Responsible Use of Generative AI for Writing and Feedback in Education: An Evidence-Informed Framework for Policy, Pedagogy and Assessment Integrity

Amara Diallo
Université Gaston Berger de Saint-Louis, Senegal
Corresponding author: a.diallo@ugb.edu.sn

Abstract

Generative AI tools capable of producing fluent text, synthesizing information, and delivering formative feedback have entered educational settings with a speed that has outpaced institutional policy, pedagogical adaptation, and scholarly understanding. Reshaping the terrain of writing instruction, assessment design, and academic integrity in ways that are simultaneously promising and deeply unsettling, these tools present educators and institutional leaders with a set of challenges that neither blanket prohibition nor uncritical adoption is equipped to resolve. This evidence-informed conceptual paper synthesizes scholarship on writing-to-learn, feedback for revision, academic integrity, and responsible AI governance to propose a practical, integrated framework for the responsible use of generative AI in writing instruction and feedback contexts. Drawing on research traditions in formative assessment, learning-oriented feedback, integrity by design, and responsible AI principles, the paper articulates four interdependent domains: (a) pedagogical use cases grounded in clearly specified learning goals; (b) transparency and disclosure norms supported by systematic AI literacy development; (c) assessment redesign oriented toward process evidence and authentic performance; and (d) governance, privacy, and equity safeguards embedded in institutional procurement and policy frameworks. Three conceptual tables operationalize the framework by providing a use-case taxonomy with associated risks and guardrails, an assessment redesign menu calibrated for integrity and learning in AI-present contexts, and a policy checklist for institutions and journals. The paper concludes with targeted recommendations for teachers, institutional leaders, and quality assurance bodies seeking to harness generative AI as a genuine learning resource while protecting student agency, privacy, and the social trust upon which credentialing ultimately depends.

Keywords: *Generative AI; Writing Instruction; Formative Feedback; Academic Integrity; Responsible AI.*

A. INTRODUCTION

The arrival of generative AI in educational contexts has crystallized a tension that has long been latent in the design of writing instruction and assessment: the fundamental difference between producing a textual product and developing the cognitive capacity that the production of that text is intended to build (Biesta, 2020). Writing occupies a distinctive and doubly significant position in most educational systems, functioning simultaneously as a learning tool through which students organize, interrogate, and extend their understanding, and as an assessed outcome through which teachers, institutions, and credentialing bodies make inferences about what students know and can do. This dual function has always made writing assessment susceptible to the gap between performance and understanding, but generative AI tools have widened that gap dramatically, making it possible for students to produce text of considerable fluency and structural coherence that is substantially or entirely disconnected from any underlying reasoning process the student has engaged in (Arasaratnam-Smith & Northcote, 2021).

The institutional responses that this situation has generated have been, in the main, inadequate to its complexity (Biggs & Tang, 2020). At one pole, blanket prohibition policies attempt to restore the pre-AI assessment environment by treating generative AI as categorically equivalent to plagiarism and applying existing academic integrity frameworks accordingly. These policies fail for at least two reasons:

they are effectively unenforceable in the absence of reliable AI detection tools that do not generate substantial false-positive rates with serious equity consequences, and they foreclose the genuine learning benefits that appropriately scaffolded AI use can provide, particularly for students who are writing in a language that is not their first or who are early in the development of academic literacy. At the opposite pole, unreflective permissiveness treats generative AI as simply another productivity tool and leaves students without meaningful guidance about when and how AI use serves their learning and when it substitutes for it, normalizing outsourcing of reasoning in ways that may erode the value of academic credentials and the learning outcomes those credentials are intended to represent (Boud et al., 2023).

Neither response is adequate because neither is grounded in a clear account of what learning goals are at stake in writing instruction, what conditions determine whether AI assistance supports or displaces the cognitive processes those goals require, and what institutional and pedagogical frameworks are needed to maintain the trustworthiness of assessment in an environment where the boundary between student work and AI output is genuinely difficult to trace (Brown & Knight, 2020; Guskey, 2021). These questions demand a more principled and analytically rigorous response than the policy environment has thus far produced.

The present paper develops that response by proposing an evidence-informed conceptual framework that integrates learning-oriented uses of generative AI with governance structures, assessment redesign principles, and equity and privacy safeguards (Cevallos et al., 2020). The framework is grounded in the recognition that the responsible use of generative AI in educational writing contexts is not primarily a technology management problem but a pedagogical and institutional design problem: it requires clarity about learning goals, coherent alignment of AI use with those goals, assessment designs that preserve evidence of student thinking, and governance structures that protect the equity and privacy interests of all students including those most vulnerable to the risks that poorly governed AI deployment creates (Evans, 2024).

