Designing Assessments for the New AI-era

Dr. Irene Tsapara, National University

Dr. Irene Tsapara is an academic leader, researcher, and educator specializing in Data Science, Artificial Intelligence, and computational learning theory. She serves as Academic Program Director for the Doctorate in Data Science at National University, where she oversees curriculum design, research supervision, and program accreditation. Dr. Tsapara holds a Ph.D. in Mathematics with specialization in Computational Learning Theory, Machine Learning, and Artificial Intelligence from the University of Illinois. Her current research focuses on the integration of AI-assisted learning frameworks, including vibe coding, into Data Science and interdisciplinary education. She develops scalable curricular models that align AI literacy, statistical reasoning, and ethical governance with foundational mathematical rigor. Her work emphasizes the use of large language models (LLMs) in promoting conceptual understanding, transparency, and reproducibility in computational practice. Dr. Tsapara has over two decades of teaching and programming experience and has contributed to the development of graduate and doctoral curricula across data science, machine learning, and applied analytics. She also collaborates with the National AI Research Institute (TILOS) and several academic-industry partnerships focusing on responsible AI, data governance, and interdisciplinary education. Her professional mission is to reimagine Data Science education as a connective discipline, one that unites computational precision, mathematical depth, and human-centered AI ethics to prepare graduates for leadership in an intelligent, data-driven world.

Andrew D'Amico, Northwestern University

Andrew D'Amico is an entrepreneur, AI systems architect, and applied data science practitioner whose work lies at the intersection of artificial intelligence, systems design, analytics, and engineering education. He is a consultant with Northwestern Analytics Partners, where he contributes to AI and analytics initiatives spanning intelligent decision support, quantitative modeling, and agent-based systems.

He holds a Master of Science in Data Science in Artificial Intelligence from Northwestern University and undergraduate degrees in Philosophy and English Literature from Loyola University Chicago. He also serves as a teaching assistant in Northwestern's Data Science program, supporting instruction in Business Process Analytics, Decision Analytics and Operations Research, Data Engineering, and Quantitative Finance in Data Science.

His current research focuses on two convergent areas. The first is AI-mediated software development, including vibe coding and collaborative programming frameworks that are transforming how programming is learned, guided, and evaluated. The second is assessment and pedagogy in the age of AI, especially questions of authorship, process-based assessment, educational validity, and the preservation of rigorous learning outcomes in environments shaped by generative AI. More broadly, his work examines how institutions can integrate AI in ways that are methodologically sound, educationally credible, and responsive to the changing structure of technical and intellectual labor.

Drawing on both technical and humanistic training, he brings a multidisciplinary perspective to the governance and institutional adoption of AI.

Designing Assessments for the New AI-Era

Irene Tsapara[1], Andrew D'Amico[2],

[1]National University, San Diego, USA, [2]Northwestern University, Evanston, USA

Abstract

The rapid integration of generative artificial intelligence (AI) into higher education has disrupted traditional paradigms of teaching, learning, and particularly assessment. As AI systems such as ChatGPT, Copilot, and Gemini increasingly mediate the production of text, code, and data analysis, conventional models of student evaluation, centered on recall, replication, and closed-ended testing, have become increasingly untenable. This paper argues that higher education must transition from an "AI-policing" stance toward the deliberate design of AI-inclusive and AI-tolerant assessments that preserve academic integrity while advancing student learning, scalability, and agency. Drawing on design frameworks from Jisc [1], UNESCO [2], and the U.S. Department of Education [3], as well as case studies from STEM education, the paper proposes a new typology of hybrid assessment. These models integrate asynchronous AI-amenable components (such as written projects, data analyses, and programming assignments) with synchronous or live performance elements (such as oral defenses, debates, hackathons, and simulations). The hybrid format ensures that students can leverage AI as a creative partner while remaining accountable for understanding, reasoning, and improvisation- skills that cannot be outsourced to generative systems. The study further introduces the concept of process-based evaluation, emphasizing the analysis of a learner's digital footprint - draft versions, reasoning logs, and iteration histories - captured through process-mining tools and learning analytics. The framework operationalizes academic integrity through a Minimum Viable Evidence (MVE) approach, requiring a bounded set of process artifacts, AI-use disclosure, documented revisions, and explicit validation checks that shift evaluation from polished outputs to demonstrable reasoning and professional judgment. This evidence of engagement and cognitive progression provides a richer, more authentic picture of competence than static end-products. Complementing this, the paper advocates for a shift from traditional rubric-based grading, which can be easily gamed by prompting AI tools, to cohort-relative assessment models that benchmark performance against dynamic group patterns rather than fixed descriptors. This approach not only mitigates volatility in grading but also restores fairness by contextualizing achievement within each learning community. Ethical, technical, and logistical implications are critically examined, including concerns about faculty workload, data privacy, and accreditation compliance. The analysis highlights that while hybrid assessments demand greater initial preparation and technological infrastructure, they ultimately foster integrity and efficiency by automating feedback and reducing grading bias. The paper concludes by outlining a roadmap for developing AI-resilient assessment ecosystems: learning environments that embrace generative technologies as collaborators rather than threats, maintain authenticity through live verification, and cultivate reflective, adaptive learners prepared for a digitally augmented world. Accordingly, this work advances a proof-of-work principle for assessment: students are evaluated not only on outputs but on the documented reasoning and iteration that produced them.

1. Introduction

The rapid adoption of generative artificial intelligence (AI) tools such as ChatGPT, Copilot, and Gemini has transformed how students produce written, computational, and analytical work in STEM education. While these tools expand access to cognitive support, they also undermine long-standing assessment assumptions, including individual authorship, observable cognition, and artifact-based evaluation, assumptions that are no longer reliably held in AI-mediated environments. Institutional responses have largely emphasized detection and restriction, approaches that are increasingly brittle, inequitable, and misaligned with professional practice in engineering and data-driven fields. This Work-in-Progress paper argues that assessment is the most vulnerable component of higher education in the AI era and that academic integrity is better preserved through assessment design than prohibition. We assume (i) AI capabilities will continue to improve, making detection increasingly unreliable, and (ii) students will have broad access to AI outside formal assessments and will use it as a routine study tool; accordingly, the paper proposes a hybrid, process-based assessment model that explicitly accommodates AI use while preserving rigor, accountability, and fairness, with early classroom implementations informing future empirical study. Against this backdrop, the paper

introduces a hybrid, process-based assessment model designed explicitly for AI-mediated learning environments. The model treats assessment as a socio-technical system shaped by delivery modality, tooling, and disciplinary practice, rather than as a static instrument applied uniformly across contexts. Assessment design must therefore be robust to ubiquitous AI assistance rather than predicated on its absence. By integrating asynchronous, AI-amenable work with targeted synchronous verification, the model operationalizes integrity through design: it preserves flexibility and access while making reasoning, judgment, and conceptual ownership visible. These framing positions hybrid assessment not as a compromise between in-person and online practices, but as a deliberate response to the limitations of surveillance-based models in an era of pervasive generative AI.

2. Literature Review

2.1. Why Traditional STEM Assessments Fail Under Generative AI

Conventional STEM assessments emphasize final artifacts such as exams, reports, or code submissions. In AI-mediated contexts, however, these artifacts can often be generated or substantially refined with minimal student engagement, causing grades to reflect tool proficiency rather than underlying reasoning or conceptual understanding. This shift produces three systemic failures. First, it erodes construct validity, meaning the assessment no longer measures the intended cognitive construct, such as problem-solving ability, conceptual mastery, or analytical reasoning, but instead captures a student's capacity to prompt, edit, or delegate work to an AI system. Second, static rubrics become vulnerable to prompt optimization, allowing students to meet surface-level criteria without engaging in the targeted learning processes. Third, scalability and trust deteriorate as reliance on detection-based enforcement increases instructional workload and fosters adversarial dynamics between students and instructors [50, 51,53]. These failures signal the need to move assessment away from verifying authorship and toward evaluating process, reasoning, and oversight, particularly in disciplines where computational tools are integral to authentic professional practice. As in mathematics education, where correct answers are not credited without demonstrated reasoning or intermediate work, AI-mediated learning environments require assessment designs that treat process evidence as the primary indicator of understanding.

The UK's Quality Assurance Agency promotes transparency and fairness through external examiner oversight [4], Australia's TEQSA frames AI as a manageable institutional risk [5], and the European Higher Education Area emphasizes validity and transparency through established quality standards [6]. In the United States, the Department of Education highlights equity and instructor workload in AI policy guidance [3], while UNESCO provides a global ethical baseline for responsible AI integration [2, 52]. Collectively, these frameworks converge on assessment redesign and risk management, reinforcing the urgency of moving beyond artifact-centered assessment toward AI-resilient, process-oriented designs. While these frameworks align on transparency, equity, and the need for redesign, they provide limited operational guidance for implementing scalable, AI-inclusive assessment practices in STEM courses, creating an implementation gap that this paper explicitly addresses.

Understanding the evolution of assessment policy clarifies why surveillance-based academic integrity models are increasingly untenable and why assessment redesign, rather than detection, must now serve as the primary mechanism for maintaining academic rigor. Notably,

empirical studies comparing traditional artifact-based assessments with hybrid or process-oriented AI-inclusive models remain limited, particularly in STEM education. This lack of comparative evidence underscores the need for theoretically grounded frameworks accompanied by evaluative research, a direction this study begins to establish.

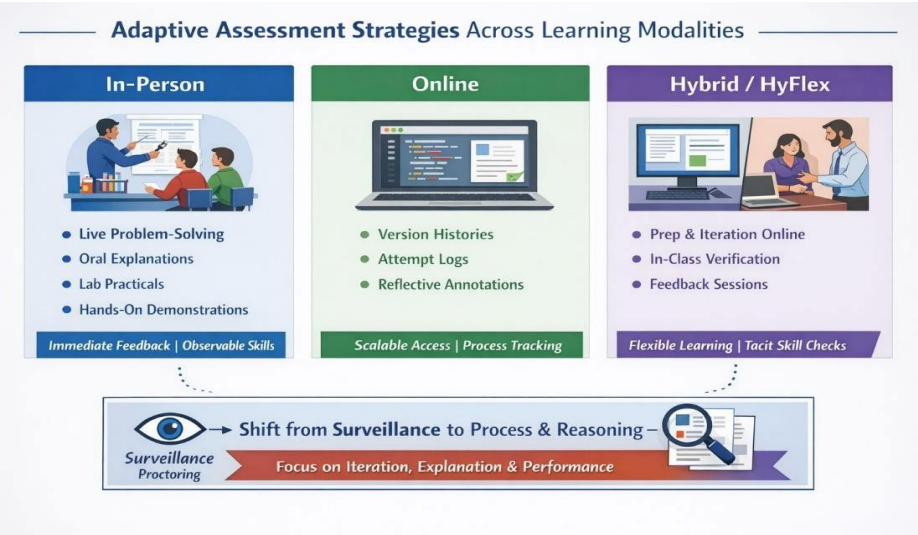
3. Current Framework

Assessment design cannot be treated as separate from the instructional context. In the age of AI, the effectiveness and integrity of assessment depend not only on what is being measured, but also on how, where, and under what conditions students demonstrate their learning.

