International Journal of Convergence Research in Engineering, Technology & Management (IJCRETM)
Volume 1, Issue 1 | July–September 2026 | *Peer Reviewed (Double-Blind)* • *Open Access* • *Quarterly*

Both Sides of the Desk: Generative Artificial Intelligence in Production and the Judgement of Student Work

Shilpa Narang[1*]

[1]*Lecturer & Curriculum and Course Development Manager, ASA Institute of Higher Education, Sydney, New South Wales, Australia*

***Corresponding Author: ORCID:** 0000-0001-8037-6822

ARTICLE INFORMATION

Manuscript ID:	IJCRETM-20260101117
How to Cite:	Narang, S. (2026). Both sides of the desk: Generative artificial intelligence in production and the judgement of student work. *International Journal of Convergence Research in Engineering, Technology & Management, 1*(1), IJCRETM-20260101117.
Copyright & License:	© 2026 The Author(s). Published by A2Z EduLearningHub, India. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0).

Abstract

Higher education has responded to generative artificial intelligence largely as a problem of student conduct. Policy, regulation and scholarship address what students may do with these tools, how such use should be declared, and how qualifications can be assured when detection is unreliable. Far less attention has been directed to the other side of the desk. Generative systems are now used to draft feedback, apply rubrics, screen submissions and support moderation, and this use is governed largely by the assertion that final responsibility rests with the human marker. The result is an asymmetry: student use is regulated by declaration, surveillance and sanction, while institutional use of the same technology in judging that work is regulated by a statement of intent. Developed through a narrative review of the assessment security, automated assessment and algorithmic accountability literatures, this conceptual paper argues that the two developments form a single phenomenon. Generative artificial intelligence now occupies three positions in the assessment relationship simultaneously, as producer of student work, as arbiter of it, and as the subject of what is assessed, and each position generates a distinct problem of validity or accountability. The paper proposes a four-dimensional framework of assessment trustworthiness comprising validity, transparency, security and equity, and finds institutional practice well developed on security, asymmetrical on transparency, under-examined on validity and largely silent on equity. Six propositions and a research agenda follow, with particular attention to non-self-accrediting providers, which are largely absent from the current evidence base.

Keywords: generative artificial intelligence; assessment trustworthiness; assessment security; automated marking; academic integrity; higher education governance

1. Introduction

Student use of generative artificial intelligence in assessed work is now close to universal. Successive sector surveys record a rise from a minority practice in 2023 to near-universal adoption within three years. The most recent evidence indicates that 95 per cent of undergraduates use artificial intelligence in at least one way, that 94 per cent use generative tools to help with assessed work, and that the proportion incorporating generated text directly into submitted work has risen from 3 per cent in 2024 to 12 per cent (Stephenson & Armstrong, 2026). Student-facing studies conducted early in that period had already found high awareness and broadly favourable disposition alongside substantial uncertainty about what was permitted (Chan & Hu, 2023), and institutional policy has struggled since to keep pace with either (Chan, 2023; Yan et al., 2024). Institutional and regulatory response has been rapid and, in the Australian case, unusually coordinated: the national regulator issued principles for assessment reform within a year of the technology's mass release (Tertiary Education Quality and Standards Agency [TEQSA], 2023) and followed them with implementation pathways two years later (TEQSA, 2025).

Institutional use of the same technology in the judgement of student work has expanded over the same period with considerably less scrutiny. Generative systems are used to draft feedback comments, to apply rubrics at scale, to generate and calibrate marking guides, to triage submissions for integrity review, and to support

moderation, and a rapidly growing body of technical evaluation has tested their performance in precisely these tasks (Flodén, 2025; Pack et al., 2024; Yavuz et al., 2025). That evaluation literature is largely separate from the scholarship on the ethical and practical governance of such systems in education, which has itself noted the disconnect (Yan et al., 2024). Some of this use is sanctioned and documented; a great deal of it is discretionary and unrecorded, a pattern consistent with earlier findings on the informal diffusion of automated assessment tools (Gardner et al., 2021). Where institutional guidance exists, it typically takes the form of a residual-responsibility clause: staff may use these tools provided the final mark and feedback remain the responsibility of the marker (Liu & Bridgeman, 2023).

