From Product to Process: A Framework and Practical Toolkit for AI-Aware University Assessment

Luis F. Rivera-Galicia *, Mónica Giménez-Baldazo , Carlos Mir-Fernández

Posted Date: 15 April 2026

doi: 10.20944/preprints202604.0989.v1

Keywords: generative artificial intelligence; higher education; assessment redesign; process-based assessment; academic integrity; AI literacy; authentic assessment

Article

From Product to Process: A Framework and Practical Toolkit for AI-Aware University Assessment

Luis F. Rivera-Galicia [1,*], Mónica Giménez-Baldazo [2] and Carlos Mir-Fernández [2]

[1] Department of Economics, Faculty of Economics, Business and Tourism, University of Alcalá, 28802 Alcalá de Henares (Madrid), Spain

[2] Department of Economics and Business, Faculty of Economics, Business and Tourism, University of Alcalá, 28802 Alcalá de Henares (Madrid), Spain

* Correspondence: luisf.rivera@uah.es

Featured Application

The framework can be used by higher education institutions to redesign essays, quantitative assignments, and project-based tasks so that assessment remains valid, transparent, and educationally meaningful when GenAI tools are available.

Abstract

The routine availability of generative artificial intelligence (GenAI) has weakened the validity assumptions behind many university assessment formats, especially those that rely on a polished final product as the main evidence of learning. This paper develops a discipline-neutral, process-based framework for AI-aware assessment that treats GenAI as a persistent feature of students' academic work rather than as a temporary anomaly. Drawing on recent scholarship on assessment validity, authenticity, transparency, and institutional governance, the paper synthesizes key risks of product-only assessment and translates them into a practical redesign logic. The framework rests on five principles: transparency of AI use, auditability through evidence-of-work, visibility of reasoning, contextual authenticity, and feasibility with inclusion. Building on these principles, the article proposes a toolkit that combines staged submissions, annotated decision logs, targeted checkpoints, short oral validations, and verification-by-design in quantitative and applied tasks. Three higher-education illustrations are used to show how the framework can be adapted to writing-intensive, quantitative, and team-based assignments. The paper argues that the central assessment question in the GenAI era should shift from authorship policing to competence verification, and it offers reusable design structures, rubric dimensions, and implementation guidance for institutions seeking to preserve academic standards while supporting responsible AI use.

Keywords: generative artificial intelligence; higher education; assessment redesign; process-based assessment; academic integrity; AI literacy; authentic assessment

1. Introduction

Generative artificial intelligence (GenAI) has moved rapidly from novelty to ordinary academic infrastructure. Students now have routine access to systems that can draft essays, summarize readings, produce code, explain concepts, generate data visualizations, and suggest solution pathways in response to highly specific prompts. This shift has expanded opportunities for scaffolding, feedback, and accessibility, but it has also exposed a structural weakness in many established assessment formats: the final submitted product is no longer a reliable proxy for the student's own reasoning.

The central problem is not simply misconduct. It is an assessment validity problem. In many university settings, essays, reports, take-home tasks, and unsupervised problem sets have historically

 doi:10.20944/preprints202604.0989.v1

functioned as reasonable indicators of achievement. Under widespread GenAI availability, however, the relationship between observable performance and underlying competence becomes more fragile. A high-quality output can now be produced with limited conceptual engagement, limited traceability, and limited opportunity for instructors to determine what the student actually understands.

This situation has generated a predictable institutional reaction: renewed concern about academic integrity and, in some settings, an overemphasis on authorship detection. Yet the literature and recent policy guidance increasingly suggest that detector-centered responses are too narrow for a long-term solution. Detection technologies remain contested in accuracy and fairness, and even when they work imperfectly, they address only one part of the problem. The more important question is how higher education can continue to assess reasoning, judgment, disciplinary understanding, and applied competence in environments where AI support is likely to remain normal rather than exceptional.

Current debates on artificial intelligence in education increasingly recognize that AI is reshaping human learning while also raising ethical, inclusive, and pedagogical challenges. That framing is especially relevant for assessment, because assessment is the point at which institutions make formal claims about what students know and can do. If those claims are no longer well supported by assessment evidence, the consequences extend beyond individual assignments to programme coherence, quality assurance, and public trust [1].

This paper argues that a more sustainable response lies in redesign rather than prohibition. More specifically, it proposes a shift from product-dominant assessment toward process-based and reasoning-centered assessment in which competence is inferred from accumulated evidence rather than from a polished artifact alone. The aim is not to eliminate AI from learning. In many contexts, that would be unrealistic and educationally unhelpful. The aim is to specify the conditions under which AI can be used without undermining the validity of academic judgment.

