

Article

A Tripartite Feedback Framework for AI-Assisted Assessment of Complex Reports in Higher Education

Demetrios T. Venetsanos 

Department of Aeronautics, Imperial College London, London SW7 2AZ, UK; d.venetsanos@imperial.ac.uk

Abstract

Assessment feedback on complex written reports remains one of the most persistent and resource-intensive challenges in higher education. However, no principled framework exists for deciding which feedback tasks might appropriately involve artificial intelligence and which must remain human responsibilities. This paper addresses that gap by proposing a tripartite feedback framework that distinguishes three analytically distinct levels: low-level structural and presentational feedback, intermediate-level factual content validation, and high-level critical evaluation and synthesis. Grounded in established feedback theory, including Hattie and Timperley's feedback model and Boud and Molloy's sustainable feedback design principles, the framework provides pedagogically justified criteria for allocating tasks between AI systems and human assessors, rather than automating whatever technology can technically perform. Five non-negotiable boundary principles govern any AI involvement at the intermediate level, preserving human oversight, academic accountability, and assessment integrity. This paper examines current technological capabilities and limitations at each level, proposes a phased implementation pathway with explicit human-in-the-loop requirements, and addresses implications for feedback literacy, student agency, equity, and security. A comprehensive mixed-methods evaluation design specifying the evidence required for empirical validation is also presented. The framework's contribution lies not in prescriptive solutions but in providing structured categories, explicit boundary conditions, and validation criteria to guide context-sensitive institutional decision-making about AI integration in assessment.

Keywords: tripartite feedback framework; AI-assisted assessment; human-in-the-loop; feedback literacy; retrieval-augmented generation; higher education

1. Introduction

The transformative potential of feedback in higher education is well established (Hattie & Timperley, 2007), yet its effective provision at scale remains persistently elusive. Academic staff struggle to deliver feedback that is simultaneously timely, comprehensive, and actionable, particularly for complex written assignments such as reports that require judgments across multiple dimensions of performance (Parkinson, 2017; Morris et al., 2021). Students, meanwhile, consistently express dissatisfaction with assessment and feedback practices, citing concerns about timeliness, usefulness, and consistency (Bell & Brooks, 2018; Deeley et al., 2019). This tension between the recognized importance of quality feedback and the practical impossibility of delivering it under current resource constraints represents not merely an operational challenge but a fundamental threat to the quality of learning in contemporary higher education (Nicol & Macfarlane-Dick, 2006; Topping et al., 2025).



Academic Editor: Lucas Kohnke

Received: 26 May 2026

Revised: 4 July 2026

Accepted: 13 July 2026

Published: 1 August 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

The recent emergence of increasingly capable artificial intelligence systems, particularly large language models, has generated both enthusiasm and anxiety about their potential role in assessment (Mollick & Mollick, 2024). Proponents suggest that AI could address challenges in providing feedback by automating routine tasks, enhancing consistency, and freeing academic staff to focus on higher-order pedagogical activities (Chen et al., 2020; Crompton & Burke, 2023). Critics, however, point to substantial limitations, including factual unreliability, an inability to exercise genuine judgment, and risks of undermining the essentially human dimensions of teaching and learning (Zawacki-Richter et al., 2019; Banihashem et al., 2024).

This paper contributes to the evolving debate by proposing a conceptual framework that systematically differentiates feedback tasks by their cognitive demands, pedagogical purposes, and suitability for technological support. This tripartite feedback framework identifies three distinct feedback levels: low (structural and presentational), intermediate (factual content validation), and high (critical analysis and synthesis). It then provides principles for determining which tasks should be handled by AI systems and which by human assessors. Crucially, this differentiation is grounded not merely in technical feasibility but in established theories of feedback and learning.

1.1. Framework Development Approach

The framework presented here emerged through iterative analysis of three intersecting domains: established feedback theory and its implications for learning; empirical evidence on AI capabilities and limitations in educational contexts; and practical challenges in report-based assessment at scale.

The development process involved five interconnected stages: First, a systematic review of the feedback theory literature to identify distinct pedagogical functions that feedback serves in learning contexts. Second, analysis of report assessment practices across disciplines to map specific feedback types to cognitive demands and epistemic status. Third, examination of AI capabilities research to determine current technological boundaries and plausible near-term constraints. Fourth, the identification of boundary conditions in which automation might be pedagogically appropriate while preserving academic judgment. Fifth, the integration of security and equity considerations as design constraints rather than post hoc additions.

This framework is intended to guide principled decision-making about AI integration in assessment by making explicit what should remain human responsibility, what might be supported by technology under specific conditions, and what evidence would be needed to validate such approaches; it is offered as an analytical tool for systematic thinking, rather than a prescriptive implementation blueprint. The value lies in providing structured categories and boundary conditions that can inform local decision-making, rather than in claiming universal applicability.

1.2. Paper Scope and Positioning

This is a conceptual paper that proposes a theoretical framework; it does not present empirical validation or proven effectiveness. This scope is deliberate: the framework establishes conceptual foundations for future implementation and evaluation, not a report on a deployed system. The benefits claimed are theoretical possibilities that require evidence from rigorous evaluation. The framework addresses complex report assessment, i.e., extended written assignments that require students to construct arguments, interpret evidence, and demonstrate disciplinary reasoning. This includes laboratory reports, research reports, design reports, case reports, and similar evidence-rich written assignments across STEM and social science disciplines. Complex reports are defined here as written

assessments that require students to construct and justify claims through the interpretation of evidence, disciplinary reasoning, and a sustained academic narrative, typically integrating multiple components, including data or source material, analytical explanations, interpretations of results, and critical evaluations.

This paper acknowledges current limitations of AI while arguing that a bounded intermediate space exists in which AI might support, but not replace, human assessment if specific conditions are met. The framework is intentionally designed to preserve human judgment and professional accountability throughout, positioning technology as a support for human expertise rather than a substitute for it.

1.3. Paper Structure

The remainder of this paper proceeds as follows: Section 2 reviews the literature on assessment feedback, feedback literacy, and AI-assisted assessment. Section 3 presents the theoretical foundations, pedagogical justification, and boundary conditions of the tripartite framework. Section 4 examines technological capabilities and necessary conditions for AI involvement. Section 5 outlines implementation approaches and safeguards. Section 6 discusses implications for pedagogy, student learning, security, and equity. Section 7 proposes evaluation approaches and validation criteria. Section 8 acknowledges limitations and outlines future research directions, while Section 9 presents the conclusions.

2. Literature Review

2.1. The Persistent Challenge of Assessment Feedback in Higher Education

Assessment and feedback practices are recognized as among the most powerful influences on student learning in higher education, yet they remain consistently problematic areas for both students and staff (Evans, 2013; Morris et al., 2021). A systematic review by Evans (2013) highlighted the persistent tension between the recognized importance of quality feedback and the practical challenges of delivering it effectively at scale. This tension is particularly evident when complex skills and knowledge must be assessed across multiple dimensions, as is typical in report-based assessments, where evaluators must make judgments about technical presentation, factual accuracy, and intellectual work simultaneously.

Despite decades of research and pedagogical development, student dissatisfaction with assessment and feedback remains stubbornly high across the sector. National Student Survey (NSS) data in the UK consistently show that assessment and feedback score lower than other aspects of the student experience across institutions (Bell & Brooks, 2018; MacKay et al., 2019). NSS qualitative data reveal student alienation from assessment processes when feedback arrives too late, lacks specificity about improvement, or fails to explain marking decisions (Bell & Brooks, 2018). This dissatisfaction represents both a quality concern and an engagement issue affecting retention and success.

From the staff perspective, academic workload pressures have intensified markedly in recent years, with assessment and feedback consistently identified as particularly demanding and time-consuming aspects of academic work (Henderson et al., 2019). Staff experience assessment fatigue, struggling to maintain consistency across large cohorts under time pressure while balancing detailed feedback against turnaround requirements. Report-based assessments compound these pressures through evaluation across structural, factual, and analytical dimensions.

2.2. Report Assessment: Distinctive Challenges and Feedback Demands

Reports are widely used in STEM disciplines to assess technical competence and intellectual development; they integrate heterogeneous components: technical presentation

(formatting, structure, referencing), factual content (data, calculations, procedures), and intellectual work (analysis, interpretation, evaluation). This multi-dimensionality challenges assessment and feedback (Parkinson, 2017; Reid & Shah, 2007).

Report assessment faces three persistent difficulties: First, technical and structural elements, e.g., citation formatting, figure captions, and structural consistency, consume substantial marking time despite being mechanical. Second, factual validation, e.g., checking calculations, verifying alignment with references, and assessing procedural adherence, requires careful attention and authoritative materials but not interpretive judgment. Third, evaluating higher-order reasoning and synthesis demands pedagogical expertise and disciplinary knowledge that vary across markers and resist standardization, particularly for open-ended problems (Parkinson, 2017).