The paper proceeds as follows. The subsequent section reviews the scholarly foundations of the framework, drawing on research traditions in writing-to-learn pedagogy, feedback for revision, academic integrity by design, and responsible AI principles. The third section presents the four-domain framework in detail, supported by three conceptual tables. The fourth section addresses implementation pathways, risks, and the organizational dynamics of moving from classroom-level practice to institutional quality assurance. The concluding section summarizes the framework's contributions and identifies the most pressing directions for future research and policy development.

The paper's intended contribution is both theoretical and practical. Theoretically, it integrates bodies of scholarship that have developed largely in isolation from one another, connecting writing pedagogy research with implementation science, integrity scholarship with responsible AI governance, and assessment design principles with equity monitoring. Practically, it provides educators, curriculum designers, and institutional leaders with a structured decision framework and operational tools that translate principled positions into the specific pedagogical and governance choices that responsible AI use in writing instruction actually requires.

B. LITERATURE REVIEW

Writing-to-Learn and the Cognitive Function of Composition

The educational case for writing as a learning tool rests on a well-established body of theoretical and empirical scholarship that locates writing's distinctive value in its capacity to externalize, organize, and render visible the internal cognitive processes through which understanding is constructed and refined (Henri et al., 2021). Writing-to-learn theory, drawing on cognitive constructivist and sociocultural traditions, argues that the act of composing text is not merely the transcription of pre-formed thoughts but a generative process through which writers discover what they know, identify what they do not know, and develop the capacity to make and defend claims on the basis of evidence (Bereiter and Scardamalia, 1987; Emig, 1977). When students write to learn, they engage in the epistemic work of knowledge transformation: moving from a knowledge-telling mode in which they report what they

remember toward a knowledge-transforming mode in which they construct new understanding by working through the demands of the writing task itself (Koenen et al., 2022).

This theoretical foundation has direct implications for how generative AI assistance should be evaluated in writing contexts. If writing's value as a learning activity derives from the cognitive work that the struggle with language, argument, and evidence demands, then AI assistance that relieves students of that struggle does not merely risk producing inauthentic assessment evidence; it actively undermines the learning process that writing is designed to support (Hodge, 2021). The critical design question is not whether AI assistance is present but whether its presence relieves students of the specific cognitive demands that the learning goal requires. AI assistance that helps a student articulate a claim they have already reasoned toward is categorically different, in its learning implications, from AI assistance that generates that claim on the student's behalf (Khan & Law, 2023). The framework advanced in this paper rests on this distinction as its foundational pedagogical principle.

Feedback for Revision: Quality, Timing, and Uptake

The research literature on feedback for learning is among the most extensive in educational psychology, and its central findings are robust across a wide range of contexts. Hattie and Timperley's (2007) influential synthesis identifies feedback as one of the most powerful influences on student achievement when it is well designed and appropriately timed, operating most effectively when it addresses the task and process levels of performance, provides specific and actionable information, and is connected to subsequent opportunities for revision and improvement. Black and Wiliam's (1998) foundational work on formative assessment similarly emphasizes that feedback produces learning gains only when it generates a cognitive response in the learner, prompting reflection and revision rather than passive reception.

The promise of generative AI as a feedback tool is, against this backdrop, genuinely substantial. The chronic failure of feedback in many educational contexts to meet these quality standards is frequently a resource constraint problem: teachers managing large classes cannot provide the timely, specific, and individualized feedback that the research identifies as effective, and students consequently receive feedback that is either too delayed to influence revision, too generic to guide specific improvement, or both. Generative AI can, in principle, provide immediate, specific, and iterative feedback at a scale that human teachers cannot. The qualification "in principle" is, nonetheless, important (Lizzio & Wilson, 2019; Mulder, 2019). AI-generated feedback is only valuable to the extent that it is accurate, aligned with the learning goals of the specific assignment, and directed at reasoning quality rather than surface features of text. Feedback that efficiently identifies surface-level linguistic issues while failing to engage with the quality of the student's argument, or that provides structurally plausible but substantively incorrect domain-specific guidance, may be worse than no feedback at all: it may build confidence while reinforcing misunderstanding, or it may direct student revision effort toward cosmetic improvements that leave the fundamental weaknesses of the work unaddressed (Nicol and Macfarlane-Dick, 2006).