That asymmetry is the problem this paper addresses. Student use is governed by declaration requirements, detection infrastructure, disciplinary procedure and, increasingly, the redesign of assessment tasks around supervision. Marker use is governed by an assertion. A student who fails to declare assistance faces an integrity allegation; a marker who does not disclose the same assistance faces nothing, and the student whose work was so marked has no means of knowing, and therefore no ground on which to contest. Whatever one concludes about the merits of either practice, the two cannot coherently be regulated by such different instruments while being described as the same technology.

The dominant framings do not accommodate this. Academic integrity scholarship construes generative artificial intelligence as a threat to the authenticity of student work, and its response has been to strengthen assurance, through detection, through supervised tasks, and through the redesign of assessment to make unauthorised assistance harder or less useful (Cotton et al., 2024; Dawson, 2021; Perkins, 2023). That response has hardened as evidence against detection has accumulated (Weber-Wulff et al., 2023) and as the case for structural rather than declaratory change has been pressed (Corbin et al., 2025). Educational technology scholarship construes the same technology as an instrument of pedagogical and administrative efficiency, and its response has been to examine accuracy, workload effects and student perceptions of automated feedback (Gardner et al., 2021; Yan et al., 2024; Zawacki-Richter et al., 2019). Each literature is internally coherent and each has been reviewed on its own terms (Bearman et al., 2023; Zawacki-Richter et al., 2019). Neither framing asks the question that arises when both developments are placed side by side: what happens to the validity of a judgement when a machine assesses machine-assisted work?

The nearest antecedent is Swiecki et al.'s (2022) synthesis of assessment in the age of artificial intelligence, which maps the roles artificial intelligence can play across the assessment cycle and identifies the risks attaching to each. That contribution is foundational to the present argument, but it was developed immediately before the mass availability of generative systems, and it treats artificial intelligence principally as an instrument deployed by institutions rather than as a participant on both sides of the assessment relationship. The subsequent literature has largely extended the student-conduct question. Bearman et al.'s (2023) critical review of discourses of artificial intelligence in higher education demonstrates how readily the field's framing choices constrain the questions it can ask; the framing this paper contests is precisely of that kind.

Three questions organise the analysis. First, how should the assessment relationship be conceptualised when generative artificial intelligence participates in both the production and the judgement of student work? Second, what does trustworthiness require beyond the security and detection metrics that dominate current institutional practice? Third, what research agenda follows, with particular reference to providers outside the well-resourced, self-accrediting institutions in which nearly all current evidence has been generated?

The contribution is conceptual. The paper reports no new empirical data. It reorganises a fragmented literature around the governance obligations that arise when the same technology sits on both sides of the desk, and specifies those obligations in a form amenable to subsequent empirical testing. Section 2 reviews the literature and sets out the three positions generative artificial intelligence now occupies in assessment. Section 3 describes the review approach. Section 4 develops the four-dimensional framework, examines each dimension in turn, and derives six propositions. Section 5 concludes.

2. Literature Review

The claim organising this review is that generative artificial intelligence occupies three positions in the assessment relationship simultaneously, each generating a distinct problem that existing governance

instruments address poorly. The three positions are set out in Sections 2.1 to 2.3 and summarised in Table 1. Section 2.4 examines the conceptual apparatus the sector has developed in response, and its limits.

2.1 Artificial intelligence as producer

The first position is the one the sector has recognised. Students use generative systems to plan, draft, revise, translate and complete assessed work, at rates that have risen steeply and show no sign of stabilising (Chan & Hu, 2023; Stephenson & Armstrong, 2026). The uses are heterogeneous, ranging from ideation and language support to wholesale generation, and policy instruments that treat them as a single category have been criticised on precisely that ground (Chan, 2023; Perkins et al., 2024a). The institutional response has nonetheless proceeded through three broadly sequential phases: prohibition, detection, and structural redesign (Cotton et al., 2024; Corbin et al., 2025).

The detection phase has effectively closed. Controlled testing of detection tools has found them neither accurate nor robust, with high false-negative rates and susceptibility to trivial evasion (Weber-Wulff et al., 2023), and academic judgement unaided performs little better, with experienced markers identifying generated text at rates close to chance (Fleckenstein et al., 2024). Combining the two does not resolve the problem, since software and human judgement fail in correlated rather than compensating ways (Perkins et al., 2024b), and the tools carry a demonstrated bias that is examined in Section 4.5 (Liang et al., 2023). The conclusion that detection cannot ground assurance is now the settled position of the Australian regulator and of the international integrity literature (Corbin et al., 2025; TEQSA, 2025; Weber-Wulff et al., 2023).