These distinct feedback demands create a practical problem: assessors must mentally shift between different modes of evaluation (rule-checking, fact verification, interpretive judgment) while working through a single document, which can be cognitively demanding and time-consuming. This observation motivates the differentiation proposed in this framework.

2.3. Feedback Literacy: Beyond Transmission Models

The concept of feedback literacy has emerged as a critical framework for understanding why feedback interventions succeed or fail, shifting attention from feedback provision to feedback reception and utilization. Carless and Boud (2018) define student feedback literacy as ‘the understandings, capacities and dispositions needed to make sense of information and use it to enhance work or learning strategies’ (p. 1315). Feedback has limited value if students cannot interpret or act upon it. Feedback literacy develops through repeated practice in seeking, interpreting, and using feedback (Carless & Boud, 2018).

Boud and Molloy (2013) distinguish two feedback paradigms: Feedback Mark 1 treats feedback as information transmission, i.e., teachers pointing out errors and providing solutions. Feedback Mark 2 positions students as active seekers and users of feedback, with teachers designing feedback opportunities rather than delivering information. This shift implies that any feedback source, human or automated, must be evaluated based on whether it supports students’ developing evaluative judgment.

Boud and Dawson (2023) empirically derived a teacher feedback literacy framework that demonstrates that effective feedback requires pedagogical knowledge beyond subject expertise: understanding student learning, feedback utilization, assessment design for evaluative judgment, and context-sensitive decisions about timing and framing. Feedback provision requires professional judgment, not technical skill.

The feedback literacy literature raises critical questions for AI-assisted assessment: Can automated systems support feedback literacy or reinforce transmission models? How do students experience feedback from multiple sources (human experts, AI systems, peers)? Does automated feedback support students’ developing evaluative judgment (Boud & Molloy, 2013)? These questions inform framework design and require empirical evaluation.

2.4. AI-Assisted Assessment: Capabilities, Limitations, and Hybrid Approaches

The use of AI in assessment, in the form of automated writing evaluation, rule-based feedback, and generative analysis, has rapidly expanded (Chen et al., 2020; Zhai & Ma, 2023). Research, however, shows mixed effectiveness, with both benefits and significant limitations depending on application.

Evidence suggests that AI-supported feedback can enhance learning in bounded contexts that target surface features, such as grammar, structure, and well-defined criteria (Mertens et al., 2022). Automated writing systems show moderate effectiveness in basic mechanics when supplemented by human assessment (Zhai & Ma, 2023). AI provides

immediate, consistent feedback at scale, addressing timeliness challenges and reducing the burden of routine error checking.

Evidence, however, also shows substantial limitations for higher-order skills, disciplinary reasoning, and evaluative judgment (Deeva et al., 2021; Banihashem et al., 2024). Recent generative AI systems (GPT-5, Claude) have intensified interest in feedback for complex writing. While producing fluent, sophisticated responses, empirical studies report significant concerns limiting appropriate educational use.

Firstly, LLMs confidently generate plausible but factually incorrect information, making them unreliable for verification without extensive oversight (Milani et al., 2025; Fedoruk, 2025), and students may trust AI feedback uncritically. Secondly, AI feedback lacks depth and disciplinary specificity, tending toward generic comments rather than targeted observations (Banihashem et al., 2024). Thirdly, opaque AI reasoning undermines transparency, as students cannot understand the rationale for feedback and teachers cannot verify its appropriateness (Deeva et al., 2021). Fourthly, AI systems may misalign with learning outcomes and assessment criteria despite prompt engineering, producing inconsistent results (Noroozi et al., 2024).

These limitations underscore the risks of AI replacing human assessment. Hybrid approaches that combine AI capabilities with human expertise are more defensible than fully automated systems (Noroozi et al., 2024; Samala et al., 2025). AI supports consistency, efficiency, and scalability for appropriate tasks; humans retain responsibility for judgment and educational decision-making. The question becomes not whether to use AI but which tasks are appropriate and under what conditions.

2.5. Gaps in Existing Frameworks

The preceding review reveals a body of literature that is theoretically rich but operationally incomplete for the challenge of AI-assisted assessment of complex reports. Three substantive gaps warrant direct attention, as each represents a consequential limitation that the tripartite framework is designed to address.

2.5.1. Gap 1: The Absence of a Feedback Taxonomy That Integrates Pedagogical Purpose with Automation Suitability

The most influential feedback frameworks in higher education, i.e., Hattie and Timperley's (2007) four-level model and Boud and Molloy's (2013) sustainable feedback design principles, provide powerful analyses of what feedback should accomplish and how it should be designed to foster student agency. However, both frameworks were developed prior to the emergence of capable AI systems and are consequently silent on a question that has become practically urgent: which specific feedback tasks are appropriately supported by technology, and which must remain human responsibilities? Hattie and Timperley's (2007) framework differentiates feedback by the level at which it operates in learning (task, process, self-regulation, self) but does not address the epistemic status of feedback tasks, whether they involve rule-based checking, factual verification, or interpretive judgment, which is precisely the distinction relevant to automation decisions. Boud and Molloy's (2013) principles foreground learner agency and sustainable assessment design but offer no criteria for deciding when technological involvement is pedagogically appropriate. The consequence of this gap is that institutional decisions about AI integration in assessment are currently made without principled guidance, defaulting either to blanket automation of whatever technology can perform or to blanket rejection driven by anxiety rather than analysis; neither position serves students or institutions well. The tripartite framework addresses this gap by providing an operational taxonomy that explicitly categorizes feedback by epistemic status and suitability for automation, complementing existing frameworks' insights into learning functions and design principles.

2.5.2. Gap 2: The Lack of Concrete, Operationalizable Boundary Conditions for AI Involvement in Assessment

The emerging literature on AI-assisted assessment recognizes the value of hybrid approaches combining AI capabilities with human expertise (Noroozi et al., 2024; Samala et al., 2025), and several scholars have argued that the central question is not whether to use AI but which tasks are appropriate and under what conditions (Chen et al., 2020; Crompton & Burke, 2023); however, this recognition has not yet produced concrete operational guidance. Existing discussions of appropriate AI use in assessment tend to remain at the level of general principles, i.e., AI should support but not replace human judgment, automated feedback should be transparent, and consistency should be monitored, without specifying what conditions must be satisfied before AI involvement is justifiable or what safeguards are non-negotiable. This gap has direct practical consequences: without explicit boundary conditions, institutions implementing AI-assisted assessment lack criteria for distinguishing responsible deployment from pedagogically inappropriate automation; they also lack accountability mechanisms when AI systems produce incorrect or misleading feedback, because the absence of defined conditions leaves no clear standard against which practice can be evaluated. The tripartite framework addresses this gap through five explicitly stated, non-negotiable boundary principles, governing knowledge base curation, human oversight authority, limitation to bounded verification, transparency and attribution, and security, that together constitute operationalizable criteria rather than aspirational guidelines.

2.5.3. Gap 3: Insufficient Critical Engagement with the Pedagogical Implications of Hybrid Human–AI Assessment Systems, Particularly for Feedback Literacy, Student Agency, and Professional Accountability

The feedback literacy literature (Carless & Boud, 2018; Boud & Dawson, 2023) has established that effective feedback provision requires more than accurate information delivery: it depends on students' capacity to interpret, evaluate, and act upon feedback from multiple sources, and on teachers' professional judgment about timing, framing, and pedagogical appropriateness. However, existing discussions of AI-assisted assessment engage only superficially with these dimensions. Critical questions remain largely unaddressed: Does automated feedback reinforce transmission models of feedback (Boud & Molloy, 2013) that position students as passive recipients rather than active evaluators of their own work? How does multi-source feedback, combining human assessment, AI verification, and automated checking, affect students' developing evaluative judgment? When routine assessment tasks are automated, how does this affect the fine-grained attention to student work that grounds academic expertise and pedagogical insight? Who bears accountability when AI systems produce feedback that disadvantages students? These questions are not merely theoretical concerns; they have direct implications for whether AI-assisted assessment systems serve or undermine the educational purposes that they are intended to support. The tripartite framework addresses this gap by systematically examining pedagogical implications at each feedback level, building student agency and feedback literacy development into the framework's design considerations rather than treating them as afterthoughts, and maintaining explicit human professional accountability as a non-negotiable design requirement rather than an optional safeguard.

Together, these three gaps justify the need for a framework that is simultaneously theoretically grounded, operationally concrete, and critically attentive to the human dimensions of assessment. The tripartite framework proposed in Section 3 is designed to address all three.

3. Tripartite Feedback Framework: Theoretical Foundations and Specification

3.1. Pedagogical Rationale for Three Feedback Levels

The framework distinguishes three levels of feedback grounded in distinct pedagogical purposes, cognitive demands, and epistemic status:

Low-level feedback addresses technical and presentational elements, i.e., formatting, referencing, grammar, and clarity, that support professional communication without engaging substantive content or intellectual work. These elements are rule-based, with explicit standards (APA, IEEE, grammatical conventions) against which work can be evaluated without interpretive judgment. This feedback fosters professional communication skills and adherence to disciplinary standards. In the framework of [Hattie and Timperley \(2007\)](#), this corresponds to task-level feedback on surface correctness.