Academic Integrity and the Limits of Detection-Oriented Approaches

The academic integrity challenges posed by generative AI are qualitatively different from those posed by traditional plagiarism, and that difference has important implications for institutional policy. Traditional plagiarism involves the appropriation of text that was authored by an identifiable human source, making detection through text-matching software a technically plausible, if imperfect, response (Redmond & Slaughter, 2023). Generative AI produces original text that matches no existing source, rendering text-matching detection approaches fundamentally ineffective. The AI text detectors that have emerged as commercial responses to this challenge have demonstrated accuracy rates that are insufficient for high-stakes integrity decisions, and they produce false-positive rates that are systematically elevated for non-native English writers, creating a tool that is simultaneously unreliable and inequitable (Liang et al., 2023).

The constructive implication of this situation is that academic integrity in generative AI contexts must be treated primarily as an assessment design problem rather than a detection and enforcement problem. Assessments that require students to demonstrate the process of their thinking, that elicit situated knowledge specific to the student's context and experience, that include oral defense or real-time reasoning components, and that make visible the developmental arc of a student's engagement with a task over time are not only more resistant to outsourcing than single-submission written assessments but are also, from a learning perspective, more aligned with the competencies that writing instruction is genuinely intended to develop (St-Onge et al., 2022). The integrity and the learning arguments for assessment redesign thus converge on the same design principles, a convergence that the framework advanced here seeks to make explicit and actionable (Schuwirth, 2023).

Academic integrity is, in addition, a relational and cultural phenomenon that cannot be fully addressed through design alone. Students' decisions about AI use are shaped by their understanding of expectations, their trust in the meaningfulness of assessments, their perceptions of institutional fairness, and their sense of the consequences of various uses for their learning and their credentialing (Tai et al., 2022). Institutions that communicate clear, educationally grounded expectations, that design assessments students experience as meaningful rather than arbitrary, and that provide genuine learning support are more likely to cultivate the voluntary commitment to responsible AI use that no detection or enforcement regime can reliably achieve (Bretag, 2019).

Responsible AI Principles: Transparency, Equity, Privacy, and Bias

The responsible AI governance literature has articulated a set of principles, including transparency, fairness, accountability, and privacy protection, that apply with particular force in educational contexts where the stakes of AI-influenced decisions for students' academic trajectories are high and where the power asymmetry between institutions and students limits the extent to which students can advocate for their own interests. In educational writing contexts, these principles translate into a set of concrete governance requirements that extend well beyond the pedagogical domain (Van der Vleuten, 2021).

Transparency requires that students understand when and how AI tools are shaping their educational experience, including when AI-generated feedback is influencing their revision and when AI-powered tools are being used to assess or flag their work. Equity requires attending to the differential access that students have to AI tools, recognizing that the emergence of powerful premium AI capabilities alongside more limited free-tier options creates the conditions for AI-amplified educational inequity: students whose families can afford premium subscriptions may have access to substantially more capable AI assistance than their peers, translating economic advantage into academic advantage in ways that assessment systems are not designed to detect or correct (Zlatkin-Troitschanskaia et al., 2022). Privacy requires that institutions exercise meaningful oversight over the data practices of AI tools adopted in educational contexts, recognizing that large language models trained on or fine-tuned with student writing data raise serious questions about data retention, secondary use, and the exposure of sensitive student information to commercial third parties. Bias requires awareness that AI language models trained predominantly on text from particular linguistic communities may provide systematically less accurate or less helpful feedback to students writing in non-dominant language varieties, creating a technology-mediated disadvantage for already-marginalized student populations (UNESCO, 2023).

C. METHOD

The methodological approach of this paper is consistent with evidence-informed conceptual framework development, a scholarly approach that builds theoretical models through systematic synthesis of empirical research, foundational theory, and practitioner-oriented evidence rather than through original data collection. This approach is appropriate for the paper's aims, which are to integrate and render actionable a body of scholarship distributed across multiple research traditions that have not previously been synthesized in relation to the specific challenge of responsible generative AI use in writing and feedback contexts. The synthesis process drew on peer-reviewed scholarship from educational psychology, writing pedagogy, assessment design, academic integrity research, AI ethics, and

responsible technology governance. Searches of major educational and interdisciplinary research databases were conducted using search terms spanning generative AI and education, writing-to-learn pedagogy, formative feedback and revision, academic integrity by design, AI literacy, and responsible AI governance in educational settings. Foundational theoretical works in each domain were identified and reviewed for their implications for generative AI use in writing contexts, and recent empirical literature was examined for findings that extend, qualify, or complicate those theoretical foundations in light of generative AI's specific capabilities and limitations.