What makes this a producer problem rather than a conduct problem is what it does to the construct. An unsupervised written task was never valued for the artefact it produced; it was valued as evidence of a cognitive process assumed to be necessary to produce that artefact, an assumption embedded in the design logic of criterion-referenced assessment since Sadler (1989). Where the artefact can be produced without the process, the inference from performance to attainment fails, not because the student cheated, but because the task no longer elicits the construct it was designed to elicit (Swiecki et al., 2022). This is a validity failure in the classical sense (Kane, 2013; Messick, 1995), and it is not remedied by establishing that a particular student did not use the tool (Dawson, 2021). The task has become uninformative in general, not compromised in the individual case, which is why the security literature's turn toward conditions of production rather than dispositions of students is the correct response even though it is an incomplete one (Corbin et al., 2025; Dawson, 2021).

2.2 Artificial intelligence as arbiter

The second position is the one the sector has not examined. Generative systems now participate in the judgement of student work: drafting feedback, applying rubric descriptors, generating marking guides, screening for integrity concerns, and supporting moderation across large cohorts (Yan et al., 2024). Institutional guidance permitting such use commonly attaches a single safeguard, namely that the final mark and the feedback remain the responsibility of the marker (Liu & Bridgeman, 2023).

The technical evidence on whether these systems can bear the weight placed on them is unsettled, and the disagreement is itself instructive. Studies applying analytic rubrics under controlled conditions have reported high internal consistency and substantial convergence with expert raters (Yavuz et al., 2025), and comparable results have been obtained for the scoring of English language learner writing (Pack et al., 2024). Studies conducted on authentic assessment at scale have reported considerably weaker performance: grading of master's-level examinations found agreement with teacher-assigned grades in a minority of cases, with a systematic tendency toward marginally higher scores (Flodén, 2025). The divergence appears to track task authenticity and rubric specificity rather than model capability, which is precisely the pattern that would be expected if these systems reproduce surface features of quality reliably and context-dependent judgement unreliably. Reviews of the wider field reach a consistent conclusion, namely that evaluations of technical performance have run well ahead of evidence about deployment, equity and accountability in real institutional settings (Yan et al., 2024; Zawacki-Richter et al., 2019).

The safeguard is weaker than it appears, for three reasons. It specifies an outcome rather than a process, and is therefore satisfied by a marker who accepts the generated judgement without alteration, a disposition that the

literature on automation in professional decision-making would predict rather than treat as aberrant (Busuioc, 2021; Yan et al., 2024). It is not observable: institutions do not, as a rule, record whether a generated mark was overridden, by how much, or how often, so the safeguard cannot be audited even in principle, and the scoring divergences documented in the technical literature are therefore invisible in practice (Flodén, 2025). And it is not disclosed: the student is not told that a generative system contributed to the judgement of their work, and consequently cannot raise that contribution as a ground of appeal, notwithstanding that the ethical frameworks the sector has adopted place transparency toward affected parties among their foundational principles (Holmes et al., 2022; Nguyen et al., 2023).

Set beside the treatment of the producer position, the asymmetry is difficult to defend. Student use attracts a declaration obligation, an evidentiary infrastructure and a disciplinary consequence; marker use attracts a statement of residual responsibility. The justification usually offered is that the marker is a professional exercising judgement while the student is a candidate demonstrating attainment. That distinction has force, but it does not follow from it that the exercise of professional judgement should be exempt from the transparency required of every other consequential administrative determination. Busuioc's (2021) analysis of algorithmic accountability establishes the general point: where determinations are automated or automation-assisted and materially affect individuals, accountability requires not merely a nominal human in the loop but a functioning route to individual contestation. A mark is such a determination. So is an integrity allegation.