This paper makes three contributions. First, it synthesizes recent work on GenAI, assessment validity, transparency, authenticity, and institutional governance into a coherent problem frame. Second, it proposes a discipline-neutral framework that integrates task design, process evidence, validation moments, and rubric logic. Third, it develops a practical toolkit with implementation guidance and applied illustrations that can be adapted by instructors and institutions. The overall claim is that the assessment question in the GenAI era should shift from "Who authored every word?" to "What evidence do we need to judge competence confidently, fairly, and consistently?"

2. Background and Related Work

2.1. Assessment Validity Under Conditions of Routine AI Support

Assessment in higher education is intended to provide valid evidence that students have achieved intended learning outcomes. In classical terms, validity concerns whether the interpretation and use of assessment results are justified by the evidence available. When the target is conceptual understanding, methodological reasoning, disciplinary judgment, or applied decision-making, the submitted artifact is only useful to the extent that it meaningfully represents those constructs. The widespread use of GenAI destabilizes that representational logic. A student may now submit an elegant response that appears sophisticated while having delegated substantial portions of idea generation, structuring, editing, coding, or interpretation to an AI system.

This does not mean that all AI-supported work is invalid. It means that the evidential chain has become less secure. In some cases, AI use may legitimately support learning in ways analogous to feedback, exemplars, or adaptive tutoring. In other cases, it may mask gaps in understanding and inflate apparent performance. The problem, therefore, is not whether AI is present, but whether the assessment design still allows instructors to make warranted claims about student competence.

 doi:10.20944/preprints202604.0989.v1

2.2. Governance, Policy, and the Limits of Prohibition

Recent policy guidance from UNESCO and the OECD frames GenAI as a structural educational issue rather than a temporary disruption [2,3]. These documents emphasize human-centered governance, transparency, institutional readiness, and the need to align technological adoption with educational purposes. In higher education, sector-level reports and policy reviews increasingly argue that the challenge cannot be managed solely through restrictive rules or misconduct procedures [4–7]. Broader syntheses likewise show that the educational opportunities and risks of GenAI are deeply intertwined, which makes assessment redesign more urgent than rulemaking alone [1]. Policy is necessary, but policy by itself does not generate valid evidence of learning.

Critical reviews of university GenAI policies point to several recurring weaknesses: vague language about acceptable use, inconsistent requirements across courses, limited attention to pedagogy, and insufficient support for students and staff. Policies often answer the question of whether AI is allowed more clearly than the question of how assessment should be designed when AI is available. This gap matters because institutions may end up with formal rules but weak assessment practice.

2.3. Why Detection-Led Responses Are Inadequate

One visible line of institutional response has focused on AI detectors and related forms of surveillance. However, the limitations of these strategies are now well documented [4,5,8]. Detector outputs are probabilistic, vulnerable to false positives and false negatives, and difficult to justify as definitive evidence for fraud detection. In addition, a detector-centered approach can shift the educational culture toward suspicion, making teachers responsible for proving authorship rather than designing stronger assessment evidence.

The issue is deeper than the technical reliability of detectors. Even a perfect detector would not solve the broader educational challenge. Students can still use AI for idea development, organization, problem solving, code debugging, translation, or partial drafting in ways that remain invisible but can significantly improve performance. A narrow focus on generated text also overlooks the wider reconfiguration of academic work brought about by multimodal, conversational, and embedded AI systems. For these reasons, recent studies on assessment increasingly argue for structural changes to assessment rather than reliance on post monitoring.

2.4. Assessment redesign, AUTHENTICITY, and Transparency

Assessment redesign literature has therefore turned toward authenticity, traceability, and reasoning visibility. Authentic assessment asks students to apply knowledge in contextualized situations, make decisions under constraints, and justify their choices in ways that are harder to outsource meaningfully. Traceability refers to evidence-of-work, such as drafts, checkpoints, process notes, data cleaning records, prompt logs, version histories, and intermediate outputs. Reasoning visibility means that what is rewarded is not only the answer or product, but also the explanation of assumptions, the interpretation of outputs, the defense of method, and the recognition of limitations.

Several recent frameworks already contribute valuable building blocks. The Artificial Intelligence Assessment Scale (AIAS) and its refined later version help educators think in graduated ways about permissible AI integration [9,10]. Other studies report teachers' emerging assessment imaginaries [11], argue for structural redesign rather than superficial policing [8], map transparency mechanisms for GenAI use [12], and propose tailored frameworks for specific disciplinary contexts [13,14]. Recent studies have also highlighted the risk that students outsource the very cognitive work that assessment is intended to reveal [15].