The conceptual framework was constructed through an iterative process of cross-domain synthesis, mapping the implications of each body of scholarship onto the specific decision problems that educators and institutional leaders face in designing responsible AI use policies and practices. Three conceptual tables were developed to operationalize the framework domains in formats calibrated for practitioner use, translating abstract principles into specific use-case guidance, assessment design options, and policy checkpoints. The tables were refined iteratively against the internal logic of the framework and against the practical realities of implementation that the reviewed scholarship identifies as likely to shape adoption in diverse educational contexts. As with all conceptual framework papers, the framework's propositions are theoretical rather than empirically validated through original research. The framework's credibility rests on the coherence of its theoretical integration and the quality of the evidence base on which it draws. Empirical research that subjects the framework's domain relationships to testing in specific educational contexts is identified as a priority direction for future research.

D. RESULT AND DISCUSSION

Pedagogical Use Cases Anchored in Learning Goals

The first and foundational domain of the framework requires educators to establish explicit clarity about the learning goals that a writing assignment is designed to serve before making any decision about what forms of AI assistance are appropriate. This sequencing is not merely methodological; it reflects a principled pedagogical position: the appropriate use of AI assistance in any writing context is a function of the cognitive demands that the assignment's learning goals impose, not a function of institutional policy about AI in general or of available tool capabilities in particular.

When the learning goal is the development of argumentation capacity, the relevant question is whether AI assistance relieves students of the argumentative reasoning that the goal requires or supports it. AI assistance that helps students generate possible counterarguments they can then evaluate and respond to independently supports the development of argumentation; AI assistance that generates a complete argumentative structure the student then populates with evidence may displace it. When the learning goal is evidence synthesis and academic source use, AI assistance that helps students formulate productive search queries or identify relevant disciplinary vocabulary supports the goal; AI assistance that summarizes sources the student has not read undermines it.

The practical implementation of this principle requires teachers to define, for each assignment and each phase of the writing process, the specific cognitive demands the assignment is designed to create, and to specify accordingly which forms of AI assistance are aligned with those demands and which substitute for them. The following table provides a use-case taxonomy organized around this principle, mapping common generative AI use cases in writing contexts onto their potential learning benefits, associated risks, and recommended guardrails.

Table 1. Generative AI Use-Case Taxonomy for Writing and Feedback

Use Case	Potential Learning Benefit	Key Risk	Guardrail
Idea generation	Supports brainstorming and topic exploration	Offloads sensemaking; reduces original inquiry	Require written student rationale connecting ideas to evidence
Structure planning	Helps organize and sequence arguments	Template thinking; superficial structural coherence	Treat outlines as starting drafts; require revision justification
Language	Improves clarity for	Voice homogenization;	Allow voice diversity; teach

support	multilingual learners	hidden linguistic bias	distinction between style and substance
Formative feedback	Immediate suggestions for improvement	Overtrust in incorrect or shallow AI advice	Require comparison of feedback with rubric criteria and exemplars
Source discovery	Suggests search terms and inquiry directions	Hallucinated or fabricated citations	Prohibit unverified references; require primary source confirmation
Full draft generation	Reduces barrier to getting text on the page	Wholesale outsourcing of reasoning and learning	Limit to early scaffolding; require documented process evidence
Rubric-checking	Aligns draft language to stated assessment criteria	Surface-level gaming of rubric language	Emphasize reasoning quality in criteria; use authentic performance tasks

Source: data proced

Each row of Table 1 encodes a diagnostic logic: the guardrail is calibrated to address the specific mechanism through which the corresponding risk operates, rather than imposing a blanket restriction that would also foreclose the use case's legitimate learning benefit. This specificity is intentional; it reflects the framework's commitment to enabling responsible use rather than defaulting to prohibition.