3.1 How Delivery Mode Shapes Assessment Design

The implications of AI differ across delivery contexts, but in all cases, they push assessment away from surveillance and toward process, reasoning, and performance. Assessment modality plays a critical role in shaping AI-resilient evaluation strategies. In-person environments provide immediate feedback and rich observational data, supporting authentic performance through problem-solving sessions, demonstrations, oral explanations, and lab practicals that emphasize reasoning and improvisation, and are difficult to outsource to AI without intrusive monitoring [47, 48]. Online environments offer scalability, access, and data-rich feedback, making them well-suited for iterative formative assessment when designed around process evidence such as version histories, logs, and reflective annotations rather than closed-book replication. In these contexts, transparent AI use and open-resource tasks are more defensible than surveillance-based proctoring, given documented privacy, bias, and efficacy concerns [7, 8]. Hybrid and HyFlex models combine the strengths of both modalities by reallocating time toward practice and feedback. Online components support preparation and iteration, while in-person components verify tacit skills through demonstrations, critiques, and walkthroughs. Across modalities, the common design principle is to shift integrity from surveillance to observable reasoning, emphasizing evidence of iteration, explanation, and performance under variation. This modality-fit approach balances equity, authenticity, and integrity without reliance on enforcement-based monitoring.

Figure 1. Adaptive Assessment Strategies



Note: Created in OpenAI- ChatGPT

3.2 Cross-Disciplinary Implications and Differences.

Across disciplines, we can see a gradient of AI vulnerability that warrants special attention. This comparative lens underscores the importance of discipline-specific assessment redesign. A one-size-fits-all solution will fail: mathematics may need to shift toward process-rich projects, engineering may need authenticity checks, and life sciences may need hybrid models integrating AI tools responsibly into lab and data-analysis tasks.

Table 1. Relative Vulnerability of Assessment Types to Generative AI

Vulnerability Level	Assessment Context	Rationale
Most vulnerable	Introductory mathematics and calculus exams	Content-heavy, summative assessments emphasizing symbolic manipulation and procedural fluency are often easily replicated by AI tools.
Moderately vulnerable	Programming labs and conceptual problem sets in computer science	AI can generate functional code, but conceptual understanding, reasoning, and oversight during debugging remain important differentiators.
Least vulnerable	Lab-intensive assessments in biology, chemistry, and field sciences	These assessments emphasize physical experimentation, observation, and tacit skills that are more difficult to outsource to generative AI.

Note: Vulnerability reflects the likelihood that assessment artifacts can be produced or optimized by AI with limited student reasoning, rather than the discipline's overall rigor or importance.

STEM education is far from monolithic. Each discipline has developed its own assessment traditions, shaped by epistemology, accreditation demands, and classroom realities. Mapping these disciplinary practices against the proposed taxonomy (topic, level, course goal, type, and theoretical/manual vs. programmatic) helps reveal both their strengths and their vulnerabilities in the era of AI. Accordingly, redesign should be discipline-calibrated: in highly AI-vulnerable contexts (e.g., introductory mathematics), require explicit reasoning traces and periodic live verification; in applied computing, emphasize AI-evaluative critique and debugging oversight; in lab sciences, integrate AI into analysis while preserving embodied performance checks. Detailed comparisons per discipline are included in Appendix B. Where Table 1 compares assessment contexts, Table 2 isolates design dimensions that drive AI susceptibility.

Table 2. Relative AI Risk Across STEM Assessment Dimensions

Assessment Dimension	Mathematics / Physics	Engineering / Computer Science	Life Sciences / Chemistry
Assessment focus (content vs. process)	High	Medium	Medium
Course level (introductory to advanced)	High	Medium	Low
Course goal (certification vs. development)	Medium	Low	Low
Assessment type (summative vs. formative/programmatic)	High	Medium	Medium

Note: Risk levels indicate the relative susceptibility of assessment designs to uncritical or opaque AI substitution. Ratings are conceptual and comparative rather than absolute and are intended to support redesign decisions for assessments. Higher risk corresponds to content-replication, introductory-level, and high-stakes summative assessments, while lower risk corresponds to process-oriented, developmental, and programmatic evaluation structures.

3.3 Why AI Makes Surveillance-Based Assessment Obsolete Toward AI-Resilient Assessment.

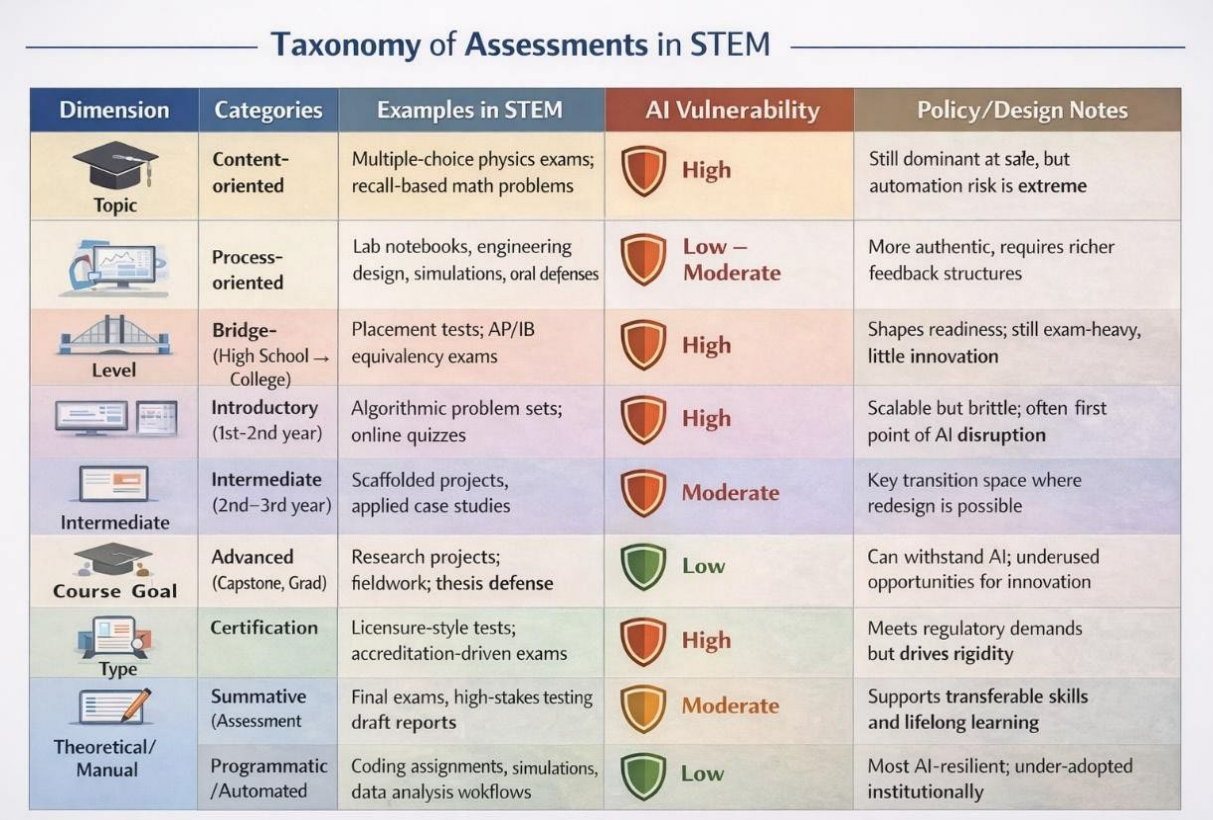
Across instructional modalities, generative AI exposes and accelerates three structural failures of surveillance-based assessment. First, detection is unreliable: even leading AI developers acknowledge limitations, and OpenAI's withdrawal of its text classifier reflects broader concerns about false positives and misuse. Second, surveillance harms equity and trust, as remote proctoring disproportionately affects marginalized students, increases anxiety, and raises unresolved privacy and regulatory concerns [9]. Third, surveillance addresses symptoms rather than design flaws, as AI tools can bypass controls faster than institutions can enforce them. This analysis motivates several actionable shifts aligned with the hybrid, process-based assessment model proposed in this paper. First, assessment should replace surveillance with structure by emphasizing frequent, low-stakes formative activities in online environments and reserving in-person contexts for performance-based verification. Second, evaluation should require process evidence, including commit histories, revision logs, prompt summaries, and reflective annotations, to make reasoning and decision-making visible. Third, assessment policies should be explicit about AI use, permitting transparent, documented support while evaluating student judgment rather than relying solely on artifact quality. Fourth, AI capability is nonstationary: models improve rapidly, and detection or restriction mechanisms will consistently lag. Finally, hybrid and HyFlex modalities should be used intentionally, assigning tacit and performative skills to in-person settings while leveraging online environments for preparation, iteration, and feedback. These shifts reposition assessment as a design challenge rather than a policing task and align academic integrity with learning rather than enforcement.

4. A Taxonomy of Assessments in the AI Era

Assessment redesign requires distinguishing between tasks based on their relationship to AI. Drawing on international policy guidance and established STEM instructional practices, this paper proposes a simplified assessment taxonomy organized into three major dimensions [4, 5, 2]. AI-vulnerable tasks are those in which the primary cognitive demand can be met through direct AI generation or refinement of outputs, such as routine problem solving, standard explanations, or formulaic code, without requiring students to demonstrate independent reasoning or decision-making [10]. AI-resilient tasks are designed so that meaningful understanding is evidenced through observable reasoning, iterative decision-making, and contextual judgment, making AI assistance insufficient without student oversight and validation [11, 12]. AI-integrated tasks explicitly incorporate AI as a supervised collaborator, requiring students to delegate subtasks, evaluate outputs, document corrections, and reflect on limitations, thereby treating AI use as part of the learning objective rather than a threat to academic integrity [13, 14].

Viewed through Bloom's taxonomy, tasks emphasizing recall, routine application, and standard explanation are most susceptible to AI substitution, as these lower-order cognitive processes align closely with current generative capabilities [7, 15, 16]. In contrast, tasks requiring evaluation, justification under constraints, error diagnosis, and creation with defensible tradeoffs engage higher-order cognitive processes that remain more resistant to automation [17, 18]. When paired with proof-of-work evidence such as intermediate drafts, validation logs, reflective annotations, and live or oral verification, these higher-order tasks become substantially more AI-resilient while remaining authentic to disciplinary practice [11, 7, 8].