A caveat is needed here: ‘Clarity’ is not a single construct; sentence-level clarity, i.e., grammar, syntax, and word choice, is genuinely surface-level and rule-based; argumentative clarity, i.e., whether an explanation coheres or whether the logical connections between concepts are evident, is often not. An unclear argument frequently signals an unclear or incomplete understanding rather than a presentational lapse. Where lack of clarity reflects lack of conceptual grasp, the underlying issue is high-level, not low-level, and correcting the prose without addressing the reasoning would leave the actual problem untouched. Low-level automation is therefore appropriate for sentence-level clarity, but flagged instances of argumentative unclarity should trigger human review rather than automated correction, since they may be diagnostic of a deeper high-level issue.

High-level feedback addresses intellectual and analytical work: reasoning quality, critical evaluation, synthesis, and disciplinary insight; it evaluates whether students genuinely understand concepts, apply them appropriately, recognize limitations, and contribute original thinking. These judgments are inherently interpretive, requiring evaluative expertise, disciplinary knowledge, and pedagogical judgment about adequate performance within specific contexts ([Boud & Dawson, 2023](#)). This corresponds to Hattie and Timperley’s process-level feedback (learning strategies) and self-regulation-level feedback (self-assessment capacity).

Intermediate-level feedback addresses factual accuracy and procedural correctness, i.e., whether calculations use specified methods correctly, stated facts align with authoritative references, and procedures follow protocols. Unlike low-level feedback, this engages substantive content rather than presentation; unlike high-level feedback, it verifies bounded claims against established knowledge rather than evaluating the quality of reasoning. The cognitive demand is verification through comparison, not interpretive judgment.

This serves task-level functions ([Hattie & Timperley, 2007](#)) but differs from low-level ones in that it engages with content rather than presentation. A calculation can be correctly presented (low-level) while arithmetically wrong (intermediate-level) and methodologically inappropriate (high-level). These three levels represent distinct dimensions, not a hierarchy.

This tripartite taxonomy addresses feedback labor allocation and epistemic status (rule-based/factual verification/interpretive evaluation), providing a basis for human–technology decisions; it differs from Hattie and Timperley’s framework, which addresses learning levels. The same content can receive feedback at multiple levels: a calculation might receive intermediate-level verification (is arithmetic correct?), high-level process feedback (was this the appropriate approach?), and self-regulation feedback (how to check future calculations?). The frameworks are complementary, serving different analytical purposes.

Consider a chemistry laboratory report: Low-level feedback flags a missing table or figure caption. Intermediate-level feedback checks whether a calculated yield matches the

correct formula and reference values, i.e., verification against a known answer. High-level feedback asks whether the student chose an appropriate method for the given reaction conditions and whether they justified that choice, i.e., evaluation with no single correct answer. The intermediate/high boundary sits between these last two: checking '*was the formula applied correctly*' is bounded verification, while checking '*was this the right formula to choose given the experimental constraints*' is interpretive judgment. The same task, method selection, can look like simple fact-checking on the surface while actually requiring disciplinary reasoning underneath. This is why Principle 3 (Section 3.2.3) requires all four bounded-verification criteria to hold simultaneously, not just surface resemblance to a factual check.

One point worth clarifying is why three levels are used: Binary distinctions obscure differences between factual verification and interpretive evaluation. Both engage with substantive content but require different cognitive processes and have different automation implications. Verification compares statements against established knowledge; evaluation requires situated judgment about quality that cannot be reduced to rules. A continuum implies smooth gradation where qualitative distinctions exist. The three levels reflect genuine differences in purpose, cognitive demands, and technological suitability while remaining tractable.

3.2. Boundary Conditions and Principles for AI-Assisted Intermediate-Level Feedback

The intermediate level represents the framework's most contested and carefully bounded space. Unlike low-level feedback, where technological maturity is established, or high-level feedback, where human judgment is unambiguously required, intermediate-level feedback occupies a conditional zone: AI support may be appropriate, but only under specific, demanding conditions that address known limitations and preserve what must not be compromised. The following five principles constitute non-negotiable boundary conditions rather than aspirational guidelines. Each is theoretically grounded, operationally specifiable, and carries defined consequences if violated. Institutions considering AI involvement at the intermediate level must demonstrate that all five conditions are satisfied simultaneously; satisfying only four is insufficient.

3.2.1. Principle 1: Assessor-Curated Knowledge Bases

The knowledge base against which any AI-assisted intermediate-level checking operates must be explicitly defined, bounded, and curated exclusively by qualified human assessors. This includes module-specific materials such as assignment briefs, rubrics, laboratory manuals, worked examples, licensed textbooks, and lecture content relevant to the assessment task. AI systems must not autonomously expand this base, access external sources without explicit assessor validation, or draw on pre-trained parametric knowledge that may be inaccurate, outdated, or misaligned with specific module objectives.

The theoretical rationale for this principle derives directly from the hallucination problem documented in the AI capabilities literature (Milani et al., 2025; Fedoruk, 2025). Large language models generate plausible but factually incorrect information with inappropriate confidence, and the model's sophistication does not mitigate this risk, as it is an architectural property of statistical pattern-matching systems. When AI operates against an open or uncontrolled knowledge base, the probability of incorrect verification feedback reaching students is unacceptably high. Constraining the system to assessor-curated materials, combined with retrieval-augmented generation techniques that ground responses in source documents rather than parametric memory, addresses this risk by bounding the domain within which errors can occur.

Violation of this principle, for instance, allowing AI systems to access external databases or rely on pre-trained knowledge to check student calculations, transforms intermediate-level feedback from bounded verification into unconstrained assertion. The practical consequence is that students may receive confident, authoritative-sounding feedback that is factually incorrect, with no mechanism to detect the error before it reaches them. This would constitute an assessment integrity failure of the most basic kind.

3.2.2. Principle 2: Human Oversight and Final Authority

Human assessors must retain full oversight and unambiguous final authority at every stage of the assessment process in which AI is involved. This means that assessors must review all AI-generated outputs before they reach students; must have absolute and unconditional discretion to override, modify, or reject any AI output; and must remain fully accountable for all feedback received from students, regardless of the technological processes involved in generating it. Any conflict between AI output and human judgment must be resolved in favor of human judgment without exception. Technology supports human expertise; it does not share, distribute, or substitute for professional accountability.

This principle is grounded in Boud and Dawson's (2023) empirical demonstration that effective feedback requires professional pedagogical judgment, including understanding of individual student development, contextual factors affecting performance, and disciplinary norms of adequate reasoning, which cannot be reduced to technical skill or replicated by pattern-matching systems. It is also grounded in the normative commitment, elaborated in Section 4.2, that professional accountability for decisions affecting students' academic standing and opportunities constitutes a core element of higher education's social and ethical purposes, not merely a practical safeguard.

Violation of this principle, for instance, allowing AI outputs to reach students without human review, or treating human oversight as optional once a system has demonstrated sufficient reliability, creates an accountability vacuum with serious consequences for students, assessors, and institutions alike. If incorrect AI-generated feedback disadvantages a student, the absence of clear human authority and accountability leaves that student without meaningful recourse and exposes institutions to both ethical and potential legal liability.

3.2.3. Principle 3: Explicit Limitation to Bounded Verification Tasks

AI involvement must be strictly limited to tasks that satisfy all of the following criteria simultaneously: the assessment criteria are unambiguous and explicitly documented; evaluation involves direct comparison against established knowledge without interpretive judgment; no assessment of alternative valid approaches is required; and there exists either a single correct answer or a well-defined, pre-specified set of acceptable alternatives. Tasks that do not satisfy all four criteria fall outside the scope of appropriate AI involvement, regardless of how routine or straightforward they may appear.

This principle operationalizes the fundamental epistemological distinction at the heart of the tripartite framework: the difference between verification and evaluation. Verification compares a student's stated claim against established knowledge and determines whether it is correct or incorrect; under appropriate architectural constraints, this process may be supported technologically. Evaluation assesses the quality of a student's reasoning, the appropriateness of their methodological choices, and the sophistication of their disciplinary thinking, a process that requires the situated professional judgment that current AI systems fundamentally cannot exercise (Banihashem et al., 2024; Deeva et al., 2021). The boundary between these processes is not always self-evident in practice; a task that appears to involve simple factual checking may contain embedded interpretive dimensions. Conservative de-

fault rules and explicit escalation protocols, in which any ambiguity automatically triggers a transfer to human assessment, are therefore essential implementation requirements.