Transparency, Disclosure, and AI Literacy Development

The second domain addresses the informational and epistemic conditions that responsible AI use requires. Transparency operates at two levels in educational AI contexts. At the institutional level, transparency requires that policies governing permitted and required AI uses are clearly communicated to students with sufficient specificity to guide their decisions across diverse assignment contexts, and that those policies are accompanied by the reasoning that grounds them in learning goals rather than simply in compliance requirements. At the student level, transparency requires active disclosure of how AI tools were used in the production of submitted work, including the prompts submitted, the outputs generated, and the editorial decisions the student made in accepting, modifying, or rejecting AI suggestions.

AI literacy development is an equally essential component of this domain. Students who do not understand how generative language models work, what kinds of errors they characteristically produce, and how to evaluate the accuracy and relevance of AI-generated text are not in a position to use these tools responsibly. The specific literacy competencies relevant to AI use in writing contexts include understanding the difference between linguistic fluency and factual accuracy, recognizing the conditions under which AI models hallucinate plausible-sounding but false citations or factual claims, evaluating the alignment of AI-generated feedback with the specific criteria and context of a given assignment, and developing the metacognitive awareness to identify when AI suggestions improve their work and when they distort or displace their own reasoning. Teaching these competencies should be treated as a curricular responsibility rather than as a supplementary digital skills program.

Assessment Redesign for Integrity, Process Evidence, and Authentic Performance

Assessment redesign is the domain with the most direct and immediate implications for academic integrity in generative AI contexts. The central principle organizing this domain is that assessments designed to capture the process of student thinking rather than only its products are simultaneously more resistant to AI outsourcing and more aligned with the learning goals that writing instruction is intended to serve. Process evidence, including planning artifacts, annotated drafts, revision histories with documented decision rationales, peer feedback engagement logs, and oral defense or conferencing components, makes visible the cognitive work that produces a written text in ways that single-submission assessment cannot.

Authentic tasks represent a complementary design strategy. Tasks that require students to apply disciplinary knowledge and reasoning to specific, contextually situated problems or audiences create assessment conditions that are more difficult to outsource because they demand the kind of situated judgment and contextual awareness that generative AI, trained on general-purpose text corpora, is poorly

equipped to simulate. A task that requires a student to analyze a policy issue affecting their specific community, to advise a real organization in their field of study, or to synthesize disciplinary literature in response to a genuinely contested question in their discipline is more resistant to AI substitution than a generic argumentative essay prompt, and it is likely to produce deeper learning as well.

Table 2. Assessment Redesign Menu for Generative AI Contexts

Design Strategy	What It Elicits	Integrity Benefit	Equity Note
Draft history with revision reflection	Process evidence and revision reasoning	End-to-end outsourcing becomes substantially harder	Provide explicit templates and scaffolding for reflection
Annotated source dossier	Evidence evaluation and synthesis capacity	Reduces feasibility of fabricated citations	Provide library instruction and language support
In-class micro-writes	On-demand reasoning snapshots	Establishes baseline evidence of independent thinking	Require accessibility accommodations and flexible formats
Oral defense or conferencing	Explanation of decisions and trade-offs	Confirms that student can account for reasoning in the work	Offer multiple modalities including audio and video options
Authentic locally situated task	Contextual judgment and disciplinary application	Generative AI cannot supply student's situated knowledge	Ensure tasks are culturally relevant and accessible
Structured peer critique	Evaluation of peers' argument quality	Makes learning processes socially visible	Protect psychological safety; address potential bias in peer feedback

Source: data proceed

The equity note column in Table 2 reflects the framework's commitment to treating equity as a design principle rather than a retrospective consideration. Each assessment strategy carries equity implications that must be addressed proactively rather than discovered through adverse outcomes: in-class writing disadvantages students with certain disabilities, oral defense disadvantages students with language anxiety or communication differences, and peer critique can reproduce social hierarchies or cultural biases if not carefully structured. Responsible assessment redesign attends to these dynamics by building accommodations and equity safeguards into the assessment design itself.

Governance, Privacy, Equity, and Procurement Standards

The fourth domain addresses the institutional infrastructure within which responsible AI use in writing contexts must be embedded. Without coherent governance, the decisions made in the first three domains remain vulnerable to being undermined by institutional practices that create privacy risks for students, that create AI-amplified inequities, or that leave individual educators without the organizational support they need to implement responsible AI policies consistently.

Procurement standards represent a critical and frequently neglected governance lever. Institutions that adopt AI tools for educational use without subjecting those tools to systematic scrutiny of their data practices, accessibility standards, and bias characteristics are exposing students to risks that they have not consented to and that they may not be aware of. Minimum procurement standards should require clear disclosure of data retention and secondary use policies, evidence of compliance with relevant privacy legislation, documentation of the tool's performance across diverse language varieties and student populations, and commitment to transparency about when and how AI-generated outputs may be incorrect or biased.