Figure 2. Taxonomy of Assessments



Note: Created in OpenAI- ChatGPT

This taxonomy clarifies how assessment can be systematically aligned with learning objectives in AI-mediated environments, shifting emphasis from output verification to the evaluation of reasoning processes, oversight capacity, and professional judgment. In Table 3. We introduce Assessment Design Dimensions, which instructors can use as a design checklist to ensure reasoning is observable.

Table 3. Assessment Design Dimensions in AI-Mediated Contexts

Design Dimension	Category	Description
Assessment focus	Product-based	Evaluation emphasizes final artifacts, such as exams, reports, or code outputs.

Design Dimension	Category	Description
AI sensitivity	Process-based	Evaluation emphasizes reasoning, revision history, validation, and explanation.
	AI-vulnerable	Assessment artifacts can be generated or optimized by AI with minimal oversight.
	AI-resilient	Assessment requires reasoning or performance that is difficult to outsource to AI.
	AI-integrated	Assessment explicitly incorporates and evaluates AI use.
Verification mode	Asynchronous	Verification occurs through artifacts, logs, reflections, and other process evidence.
	Synchronous	Verification occurs live through oral explanations or problem-solving.
	Hybrid	Verification combines asynchronous work with live verification.

Note: Across categories, the common redesign move is to require proof-of-work: structured evidence documenting reasoning, iteration, and validation

AI-vulnerable assessments rely exclusively on static artifacts. AI-resilient assessments emphasize reasoning, explanation, and adaptation. If assessment is to evolve in the era of AI, we need a clearer framework for analyzing the types of assessments currently in use and how they map to institutional purposes. A useful way to do this is through a taxonomy that classifies assessments across three dimensions: topic, level, and course goal. This structure highlights not only the diversity of assessment practices but also where they remain repetitive, outdated, or vulnerable to disruption by AI [4, 19]. Detailed explanation of the assessments by topic can be found in Appendix C.

Table 4. Taxonomy of Assessments

Dimension	Categories	Examples in STEM	AI Vulnerability	Policy/Design Notes
Topic	Content-oriented	Multiple-choice physics exams. Recall-based math problems	High	Still dominant at scale, but automation risk is extreme
	Process-oriented	Lab notebooks, engineering design, simulations, oral defenses	Low–Moderate	More authentic, requires richer feedback structures
Level	Bridge (High School → College)	Placement tests. AP/IB equivalency exams	High	Shapes readiness; still exam-heavy, little innovation
	Introductory (1st–2nd year)	Algorithmic problem sets; online quizzes	High	Scalable but brittle; often the first point of AI disruption
	Intermediate (2nd–3rd year)	Scaffolded projects, applied case studies	Moderate	Key transition space where redesign is possible

	Advanced (Capstone, Grad)	Research projects, fieldwork, thesis defense	Low	Can withstand AI; underused opportunities for innovation
Course Goal	Certification	Licensure-style tests. accreditation-driven exams	High	Meets regulatory demands but drives rigidity
	Development	Portfolios, peer evaluations, reflective assignments	Low	Supports transferable skills and lifelong learning
Type	Summative (Assessment of Learning)	Final exams, high-stakes testing	High	Overemphasized in STEM, easy AI targets
	Formative (Assessment for Learning)	Iterative quizzes with feedback; draft reports	Moderate	Promotes learning; needs a cultural shift to scale
	Self-regulatory (Assessment as Learning)	Learning journals, self-reflection, peer calibration	Low	Most AI-resilient; under-adopted institutionally
Mode	Theoretical/Manual	Written proofs; hand-solved equations; in-class essays	High	Easily replicated by AI and digital tools
	Programmatic/Automated	Coding assignments, simulations, and data analysis workflows	Moderate	AI can assist, but originality, process, and debugging remain defensible

5. Hybrid Assessment Design (Core Model)

Assessment design is inseparable from the delivery context. Over the past two decades, higher education has undergone successive modality shifts that have reshaped how assessments are administered, monitored, and interpreted. Generative AI is not an isolated disruption but the latest inflection point in this trajectory, underscoring why surveillance-based integrity models are increasingly untenable and why academic rigor must now be carried out by assessment design rather than detection. The central contribution of this work is a hybrid assessment structure that explicitly integrates generative AI into assessment design while preserving academic rigor and integrity. Rather than attempting to restrict or detect AI use, the model pairs AI-amenable asynchronous work with low-overhead synchronous verification, ensuring that students can leverage AI as a cognitive tool while remaining accountable for understanding, reasoning, and judgment. This approach aligns with emerging guidance from educational policy bodies that emphasize redesign over enforcement and reflects established findings in engineering education that performance-based and process-oriented assessments more accurately capture higher-order learning outcomes than static, product-focused evaluations [1, 3]. Please consult Appendix A for a complete historical timeline of modality.

5.1 Asynchronous, AI-Amenable Components

Asynchronous assessment components are designed to reflect authentic professional practice, where engineers, analysts, and data scientists routinely rely on computational tools, documentation, and iterative workflows. Within this model, AI use is explicitly permitted and

expected. Assessment emphasis shifts from whether AI is used to how students engage with, validate, and refine AI-supported outputs.

1. **AI-use disclosure**
   A short statement identifying where AI tools were used and for what purpose (for example, brainstorming, code generation, debugging, or explanation).
2. **Revision evidence (3–5 entries)**
   Brief revision log entries describing what changed, why the change was made, and how feedback, errors, or validation informed the revision.
3. **Validation checks**
   At least one explicit validation step demonstrating independent verification of results (for example, unit tests, alternative calculations, sanity checks, or comparison against expected behavior).

Typical asynchronous tasks include programming and debugging assignments, data analysis and modeling, design drafts, technical reports, and written explanations. Rather than evaluating final artifacts in isolation, students are required to submit process evidence alongside completed work. Such evidence may include iterative drafts or version histories (for example, code commits or document revisions), summaries of AI interactions or prompt strategies, annotated corrections, error analyses, and validation steps. These artifacts function as primary assessment data, enabling evaluation of reasoning, revision, and decision-making rather than surface correctness alone. This approach aligns with findings from the learning sciences showing that documenting reasoning, self-explanation, and revision provides a more valid measure of learning than static end products, particularly for complex problem-solving tasks [20, 21]. In AI-mediated contexts, process evidence also serves an integrity function. By requiring transparent documentation of AI use, assessment discourages concealment while rewarding critical judgment, error detection, and responsible tool use. Finally, this model supports broader calls from organizations such as UNESCO and Jisc to adopt AI-inclusive assessment designs that emphasize ethical engagement and learning processes rather than prohibition. A related research direction emerging from this work is the need for a faculty workload measurement study to evaluate how MVE-based asynchronous assessment impacts grading time, feedback quality, and instructional sustainability at scale.

5.2 Lightweight Verification for Ownership, Scalability, and Fairness

5.2.1 Scalability and Fairness Mechanisms

To ensure scalability and equitable treatment across students, verification is applied to MVE artifacts rather than to entire submissions. Verification may be implemented through random sampling, rotating checks, or standardized prompt banks, minimizing evaluation variance while controlling faculty workload. Because all students submit the same MVE components, selective verification does not penalize individuals or introduce hidden standards. Instead, it reinforces transparency and consistency while discouraging concealment of AI use. This approach supports fairness by avoiding blanket surveillance and reducing disproportionate scrutiny of specific students or groups.

5.2.2 Hybrid Verification Components

To complement asynchronous, AI-amenable work, the hybrid model incorporates brief synchronous verification components that operate explicitly as confirmation of MVE evidence,

not as independent high-stakes assessments. These verification activities are intentionally lightweight and time-bounded and may include short oral walkthroughs of submitted solutions, live problem-solving or debugging tasks, and brief code, model, or design explanation sessions.

Rather than replicating traditional high-stakes oral examinations, live verification functions as a targeted conceptual check (TCC), assessing whether students can articulate assumptions, respond to variations, and justify decisions without AI mediation. From an assessment validity perspective, these interactions provide evidence of transfer, flexibility, and conceptual coherence, which are widely recognized indicators of deep learning in engineering and STEM education [22, 23]. Embedding verification within assessment design also reduces reliance on surveillance-based technologies such as remote proctoring. Prior research has shown that such technologies introduce privacy risks, exacerbate student anxiety, and raise equity concerns, while offering limited evidence of improved learning outcomes or academic integrity [9, 7, 8]. By contrast, MVE-based verification prioritizes design transparency and instructor judgment over enforcement, aligning with national guidance on AI-inclusive assessment [3]. However, additional empirical work is needed. Future research should examine reliability across instructors and sections, as well as the impact of lightweight verification on student anxiety and perceived fairness, to ensure these mechanisms enhance learning without introducing unintended burdens [3].

5.3. Process-Based Evaluation and Learning Footprints

A defining feature of the proposed hybrid model is its shift from product-centered grading toward process-based evaluation. Rather than treating the final artifact as the primary indicator of learning, evaluation focuses on the reasoning, iteration, and verification activities that lead to that artifact. We formalize this approach through a proof-of-work principle, in which academic credit is earned through documented reasoning, revision, and validation steps, rather than solely through polished or correct outputs. This principle operationalizes learning as an observable process, making cognitive engagement, judgment, and decision-making explicit and assessable. Process-based evaluation is grounded in established research demonstrating that evidence of self-explanation, iterative refinement, and feedback use provides a more valid measure of learning than end products alone, particularly for complex problem-solving tasks [20, 21]. In applied STEM and engineering contexts, documenting how a solution was developed offers stronger evidence of conceptual understanding, transfer, and adaptability than correctness in isolation [22, 23, 24]. Within AI-mediated learning environments, the proof-of-work principle also serves an integrity function. By requiring transparent documentation of reasoning and verification, the model discourages concealment of AI use while rewarding critical evaluation, error detection, and responsible tool engagement. This aligns with emerging guidance that emphasizes assessment designs centered on learning processes and ethical AI use rather than surveillance or prohibition [2, 3].

5.3.1 What Is Evaluated

The core evaluation dimensions are included and explained in Table 5.