Currently, a complementary line of work is emerging from the practice of higher education specific to each discipline. The edited volume Teaching Innovations in Economics brings together conceptual, empirical, and classroom-based contributions on AI and emerging technologies in economics and related fields [16]. Within that volume, Mir and Pablo-Martí develop a critical

 doi:10.20944/preprints202604.0989.v1

framework for pedagogical evaluation in generative environments that repositions the educator as an epistemic mediator [17], Cabrera et al. examine how AI can support teaching and research in welfare economics, inequality, and poverty [18], and Giménez Baldazo reports applied business education experiences that combine AI literacy, ethical reflection, and active learning [19]. Together, these contributions reinforce the broader movement away from binary positions and toward structured, context-sensitive integration of AI into university teaching and assessment.

2.5. Positioning of the Present Study

This paper is located at that intersection. It does not seek to provide a universal ban-or-allow taxonomy, nor does it assume that one assessment format will solve the problem across disciplines. Instead, it advances a process-based model in which competence is verified through a combination of contextual task design, traceable evidence, and targeted validation moments. The purpose is to help instructors and institutions preserve meaningful assessment judgments without requiring unrealistic levels of surveillance or standardization.

3. Why Product-Only Assessment Has Become Fragile

Product-only assessment refers here to formats in which grading depends predominantly on the final product submitted and where little structured evidence exists regarding how that creation was made. Such formats were always imperfect, but GenAI magnifies their weaknesses in at least five ways.

3.1. Outsourcing of Cognitive Work and Unearned Fluency

GenAI can now produce coherent prose, plausible analysis, structured argumentation, polished summaries, and executable code in seconds. Students may therefore submit work that looks fluent and academically mature without engaging fully in the conceptual labor that the assessment was intended to generate. The risk is not simply cheating in a traditional sense. It is that fluency does not correspond to a better understanding. Grades can then reward polished performance rather than actual learning.

3.2. Hallucinated Content and Fabricated Academic Signals

Another challenge is the production of fabricated but credible-looking content, including invented references, inaccurate citations, spurious interpretations, and overconfident explanations. In product-only workflows, instructors may not have enough visibility into source selection, verification practices, or interpretive reasoning to identify these issues systematically. Practice-oriented discussion in the Spanish higher education context has highlighted exactly this problem, warning that apparently well-referenced submissions may contain nonexistent sources generated to simulate academic credibility [20]. This creates a danger that some academic signals can be mistaken for genuine competence.

3.3. Shortcut Solutions in Quantitative and Computational Tasks

In statistics, programming, economics, finance, engineering, and related disciplines, GenAI can produce apparently correct formulas, code, or outputs while leaving the student unable to explain assumptions, interpret results, debug errors, or justify model choices. The answer may be numerically plausible, yet the underlying procedural and conceptual knowledge remains weak. This problem is especially serious when courses seek to assess transfer, judgment, or problem formulation rather than only final calculation.

 doi:10.20944/preprints202604.0989.v1

3.4. Equity, Access, and Inconsistency

GenAI use also raises fairness issues. Students differ in access to paid tools, familiarity with prompting, prior digital literacies, language backgrounds, and availability of support. A system that silently rewards better AI orchestration may measure social and technological advantages rather than course outcomes. At the same time, inconsistent instructor responses can generate inequitable enforcement, especially when some courses tolerate AI heavily while others penalize similar practices.

3.5. Erosion of Construct Validity

A central risk is that assessment may cease to capture the construct it is intended to measure. When assessment conditions change while task design remains unchanged, grades may begin to reflect unintended dimensions, such as prompting ability, editorial fluency, or differential access to more sophisticated tools. The concern, therefore, extends beyond academic integrity alone. The validity of the assessment itself is at stake, since the institution may no longer be evaluating the knowledge, understanding, or reasoning it claims to assess. This threat becomes more pronounced where tasks rely heavily on the production of a polished final output and provide limited evidence of the student's underlying reasoning or process.

3.6. Design Requirements Derived from These Failures

Taken together, these vulnerabilities imply a set of redesign requirements. First, assessment must generate traceable evidence of student work. Second, important judgments should be verified through moments in which the student explains, defends, or applies their work under limited conditions. Third, tasks should increase contextual specificity and require applied judgment rather than generic exposition. Fourth, AI-use rules should be explicit enough to support fairness and shared expectations. Finally, redesign must remain feasible for real teaching contexts rather than assuming unlimited instructor time.