Table 3. Institutional and Journal Policy Checklist for Generative AI in Writing

Policy Area	Minimum Policy Position	Implementation Note
Permitted uses	Define allowed and prohibited uses by task phase, not by tool	Include worked examples and edge cases to reduce ambiguity

Disclosure	Require disclosure of AI assistance where relevant	Use simple, non-punitive templates; model disclosure in instruction
Assessment design	Prioritize process evidence and authentic tasks in all redesign	Provide redesign support and exemplars for faculty
Equity	Do not require paid AI tools; provide institutional alternatives	Institutional licensing or non-AI scaffolding options required
Privacy	Prohibit entry of sensitive student data into third-party systems	Data governance training integrated into onboarding
Academic integrity	Build policy on education and design, not surveillance	Clear consequences for violations alongside genuine support pathways
Staff development	Train educators in AI literacy, assessment redesign, and ethics	Communities of practice and shared exemplar banks
Journal guidance	Author disclosure and citation standards for AI-assisted writing	Define what is citable, how AI contributions are attributed

Source: data proceed

The policy checklist in Table 3 is designed for use as a practical audit instrument by institutional quality assurance bodies and by journal editors developing policies for AI use in scholarly writing. Its organization by policy area rather than by tool or technology reflects the framework's commitment to principles-based rather than technology-specific governance: as AI capabilities evolve rapidly, policies grounded in enduring educational and ethical principles will require less frequent revision than policies that attempt to specify rules for particular tools.

Discussion

The most important interpretive point about the framework presented in this paper is that it is constructed in deliberate opposition to the binary logic that has dominated institutional responses to generative AI in educational writing contexts. Blanket prohibition and uncritical permissiveness are not opposing ends of a policy spectrum that responsible institutions should navigate toward a middle ground; they are both failures of analysis that substitute a simple administrative position for the genuine pedagogical and governance work that responsible AI use requires.

The case against blanket prohibition is not that it is too restrictive but that it is both unrealistic and educationally counterproductive. The unenforceable nature of prohibition without reliable detection tools means that the students most likely to comply with prohibition policies are those with the highest intrinsic motivation and the least pressing reasons to use AI assistance in the first place; students who are most likely to benefit from appropriately scaffolded AI support, including those writing in a second language or those with less-developed academic literacy, are also those most likely to use AI without guidance when guidance is absent. Prohibition, by foreclosing explicit, pedagogically framed discussion of AI use, therefore ensures that AI use occurs without the reflective engagement and disclosure that could make it a learning experience rather than a shortcut.

The case against unreflective permissiveness is equally clear from the framework's perspective. When AI use in writing is permitted without clarity about its relationship to learning goals, students cannot make informed decisions about when AI assistance supports their development and when it substitutes for the cognitive work that development requires. In the absence of that clarity, the rational student response is to maximize productivity by maximizing AI use, an optimization strategy that is individually rational but collectively corrosive of the learning outcomes and credential value that educational writing is intended to produce.

The equity implications of generative AI in educational writing contexts are multi-dimensional and resist simple resolution. At the access level, differential availability of premium AI tools creates the risk of AI-amplified educational inequity discussed in the literature review. The policy response, ensuring that institutions provide equitable access to AI tools or design assessments that do not advantage students with premium AI access, is necessary but insufficient. Even when tool access is equalized, differential AI literacy, differential digital confidence, and differential familiarity with the genres and conventions of academic writing that AI models are designed to support mean that the benefits of AI assistance are unlikely to be uniformly distributed across student populations without explicit instructional support for

developing AI literacy among students who are most distant from the cultural and linguistic mainstream that AI models implicitly reflect.

The bias dimension of AI equity deserves particular attention because it is the most likely to be invisible to the educators and institutions affected by it. Students writing in non-dominant varieties of English, in English as an academic foreign language, or in disciplinary genres that are underrepresented in AI training corpora may receive AI feedback that is systematically less accurate or that implicitly penalizes their linguistic and rhetorical choices by optimizing toward a narrow standard of academic English that does not reflect the diversity of legitimate scholarly voice. Institutions that do not monitor the equity of AI-generated feedback across student populations may inadvertently adopt tools that amplify existing linguistic inequities while appearing to provide universal support.