Table 5. Process-Based Assessment Criteria and Evidence

Assessment Focus	What Is Evaluated	Mapped MVE Evidence	Rationale and Alignment

Error identification and correction	Students' ability to recognize inaccuracies, inconsistencies, or inappropriate AI-generated outputs and correct them using disciplinary knowledge	Revision log entries; annotated corrections; validation notes	Distinguishes superficial AI use from informed oversight and reflects core engineering judgment and professional accountability
Justification of revisions	Conceptual motivation behind revisions, as evidenced through revision histories, annotations, or logs	3–5 revision log entries describing what changed and why	Ensures that changes are principled rather than cosmetic and demonstrates understanding of underlying assumptions, constraints, and tradeoffs
Conceptual explanation of decisions.	Clarity and coherence of reasoning behind modeling choices, algorithm selection, parameter tuning, or design decisions, conveyed through written or oral explanations.	AI-use disclosure; written explanation or short oral walkthrough.	Provides evidence of conceptual understanding, transfer of knowledge, and the ability to communicate technical reasoning.
Reflection on AI limitations and misuse	Students' ability to identify when AI assistance was inadequate, misleading, or inappropriate, and to explain why.	AI-use disclosure; reflective annotation or brief narrative	Supports metacognitive awareness and ethical reasoning, aligning with accreditation and policy guidance on responsible AI use [2, 3]
Validation practice and robustness checks	Students' use of explicit verification steps, such as alternative calculations, test cases, sanity checks, sensitivity analysis, or comparison against expected behavior	Explicit validation check artifact	Operationalizes the proof-of-work principle by requiring independent confirmation of results, strengthening assessment validity, and demonstrating epistemic responsibility beyond surface correctness

Note: These criteria shift the assessment from evaluating what was produced to evaluating how and why it was produced. Assessment emphasizes reasoning, judgment, and reflective practice rather than artifact quality alone, positioning AI as a tool for learning rather than a substitute for understanding.

This approach aligns with research demonstrating that reflection, self-explanation, and revision are strong indicators of deep learning and self-regulated problem solving [20, 21].

5.3.2 Learning Footprints as Evidence of Engagement

Digital learning environments generate extensive trace data as a byproduct of normal activity. Examples include document version histories, code commits, execution logs, prompt-

response iterations, and submission timestamps. Rather than treating these traces as surveillance data, the proposed model treats them as learning evidence when interpreted transparently and ethically. When used appropriately, learning footprints may reveal patterns of iteration and persistence, distinguish productive struggle from rapid AI substitution, support formative feedback at scale, and reduce ambiguity in grading decisions. Most importantly, the model emphasizes student-visible use of process data, with expectations clearly communicated and students understand how their learning traces contribute to evaluation. This transparency supports trust and reduces the adversarial dynamics associated with hidden monitoring or automated misconduct detection [9, 7, 8]. Instructors should practice data minimization, provide visibility into collected traces, and offer alternative documentation pathways.

5.3.3 Work-in-Progress: Analytics Integration

As a Work-in-Progress, this paper outlines future integration of lightweight learning analytics and process-mining techniques to analyze engagement patterns across cohorts. Rather than functioning as detectors or enforcement tools, these analytics are intended to support instructional decision-making and formative feedback. Planned analytic approaches are focusing on: Aggregated analysis of revision depth and frequency, Identification of common error-correction pathways, Comparison of engagement patterns across assessment modalities, and Cohort-level benchmarking to contextualize individual performance.

Analytics outputs are not used for punitive decisions or automated flagging of misconduct. This cohort-relative perspective aligns with emerging guidance that analytics should be used to improve learning design, not to police behavior [1]. Future empirical work will examine the relationships among process-based indicators, learning outcomes, and students' perceptions of transparency, fairness, and workload. These studies will inform model refinement and support broader adoption across STEM contexts.

5.3.4 Metacognitive Regulation in AI-Mediated Assessment

A critical dimension of process-based evaluation in AI-mediated learning is the assessment of metacognitive regulation, defined as a learner’s ability to plan, monitor, and evaluate their own thinking and tool use [46, 47, 48]. When generative AI systems are available, metacognition becomes especially salient: effective learners must decide when to rely on AI, how to interrogate its outputs, and whether its responses align with disciplinary constraints. In the proposed model, metacognition is assessed through structured reflective artifacts, such as reasoning logs, annotated revisions, and short post-task reflections, that prompt students to justify AI decision-making, identify errors introduced by AI, and articulate corrective strategies.

These artifacts provide observable evidence of planning (e.g., selecting a prompt strategy), monitoring (e.g., detecting implausible or inconsistent outputs), and evaluation (e.g., post hoc reflection on AI limitations), which are core components of self-regulated learning [25, 21]. Metacognitive assessment reshapes both grading and instruction. From an assessment perspective, reflective quality and regulation strategies are weighted alongside technical correctness, reducing the incentive for uncritical AI substitution [41, 45]. From a teaching perspective, aggregated metacognitive patterns inform instructional design decisions, such as where additional scaffolding, conceptual clarification, or modeling of expert reasoning is needed. Prior research demonstrates that explicit attention to metacognition improves transfer, adaptability, and long-term learning outcomes, particularly in complex problem-solving

domains like engineering and data science [26, 20]. In AI-rich environments, assessing metacognition thus functions as both an integrity mechanism and a pedagogical lever, aligning responsible AI use with deeper learning rather than compliance.

5.3.5 Beyond Static Rubrics: Cohort-Relative Evaluation

Static rubrics have long been valued for promoting transparency and consistency in assessment. In AI-mediated learning environments, however, they are increasingly susceptible to optimization rather than understanding. When performance descriptors are fixed and fully knowable in advance, generative AI systems can produce responses that satisfy rubric criteria without requiring meaningful conceptual engagement. This shifts assessment toward surface compliance and increases construct-irrelevant variance, particularly in large or heterogeneous cohorts where patterns of AI use and revision behavior differ substantially [46, 47, 48].

To address these limitations, we propose cohort-relative evaluation as a complementary interpretive layer within the hybrid assessment model. Rather than relying exclusively on static descriptors, student performance is interpreted in relation to empirically observed cohort-level patterns of reasoning, process evidence, and engagement. Assessment remains criterion-informed but is contextualized by evidence of how tasks are completed in AI-rich environments [49]. Importantly, cohort-relative interpretation is not norm-referenced grading and does not impose grade quotas; instead, it supports instructional judgment by situating individual work within documented patterns of practice [43, 44].

Cohort-relative evaluation operates through three principles. First, student work is interpreted relative to cohort-level patterns of documented process evidence, enabling instructors to distinguish typical engagement, high-evidence patterns, and low-evidence patterns based on revision histories, validation checks, and explanatory depth. Second, departures from cohort norms prompt targeted human review rather than automatic penalization [48, 49, 51]. High-evidence patterns may indicate advanced mastery or unusually strong metacognitive regulation, while low-evidence patterns may signal disengagement, excessive reliance on AI, or insufficient validation, without presuming misconduct. Third, instructor judgment is intentionally re-centered while fairness is preserved through transparent criteria, bounded documentation requirements via the Minimum Viable Evidence (MVE) framework, and consistent interpretive standards applied across students.

Grounded in established work in educational measurement and learning analytics, contextualized interpretation has been shown to improve construct validity and reduce noise in assessment data by accounting for variation in problem-solving processes rather than relying solely on final artifacts [27, 28]. In AI-mediated contexts, where traditional assumptions about authorship, effort, and linear workflows no longer reliably hold, cohort-relative evaluation provides a principled mechanism for distinguishing genuine learning from optimized artifact production. This approach does not replace rubrics; rather, it extends them. Rubrics establish baseline expectations, while cohort-relative interpretation accounts for emergent AI-driven behaviors and evolving engagement patterns. Together, these mechanisms support fair, scalable, and integrity-preserving assessment designs aligned with contemporary learning science and responsible AI-use guidance. A detailed illustrative example is provided in Appendix D.

5.3.6 Example: Rubric-Based Assessment (AI-Evaluative Focus)

To illustrate how cohort-relative evaluation operates in an AI-mediated assessment, consider a data science or programming assignment in which students are required to critically evaluate and extend AI-generated solutions rather than submit AI-assisted outputs as final products.

Course context.

The proposed hybrid, process-based assessment model was piloted in an upper-division graduate data science course focused on predictive modeling. Students in the course had prior exposure to machine learning workflows, evaluation metrics, and ethical considerations related to data use. Generative AI tools were permitted for coursework, provided that their use was transparent and documented.

Assessment design.

The primary assessment was a predictive modeling project in which students were tasked with developing, validating, and interpreting a model for a real-world dataset. Rather than prohibiting AI assistance, the assignment was explicitly structured to require critical evaluation of AI-generated outputs.

Evaluation

1. ***An AI-assisted predictive model***

   Students developed an initial model using AI-supported coding or analysis tools. Model performance alone was not the primary grading criterion.

2. ***A revision log documenting corrections and decisions***

   Students maintained a structured revision log that captured AI prompts or tool interactions, identified inaccuracies or omissions in AI-generated outputs, documented missing or inappropriate evaluation metrics, and explained subsequent corrections. Students were also required to note alternative methods suggested by the AI and justify their final methodological choices.

3. ***A brief live explanation of model assumptions and decisions***

   Each student participated in a short live session in which they explained key modeling assumptions, justified metric selection, and discussed limitations of both the AI-generated and final models. These sessions functioned as conceptual verification rather than high-stakes oral exams.

Table 6. Cohort-Relative Evidence Patterns and Assessment Interpretation

Engagement Pattern	Observable Student Behaviors	Assessment Interpretation
Typical pattern	Identifies surface-level AI errors, adds missing evaluation metrics, and makes modest refinements to AI-generated outputs	Indicates adequate engagement and baseline mastery; meets rubric expectations
High-Evidence pattern Positive outlier	Conducts deep critique across multiple AI systems, identifies structural modeling flaws, and proposes alternative approaches grounded in the domain knowledge	Signals advanced reasoning and conceptual mastery; may warrant recognition of higher- level achievement

Low-Evidence Pattern Negative outlier	Reproduces AI outputs with minimal critique, limited questioning, and no substantive additions, despite polished results	Indicates superficial engagement; prompts follow-up verification or targeted feedback rather than automatic penalization

Note: Rather than triggering automatic penalties, negative outliers prompt follow-up verification (e.g., short oral explanation), while positive outliers provide evidence of advanced reasoning and mastery.