Violation of this principle, either by extending AI involvement beyond bounded verification or by failing to implement conservative escalation defaults, risks conflating verification with evaluation; the consequence is that students receive AI-generated feedback on tasks requiring professional judgment, with no reliable mechanism to ensure that the feedback is pedagogically appropriate, disciplinarily accurate, or responsive to the individual student's reasoning process.

3.2.4. Principle 4: Transparency and Attribution

Students must have a clear, accessible understanding of the provenance, nature, and limitations of all feedback that they receive. Where feedback originates from different sources, e.g., human assessment, AI-assisted verification, or automated rule-checking, those sources must be clearly distinguished in the feedback itself. Students must be informed what AI was involved in their assessment, for which specific tasks, under what conditions, and subject to what limitations. The primacy of human judgment must be explicitly communicated. Assessment criteria must make clear which dimensions of the work involve verification and which involve evaluative judgment.

This principle is grounded in two distinct theoretical foundations: The first is [Carless and Boud's \(2018\)](#) feedback literacy framework, which holds that students' capacity to interpret and act on feedback depends on understanding where it comes from, how it was generated, and how much weight to give it. Feedback from an AI verification system checking arithmetic carries a different epistemic status from a human assessor evaluating the quality of a student's analytical reasoning; students require explicit support in making this distinction, rather than encountering undifferentiated feedback of mixed provenance. The second foundation is the broader ethical commitment to transparency in assessment that is fundamental to educational trust. Students have a legitimate interest in understanding the processes by which their work is evaluated, and opacity about AI involvement, even well-intentioned opacity aimed at reducing confusion, undermines the relational and institutional trust on which assessment depends.

Violation of this principle, presenting AI-generated feedback without clear attribution, or obscuring the role of technology in assessment processes, risks students placing inappropriate confidence in AI outputs, developing mistaken beliefs about how their work is being evaluated, and being unable to develop the critical engagement with feedback sources that constitutes genuine feedback literacy; it also raises straightforward ethical concerns about institutional honesty with students regarding assessment processes that directly affect their academic outcomes.

3.2.5. Principle 5: Security and Integrity by Design

Robust security against manipulation and prompt injection must be a foundational design requirement, not a retrospective addition. This entails comprehensive input sanitization capable of detecting multiple forms of embedded instruction, including content concealed in white or small text, encoded text (hexadecimal, base64, Unicode), instructions embedded within images, and content hidden in document metadata or comments ([Gulyamov et al., 2026](#); [Sahoo et al., 2025](#)). It further requires continuous monitoring of suspicious submission patterns; clear, enforced academic integrity policies that explicitly address AI system manipulation and impose proportionate sanctions; and regular independent security audits by qualified personnel. Critically, security design must proceed from the assumption of adversarial use: systems must be designed to remain reliable under deliberate attempts at exploitation, not only under good-faith use.

This principle is grounded in empirical evidence that, within large student populations, a meaningful minority will view automated assessment systems as technical challenges to be exploited rather than as educational tools to be engaged with honestly (Gruenhagen et al., 2024; Sahoo et al., 2025). Motivations vary, e.g., academic pressure, intellectual curiosity, competitive peer dynamics, but the consequence is consistent: successful circumvention techniques spread rapidly through student networks, potentially compromising assessment integrity at scale before countermeasures can be deployed. Security is therefore not a peripheral technical concern but a central condition of viability. If a system cannot be made sufficiently secure against adversarial use, the efficiency and consistency benefits of AI involvement are negated by the costs to assessment integrity.

Violation of this principle, by deploying AI-assisted assessment without robust adversarial security design, creates vulnerabilities that motivated students will inevitably discover and exploit. The consequence is not merely individual academic misconduct but also a potential systemic compromise of assessment validity, with a disproportionate impact on students who engage honestly and are disadvantaged relative to those who successfully manipulate the system.

Taken together, these five principles constitute the minimum necessary conditions for pedagogically responsible AI involvement at the intermediate feedback level. They are mutually reinforcing: a system that satisfies Principles 1 and 3 but neglects Principle 5 remains vulnerable to integrity failure; a system that satisfies Principles 1 through 4 but concentrates authority in AI outputs rather than human assessors violates the accountability commitments that underpin the entire framework. The principles should therefore be understood as a set rather than a menu, and institutional decisions about AI-assisted assessment should be evaluated against all five simultaneously.

3.3. Relationship to Existing Frameworks

Situating the tripartite framework within the existing literature requires clarity about what different frameworks are designed to do, since frameworks serving different analytical purposes are complementary rather than competing. Table 1 provides a systematic comparison across six dimensions relevant to the AI-assisted assessment of complex reports: the primary analytical purpose of each framework; whether it addresses the epistemic status of feedback tasks (rule-based, verificational, or interpretive); whether it provides criteria for task allocation between human and automated agents; whether it addresses technological support or AI involvement explicitly; whether it engages with hybrid human–AI accountability; and whether it addresses pedagogical implications such as feedback literacy and student agency in hybrid assessment contexts.

The pattern that emerges from Table 1 is consistent and consequential. Established feedback theories, such as those of Hattie and Timperley (2007), Boud and Molloy (2013), Carless and Boud (2018), and Boud and Dawson (2023), provide the field’s most powerful analytical tools for understanding what feedback should accomplish and how it should be designed for learning. However, none of these frameworks addresses the epistemic status of feedback tasks, which is the dimension most directly relevant to decisions about automation. A framework that explains where feedback operates in learning (Hattie & Timperley, 2007) does not, by itself, tell us whether a particular feedback task involves rule-based checking or interpretive judgment, yet this is precisely the distinction that determines whether technological support is pedagogically appropriate.

Table 1. Comparison of the tripartite feedback framework with established feedback and assessment frameworks across dimensions relevant to AI-assisted assessment.

Framework	Primary Purpose	Addresses Epistemic Status of Tasks	Criteria for Human–AI Task Allocation	Addresses AI/Technology Explicitly	Addresses Hybrid Accountability	Addresses Feedback Literacy in Hybrid Contexts
Hattie and Timperley (2007)	Explains where feedback operates in learning (task, process, self-regulation, self-levels)	No	No	No	No	Partially (self-regulation level)
Boud and Molloy (2013)	Reconceptualizes feedback design for sustainable learning and student agency	No	No	No	No	Yes (central concern)
Carless and Boud (2018)	Defines feedback literacy and conditions for its development	No	No	No	No	Yes (primary focus)
Boud and Dawson (2023)	Empirically derives teacher feedback literacy framework	No	No	No	Partially (professional judgment)	Yes
Chen et al. (2020); Crompton and Burke (2023)	Reviews AI capabilities and potential in educational assessment	Partially	No	Yes	No	No
Noroozi et al. (2024); Samala et al. (2025)	Advocates hybrid human–AI assessment approaches	No	Partially (general principles)	Yes	Partially	No
Tripartite Feedback Framework (this paper)	Provides operational criteria for pedagogically principled task allocation in hybrid human–AI assessment	Yes (central distinction)	Yes (five boundary principles)	Yes (at each level)	Yes (non-negotiable)	Yes (design requirement)

The emerging AI-assisted assessment literature (Chen et al., 2020; Crompton & Burke, 2023; Noroozi et al., 2024; Samala et al., 2025) engages directly with technology but in a different register: it describes capabilities, identifies risks, and argues for hybrid approaches in general terms; what it does not yet provide is an operational framework specifying which tasks are appropriate for which agents under which conditions, i.e., criteria that institutions actually need when making deployment decisions. General advocacy for hybrid approaches, without boundary conditions that define what makes a particular task appropriate for AI involvement, leaves institutions without principled guidance and accountability mechanisms when deployments produce poor outcomes.

The tripartite framework is designed to operate at the intersection of these two bodies of work; it accepts the pedagogical insights established by feedback theory about learning functions, sustainable design, and feedback literacy development, and then it asks the operational question that those theories were not designed to answer: Given these pedagogical commitments and current technological realities, which feedback tasks can appropriately involve AI support, under what conditions, and how would we evaluate effectiveness? In this sense, the framework is positioned not as a replacement for existing models but as a necessary complement that addresses the task allocation question that they leave open.

Two clarifications are worth making explicit. First, the tripartite taxonomy, i.e., low, intermediate, and high levels, is analytically distinct from Hattie and Timperley's (2007) four-level model, which addresses where feedback operates in learning rather than the epistemic status of feedback tasks. The frameworks address different questions and can be used simultaneously without conflict: a single piece of student work can receive intermediate-level verification feedback (is this calculation correct?) and high-level process feedback (was this the appropriate analytical approach?), each serving different pedagogical purposes and requiring different agents. Second, the tripartite framework does not claim that its three levels map neatly onto all disciplines or all report genres. As acknowledged in Section 8, the boundary between intermediate- and high-level feedback may be harder to operationalize in some disciplinary contexts than the taxonomy implies, and empirical work will be needed to test whether the distinctions hold in practice across the range of complex report types that the framework is intended to address.