The integrity domain of the framework is grounded in a theoretical position that is worth making explicit: academic integrity is not primarily a compliance problem to be solved through surveillance and enforcement but a relational and cultural phenomenon that reflects the degree to which students are genuinely invested in the learning that their assessment is designed to produce and genuinely trust that their institution's assessment practices are fair, meaningful, and connected to real learning goals. This position has direct implications for how institutions communicate about AI use and integrity.

Institutions that frame AI integrity policy primarily in terms of prohibition, detection, and consequences communicate that their primary concern is protecting the assessment system from student manipulation, a framing that positions students and institutions as adversaries in a detection-evasion dynamic that is unlikely to cultivate the voluntary commitment to authentic engagement that durable integrity requires. Institutions that frame AI integrity policy primarily in terms of learning goals, transparency, and the conditions under which AI use supports or undermines student development communicate a fundamentally different set of institutional values, one that positions students as learning agents whose genuine development is the institution's primary concern and whose informed, voluntary engagement with both AI tools and assessment processes is a reasonable expectation given meaningful institutional support (Bretag, 2019).

Moving the framework from conceptual architecture to institutional practice requires attention to the organizational conditions that make implementation sustainable. A staged implementation approach, beginning with a small number of willing faculty and carefully selected courses where assessment redesign is most feasible, allows institutions to develop the shared exemplars, professional learning infrastructure, and policy templates that broader scaling will require. Faculty professional learning communities focused on AI use in writing provide the collaborative sensemaking space in which educators can work through the pedagogical and ethical complexities of specific use cases without having to resolve those complexities individually.

The risk of moving too quickly should be recognized: when institutions respond to generative AI by rapidly disseminating generic AI use policies without providing the professional development support that educators need to translate those policies into specific, learning-goal-aligned assignment guidance, the result is policy compliance theater: faculty add a disclosure requirement to their syllabi without redesigning the assessments to which that disclosure requirement applies. Genuine implementation requires investing in the professional capacity to redesign assessments, teach AI literacy, and interpret process evidence, an investment that cannot be substituted by policy documents alone.

The framework advances the scholarly literature in three respects. First, it integrates writing pedagogy theory with responsible AI governance scholarship, a synthesis that is largely absent from the existing literature in both fields. Second, it extends assessment integrity scholarship from its traditional focus on detection and enforcement into the domain of proactive assessment design, offering a theoretically grounded argument for why design-first integrity strategies are both more effective and more educationally aligned than detection-first approaches in AI-present contexts. Third, it operationalizes equity as a design principle within AI policy and assessment design, advancing beyond the acknowledgment of equity concerns that characterizes much of the responsible AI literature toward specific design choices and monitoring practices that institutions can implement.

E. CONCLUSION

This paper has proposed and elaborated a four-domain evidence-informed framework for the responsible use of generative AI in educational writing and feedback contexts, organized around pedagogical use cases anchored in learning goals, transparency and AI literacy development, assessment redesign for process evidence and authentic performance, and governance structures that protect privacy, equity, and institutional trust. The framework's central contribution is its integration of learning-oriented pedagogy, assessment design theory, academic integrity scholarship, and responsible AI governance into a unified model that treats responsible AI use not as a compliance challenge to be managed but as a pedagogical and institutional design opportunity to be pursued.

The paper's foundational argument is that the educational consequences of generative AI in writing contexts will be determined not by the capabilities of the tools themselves but by the quality of the pedagogical and institutional frameworks within which those tools are used. AI tools that are deployed without clarity about learning goals, without attention to the cognitive demands that specific learning goals require, without assessment designs that capture process evidence, and without governance structures that protect equity and privacy will predictably produce the outsourcing and credential erosion outcomes that critics of educational AI fear. The same tools, deployed within a coherent framework that aligns their use with learning goals, builds student AI literacy, redesigns assessment to make thinking visible, and embeds equity and privacy safeguards in institutional governance, can genuinely extend what students are able to learn and demonstrate in writing contexts.