5.3.7 Assessment Focus and Preliminary Observations

Evaluation emphasizes validation, interpretability, and reasoning, rather than model performance alone. Grading criteria prioritized the student's ability, anchored to the following four priorities: a) Identify and correct AI-generated errors, b) Select and justify appropriate evaluation metrics, c) Explain modeling decisions and tradeoffs, d) Reflect on limitations of AI assistance

This shift aligned assessment with professional data science practice, where oversight and interpretability are essential competencies. Although formal empirical analysis is ongoing, early instructional observations suggest several promising outcomes [[46, 47, 48]. Students demonstrated improved conceptual articulation during live explanations, particularly when discussing evaluation metrics and model assumptions. Revision logs revealed more frequent error identification and deeper engagement with validation strategies than in previous offerings. Additionally, grading disputes were reduced, as process evidence provided clearer justification for evaluation decisions. These early indicators motivate systematic study in future iterations, including analysis of learning outcomes, student perceptions of fairness, and faculty workload. As a Work-in-Progress, this vignette illustrates the feasibility and instructional value of process-based, AI-inclusive assessment in STEM education.

5.4 Limitations and Boundary Conditions

While the proposed hybrid, process-based assessment model offers important advantages in AI-mediated learning environments, several limitations and boundary conditions must be acknowledged. First, cohort-relative evaluation relies on a sufficient cohort size to establish meaningful patterns of engagement. In very small classes, cohort-level signals may be weaker, limiting the interpretive value of pattern-based comparison. In such contexts, instructors may need to rely more heavily on individual longitudinal evidence, direct verification, or formative dialogue rather than cohort-relative interpretation alone. This aligns with measurement theory, emphasizing that the validity of interpretive inferences depends on the evidentiary context in which they are made [27].

Second, effective implementation assumes instructor familiarity with AI-mediated workflows, disciplinary norms, and common failure modes of generative systems. Without calibration, there is a risk of inconsistent interpretation across instructors or sections. Prior work in engineering education and assessment design highlights the importance of shared exemplars, professional judgment, and instructional alignment in supporting reliable evaluation of complex problem-solving [22, 23, 28].

Third, student perceptions of fairness and transparency must be actively managed. Although cohort-relative evaluation is explicitly not norm-referenced grading, unfamiliar interpretive mechanisms may be misconstrued as curving or peer comparison if not clearly communicated. Research on assessment transparency and student self-regulation emphasizes that clarity of expectations and feedback processes is essential for maintaining trust and

supporting learning [21]. Explicit framing of Minimum Viable Evidence (MVE) requirements and the role of instructor judgment is therefore critical.

Finally, the proposed framework does not eliminate the need for instructor judgment, nor does it detect all forms of inappropriate AI reliance. Instead, it reduces construct-irrelevant variance by shifting evaluation away from polished artifacts toward documented reasoning, revision, and validation practices. This design choice aligns with broader calls to move away from surveillance-based integrity mechanisms, which have been shown to introduce equity, privacy, and anxiety concerns without reliably improving learning outcomes [9, 7, 8], and toward AI-inclusive assessment models that emphasize ethical engagement and learning processes [2, 1, 3]

5.5 Planned Empirical Evaluation

The assessment model described in this section is intended as a design proposal supported by preliminary instructional observations. To establish empirical validity and inform future refinement, several evaluation studies are planned, consistent with design-based research approaches in assessment and learning analytics [27, 28]. Planned analyses include examination of inter-rater reliability to assess the consistency of instructor judgments when applying process-based, cohort-relative evaluation criteria anchored in Minimum Viable Evidence (MVE). This line of inquiry is critical given the central role of instructor judgment in evaluating complex reasoning, revision behavior, and validation practices, and aligns with prior work emphasizing the importance of reliability and calibration in authentic and performance-based assessment [22, 23].

Additional studies will investigate student perceptions of fairness, transparency, and assessment-related anxiety, particularly given the unfamiliarity of cohort-relative interpretation in traditional grading contexts. Prior research has shown that assessment designs emphasizing transparency, feedback, and self-regulation are associated with improved student trust and engagement, while opaque or surveillance-oriented approaches can negatively impact student experience [21, 9, 7, 8]. Finally, learning outcomes will be evaluated by comparing conceptual understanding, transfer performance, and explanation quality across course offerings that employ traditional rubric-based assessment versus the proposed hybrid model. These analyses will examine whether an emphasis on documented reasoning, revision behavior, and validation practices is associated with improved learning outcomes, greater conceptual coherence, and reduced grading disputes. This focus aligns with evidence that learning is more robustly measured through explanation, transfer, and iterative problem-solving than through static end products alone [20, 21]. Results from these evaluations will inform future iterations of the framework and support evidence-based guidance for adoption in AI-rich learning environments, in alignment with emerging policy recommendations that call for AI-inclusive, process-oriented assessment practices grounded in transparency, ethics, and learning science [2, 1, 3].

6. Ethical, Logistical, and Accreditation Considerations

Hybrid, process-based assessment models raise legitimate concerns about faculty workload, student privacy, and alignment with accreditation standards, particularly during initial implementation; however, while redesign requires upfront investment, such models can reduce long-term grading ambiguity by providing clearer evidence of student reasoning and learning progression [21, 22]. From a privacy and equity perspective, process-based assessment

offers a principled alternative to surveillance-driven integrity mechanisms, as remote proctoring and automated detection tools have been criticized for privacy intrusion, algorithmic bias, accessibility barriers, and false-positive misconduct flags, whereas transparency-based approaches rely on student-visible artifacts to preserve accountability without covert monitoring [9, 7, 8]. AI-inclusive, process-based assessment aligns with accreditation expectations, emphasizing ethical reasoning, professional responsibility, and lifelong learning, supporting a transition from enforcement-based integrity to design-based accountability and positioning AI as a catalyst for deeper learning rather than a compliance risk [24, 3, 1, 2].

7. Work-in-Progress Status and Next Steps

This paper presents an early-stage design framework supported by preliminary classroom use and illustrative examples rather than comprehensive empirical validation. The proposed hybrid assessment model is grounded in a proof-of-work principle, in which academic credit is earned through documented reasoning, revision, and validation practices rather than polished end products alone. Central to this approach is the use of Minimum Viable Evidence (MVE), a bounded set of required process artifacts that includes an AI-use disclosure, a small number of revision log entries, and at least one explicit validation or robustness check. Initial implementations demonstrate the feasibility of integrating AI-amenable asynchronous work with lightweight synchronous verification in upper-division STEM contexts. Rather than policing AI use, the model reframes assessment around students' ability to critique, validate, and extend AI-generated outputs using disciplinary knowledge and professional judgment. Process evidence and brief live explanations function as primary indicators of conceptual ownership, adaptability, and ethical engagement, aligning assessment practices more closely with authentic engineering and data science workflows.

As generative and agentic artificial intelligence becomes embedded in STEM education, assessment approaches that rely on detection or surveillance-based integrity mechanisms are increasingly brittle and misaligned with professional practice. In contrast, the proposed model emphasizes transparency, accountability, and metacognitive regulation by making reasoning processes visible and assessable [46]. By shifting evaluation from static artifacts to documented engagement and verification, the framework reduces grading volatility, discourages superficial optimization of AI outputs, and supports integrity through design rather than enforcement.

7.1 Next Steps and Ongoing Work

Ongoing work focuses on extending and empirically validating the proposed framework across instructional contexts. Planned efforts include:

(a) pilot implementations across multiple STEM courses and modalities.
(b) systematic study of student perceptions of fairness, transparency, and workload under process-based, AI-inclusive assessment;
(c) analysis of faculty adoption, calibration, and scalability, particularly in larger or multi-section courses; and
(d) development of lightweight, analytics-supported feedback tools to assist instructors in identifying evidence patterns without increasing grading burden.

Future work will also examine the relationship between proof-of-work evidence density (as operationalized through MVE artifacts) and learning outcomes, including conceptual understanding, transfer performance, and explanation quality. These studies will inform

refinement of evidence requirements, verification strategies, and cohort-relative interpretation practices. As a Work-in-Progress, this paper establishes the design rationale, core mechanisms, and early instructional feasibility of a hybrid, process-based assessment model for AI-rich learning environments. Results from ongoing empirical studies will be reported in a future full paper, contributing evidence-based guidance for assessment practices that preserve rigor, support scalability, and align with ethical and professional expectations in contemporary STEM education.

8. Conclusion

In the new AI era, assessment cannot rely on artifact-based evidence, fixed rubrics, or detection-centered integrity models. As generative and agentic systems increasingly mediate student work, the central task of assessment shifts from verifying authorship to evaluating reasoning, judgment, validation, and oversight.

This paper argues for hybrid, process-based, AI-inclusive designs that preserve rigor while aligning more closely with authentic disciplinary practice. The proposed framework, including live verification, Minimum Viable Evidence, and cohort-relative interpretation, offers a practical path toward more resilient and equitable assessment. Ultimately, the future of assessment in higher education depends not on excluding AI, but on redesigning evaluation so that human understanding, reflection, and professional accountability remain visible and central.

References

[1] Jisc. (2023). Assessment reform for the age of artificial intelligence. Jisc.

[2] UNESCO. (2023). Guidance for generative AI in education and research. United Nations Educational, Scientific and Cultural Organization.

[3] U.S. Department of Education. (2023). Artificial intelligence and the future of teaching and learning. Office of Educational Technology.

[4] QAA. (2023). The rise of artificial intelligence in education: A sector perspective on generative AI. Quality Assurance Agency for Higher Education.

[5] TEQSA. (2023). Generative AI and academic integrity: Information and advice. Tertiary Education Quality and Standards Agency.

[6] ENQA. (2015). Standards and guidelines for quality assurance in the European Higher Education Area (ESG). European Association for Quality Assurance in Higher Education.

[7] Coghlan, S., Miller, T., & Paterson, J. (2021). Good proctor or "Big Brother"? Ethics of online exam proctoring technologies. Open Praxis, 13(2), 1–15.

[8] Coghlan, S., Miller, T., Paterson, J., & Vamplew, P. (2021). Will AI replace human proctors? Examining automated proctoring tools. Computers and Education: Artificial Intelligence, 2, 100016. https://doi.org/10.1016/j.caeai.2021.100016

[9] Swauger, S. (2020). Our bodies encoded: Algorithmic test proctoring in higher education. The Scholarly Kitchen.

[10] Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). ChatGPT: Implications for assessment in higher education. Innovations in Education and Teaching International. Advance online publication. https://doi.org/10.1080/14703297.2023.2190148

[11] Bearman, M., Dawson, P., Ajjawi, R., Tai, J., & Boud, D. (2022). Reconsidering assessment for learning in a time of artificial intelligence. Assessment & Evaluation in Higher Education, 47(1), 93–104. https://doi.org/10.1080/02602938.2021.1953355

[12] Dawson, P., Bearman, M., Boud, D., Tai, J., & Ajjawi, R. (2023). Assessment reform for the age of artificial intelligence. Teaching in Higher Education. Advance online publication. https://doi.org/10.1080/13562517.2023.2208609

[13] Dwivedi, Y. K., et al. (2023). So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges, and implications of generative AI for research, practice, and policy. International Journal of Information Management, 71, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

[14] Bughin, J., Hazan, E., Ramaswamy, S., Chui, M., Allas, T., Dahlström, P., Henke, N., & Trench, M. (2018). Skill shift: Automation and the future of the workforce. McKinsey Global Institute.