4. A Process-Based Framework for AI-Aware Assessment

4.1. Overview

The framework proposed in this paper shifts the center of assessment from the final product to an evidential sequence. The claim is not that products no longer matter; they still matter. Rather, products should be interpreted alongside process evidence and validation moments that make reasoning visible. The framework is discipline-neutral in the sense that its principles apply across fields, although the concrete patterns used will differ by learning outcome and task type.

4.2. Core principles

Principle 1: Transparency of AI use. Students should know whether AI use is prohibited, permitted, or expected, for which purposes, and with what disclosure requirements. Transparency reduces ambiguity, supports fairness, and encourages responsible engagement.

Principle 2: Auditability through evidence-of-work. Important judgments should be supported by intermediate traces such as plans, drafts, annotated calculations, data decisions, prompt-use summaries, code snapshots, or design rationales. Auditability does not require exhaustive monitoring; it requires enough structured visibility to make assessment claims more defensible.

Principle 3: Visibility of reasoning. Rubrics should reward explanation, justification, interpretation, and methodological defense rather than mere appearance. This principle is fundamental for distinguishing well-founded understanding from generated fluency.

Principle 4: Contextual Authenticity. Tasks should include local data, course-specific constraints, applied scenarios, personal decisions, and the use of personal or professional contexts that require

 doi:10.20944/preprints202604.0989.v1

contextualized judgment. Authenticity does not eliminate the use of AI, but it reduces the value of generically generated responses and increases the need for genuine understanding.

Principle 5: Feasibility and inclusion. Assessment reform must work within realistic constraints of class size, staff time, accessibility needs, and disciplinary practice. A theoretically elegant system that cannot be implemented consistently will not strengthen assessment.

4.3. The process-Based Assessment Cycle

The framework can be represented as a four-stage cycle:

(1) Task specification with AI support. The task instructions establish the learning outcomes, acceptable uses of AI, required work evidence, validation conditions, and rubric logic.

(2) Iterative production. Students produce the task through visible stages, such as proposal, data plan, outline, draft, analysis log, prototype, or progress review.

(3) Validation moments. Instructors incorporate short opportunities for students to explain and defend key decisions. These may take the form of mini-exams, in-class progress review, spot questions, demonstrations, reflections, or oral discussions.

(4) Judgment using reasoning-focused rubrics. Grading combines the quality of the final output with the quality of reasoning, verification, transparency, and process evidence.

4.4. Redesign Patterns Toolkit

Six patterns are especially reusable across disciplines.

Pattern A: Staged submissions. Students submit short but meaningful milestones, such as topic justification, preliminary method, draft interpretation, or prototype output. These stages create evidence and distribute effort over time.

Pattern B: Annotated reasoning and decision logs. Students explain why they selected a method, accepted or rejected AI suggestions, verified sources, or revised a model. This pattern converts hidden decision-making into assessable evidence.

Pattern C: In-class or synchronous progress review. Brief supervised moments are used to test understanding of the ongoing task. These need not dominate the assessment, but they stabilize inference about authorship and comprehension.

Pattern D: Mini-exams or oral validation. Short oral defenses are highly efficient for confirming whether the student understands what was submitted. These can be individual or small-group and targeted rather than exhaustive.

Pattern E: Contextualized and data-bound tasks. Assessment is anchored to local cases, class-generated materials, unique datasets, personal choices, or discipline-specific constraints, making generic AI responses less useful.

Pattern F: Verification-by-design. Particularly important in quantitative subjects, this pattern requires students to check assumptions, interpret anomalies, compare outputs, identify limitations, or debug a deliberately defective solution.

4.5. Failure Modes and Redesign Responses

The toolkit is not meant to be used identically in every course. Instead, each pattern addresses a specific risk profile. For example, outsourced prose is best mitigated through staged drafting and oral defense; fabricated references through verification logs and source annotation; quantitative shortcutting through interpretive checkpoints and debugging tasks. The framework therefore provides a mapping logic rather than a single template.

Table 1. Failure modes and redesign responses.

Failure mode	Threat to validity	High-value redesign response	Typical evidence

 doi:10.20944/preprints202604.0989.v1

Outsourced prose and unearned fluency	Product quality no longer reflects underlying reasoning	Staged drafting + oral validation	Issue framing, outline, draft notes, mini exams
Fabricated references or claims	False academic signals may be rewarded	Source annotation + verification log	Annotated bibliography, checking notes, correction memo
Quantitative shortcutting	Correct-looking output without interpretive competence	Verification-by-design + checkpoint	Commented code, model choice note, diagnostic explanation
Silent AI orchestration advantages	Grades may reflect AI literacy or access rather than outcomes	Transparent AI-use rules + rubric alignment	Disclosure note, prompt summary, decision log
Construct drift	Assessment begins to measure unintended constructs	Rebalance toward process evidence and contextual performance	Checkpoint artefacts, targeted questioning, applied response

5. Practical Toolkit: Prompts, Rubrics, and Transparency Mechanisms

5.1. Designing an AI-Aware Task Brief

An AI-aware task brief should answer five questions clearly. What capability is being assessed? What kinds of AI use are allowed, restricted, or required? What intermediate evidence must be provided? How will understanding be verified? What will the rubric reward? When these questions are answered explicitly, students are less likely to rely on guesswork or inconsistent assumptions.