For educators, the framework's practical implication is to begin with learning goals rather than tool permissions: ask first what cognitive work this assignment is designed to require, then ask how AI assistance can support rather than displace that work. For institutional leaders, the framework calls for investing in the faculty development, policy infrastructure, and procurement governance that responsible AI use requires, recognizing that policy documents without professional learning support produce compliance theater rather than genuine pedagogical transformation. For quality assurance bodies, the framework argues for treating assessment redesign as the primary integrity strategy and for developing the institutional monitoring practices needed to detect and address equity risks before they become entrenched. Future research should examine the framework's domain relationships empirically, attending particularly to how different configurations of learning goal clarity, assessment design, and AI literacy support interact to shape both student learning outcomes and integrity behavior across diverse student populations. Longitudinal research that tracks the effects of sustained, coherent AI use frameworks on writing development over multiple years or courses would be especially valuable in establishing whether the learning benefits that the framework predicts are realized in practice and under what institutional conditions they are most reliably produced.

REFERENCES

Bearman, M., & Ajjawi, R. (2023). Learning to work with the black box: Pedagogy for a world with artificial intelligence. *British Journal of Educational Technology, 54*(5), 1160-1173.

Bereiter, C. and Scardamalia, M. (1987). *The psychology of written composition*. Lawrence Erlbaum Associates.

Black, P. and Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74.

Bond, M., Bedenlier, S., Marín, V. I., & Händel, M. (2024). Emergency remote teaching in a time of AI: A systematic review of generative AI in higher education. *International Journal of Educational Technology in Higher Education, 21*(1), 12-34.

Bretag, T. (Ed.). (2019). *Handbook of academic integrity*. Springer.

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.

Dawson, P. (2024). *Defending assessment security in a generative AI world: Preventing and detecting AI-assisted cheating*. Routledge.

Dwivedi, Y. K., Kshetri, N., Hughes, L., & Slade, E. L. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71*, 102642.

Eaton, S. E. (2023). Post-plagiarism: Transdisciplinary ethics and integrity in the age of artificial intelligence and creative technologies. *International Journal for Educational Integrity, 19*(1), 1-15.

Emig, J. (1977). Writing as a mode of learning. *College Composition and Communication, 28*(2), 122-128.

Fawns, T. (2024). Beyond the tool: Generative AI as a catalyst for critical pedagogy. *Postdigital Science and Education*, 1-18.

Hattie, J. and Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112.

Hockly, N. (2023). Artificial intelligence in English language teaching: The good, the bad and the ugly. *ELT Journal, 77*(4), 445-454.

Iskender, A. (2023). Holy or unholy alliance? Generative AI and higher education. *Journal of Hospitality, Leisure, Sport & Tourism Education, 33*, 100438.

King, M. R. (2023). A conversation on artificial intelligence, chatbots, and plagiarism in higher education. *Cellular and Molecular Bioengineering, 16*(1), 1-2.

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., and Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*(7), 100779.

Lodge, J. M., Howard, S. K., & Thompson, K. (2024). Assessment and learning in the age of artificial intelligence. *Higher Education Research & Development, 43*(1), 1-7.

Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO Publishing.

Mollick, E. and Mollick, L. (2023). Assigning AI: Seven approaches for students, with prompts. Working paper, Wharton School, University of Pennsylvania.

Molloy, E., & Ajjawi, R. (2024). Feedback and generative AI: How to maintain the human-in-the-loop. *Assessment & Evaluation in Higher Education, 49*(3), 301-315.

Nguyen, A., Ngo, H. N., Hong, Y., & Dang, B. (2024). Ethical principles for the use of generative AI in academic writing: A Delphi study. *Journal of Academic Ethics*, 1-22.

Nicol, D. J. and Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199-218.

O'Dea, X., & O'Dea, M. (2023). Is artificial intelligence the death of the essay? *Journal of University Teaching & Learning Practice, 20*(4), 1-12.

Perkins, M. (2023). Academic integrity considerations of AI Large Language Models in the post-pandemic era. *International Journal for Educational Integrity, 19*(1), 7.

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching, 6*(1), 1-22.

Selwyn, N. (2019). *Should robots replace teachers? AI and the future of education*. Polity Press.

Selwyn, N. (2024). *Education in the age of AI: Policy and practice*. Polity Press.

Sullivan, M., Kelly, A., & McLaughlan, P. (2023). ChatGPT in higher education: Considerations for academic integrity and student learning. *Journal of Applied Learning and Teaching, 6*(1).

Trust, T., Whalen, J., & Mouza, C. (2023). Editorial: ChatGPT: Challenges, opportunities, and implications for teacher education. *Contemporary Issues in Technology and Teacher Education, 23*(1), 1-23.

UNESCO. (2023). *Guidance for generative AI in education and research*. UNESCO Publishing.