[15] Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. Longmans, Green.

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

[17] Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

[18] Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2022). Artificial intelligence and education: Promises and implications for teaching and learning. OECD Publishing.
[19] Redecker, C. (2013). The use of ICT for the assessment of key competences. Joint Research Centre, European Commission.
[20] Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. Educational Psychologist, 49(4), 219–243.
[21] Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning. Studies in Higher Education, 31(2), 199–218.
[22] Wiggins, G. (1998). Educative assessment: Designing assessments to inform and improve student performance. Jossey-Bass.
[23] Prince, M. J., & Felder, R. M. (2006). Inductive teaching and learning methods. Journal of Engineering Education, 95(2), 123–138.
[24] ABET. (2022). Criteria for accrediting engineering programs. ABET Engineering Accreditation Commission.
[25] Zimmerman, B. J. (2002). Becoming a self-regulated learner. Theory Into Practice, 41(2), 64– 70.
[26] Schraw, G., Crippen, K. J., & Hartley, K. (2006). Promoting self-regulation in science education: Metacognition as part of a broader perspective. Metacognition and Learning, 1(1), 111–125.
[27] Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., & Levy, R. (2014). Design and discovery in educational assessment. Journal of Educational Measurement, 51(4), 409–427.
[28] Knight, S., Shum, S. B., & Littleton, K. (2017). Epistemology, pedagogy, assessment and learning analytics. Journal of Learning Analytics, 4(2), 5–18.
[29] Allen, I. E., & Seaman, J. (2013). Changing course: Ten years of tracking online education in the United States. Babson Survey Research Group.
[30] Beatty, B. J. (2019). Hybrid-Flexible course design: Implementing student-directed hybrid classes. EdTech Books.
[31] National Center for Education Statistics. (2022). Distance education in college: Fast facts. U.S. Department of Education.
[32] OpenAI. (2023). Update on AI text classifier. https://openai.com
[33] Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. Computers & Education, 57(4), 2333–2351.
[34] Hora, M. T., & Ferrare, J. J. (2014). Remeasuring postsecondary teaching: How classroom interactions shape STEM student success. Wisconsin Center for Education Research.
[35] Kasneci, E., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274.
[36] Dochy, F., Segers, M., Gijbels, D., & Struyven, K. (2007). Assessment engineering: On the relationship between assessment, learning, and instruction. In Rethinking assessment in higher education (pp. 87–100). Routledge.
[37] Eaton, S. E. (2023). Artificial intelligence and academic integrity: Rethinking assessment in higher education. International Journal for Educational Integrity, 19(1), 1–14.

[38] Sadler, D. R. (2009a). Indeterminacy in the use of preset criteria for assessment and grading. Assessment & Evaluation in Higher Education, 34(2), 159–179.
[39] Shavelson, R. J., Zlatkin-Troitschanskaia, O., Beck, K., Schmidt, S., & Marino, J. (2019). Assessment of university students' critical thinking. International Journal of Testing, 19(4), 337–362.
[40] Earl, L. (2003). Assessment as learning: Using classroom assessment to maximize student learning. Corwin Press.
[41] Biggs, J., & Tang, C. (2011). Teaching for quality learning at university (4th ed.). McGrawHill Education.
[42] Bode, M., Ross, J., & Tsapara, I. (2018). Online grading platforms improve quality and efficiency of grading! In EDULEARN18 Proceedings (p. 10943). IATED Academy. https://doi.org/10.21125/edulearn.2018.2589
[43] Boud, D., & Falchikov, N. (Eds.). (2007). Rethinking assessment in higher education: Learning for the longer term. Routledge.
[44] Boud, D., & Molloy, E. (2013). Feedback in higher and professional education: Understanding it and doing it well. Routledge.
[45] Sadler, D. R. (2009b). Transforming holistic assessment and grading into a vehicle for complex learning. Assessment & Evaluation in Higher Education, 34(1), 1–19.
[46] Xia, Q., Weng, X., Ouyang, F., Chen, J., & Xie, H. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. International Journal of Educational Technology in Higher Education, 21, Article 40. https://doi.org/10.1186/s41239-024-00468-z
[47] Brookhart, S. M. (2013). How to create and use rubrics for formative assessment and grading. ASCD.
[48] Andrade, H. G. (2005). Teaching with rubrics: The good, the bad, and the ugly. College Teaching, 53(1), 27-31.
[49] Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. Educational Research Review, 9, 129-144.
[50] Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1-73. https://doi.org/10.1111/jedm.12000
[51] Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-103). American Council on Education/Macmillan.
[52] AACSB International. (2020). 2020 guiding principles and standards for business accreditation. AACSB International.
[53] Kamalov, F., Calonge, D. S., Smail, L., Azizov, D., Thadani, D. R., Kwong, T., & Atif, A. (2025). Evolution of AI in education: Agentic workflows. arXiv. https://doi.org/10.48550/arXiv.2504.20082

Appendix A.

Assessment Modality in the AI Era: In-Person, Online, and Hybrid Contexts

Assessment design cannot be decoupled from the delivery context. Over the past two decades, higher education has moved through successive modality shifts that have reshaped how assessments are administered, monitored, and interpreted. Generative and Agentic AI represent not an isolated disruption but the latest inflection points in a longer trajectory of modality-driven assessment change.

A.1 Early Scale-Up of Online Learning (2005–2012)

Throughout the 2000s, U.S. higher education steadily expanded distance education offerings. Longitudinal national surveys documented sustained growth in students taking at least one online course and a consolidation of online enrollments among a subset of institutions, signaling a shift from experimentation to infrastructure-level adoption [29].

Despite this growth, assessment practices changed minimally. Most online courses digitized existing classroom assessments by deploying timed quizzes and exams within learning management systems. Academic integrity was primarily addressed through proctoring, either in physical testing centers or via early remote proctoring tools, rather than through fundamental assessment redesign. Assessment remained product-focused and surveillance-dependent.

A.2 Mainstreaming of Blended and HyFlex Models (2013–2019)

During the 2010s, blended and hybrid instructional models matured, and Hybrid-Flexible (HyFlex) designs gained traction, allowing students to choose among in-person, synchronous online, or asynchronous participation modes [30]. These models encouraged pedagogical shifts such as flipped classrooms, increased formative feedback, and iterative practice.

Assessment innovation, however, was uneven. While some courses adopted project-based and low-stakes formative assessments, many large STEM courses retained high-stakes, summative exams, particularly at introductory levels. Assessment design often lagged behind instructional innovation, with integrity concerns still addressed primarily through control mechanisms rather than learning-centered redesign.

A.3 Pandemic Pivot and the Rise of Surveillance Technologies (2020–2021)

The COVID-19 pandemic triggered an unprecedented expansion of remote instruction. National data show that the percentage of undergraduates taking at least one distance education course increased from 36% in 2019 to 75% in 2020, before stabilizing at 61% in 2021 [31]. Exclusive distance enrollment rose from 15% to 44% during the same period.

In response, remote proctoring technologies scaled rapidly. This expansion exposed significant flaws, including privacy intrusions, algorithmic bias, accessibility barriers, and false-positive misconduct flags. Scholarly analyses and public reporting documented psychological stress, equity harms, and limited efficacy, prompting calls for alternatives to surveillance-based integrity models [7, 8, 9].

A.4 Post-Pandemic Hybrid Governance and the AI Shock (2022–2025)

Following the emergency pivot, many institutions retained hybrid and HyFlex capacity as a resilience and access strategy, supported by emerging governance frameworks and design playbooks. At the same time, generative AI fundamentally destabilized detection-led integrity strategies. In 2023, OpenAI discontinued its AI-generated text classifier due to low accuracy, underscoring the unreliability of automated detection tools and reinforcing concerns about false positives and misuse [32].

This convergence of hybrid delivery and AI capability marks a decisive break: assessment integrity can no longer be ensured through surveillance or artifact authentication alone. Instead, integrity must be designed into assessment structures themselves.

Appendix B.

B.2 Mathematics and Physics.

Mathematics and much of physics still rely heavily on theoretical/manual, content-oriented summative assessments: written proofs, symbolic manipulations, and time-limited exams. These practices emphasize certification of individual mastery and scale efficiently to large cohorts, especially in introductory courses. However, they are also among the most AI-vulnerable, as large language models and symbolic solvers can automate proofs, derivations, and problem sets with high accuracy [46] Advanced courses sometimes shift toward open-ended modeling tasks or research projects, but even here, summative exams remain dominant [42].

B.3 Engineering and Computer Science.

Engineering programs tend to emphasize process-oriented and programmatic assessments, including design projects, prototyping, coding labs, and team-based capstones. These align with development-oriented course goals and formative "assessment as learning" approaches. While AI poses risks to routine programming assignments, where tools like GitHub Copilot or ChatGPT can produce working code, process-focused evaluations (design notebooks, oral defenses, collaborative projects) remain relatively resilient [46, 47]. The challenge for these disciplines lies not in obsolescence but in redesigning authenticity tests that distinguish between AI-supported work and genuine student design reasoning.

B.4 Life Sciences and Chemistry.

Life sciences blend content-oriented memorization (terminology, pathways, anatomy) with process-oriented laboratory practice. Summative multiple-choice exams continue to dominate at the introductory level, especially in biology and chemistry "gateway" courses, creating a tension: AI can easily generate flashcards, answer bank-style questions, and simulate problem-solving. Yet lab-based assessments, such as maintaining a notebook, recording observations, and executing experiments, remain AI-resistant because they involve embodied practice and physical environments. Accreditation frameworks (e.g., AAC&U VALUE rubrics for scientific inquiry) already emphasize this process dimension, but scaling authentic lab assessments remains a resource bottleneck [52, 52, 46]

Appendix C.

C.1 By Topic: Content-Oriented vs. Process-Oriented

The most common distinction is between content-oriented assessments, which measure mastery of subject knowledge (e.g., problem sets in calculus, exams testing formula recall), and process-oriented assessments, which evaluate skills such as modeling, experimental design, communication, or teamwork [33]. In STEM fields, summative assessments overwhelmingly favor content-oriented tasks because they are easier to grade reproducibly across large cohorts [34]. Yet these are also the most susceptible to generative AI, since recall and routine problem-solving can often be automated [35]. In contrast, process-oriented assessments, such as design projects, lab notebooks, or oral defenses, demand higher-order reasoning and context-specific application, making them more resilient.