4. Current Technological Capabilities: Necessary Conditions for Appropriate AI Involvement

This section examines what current AI systems can and cannot reliably achieve in assessment contexts, identifying necessary (though not necessarily sufficient) conditions for appropriate technological involvement at each feedback level.

4.1. Low-Level Feedback: Technologically Mature Domain

For low-level feedback, technological solutions have reached maturity and demonstrated reliability. Tools such as Grammarly (web-based Grammarly Editor, <https://app.grammarly.com/>, accessed on 25 May 2026), Automated Feedback by Feedback-Fruits (<https://feedbackfruits.com/solutions/automated-feedback>, accessed on 25 May 2026), and Zotero 9 (<https://www.zotero.org/>)/EndNote 2025 (<https://endnote.com/>, accessed on 25 May 2026) reliably address grammar, structural requirements, and referencing, respectively, through algorithmic rule-checking with high accuracy.

Main considerations concern institutional implementation, not technological capability, i.e., licensing costs, Virtual Learning Environment (VLE)/Learning Management System (LMS) integration, training, and ensuring that automated feedback complements rather than displaces high-level human feedback. There is a risk that an emphasis on readily automated surface features could distort assessment priorities if not thoughtfully implemented.

4.2. High-Level Feedback: Human Expertise Remains Essential

For high-level feedback, current AI systems present fundamental challenges. Large language models are statistical pattern-matching systems lacking capabilities that are essential for educational feedback: genuine disciplinary understanding, contextual judgment about adequate reasoning at specific educational levels, professional accountability when affecting students' standing and opportunities, and capacity for educational relationships that respond to individual development over time.

Empirical studies consistently find significant limitations in depth, disciplinary specificity, and pedagogical appropriateness (Banihashem et al., 2024; Fedoruk, 2025). While generating fluent, sophisticated commentary, AI feedback tends toward generic observations that are applicable to many submissions rather than targeted, expert, pedagogically informed responses addressing individual student needs.

Given current LLM architectures and near-term deployment constraints in high-stakes assessment, high-level feedback should remain a human responsibility for the foreseeable future. This reflects both technical assessment (current systems cannot reliably exercise the situated professional judgment required for valid evaluation) and normative commitments (professional accountability, pedagogical relationship, and evaluative expertise constitute core elements of higher education's social and ethical purposes that should not be delegated algorithmically, regardless of technical capability). Whether future AI paradigm shifts might change this technical assessment, is outside the scope of the present study. The normative commitment to human professional judgment should remain constant regardless of technical progress.

4.3. Intermediate-Level Feedback: Conditional Viability Under Stringent Constraints

Intermediate-level feedback, i.e., factual content validation through bounded verification, represents the contested technological space where AI systems might conceivably support human assessment, but only under specific, demanding conditions that address known limitations and risks. Major challenges and potential mitigation measures are presented in the following paragraphs.

4.3.1. Challenge 1: Hallucination and Factual Reliability

Large language models generate plausible yet factually incorrect information with unwarranted confidence, a phenomenon known as 'hallucination' (Milani et al., 2025). This occurs because LLMs may misrepresent document content or generate confident explanations without acknowledging uncertainty.

To mitigate this risk, architectural and procedural constraints can be implemented, such as restricting systems to assessor-curated knowledge bases rather than pre-trained knowledge, introducing retrieval-augmented generation to ground responses in source materials, and using multi-pass verification, in which systems check their outputs.

Retrieval-augmented generation (RAG) reduces hallucinations but does not eliminate them. Retrieval can fail: relevant passages may be missed, or irrelevant ones retrieved and treated as authoritative. Even with correct retrieval, models can misread or misapply source content, producing errors that look grounded but are not. The source material itself may contain ambiguities or gaps that retrieval cannot resolve. RAG therefore narrows the space in which errors occur, but it does not guarantee correctness. This is why Principle 2 (human final authority) and Principle 3 (bounded verification only) remain necessary safeguards even where RAG is implemented, not redundant ones.

4.3.2. Challenge 2: Verification Capability Limitations

LLMs generate probable token sequences based on statistical patterns rather than performing logical verification, creating risks such as accepting incorrect work that resembles correct patterns, rejecting unusual correct approaches, or missing uncommon errors.

To mitigate this risk, AI can be framed as flagging issues for human review rather than as authoritative verification; implement hybrid architectures that combine rule-based checks (arithmetic, units, consistency) with language models; maintain humans as final authorities; and design interfaces that surface AI reasoning for review, not verdicts.

4.3.3. Challenge 3: Verification–Interpretation Boundary Ambiguity

The boundary between verification and interpretation, while analytically distinct, can be ambiguous in practice. Tasks that appear to involve straightforward factual checking may contain embedded interpretive dimensions that AI systems lack the contextual judgment to recognize, potentially generating feedback beyond their appropriate scope.

To mitigate this risk, one can develop explicit escalation rules, implement conservative defaults that automatically escalate ambiguous cases to humans, create training with clear versus ambiguous examples, and build systems that express uncertainty and defer.

4.3.4. Challenge 4: Security Vulnerabilities (Prompt Injection)

Students might manipulate AI systems through crafted inputs in submissions (Gulyamov et al., 2026; Sahoo et al., 2025) to include white/small text, content in images, encoded instructions (hexadecimal, base64, Unicode), or metadata/comments.

Multiple countermeasures can address this challenge, such as comprehensive input sanitization with multiple defense layers, monitoring for suspicious patterns, clear academic integrity policies that address manipulation with sanctions, regular security auditing, and design that assumes adversarial use.

Based on the above challenges and potential mitigations, intermediate-level AI support appears viable, provided these demanding conditions can be satisfied simultaneously.

5. Implementation Considerations: Approaches, Safeguards, and Acknowledged Uncertainties

5.1. Phased Implementation Pathway

A systematic implementation approach would follow a staged progression with explicit decision points, as described below.

5.1.1. Stage 1: Low-Level Automation Using Mature Tools

This stage would require deploying commercially available, well-tested tools for structural and presentational feedback. Though relatively straightforward given the technological maturity, this stage requires attention to institutional procurement, integration with existing systems, training for staff and students, and careful communication about what is and is not being automated.

5.1.2. Stage 2: AI-Supported Intermediate-Level Checking Under Human Oversight

This stage concerns the development or adaptation of AI-supported systems for preliminary checking of factual content, explicitly framed as flagging potential issues for expert human review rather than generating autonomous feedback. This stage requires substantial development effort, such as creating and maintaining assessor-curated knowledge bases, implementing security measures, designing interfaces that surface AI reasoning for human review, and establishing clear protocols for human oversight. Crucially, this stage begins

with high-frequency human oversight of all AI outputs, scaling back only as evidence demonstrates acceptable reliability and security.

5.1.3. Stage 3: Human-Generated High-Level Feedback Using Enhanced Tools

In this stage, academic staff will provide critical evaluation, synthesis assessment, and pedagogical feedback using existing VLE/LMS tools and their professional expertise. At this stage, it is possible to consider whether time saved through Stages 1 and 2 (if any) can be strategically reinvested to enhance the depth and quality of high-level feedback, or whether other pedagogical priorities should receive this time.

5.1.4. Stage 4: Integrated Feedback Delivery with Clear Attribution

This stage concerns integrating feedback from all stages into coherent assessment responses delivered to students, with clear attribution of sources and an explanation of the purposes and limitations of different types of feedback. Notably, the entire approach should support student feedback literacy development by making explicit what different feedback sources can and cannot tell them about their work.

A flowchart that seamlessly implements the proposed procedure for assessing and providing feedback on a report is shown in Figure 1.