A robust task brief normally contains: a short statement of purpose; a list of permitted and non-permitted AI uses; an evidence-of-work requirement; a note about validation or follow-up questioning; and a rubric that identifies reasoning, verification, and disclosure as part of quality.

5.2. Reasoning-Focused Rubric Architecture

In many existing rubrics, presentation quality and final correctness overshadow underlying judgment. In an AI-aware design, the rubric should re-balance these dimensions. A useful generic rubric can include: conceptual understanding of the issue; appropriateness of method or approach; transparency of process and AI use; quality of interpretation and justification; verification of claims or outputs; and effectiveness of the final communication.

The aim is not to penalize students for using AI per se. The aim is to ensure that AI support does not replace the evidence needed for academic judgment. Transparency should therefore be treated as a quality dimension rather than only as a compliance issue.

5.3. Transparency Mechanisms

Transparency mechanisms should be proportionate and educationally purposeful. They may include a brief AI-use statement, a prompt summary, an annotated bibliography explaining source verification, a data-cleaning note, a code-comment trail, or a reflective memo describing which suggestions were adopted or rejected. Recent scoping work on transparency mechanisms shows that

 doi:10.20944/preprints202604.0989.v1

many institutions are converging on disclosure-based strategies, but disclosure alone is insufficient unless it is connected to rubric logic and verification.

5.4. Moderation, Workload, and Reliability

A common concern is that process-based assessment will impose excessive workload. That risk is real if redesign is interpreted as "collect everything and read everything". A more feasible approach is selective evidence and sampled validation. Instructors do not need to inspect every intermediate trace with equal intensity. Instead, they can build small but high-value checkpoints, use structured templates, and sample follow-up questions strategically. Reliability can also be improved through shared prompts, common mini-viva question banks, and moderation notes for recurring rubric dimensions.

5.5. Minimal Viable Redesign

For instructors beginning this transition, a minimal viable redesign can be achieved through three changes: first, add one required evidence-of-work product; second, add one short validation moment; third, revise the rubric so that reasoning and verification havemeaningful marks. Even this modest shift significantly improves the defensibility of assessment judgments.

Table 2. Generic reasoning-focused rubric architecture.

Criterion	What excellent performance shows	Why it matters in AI-aware assessment	Suggested weighting
Conceptual understanding	Accurate framing of the problem, concepts, and relevant theory	Reduces the risk that fluency masks misunderstanding	20-25%
Method or approach	Appropriate selection and justified use of method, evidence, or procedure	Shows judgment rather than copied procedure	15-20%
Transparency of process and AI use	Clear account of stages, AI support, and author decisions	Makes the evidential chain visible	10-15%
Interpretation and justification	Explains results, assumptions, trade-offs, and limitations	Rewards reasoning rather than surface polish	20-25%
Verification behavior	Checks outputs, sources, calculations, or claims systematically	Counters hallucinations and shortcutting	15-20%
Communication and disciplinary quality	Presents the work coherently and appropriately for the field	Maintains standards without over-rewarding polish alone	10-15%

6. Applied Illustrations from Higher Education

The following illustrations are not presented as empirical intervention studies. They are worked examples designed to show how the framework can be instantiated across common task types in higher education.

6.1. Illustration 1: Writing-Intensive Essay in Economics or Business

Original task. Students submit a 2,500-word essay discussing the likely effects of a recent policy intervention on inflation, employment, or tourism demand.

 doi:10.20944/preprints202604.0989.v1

Risk profile. A language model can produce a coherent structure, plausible arguments, and even fabricated references with little visibility into the student's own reasoning. The essay may appear sophisticated even when the student cannot justify causal assumptions, weigh competing explanations, or evaluate the quality of evidence.

Redesigned task. The essay is divided into three assessed components: (a) a 300-word issue framing and source plan, (b) a structured argument map identifying the main claim, counterclaim, and evidence strategy, and (c) the final essay with an AI-use disclosure note. After submission, each student completes a five-minute oral validation focused on two decisions made in the essay.