C.2 By Level: Introductory vs. Advanced

Assessments also vary by level of study. At the introductory level, assessments tend to be standardized and summative (multiple-choice exams, algorithmic homework), intended to certify foundational knowledge and to scale across hundreds of students [21]. At advanced undergraduate or graduate levels, assessments often shift toward open-ended inquiry, projects, and research tasks [36]. However, this distinction is not always consistently applied, and many advanced STEM courses still default to exam-heavy models [34]. AI challenges this imbalance by making entry-level assessments especially vulnerable, while simultaneously offering opportunities to redesign them as authentic learning experiences that better prepare students for advanced work [37].

C.3 By Course Goal: Certification vs. Development

Finally, assessments can be aligned with the goals of the courses themselves. Some courses are designed primarily for certification - to demonstrate mastery of core competencies required by professional standards or accreditation frameworks [38]. Others are intended for development, helping students build transferable skills such as problem-solving, critical thinking, or collaboration. Certification-oriented assessments (e.g., professional licensure-style exams) have historically shaped much of STEM education [39], but they also contribute to the rigidity and repetition described earlier. Development-oriented assessments, however, align more naturally with formative and “assessment as learning” approaches, particularly when supported by reflective tasks, portfolios, or collaborative projects [40].

Appendix D.

D.1 Example: Rubric-Based Assessment with Cohort-Relative Overlay (AI-Evaluative Focus)

To illustrate how cohort-relative evaluation operates in an AI-mediated assessment, consider a data science or programming assignment in which students are required to critically evaluate and extend AI-generated solutions rather than submit AI-assisted outputs as final products.

D.2 Baseline Rubric (Criterion-Referenced)

The assignment is evaluated using a transparent rubric, as outlined in Table D1.

Table D1. Rubric Criteria

Criterion	Description
Technical correctness requirements	The final model or solution executes correctly and meets the stated goals.
AI evaluation and completeness	Students evaluate AI-generated outputs for correctness, critique, completeness, and appropriateness.
Validation and metrics	ıdent selects, justifies, and applies appropriate evaluation metrics.
Methodological reasoning	Student explains modeling choices, functions used, and alternatives considered.
Reflection and extension	Student identifies gaps in AI output and articulates original contributions.

D.3 Student Responsibilities (Explicit AI-Evaluative Tasks)

Students are required to submit both a final solution and a structured AI evaluation dossier documenting the following:

- **Review and citation of AI outputs**
  Students cite the AI-generated code, analysis, or explanations they used, clearly distinguishing AI contributions from their own additions.
- **Identification of inaccuracies and omissions**
  Students identify what the AI output got wrong, what was incomplete, and which assumptions were implicit or unjustified. This includes missing edge cases, oversimplified logic, or inappropriate defaults.
- **Clarification and inquiry process**
  Students document the follow-up questions they posed to the AI, including requests for clarification, justification, or alternative approaches. This demonstrates iterative interrogation rather than one-shot prompting.
- **Metrics and validation critique**
  Students analyze which evaluation metrics the AI proposed, which metrics were missing or inappropriate, and justify the inclusion of additional metrics aligned with the problem context.

- **Function and method analysis**
  Students explain which functions, libraries, or algorithms the AI selected, why they were (or were not) suitable, and what alternatives could have been used.
- **Comparison across AI systems**
  Students compare outputs from two AI systems (e.g., ChatGPT and Copilot, or Gemini) and identify differences in assumptions, recommendations, and failure modes.
- **Student-added contributions**
  Finally, students articulate what they contributed beyond AI outputs, including improved validation strategies, alternative models, domain-specific constraints, and refined interpretations of results.

D.4 How Grading Decisions Are Affected

The rubric scores remain the primary basis for grading, but the rubrics will remain generic when shared with students; cohort-relative patterns provide context for interpreting AI evaluative depth. The instructor's judgment is guided by documented evidence of critique, comparison, and extension rather than artifact polish. This approach reduces grading volatility by rewarding disciplined skepticism, validation, and original contribution, even when AI-generated artifacts appear similar across submissions.

D.5 Pedagogical Impact

From an instructional perspective, aggregated cohort data reveal where students struggle to effectively interrogate AI outputs. For example, widespread failure to critique metrics may signal the need for targeted instruction on evaluation strategies. In this way, assessment doubles as a diagnostic tool for improving AI-integrated teaching. This example demonstrates how assessment can shift from detecting AI to evaluating AI oversight. By requiring students to critique, compare, and extend AI-generated work, the assessment design transforms generative AI from a shortcut into a catalyst for deeper learning, metacognitive regulation, and professional judgment.

Appendix E.

E1. Example Assignment: Titanic Survival Classification with Random Forest

Overview

In this assignment, you will participate in the Titanic: Machine Learning from Disaster project by building a classification model that predicts whether a passenger survived the 1912 Titanic disaster. You will use the provided dataset, train.csv, and develop a complete modeling workflow centered on a Random Forest classifier.

This assignment is designed to strengthen your ability to frame a practical analytical problem, prepare data for modeling, evaluate model quality, and reflect critically on your use of Generative AI as a learning partner.

Written Analysis Requirement

Include a 300–500 word discussion of the project and its broader significance. Your response should:

- provide an in-depth analysis,
- include in-text citations and references.

Technical and Presentation Requirements

- *Stage 1: Setup and Reproducibility*

Document AI use, add a revision log entry, and validate that dataset shape, missingness, and target distribution look reasonable.

- *Stage 2: Problem Framing and Initial EDA*

Record a revision log entry and include a validation check for unrealistic values, impossible ranges, and early data risks such as missingness or leakage proxies.

- *Stage 3: Expanded EDA and Feature Plan*

Add a revision log entry and confirm that any previews or transformations do not improperly use the target variable.

- *Stage 4: Cleansing, Preprocessing, and Feature Engineering*

Record preprocessing decisions and validate that all preprocessing is done inside cross-validation using a proper pipeline, not before CV.

- *Stage 5: Baseline Model with Cross-Validation*

Update the AI-use disclosure, justify the validation approach, and confirm that cross-validation is correctly set up without leakage.

- *Stage 6: Hyperparameter Tuning*

Add a revision log entry and verify that tuning is done only on the training data within CV, not on the unseen test set.

- *Stage 7: Model Testing and Improvement Plan*

Include one additional validation check, such as comparing two pipelines or two metric interpretations, to show independent verification.

- *Stage 8: Kaggle Submission*

Optional and not central to the dossier, except as possible supporting evidence of final model performance.

- *Stage 9a: Validation, Metacognition, and AI Dossier*

Part A: Validation Plan (maximum 200 words)

Before model training, write 6–10 sentences addressing:

- The top 3 ways your pipeline could be wrong,
- How you will detect each risk,
- What "suspiciously good performance" would look like,
- What would you check first if that occurs?

Part B: Three Required GAI Dialogues (maximum 400 words total)

Complete three structured dialogues with a GAI tool. These must show iterative reasoning, not one-shot prompting. Each dialogue must lead to a visible action in your notebook.

- ***Dialogue 1: Leakage and Cross-Validation Integrity***
  - Identify leakage risks.
  - Refactor the workflow into a proper pipeline-based CV design.
  - Include:
    - AI summary,
    - your critique,
    - action taken,
    - evidence that the change helped.
- ***Dialogue 2: Metric Critique and Threshold Thinking***
  - Justify your primary metric.
  - Add at least one additional metric.
  - Explain what could be misleading.
  - Include:
    - one metric you adopted,
    - one metric you rejected,
    - your reasoning.
- ***Dialogue 3: Method and Hyperparameter Reasoning***
  - Demonstrate understanding of what each tuned parameter controls.
  - Explain how parameter choices affect overfitting and bias-variance tradeoffs.
  - Include:
    - one concrete tuning change you made because your understanding improved.

Part C: Independent Validation Check (maximum 200 words)

Choose at least one:

- majority-class baseline sanity check,
- two-pipeline comparison with different preprocessing choices but the same model,
- another validation step that independently verifies a claim beyond the AI's suggestion.

Part D: Post-Task Reflection (200–300 words)

Answer all of the following:

- What did you expect to work?

- What evidence forced you to revise your approach?
- Where did GAI help most?
- Where did it mislead or oversimplify?
- Identify one AI output that was incomplete or wrong and explain how you corrected it.
- What will you do differently next time to validate earlier?

- *Stage 9b: AI Dossier*

  Required AI Evaluation Dossier

  Include a clearly labeled section titled Required AI Evaluation Dossier. It must contain:

  - AI-use disclosure: where AI tools were used and for what purpose,
  - Revision evidence: 3–5 entries describing what changed, why, and what evidence triggered the change,
  - Validation checks: at least one explicit validation step,
  - Metrics and validation critique: which metrics AI suggested, which were missing or inappropriate, and what you added,
  - Function and method analysis: what functions, libraries, or algorithms were used, why they were suitable, and what alternatives exist,
  - Student-added contributions: what you contributed beyond AI outputs, such as stronger validation, alternative modeling choices, or refined interpretation.

  You should also retain and submit the dialogue logs generated during your AI interactions as supporting artifacts.

Appendix F.

F1. Validation Template Purpose

This appendix includes a generic Validation and Metacognition Log template for data science modeling assignments. The template is intended to support transparent documentation of both model validation practices and the reflective thinking that accompanies the development process. In addition to prompting students to identify and test possible weaknesses in their pipeline, it requires structured engagement with AI-assisted dialogue, critical evaluation of suggested methods, and reflection on revisions made during the assignment. By pairing technical validation with metacognitive analysis, the template helps make the reasoning behind modeling choices explicit and supports more rigorous, accountable, and self-aware data science practice. While the template is intended to be typed and appended to a written paper, Part D may also be submitted in **handwritten form**, provided the required components are presented clearly and completely.

The Template is provided to the students in the MD form to be appended to their code.