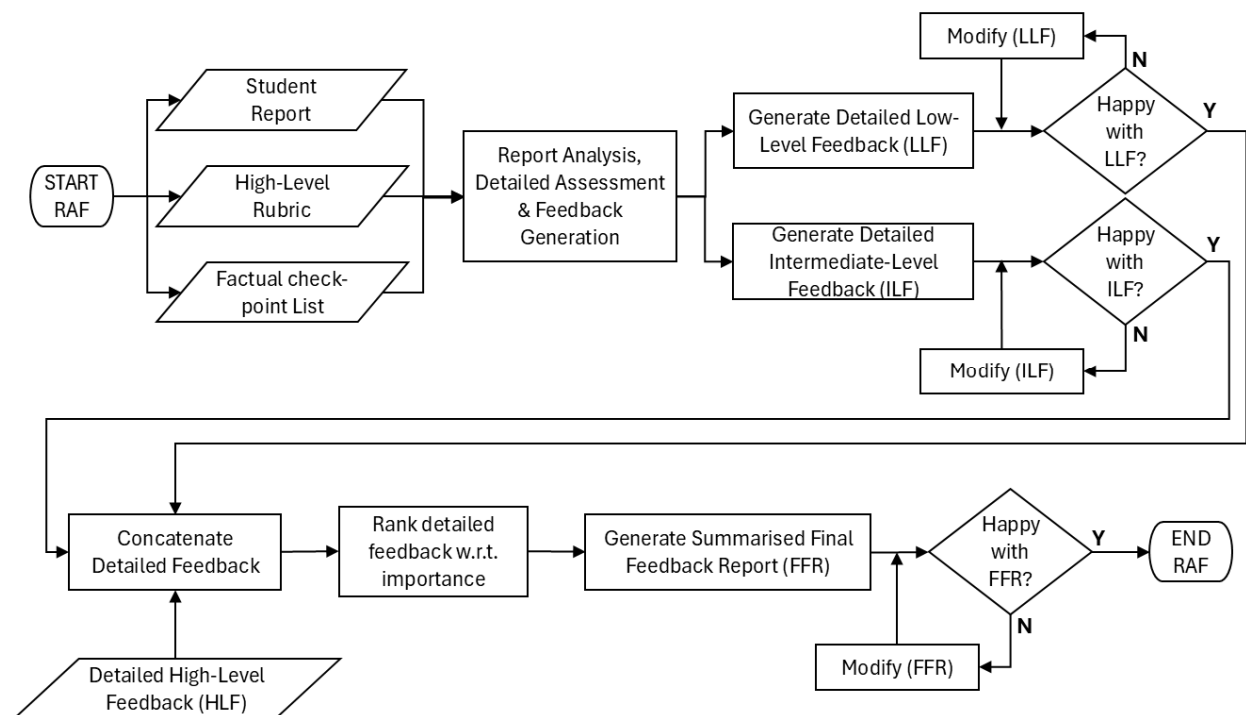


Figure 1. Flowchart for the proposed Report Assessment & Feedback (RAF) procedure.

The main steps are as follows:

- The Report Assessment & Feedback (RAF) procedure commences as follows:
 - Defining the list of points (along their relative importance) that the system will factually check;
 - Setting the rubric, which a human assessor will use to generate high-level feedback;

Sending the student report to the system for analysis.

- The system analyses the student report and generates one detailed low-level feedback (LLF) report and one detailed intermediate-level feedback (ILF) report.

- The human assessor evaluates the LLF and ILF reports generated and can modify them to their satisfaction.
- In the sequel, the human assessor provides their own high-level feedback.
- At this stage, the system has three feedback reports available (i.e., LLF, ILF, and HLF), processes them, ranks the available feedback comments with respect to their importance, and generates a student-facing Final Feedback Report (FFR).
- The human assessor can modify the FFR to their satisfaction and approve it for release.

For multiple reports, it is possible to use the flowchart for each report separately, in sequence; alternatively, it is possible to process the reports in batches in parallel. In all cases, the human assessor is included in the loop, deciding whether to accept, reject, or modify the low- and intermediate-level feedback generated. Of importance are both the frequency and the type of action that the human-in-the-loop (HITL) takes.

5.2. Human-in-the-Loop: Balancing Oversight and Efficiency

A crucial implementation design decision concerns the frequency and nature of human-in-the-loop (HITL) involvement in AI-assisted processes.

High-frequency HITL, i.e., reviewing every AI output before student delivery, provides strong quality control, enables rapid error detection and correction, maintains human accountability, and supports ongoing calibration and system improvement; however, it may significantly reduce the efficiency gains that motivated AI integration, create workflow bottlenecks during peak assessment periods, and be resource-intensive at scale.

Low-frequency HITL, i.e., spot-checking samples and reviewing only flagged cases, scales better with increasing assessment volumes, allows faster feedback turnaround, and may be more sustainable in the long term; however, it introduces risks that errors may propagate undetected across multiple student submissions, weakens accountability and quality assurance, and may create equity concerns if some students receive more thorough human review than others.

These considerations represent a fundamental tension between oversight quality and operational efficiency. Rather than prescribing universal solutions, the framework requires making this trade-off explicit and deliberate, based on context-specific factors such as disciplinary norms and error tolerances, the stakes of the assessment (i.e., formative versus summative, progression-critical versus developmental), cohort size, and institutional resources.

As a default position for initial implementation, the framework recommends beginning with high-frequency HITL oversight and reducing the frequency once substantial evidence from systematic evaluation demonstrates acceptable reliability, security, and fairness. The burden of proof should rest with demonstrating that reduced oversight is safe and appropriate, not with demonstrating that high oversight is necessary.

6. Broader Implications and Critical Considerations

6.1. Pedagogical Implications: Centering Student Learning

The fundamental question that must guide any AI integration in assessment is whether such approaches demonstrably support student learning as well as or better than current practice. Efficiency gains and workload reduction, while important for sustainability, must remain secondary to learning outcomes. This learning-centered perspective requires critical examination of how different forms of automated feedback affect student development.

For low-level feedback on structural and presentational elements, key questions include the following: Does automated correction of formatting and grammar errors help students develop these skills, or does it merely produce correctly formatted documents without building understanding of professional communication standards? Does immediate automated feedback support learning from mistakes, or does it allow students to

submit work without developing attention to detail? How can low-level feedback systems be designed to support learning rather than just error correction?

For intermediate-level factual verification, critical considerations include the following: Does automated checking of calculations and factual accuracy help students develop self-verification habits and numeracy skills, or does it create dependency on external validation that undermines their developing autonomy? When students receive verification feedback from AI systems, do they engage with understanding why errors occurred and how to avoid them, or do they correct flagged errors without deeper learning? Does the presence of automated checking change student approaches to completing work, i.e., do they invest less effort in careful checking because they expect that systems will catch errors?

These questions directly engage with Boud and Molloy's (2013) distinction between Feedback Mark 1 (transmission and correction) and Mark 2 (learner agency and sustainable assessment). There is a risk that automated feedback, even when accurate and timely, reinforces transmission models by providing answers rather than promoting students' capacity to assess the quality of their own work. This risk must be actively addressed through deliberate design that positions automated feedback as supporting, rather than replacing, the development of students' self-assessment and evaluative judgment.

6.2. *Impacts on Academic Work, Professional Identity, and Student Agency*

6.2.1. Academic Expertise and Work Transformation

Automation in professional domains may reconfigure expertise in complex ways. When routine elements of skilled practice are automated, there is concern that practitioners may lose the fine-grained familiarity with detail that informs higher-level judgment. Applied to assessment, if routine verification tasks are automated, will academic staff maintain the detailed attention to student work that grounds pedagogical insight?

Additionally, accountability frameworks must address the following question: If an AI system provides incorrect, biased, or pedagogically inappropriate feedback that disadvantages students, who bears responsibility? The academic assessor who configured the system? The institution that deployed it? The vendor who provided the technology? Current legal and ethical accountability frameworks have not kept pace with technological capabilities, creating ambiguity about responsibility that may expose all parties to risk and leave students without clear recourse.

6.2.2. Student Agency and Feedback Literacy

The framework's impact on students' development of feedback literacy requires careful attention. Potential benefits include immediate feedback on some elements enabling faster correction, exposure to multiple perspectives on their work, and explicit transparency about different feedback sources potentially supporting critical evaluation of feedback staff (Jin et al., 2025).

However, significant concerns warrant monitoring: Do students learn to distinguish reliable from unreliable feedback sources and develop appropriate critical skepticism, or do they uncritically accept AI-generated feedback as authoritative? Does automation create dependency that reduces students' development of self-assessment capabilities and evaluative judgment? Do students engage less deeply with feedback that they perceive as algorithmically generated compared to a personalized human response? Does multi-source feedback create confusion about conflicting messages, particularly for students with less developed academic skills?

Furthermore, if students learn that low-level and intermediate-level feedback is automated, they may adapt in ways contrary to their learning objectives: focusing effort on 'gaming' automated systems through strategic choices rather than genuine learning;

discounting the importance of dimensions that they perceive as merely ‘checked’ rather than genuinely evaluated; or developing strategic approaches to assessment that prioritize algorithmic optimization over intellectual development and authentic engagement with disciplinary content.

6.3. Security, Integrity, and Equity Considerations

6.3.1. Security and Manipulation Risks

Empirical evidence suggests that, within large student populations, a small but significant minority may view AI-assisted assessment systems as technical challenges to be exploited rather than educational tools to be engaged with honestly (Gruenhagen et al., 2024; Sahoo et al., 2025). Motivations for system manipulation vary considerably, including academic pressure to achieve high marks, intellectual curiosity about system limitations (viewing circumvention as an engaging puzzle), retaliatory behavior following unsatisfactory assessment outcomes, and peer pressure or competition where circumvention techniques spread through student networks.

Successful circumvention techniques can spread rapidly through student communities via social media, online forums, and direct communication, potentially compromising assessment integrity at scale. Security concerns may ultimately limit system viability more than technological capabilities, i.e., if maintaining adequate security requires such extensive monitoring, testing, and human intervention that efficiency gains disappear or reverse, the business case for AI integration collapses entirely.

