549–4576. https://doi.org/10.63371/ic.v4.n3.a355