F2. Validation and Metacognition Template for Students

```
## Validation and Metacognition Log

> **Required:** This section must appear once in your RMD/notebook and be clearly labeled.
> **Purpose:** Show your planning, monitoring, and evaluation through structured GAI dialogue and reflection.
> **Important:** Do not duplicate your Dossier here. The **AI Evaluation Dossier** will contain your AI-use disclosure, revision log entries, and validation check documentation.
> In this Log, you will provide **the plan, the dialogue summaries with critique, and the reflection**. For any concrete validation results or final change summaries, point to the relevant subsection in the Dossier.

---

### Part A: Validation plan (complete before model training) - 200 words

Write **6–10 sentences** answering:

1) **Top 3 ways my pipeline could be wrong** (examples: leakage, incorrect CV setup, wrong metric choice, encoding errors):
- Risk 1:
- Risk 2:
- Risk 3:

2) **How I will detect each risk** (specific checks I will run):
- Risk 1 detection method(s):
- Risk 2 detection method(s):
- Risk 3 detection method(s):

3) **What "suspiciously good performance" would look like** and what I would check first:
- Suspicious pattern(s):
- First check(s):

---

### Part B: Three required GAI dialogues (exactly 3) - 500 words

> **Rule:** These dialogues must show iterative interrogation, not one-shot prompting.
> Do not paste assignment instructions into GAI.
> Include summaries, not full transcripts.
> For each dialogue: include **Prompt summary, AI summary, My critique, Action I took, What I will validate (plan)**.
> Detailed revision entries and validation results belong in the **AI Evaluation Dossier**.
```

Dialogue 1: Leakage and cross-validation integrity
Goal: Identify leakage risks and refactor into a Pipeline-based CV design.

- **GAI tool used:**
- **Prompt summary (2–6 lines):**
- **AI summary (3–6 lines):**
- **My critique (2–6 lines):**
  - What I trusted:
  - What I did not trust immediately (and why):
- **Action I took (1–4 lines):**
  - What I changed in preprocessing/CV structure:
- **What I will validate next (1–3 lines):**
  - How I will confirm leakage is reduced:
- **Pointer to Dossier section:** (e.g., “Dossier 2: Revision Entry #1” and “Dossier 3: Validation Check #1”)

Dialogue 2: Metric critique and threshold thinking
Goal: Justify my primary metric, add at least one additional metric, and identify what could be misleading.

- **GAI tool used:**
- **Prompt summary (2–6 lines):**
- **AI summary (3–6 lines):**
- **My critique (2–6 lines):**
  - Metric suggestion I adopted (and why):
  - Metric suggestion I rejected (and why):
- **Action I took (1–4 lines):**
  - What metric(s) I added/changed:
- **What I will validate next (1–3 lines):**
  - What evidence I will use to confirm the metric choice is appropriate:
- **Short threshold note (2–4 sentences):**
  - How threshold choice changes conclusions about performance:
- **Pointer to Dossier section:** (e.g., “Dossier 4: Metrics and validation critique”)

Dialogue 3: Method and hyperparameter reasoning
Goal: Show I understand what tuned parameters control and how they affect overfitting and bias-variance.

- **GAI tool used:**
- **Prompt summary (2–6 lines):**
- **AI summary (3–6 lines):**
- **My critique (2–6 lines):**
  - What clarified something for me:
  - What remained uncertain and required my judgment:
- **Action I took (1–6 lines):**
  - One concrete tuning change I made because I understood the parameter better:
- **Bias–variance note (3–5 sentences):**
  - How my change likely affects bias and variance, and why:
- **What I will validate next (1–3 lines):**
  - What evidence I will use to check for overfitting or instability:
- **Pointer to Dossier section:** (e.g., “Dossier 2: Revision Entry #4” and “Dossier 3: Validation Check #1”)

Part C: Independent validation check (choose at least one) - 200 words

> **Do not paste results here.**
> In this Log, specify which check you chose and why.
> Put the actual procedure, results, and interpretation in the **AI Evaluation Dossier, Section 3: Validation checks**.

- **Chosen validation option (select one):**

  - [ ] Baseline sanity check (majority class baseline)
  - [ ] Two-pipeline comparison (same model, same CV)

- **Why I chose this check (3–6 sentences):**
  [Write here.]

- **Pointer to Dossier section:** (e.g., “Dossier 3: Validation Check #1”)

Part D: Post-task reflection (200–300 words)

Write **200–300 words** answering:

1) What did I assume would work, and what evidence forced me to revise?
2) Where did GAI help most, and where did it mislead or oversimplify?
3) Identify one AI output that was incomplete or wrong, and how I corrected it.
4) What will I do differently next time to validate earlier?

Reflection:
[Write your 200–300 words here.]

Optional: Navigation pointers (recommended)
- Where I revised my pipeline/CV design:
- Where I added/changed evaluation metrics:
- Where I adjusted tuning strategy:
- Where my Dossier begins:

Appendix G

G1. AI Dossier Template Purpose

This appendix presents the **AI Evaluation Dossier**, which is intended to provide a clear and organized account of how generative AI informed the development of the assignment. The dossier documents the purposes for which AI was used, the revisions made during the modeling process, the validation checks applied, and the reasoning behind key methodological choices. It also emphasizes student ownership by requiring reflection on what was changed, what was rejected, and what contributions extended beyond AI output. In this way, the dossier supports transparency, accountability, and critical evaluation in AI-assisted data science work.

G2. AI Dossier Template for students.

```
## AI Evaluation Dossier

> **Purpose:** This section documents how you used GAI (if you did), what you changed as you iterated, and how you validated your work.
> **Requirement:** Complete all subsections below. Keep entries concise but specific.

---

### 1) AI-use disclosure (required)

**GAI tools used (if any):**
- Tool(s): (e.g., ChatGPT, Copilot, Gemini)
- Dates used:
- Primary purposes (check all that apply):
  - [ ] Brainstorming / planning
  - [ ] Explaining concepts
  - [ ] Debugging / error resolution
  - [ ] Code refactoring
  - [ ] Metric selection / interpretation
  - [ ] Hyperparameter tuning strategy
  - [ ] Other:

**Where I used GAI (point to notebook sections):**
- Stage / Section:
  - What I asked for:
  - What I changed after:
- Stage / Section:
  - What I asked for:
  - What I changed after:

**What I did NOT use GAI for:**
- [ ] Copying assignment instructions into GAI
- [ ] Generating an entire end-to-end notebook solution
- [ ] Writing my final management/research response verbatim

**One AI suggestion I rejected or corrected (required):**
- Suggestion:
- Why I rejected/corrected it:
- What I did instead:

---

### 2) Revision evidence (3–5 entries required)

> Each entry must include: **What changed, Why, Trigger/Evidence, What I did next, Outcome/Learning.**
> Aim for 5–10 lines per entry.

#### Revision Log Entry 1
- **What changed:**
- **Why it changed:**
- **Trigger/Evidence:** (error message, metric comparison, stability issue, EDA insight, validation result)
- **What I did next:**
- **Outcome + what I learned:**
```

Revision Log Entry 2
- **What changed:**
- **Why it changed:**
- **Trigger/Evidence:**
- **What I did next:**
- **Outcome + what I learned:**

Revision Log Entry 3
- **What changed:**
- **Why it changed:**
- **Trigger/Evidence:**
- **What I did next:**
- **Outcome + what I learned:**

Revision Log Entry 4 (optional)
- **What changed:**
- **Why it changed:**
- **Trigger/Evidence:**
- **What I did next:**
- **Outcome + what I learned:**

Revision Log Entry 5 (optional)
- **What changed:**
- **Why it changed:**
- **Trigger/Evidence:**
- **What I did next:**
- **Outcome + what I learned:**

---

3) Validation checks (at least one required)

> Choose at least one check and document: **What you checked, How you checked it, Result, What it means.**
> Your validation must be independent (not just "the AI said it is correct").

Validation Check 1 (required)
- **Type (choose one):**
  - [ ] Baseline sanity check (majority class, simple model)
  - [ ] Two-pipeline comparison (different preprocessing, same model, same CV)
  - [ ] Stability check (different seeds/splits)
  - [ ] Feature ablation / permutation importance
  - [ ] Other:
- **What I checked:**
- **How I checked it:** (brief steps)
- **Result:** (metric values, plot summary, or comparison)
- **What it means:** (interpretation, not just numbers)

Validation Check 2 (optional)
- **Type:**
- **What I checked:**
- **How I checked it:**
- **Result:**
- **What it means:**

---

4) Metrics and validation critique (required)

Primary metric used: (e.g., ROC AUC, accuracy, F1)
Why this metric fits this problem: (3–6 sentences)

Metrics I reported (list):
-
-

What the AI suggested (if applicable):
- Suggested metrics:
- Suggested evaluation approach:

What I added or changed and why:
- Added/changed:

- Reason:

One metric or practice I avoided and why (required):
- Avoided:
- Reason:

---

5) Function and method analysis (required)

Core modeling approach (brief):
- Models used: (RandomForestClassifier, ExtraTreesClassifier, GradientBoostingClassifier, etc.)
- Cross-validation design: (StratifiedKFold, folds, random_state)
- Preprocessing approach: (Pipeline, ColumnTransformer, encoders, imputers)

Key functions/classes I used and why (3–8 bullets):
- `Pipeline`:
- `ColumnTransformer`:
- `OneHotEncoder`:
- `SimpleImputer`:
- `StratifiedKFold`:
- `GridSearchCV` or `RandomizedSearchCV`:
- Other:

Alternatives I considered but did not use (2–4 bullets):
- Alternative:
  - Why I did not use it:
- Alternative:
  - Why I did not use it:

---

6) Student-added contributions beyond AI output (required)

> Describe what you contributed that demonstrates ownership and judgment.

- Improvements I made beyond AI output (3–6 bullets):
  -
  -
  -

- Most important decision I made (and why): (4–8 sentences)

---

7) Lightweight verification for ownership (policy notice)

> You may be asked to complete a brief verification check based on your MVE artifacts.
> If selected, you should be able to explain:
> - Where preprocessing occurs and why
> - Why you chose your metric
> - One tradeoff you made
> - One AI suggestion you rejected or corrected
> - One validation check you trust most and why

Biographies

Irene Tsapara, National University & TILOS NSF-sponsored AI-Institute, Chicago, Illinois

Dr. Irene Tsapara is an academic leader, researcher, and educator specializing in Data Science, Artificial Intelligence, and computational learning theory. She serves as the Academic Program Director for the Doctorate in Data Science at National University, overseeing curriculum design, research supervision, and program accreditation. She holds a faculty position at Northwestern University and is also a faculty member of the TILOS NSF-sponsored AI Institute. Dr. Tsapara holds a Ph.D. in Mathematics, specializing in Computational Learning Theory, Machine Learning, and Artificial Intelligence, from the University of Illinois, and has extensive experience in the Financial and Education Sectors. Her experience includes serving as an Investment Analyst at Hedge Fund Research and a Data Scientist at Goldman Sachs. Her current research focuses on integrating AI-assisted learning frameworks, including vibe coding, into Data Science and interdisciplinary education. She develops scalable curricular models that align AI literacy, statistical reasoning, and ethical governance with a foundation in mathematical rigor. Her work emphasizes the use of large language models (LLMs) in promoting conceptual understanding, transparency, and reproducibility in computational practice.