6.3.2. Equity and Fairness Concerns

Algorithmic consistency does not automatically ensure equity or fairness. AI systems introduce multiple equity concerns requiring proactive attention. Training data biases may systematically disadvantage particular student demographic groups in ways that are difficult to detect. Systems may perform less reliably on non-standard expression styles, potentially penalizing creativity, linguistic diversity, or culturally specific argumentation. Students with disabilities may be differentially impacted without appropriate accessibility design. Additionally, unequal access to AI-assisted feedback across institutions and variation in students’ capacity to use multi-source feedback effectively create further fairness concerns.

Addressing these concerns requires systematic fairness evaluation through explicit design requirements. The framework mandates regular auditing to identify whether specific demographic groups experience disadvantage, transparent documentation of system operations and decision-making processes, clear mechanisms for students to challenge feedback that they believe to be incorrect or unfair, and ongoing evaluation examining equity outcomes rather than only aggregate performance measures. Equity must be embedded as a central design consideration from the outset, not treated as an afterthought or optional enhancement.

7. Evaluation Framework: Necessary Evidence for Validation

The framework presented here is a theoretical proposition. Practical adoption requires rigorous empirical evidence demonstrating that it supports learning, provides fair assessment, delivers actual efficiency benefits, maintains quality, and introduces no unacceptable risks regarding accuracy, security, or equity.

Comprehensive Mixed-Methods Evaluation Design

Valid evaluation requires combining quantitative indicators of efficiency and consistency with qualitative evidence of learning impacts and stakeholder experiences.

Quantitative evaluation should include efficiency metrics such as feedback turnaround time, staff time per submission, and total assessment workload (comparing framework and control modules). Consistency metrics should assess inter-marker reliability for high-level feedback, feedback consistency at the low and intermediate levels, and AI error frequency. Quality metrics should capture student satisfaction through module evaluations and NSS, alongside independently rated feedback specificity and actionability. The evaluation of learning outcomes should examine subsequent assessment performance, self-assessment and feedback literacy development (using validated instruments), and learning gains when comparing framework and control conditions.

Qualitative evaluation should examine student experiences through focus groups exploring feedback clarity, usefulness, and trustworthiness; feedback literacy development in multi-source environments; and patterns of feedback uptake. Staff experiences should be explored through interviews examining assessment workload changes, judgment processes, and professional identity shifts, as well as confidence in AI outputs, oversight evolution, how time saved was used, and whether anticipated benefits materialized. Process evaluation should document implementation challenges, necessary adaptations, and contextual factors, while systematically analyzing AI errors, security incidents and their remediation, and equity impacts through disaggregated demographic analysis.

Rigorous evaluation requires controlled comparison between modules adopting the framework and matched control modules with equivalent characteristics (discipline, level, cohort size, marking team experience). While randomized controlled trials assigning students or sections to framework-assisted versus traditional feedback would provide the strongest causal inference, practical and ethical constraints may limit their feasibility in many institutional contexts. Where experimental designs are not viable, pragmatic alternatives include pre–post comparisons across academic years (with careful attention to potential confounds) or mixed cohort designs where some assignments use the framework while others do not, allowing within-subject comparison.

The evaluation design described above sets the target for full validation; not every institution can resource it immediately. Full mixed-methods work needs infrastructure, staff time, and ethical clearance that build up over time. Randomized controlled trials (RCTs) provide the strongest evidence, but access and cohort size can limit their use in some settings; a staged approach helps institutions work toward the full design. Early stages can use low-cost proxies, e.g., turnaround time, flagged error rates, and human–AI agreement in spot checks; later stages add the fuller instrument battery. Smaller pilots remain valuable as steps toward comprehensive validation, provided they are reported as partial rather than conclusive.

8. Limitations and Future Research Directions

As a theoretical proposition, the tripartite framework carries limitations that require explicit acknowledgement and that define the empirical research agenda necessary for its validation.

The taxonomy's boundaries may be harder to operationalize in practice than the framework implies. The distinction between intermediate-level factual verification and high-level interpretive evaluation is analytically clear but may prove contextually ambiguous in practice. In some disciplinary contexts, for instance, reports involving open-ended experimental design, contested methodological choices, or data with multiple valid interpretations, tasks that appear to involve straightforward factual checking may contain embedded interpretive dimensions that are not immediately apparent. The framework's conservative default position, i.e., escalating all ambiguous cases to human assessment, addresses this risk procedurally, but it does not eliminate the underlying difficulty. Empirical

work testing where the intermediate–high boundary falls in specific disciplinary contexts, and how assessors and AI systems handle genuinely ambiguous cases, is needed before the taxonomy can be considered validated for those contexts.

The five boundary principles have not been tested for simultaneous feasibility. Each principle is theoretically justified and practically specifiable, but the framework requires that all five be satisfied simultaneously for AI involvement at the intermediate level to be considered appropriate. Whether this is achievable within realistic institutional resource constraints, particularly regarding assessor time for knowledge base curation, the infrastructure costs of robust security implementation, and the workflow demands of high-frequency human oversight, is an open empirical question. There is a risk that the conditions are so demanding that intermediate-level AI support proves impractical in most real institutional contexts, which would limit the framework's operational value. Pilot implementations with honest reporting of resource requirements and practical challenges are needed to test this.

The framework was developed primarily with STEM and social science report contexts in mind. Complex reports, as defined in Section 1.2, i.e., laboratory reports, research reports, design reports, and case reports, share structural and epistemic features that motivate the tripartite taxonomy. Whether the framework translates to other written assessment genres, such as essays, portfolios, dissertations, or reflective writing, is not addressed here and should not be assumed. These genres may involve different distributions of verification and evaluative tasks, different disciplinary norms about what constitutes a bounded versus interpretive claim, and different student populations with different feedback literacy profiles. Extension to other genres requires separate theoretical and empirical work.

The framework's pedagogical assumptions require empirical validation. The claim that differentiating feedback by level and source supports rather than undermines student feedback literacy development is theoretically grounded in the work of Carless and Boud (2018) and Boud and Molloy (2013), but it is an assumption rather than a demonstrated finding. It is equally plausible that multi-source feedback, combining human assessment, AI verification, and automated checking, creates confusion, undermines trust in assessment processes, or reinforces transmission models of feedback by multiplying the volume of information delivered to students rather than developing their capacity to seek and use feedback autonomously. Whether the framework's transparency and attribution requirements (Principle 4) are sufficient to mitigate these risks is a question that only empirical investigation can answer.

Security considerations may ultimately constrain viability more than technological capabilities. The framework's security principle assumes that robust countermeasures against prompt injection and system manipulation are implementable within realistic institutional contexts. This assumption is theoretically reasonable but empirically unverified. If the adversarial landscape evolves faster than institutional security responses, which the general cybersecurity literature suggests is a real risk, then the practical viability of AI-assisted intermediate-level feedback may be limited by security constraints regardless of technological capability improvements. This possibility should be monitored as part of any pilot implementation and explicitly evaluated in the research agenda proposed in Section 7.

A further limitation is the absence of an implemented pilot. This paper is deliberately conceptual, as stated in Section 1.2, and does not test the framework against a working system. This leaves open questions that only implementation can answer: which LLM is used, how it is prompted, how the RAG pipeline is configured, and how these choices affect the accuracy and consistency of intermediate-level feedback. Different models and configurations may satisfy the five boundary principles to varying degrees, particularly Principles 1 (knowledge base grounding) and 3 (bounded verification). A follow-up

implementation study that benchmarks specific LLM and RAG configurations against human-verified ground truth on real student reports is a necessary next step and is proposed as a priority in the research agenda above.

The equity implications require deeper empirical investigation than the framework can provide. Section 6.3.2 identifies equity concerns including training data bias, differential impacts on non-standard expression styles, accessibility for students with disabilities, and unequal institutional access to AI-assisted feedback. The framework mandates equity auditing as a design requirement, but it cannot pre-specify what equity problems will emerge in practice or how serious they will be. There is a real possibility that AI-assisted assessment, even when implemented according to the framework's principles, produces systematically differential outcomes for identifiable student groups in ways that are difficult to detect without sustained, disaggregated outcome monitoring. This must be treated as a live risk rather than a solved problem.

The framework's core claims also invite critical scrutiny. Firstly, Sections 3.3 and 4.3.3 already note that verification and evaluation can blend in practice. A critic could push this further, questioning how cleanly the three levels separate in some contexts. This is precisely why Section 8 calls for boundary-testing studies rather than assuming that the taxonomy holds universally. Furthermore, the five principles require knowledge base curation, security auditing, and frequent human oversight; this may shift staff effort rather than reduce it, resulting in lower efficiency gains once implementation costs are accurately accounted for. Whether the time saved truly offsets this investment is an open empirical question. In addition, a bounded framework could still ease future expansion of AI's role in assessment. The non-negotiable principles are designed as a standing check against this, not a one-off safeguard. Lastly, an alternative taxonomy based on task type, i.e., numeric, textual, or visual, is worth testing alongside the epistemic status distinction used here. These are open questions that the framework is designed to surface.