Assessment logic. The first stage checks whether the student can define a question and identify appropriate sources. The argument map makes reasoning visible before stylistic polish enters the process. The oral validation checks whether the student can defend causal logic and discuss why certain evidence was included or excluded. The final essay remains important, but the grade now rests on a broader evidential base.

6.2. Illustration 2: Quantitative Statistics Assignment

Original task. Students analyze a dataset, estimate a regression model, and interpret the results in a short report.

Risk profile. GenAI can generate statistical code and generic interpretations that look credible but may contain methodological errors. Students may submit correct-looking outputs while lacking understanding of variable coding, assumptions, diagnostics, or the practical meaning of coefficients.

Redesigned task. Students submit (a) a pre-analysis note with variable definitions and hypotheses, (b) a commented code file or analysis log, (c) a one-page interpretation report, and (d) a short in-class review in which they explain one diagnostic choice and one limitation of their model.

Assessment logic. The pre-analysis note captures problem formulation before full automation takes over. The code comments or analysis log show the sequence of decisions. The review tests whether students understand why the model was specified as it was and whether they can interpret outputs beyond repeating software-generated language. A verification-by-design element may be added by asking students to critique a deliberately wrong model or compare two model specifications.

6.3. Illustration 3: Team-Based Applied Project in Tourism or Management

Original task. Student teams produce a consultancy-style report and presentation for a local tourism challenge, such as seasonality management or digital marketing strategy.

Risk profile. Teams may rely heavily on AI for ideation, writing, slide production, and market analysis, making it difficult to know how decisions were made and whether all team members developed the intended competencies. Group settings also increase the risk that one member manages the AI workflow while others disengage.

Redesigned task. Teams submit (a) a project charter with role allocation and AI-use expectations, (b) a decision log documenting major design choices, (c) the final report and presentation, and (d) an individual reflective defense in which each student explains one key contribution, one rejected option, and one limitation of the proposal.

Assessment logic. The project charter clarifies expectations and prevents AI use from remaining tacit. The decision log captures the evolution of the project. The individual defense preserves accountability for team learning. The final report is thus interpreted alongside evidence of collaboration, judgment, and reflective ownership.

6.4. Cross-Case Lesson

Across these three illustrations, we find the same design logic. Each redesign introduces a limited number of process traces, at least one validation moment, and a rubric that rewards explanation rather than the result alone. The framework therefore scales not because it eliminates

 doi:10.20944/preprints202604.0989.v1

disciplinary variation, but because it preserves a consistent evidential philosophy across different tasks.

Table 3. Summary of applied assessment illustrations.

Task type	Main GenAI risk	Redesign pattern	Validation moment	Main competence preserved
Policy essay	Generated argument and fabricated references	Issue framing + argument map + final essay	Short oral defence	Causal reasoning and evidence evaluation
Statistics report	Generated code and shallow interpretation	Pre-analysis note + code log + report	In-class model checkpoint	Method selection and result interpretation
Team consultancy project	Uneven participation hidden by AI-assisted production	Project charter + decision log + report	Individual reflective defence	Applied judgment and accountable collaboration

7. Institutional Integration and Policy Alignment

Assessment redesign cannot succeed as a purely individual instructor strategy. If institutional policy remains ambiguous, students will encounter contradictory expectations and staff will bear reform costs unevenly. For this reason, programme-level and institution-level alignment matters.

First, institutions should adopt a small number of shared principles. These might include clarity about permissible AI use, expectation of proportional transparency, support for assessment redesign, and recognition that validity and fairness are more important than blanket prohibition. Second, programme teams should map where and how AI may appropriately enter learning and assessment across the degree. Not every task should respond in the same way. Some assessments may restrict AI substantially because foundational skills are being built; others may integrate AI openly because the learning outcome includes responsible professional use.

Third, institutions should invest in staff capability. Many instructors understand the risk of GenAI but are unsure how to redesign tasks efficiently. Practical toolkits, exemplars, moderation resources, and shared rubric language can reduce reinvention. Fourth, students need explicit AI literacy for assessment. This includes not only technical skill, but also understanding of verification, bias, hallucination, source checking, and the ethical implications of delegated work.

Finally, policy should distinguish between disclosure, misuse, and invalid evidence. Overly punitive models may produce concealment rather than learning. A better approach is to make expectations explicit, require evidence proportionate to the stakes, and reserve misconduct procedures for cases in which students knowingly violate published conditions.

8. Discussion

The framework proposed here responds to a central dilemma in contemporary higher education: institutions cannot assume that AI is absent, but neither can they allow assessment judgments to drift away from actual competence. The most defensible path is therefore not full permissiveness and not exhaustive surveillance, but structured assessment design.

The first conceptual implication is that academic integrity should increasingly be understood as a design issue rather than only a detection issue. Integrity remains important, but it is better supported when tasks generate interpretable evidence than when institutions rely on weak after-the-fact signals. The second implication is that GenAI should not be framed only as a threat. In some contexts, it can support feedback, language access, idea generation, and iterative refinement. What

 doi:10.20944/preprints202604.0989.v1

matters is whether the assessment still captures the learning outcome that the course intends to assess.

The third implication concerns student agency. A process-based model can support more reflective and responsible AI use by asking students to explain how they used AI, why they accepted or rejected suggestions, and how they verified outputs. In this sense, the framework does not simply adapt to AI; it can also educate students for professional environments in which AI-mediated work will be normal but accountability will remain human.

There are, however, limits to what a conceptual framework can claim. The approach developed here still requires empirical testing across disciplines, class sizes, and institutional contexts. Different implementations may produce different workload profiles. In some programmes, oral validation will be highly efficient; in others, structured written checkpoints may be preferable. Accessibility must also remain central. Process evidence and validation moments should be designed so that they do not create unnecessary disadvantage for students with language, disability, or time-related constraints. Inclusion is therefore not an add-on to assessment reform; it is part of its legitimacy.

A further caution concerns overcorrection. The goal is not to burden every assessment with excessive documentation. Poorly designed process requirements can become bureaucratic and reduce time for substantive learning. The stronger interpretation of this framework is therefore selective rather than maximalist: build enough evidence to support valid judgment, but no more than is educationally useful.

9. Conclusions

Generative artificial intelligence has changed the conditions under which university assessment operates. In a context where students can generate polished academic products with unprecedented ease, the traditional reliance on final artifacts as proxies for competence becomes increasingly fragile. This paper has argued that the most sustainable response is not a return to prohibition or a dependence on uncertain detectors, but a shift toward process-based, reasoning-centered, and transparently AI-aware assessment.

The proposed framework contributes a practical bridge between assessment validity, redesign patterns, and institutional implementation. Its central claim is simple: the assessment question should move from policing authorship to verifying competence. To make that shift operational, the paper has outlined five guiding principles, a four-stage process cycle, a toolkit of reusable redesign patterns, reasoning-focused rubric dimensions, and three applied illustrations. Together, these elements offer instructors and institutions a workable pathway for preserving academic standards while engaging constructively with the realities of AI-enabled learning.

Future research should test the framework empirically across disciplines and levels of study, examine student and staff perceptions of fairness and workload, and explore how programme-level assessment ecosystems can be made coherent under routine AI use. Even so, one conclusion is already clear: if GenAI is here to stay, then valid assessment cannot depend on invisible assumptions about how academic work gets produced. It must be designed for the world students now inhabit.

Author Contributions: Conceptualization, L.F.R.-G. , M.G.-B. and C.M.-F.; methodology, L.F.R.-G.; formal analysis, L.F.R.-G., M.G.-B. and C.M.-F.; writing—original draft preparation, L.F.R.-G, M.G.-B. and C.M.-F..; writing—review and editing, L.F.R.-G., M.G.-B. and C.M.-F.; supervision, L.F.R.-G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analysed in this study. Data sharing is not applicable to this article.