These limitations do not undermine the framework's theoretical contribution. Identifying principled distinctions, specifying boundary conditions, and proposing validation criteria are valuable scholarly contributions independent of whether every implementation challenge proves soluble. They do, however, define a substantial and specific research agenda: disciplinary boundary-testing studies, pilot implementations with honest resource accounting, longitudinal feedback literacy investigations, adversarial security evaluations, and equity-focused outcome monitoring. The framework is offered as a starting point for that agenda, not as its conclusion.

9. Conclusions

This paper has proposed and theoretically examined a tripartite feedback framework for the AI-assisted assessment of complex reports in higher education, distinguishing low-level structural and presentational feedback, intermediate-level factual content validation, and high-level critical evaluation and synthesis. The framework provides pedagogically grounded criteria for allocating feedback tasks between AI systems and human assessors, not on the basis of what technology can technically perform, but on the basis of what pedagogical purposes require and what epistemic status different feedback tasks carry.

Three contributions are offered to the assessment and feedback literature: First, the framework provides an operational taxonomy that complements established feedback theories by addressing the task allocation question that they were not designed to answer. Second, it specifies five non-negotiable boundary principles that transform general advocacy for hybrid human–AI approaches into concrete, accountable criteria for responsible deployment. Third, it proposes a rigorous mixed-methods evaluation design that specifies

the evidence needed for empirical validation, offering a structured research agenda rather than leaving validation unspecified.

The framework's value lies not in prescriptive solutions but in providing structured categories, explicit boundary conditions, and validation criteria to guide context-sensitive institutional decision-making. Technology may contribute to addressing the persistent challenge of feedback provision at scale in higher education, but only when rigorously subordinated to pedagogical purposes, human professional accountability, and the equitable treatment of all students. The central question is not whether AI can help with assessment; it is whether, under what conditions, and with what safeguards it can do so in ways that genuinely serve student learning. This framework is offered as a principled foundation for pursuing that question empirically.

Funding: This research received funding from Imperial College London through the Digital Innovation Fund (G74038).

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: During the preparation of this manuscript, the author used Claude Opus 4.6 (Anthropic, claude.ai) and Grammarly (web-based version, <https://app.grammarly.com/>, accessed on 25 May 2026) for the purposes of grammar checking and syntax improvement. The author would also like to thank the anonymous reviewers of this journal, who reviewed this manuscript, for their constructive comments.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
APA	American Psychological Association
FFR	Final Feedback Report
GPT	Generative Pre-Trained Transformer
HITL	Human-in-the-Loop
HLF	High-Level Feedback
IEEE	Institute of Electrical and Electronics Engineers
ILF	Intermediate-Level Feedback
LLF	Low-Level Feedback
LLM	Large Language Model
LMS	Learning Management System
NSS	National Student Survey
RAF	Report Assessment & Feedback
RAG	Retrieval-Augmented Generation
STEM	Science, Technology, Engineering, and Mathematics
VLE	Virtual Learning Environment

References

- Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education*, 21(1), 23. [\[CrossRef\]](#)
- Bell, A. R., & Brooks, C. (2018). What makes students satisfied? A discussion and analysis of the UK's national student survey. *Journal of Further and Higher Education*, 42(8), 1118–1142. [\[CrossRef\]](#)
- Boud, D., & Dawson, P. (2023). What feedback literate teachers do: An empirically-derived competency framework. *Assessment & Evaluation in Higher Education*, 48(2), 158–171. [\[CrossRef\]](#)

- Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment & Evaluation in Higher Education*, 38(6), 698–712. [\[CrossRef\]](#)
- Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, 43(8), 1315–1325. [\[CrossRef\]](#)
- Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access*, 8, 75264–75278. [\[CrossRef\]](#)
- Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education*, 20(1), 22. [\[CrossRef\]](#)
- Deeley, S. J., Fischbacher-Smith, M., Karadzhev, D., & Koristashevskaya, E. (2019). Exploring the “wicked” problem of student dissatisfaction with assessment and feedback in higher education. *Higher Education Pedagogies*, 4(1), 385–405. [\[CrossRef\]](#)
- Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., & De Weerd, J. (2021). A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education*, 162, 104094. [\[CrossRef\]](#)
- Evans, C. (2013). Making sense of assessment feedback in higher education. *Review of Educational Research*, 83(1), 70–120. [\[CrossRef\]](#)
- Fedoruk, B. D. (2025). Cross-evaluation of large language model assessment behaviours in educational tasks by cognitive level. *Journal of Educational Informatics*, 6(1), 2–23. [\[CrossRef\]](#)
- Gruenhagen, J. H., Sinclair, P. M., Carroll, J.-A., Baker, P. R., Wilson, A., & Demant, D. (2024). The rapid rise of generative AI and its implications for academic integrity: Students’ perceptions and use of chatbots for assistance with assessments. *Computers and Education: Artificial Intelligence*, 7, 100273. [\[CrossRef\]](#)
- Gulyamov, S., Gulyamov, S., Rodionov, A., Khursanov, R., Mekhmonov, K., Babaev, D., & Rakhimjonov, A. (2026). Prompt injection attacks in large language models and AI agent systems: A comprehensive review of vulnerabilities, attack vectors, and defense mechanisms. *Information*, 17(1), 54. [\[CrossRef\]](#)
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. [\[CrossRef\]](#)
- Henderson, M., Ryan, T., & Phillips, M. (2019). The challenges of feedback in higher education. *Assessment & Evaluation in Higher Education*, 44(8), 1237–1252. [\[CrossRef\]](#)
- Jin, F. J.-Y., Dai, W., Maheshi, B., Martinez-Maldonado, R., Gašević, D., & Tsai, Y.-S. (2025). Feedback in K-12 and higher education: Educators’ perspectives. *Teaching and Teacher Education*, 156, 104933. [\[CrossRef\]](#)
- MacKay, J. R. D., Hughes, K., Marzetti, H., Lent, N., & Rhind, S. M. (2019). Using national student survey (NSS) qualitative data and social identity theory to explore students’ experiences of assessment and feedback. *Higher Education Pedagogies*, 4(1), 315–330. [\[CrossRef\]](#)
- Mertens, U., Finn, B., & Lindner, M. A. (2022). Effects of computer-based feedback on lower- and higher-order learning outcomes: A network meta-analysis. *Journal of Educational Psychology*, 114(8), 1743–1772. [\[CrossRef\]](#)
- Milani, A., Franzoni, V., Florindi, E., Omarbekova, A., Bekmanova, G., & Yergesh, B. (2025). When AI is fooled: Hidden risks in LLM-assisted grading. *Education Sciences*, 15(11), 1419. [\[CrossRef\]](#)
- Mollick, E. R., & Mollick, L. (2024). Instructors as innovators: A future-focused approach to new AI learning opportunities, with prompts. *SSRN Electronic Journal*. [\[CrossRef\]](#)
- Morris, R., Perry, T., & Wardle, L. (2021). Formative assessment and feedback for learning in higher education: A systematic review. *Review of Education*, 9(3), e3292. [\[CrossRef\]](#)
- Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199–218. [\[CrossRef\]](#)
- Noroozi, O., Soleimani, S., Farrokhnia, M., & Banihashem, S. K. (2024). Generative AI in education: Pedagogical, theoretical, and methodological perspectives. *International Journal of Technology in Education*, 7(3), 373–385. [\[CrossRef\]](#)
- Parkinson, J. (2017). The student laboratory report genre: A genre analysis. *English for Specific Purposes*, 45, 1–13. [\[CrossRef\]](#)
- Reid, N., & Shah, I. (2007). The role of laboratory work in university chemistry. *Chemistry Education Research and Practice*, 8(2), 172–185. [\[CrossRef\]](#)
- Sahoo, D., Majhi, V., Neekhara, A., Sinha, Y., Mandal, M., & Kumar, D. (2025). How to trick your AI TA: A systematic study of academic jailbreaking in LLM code evaluation. *arXiv*. [\[CrossRef\]](#)
- Samala, A. D., Rawas, S., Wang, T., & Reed, J. M. (2025). Unveiling the landscape of generative artificial intelligence in education: A comprehensive taxonomy of applications, challenges, and future prospects. *Education and Information Technologies*, 30(3), 3239–3278. [\[CrossRef\]](#)
- Topping, K. J., Gehringer, E., Khosravi, H., Gudipati, S., Jadhav, K., & Susarla, S. (2025). Enhancing peer assessment with artificial intelligence. *International Journal of Educational Technology in Higher Education*, 22(1), 3. [\[CrossRef\]](#)

- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—Where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1), 39. [\[CrossRef\]](#)
- Zhai, N., & Ma, X. (2023). The effectiveness of automated writing evaluation on writing quality: A meta-analysis. *Journal of Educational Computing Research*, 61(4), 875–900. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.