 doi:10.20944/preprints202604.0989.v1

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT (OpenAI) for structural support and language refinement. The authors reviewed and edited all generated material and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Giannakos, M.; Cukurova, M.; Papamitsiou, Z.; et al. The promise and challenges of generative AI in education. Behav. Inf. Technol. 2025, 44(11), 2518-2544. https://doi.org/10.1080/0144929X.2025.2456802.
2. UNESCO. Guidance for Generative AI in Education and Research; UNESCO: Paris, France, 2023. https://doi.org/10.54675/EWZM9535.
3. OECD. OECD Digital Education Outlook 2023: Towards an Effective Digital Education Ecosystem; OECD Publishing: Paris, France, 2023. https://doi.org/10.1787/c74f03de-en.
4. Lodge, J.M.; Howard, S.; Bearman, M.; Dawson, P.; and Associates. Assessment Reform for the Age of Artificial Intelligence; Tertiary Education Quality and Standards Agency (TEQSA): Melbourne, Australia, 2023.
5. Lodge, J.M. The Evolving Risk to Academic Integrity Posed by Generative Artificial Intelligence: Options for Immediate Action; Tertiary Education Quality and Standards Agency (TEQSA): Melbourne, Australia, 2024.
6. Ullah, M.; Bin Naeem, S.; Kamel Boulos, M.N. Assessing the Guidelines on the Use of Generative Artificial Intelligence Tools in Universities: A Survey of the World's Top 50 Universities. Big Data Cogn. Comput. 2024, 8, 194. https://doi.org/10.3390/bdcc8120194.
7. Luo, J. A critical review of GenAI policies in higher education assessment: a call to reconsider the "originality" of students' work. Assessment & Evaluation in Higher Education, 2024, 49:5, 651-664, https://doi.org/10.1080/02602938.2024.2309963.
8. Corbin, T.; Dawson, P.; Liu, D. Talk is cheap: Why structural assessment changes are needed for a time of GenAI. Assessment & Evaluation in Higher Education 2025, 50, 1087-1097. https://doi.org/10.1080/02602938.2025.2503964.
9. Perkins, M.; Furze, L.; Roe, J.; MacVaugh, J. The Artificial Intelligence Assessment Scale (AIAS): A framework for ethical integration of generative AI in educational assessment. Journal of University Teaching & Learning Practice. 2024, 21, Article 6. https://doi.org/10.53761/q3azde36.
10. Perkins, M.; Roe, J.; Furze, L. Reimagining the Artificial Intelligence Assessment Scale (AIAS): A refined framework for educational assessment. Journal of University Teaching & Learning Practice. 2025, 22, Article 7. https://doi.org/10.53761/rrm4y757.
11. Karunaratne, T.; Linblad, L. Imagining assessment futures through artificial intelligence in higher education teachers' perspectives. Discover Education 2025, 4, 532. https://doi.org/10.1007/s44217-025-00987-5.
12. Perez-Perez, I.; Gonzalez-Afonso, M.C.; Plasencia-Carballo, Z.; Perez-Jorge, D. Transparency mechanisms for generative AI use in higher education assessment: A systematic scoping review (2022-2026). Computers 2026, 15 (2), 111. https://doi.org/10.3390/computers15020111.
13. Ilieva, G.; Yankova, T.; Ruseva, M.; Kabaivanov, S. A framework for generative AI-driven assessment in higher education. Information 2025, 16 (6), 472. https://doi.org/10.3390/info16060472.
14. Ahangama, N. Designing assessments in the generative AI era: A tailored assessment framework for ICT tertiary education. International Journal of Educational Technology in Higher Education. 2026, 23, Article 9. https://doi.org/10.1186/s41239-026-00582-0.
15. Chase, A.-M.; Galvin, K. Thinking to learn: Managing the risks of outsourcing to GenAI. Assessment & Evaluation in Higher Education. 2026, 1-20. https://doi.org/10.1080/02602938.2026.2620055.
16. Rivera-Galicia, L.F.; Montero, J.-M.; García-Pérez, C.; Senra-Díaz, E., Eds. Teaching Innovations in Economics: Integrating Artificial Intelligence and Emerging Technologies; Springer: Cham, Switzerland, 2026. https://doi.org/10.1007/978-3-032-08213-8.

 doi:10.20944/preprints202604.0989.v1

17. Mir Fernández, C.; Pablo Martí, F. A critical framework for pedagogical evaluation in generative environments: Integrating heuristic serendipity and assisted materiality in higher education. In Teaching Innovations in Economics: Integrating Artificial Intelligence and Emerging Technologies; Rivera-Galicia, L.F.; Montero, J.-M.; García-Pérez, C.; Senra-Díaz, E., Eds.; Springer: Cham, Switzerland, 2026; pp. 3-23. https://doi.org/10.1007/978-3-032-08213-8_1.
18. Cabrera, A.; García-Pérez, C.; Rivera-Galicia, L.F.; Senra-Díaz, E. Artificial intelligence applied to teaching and research in welfare economics, inequality, and poverty. In Teaching Innovations in Economics: Integrating Artificial Intelligence and Emerging Technologies; Rivera-Galicia, L.F.; Montero, J.-M.; García-Pérez, C.; Senra-Díaz, E., Eds.; Springer: Cham, Switzerland, 2026; pp. 85-102. https://doi.org/10.1007/978-3-032-08213-8_4.
19. Giménez Baldazo, M. When AI takes a seat at the desk: Innovating business education. In Teaching Innovations in Economics: Integrating Artificial Intelligence and Emerging Technologies; Rivera-Galicia, L.F.; Montero, J.-M.; García-Pérez, C.; Senra-Díaz, E., Eds.; Springer: Cham, Switzerland, 2026; pp. 347-364. https://doi.org/10.1007/978-3-032-08213-8_16.
20. Mir, C.; Pablo-Martí, F. Evaluación en tiempos de IA. Perspectivas SCCS 2025, 2505, July. 10.13140/RG.2.2.17473.47201.