2.3 Artificial intelligence as subject

The third position is the newest and the least stable. Capability with generative artificial intelligence is increasingly specified as a graduate attribute, embedded in course learning outcomes and, in professionally accredited fields, written into competency standards. Where that is so, the technology is no longer only a threat to assessment or an aid to it; it is part of what is being assessed. Frameworks such as the Artificial Intelligence Assessment Scale (Perkins et al., 2024a) and institutional policy typologies (Chan, 2023) were developed to give this specification operational form, distinguishing levels of permitted use task by task.

The problem this position generates is a direct collision with the second. Assessing capability with a tool requires evidence of how the student worked with it: the quality of prompting, the critical evaluation of output, the judgement about when to accept and when to reject. That is process evidence, and it is generated in exactly the open, unsupervised conditions that assurance regimes are pushing assessment away from. Bearman and Ajjawi (2023) argue that pedagogy for a world of opaque systems must develop students' capacity to work with and interrogate the black box, and Tai et al. (2018) show that evaluative judgement of this kind develops through iterative, disclosed practice rather than through terminal demonstration. An institution that responds to the producer problem by relocating its assurance into supervised conditions therefore forecloses much of its capacity to assess the very capability it has undertaken to develop.

Table 1. The three positions of generative artificial intelligence in the assessment relationship

Position	Problem generated	Governance gap
Producer (Section 2.1)	The task no longer elicits the construct it was designed to elicit	Assurance decisions are not recorded as validity decisions
Arbiter (Section 2.2)	Judgement is automation-assisted without disclosure or recorded override	Residual-responsibility clauses are unobservable and unauditable
Subject (Section 2.3)	Capability with the tool requires process evidence from open conditions	Assurance regimes relocate assessment away from those conditions

Source: Author's compilation

2.4 From academic integrity to assessment security: the concept and its limits

The organising concept of the sector's response has moved, over roughly a decade, from academic integrity to assessment security. Integrity framed the problem as one of student values, addressed through education, declaration and honour codes. Security, developed most fully by Dawson (2021), reframed it as a design problem: assessment should be constructed so that the inference from performance to attainment holds even where some students are motivated to cheat, and this is achieved by controlling the conditions of production rather than by appealing to disposition. The reframing was a substantial advance and it is the foundation of

current practice, including the two-lane secure and open model now widely adopted in Australia (Liu & Bridgeman, 2023; TEQSA, 2023).

Two limitations of that conceptualisation matter here. The first is that it locates the source of contamination exclusively in the student. Every instrument the security framing generates, including supervision, verification, declaration and detection, is directed at the production side of the relationship (Dawson, 2021; TEQSA, 2023; Weber-Wulff et al., 2023). The framing has no vocabulary for contamination introduced by the institution's own use of the same technology in judging what was produced, and so does not see it, even though the reliability of that use is contested in the adjacent literature (Flodén, 2025; Pack et al., 2024). The second is that the framing was developed, tested and elaborated in large, self-accrediting research universities with dedicated learning-innovation units, invigilation infrastructure, institutional research capacity and academic development staff. Non-self-accrediting providers operate under identical threshold standards with a small fraction of that capability. Guidance calibrated to the first setting is not neutral with respect to the second; it prescribes remedies that the second cannot implement, and the resulting divergence between policy and practice is at present entirely undocumented.

3. Research Methodology

This is a conceptual paper developed through a narrative review of literature at the intersection of three fields: academic integrity and assessment security; automated assessment, feedback and learning analytics; and algorithmic accountability and administrative fairness. Sources were identified through purposive searching of the citation networks surrounding foundational contributions in each field, namely Boud and Associates (2010) and Boud and Soler (2016) for assessment reform, Dawson (2021) for assessment security, Swiecki et al. (2022) for artificial intelligence in assessment, and Busuioc (2021) and Peeters and Widlak (2018) for algorithmic accountability. These were supplemented by forward citation tracing and by the reference lists of recent reviews (Bearman et al., 2023; Zawacki-Richter et al., 2019). Regulatory and sector documents were included where they constitute the governance instruments the paper analyses rather than evidence about their effects.

The purpose of the review is theoretical integration rather than exhaustive enumeration, and the paper accordingly makes no claim to systematic coverage. Following Snyder's (2019) distinction between review types, a narrative approach is appropriate where the objective is to develop a conceptual model by reading bodies of work against one another that have not previously been so read, rather than to summarise the state of evidence on a settled question. The empirical illustrations drawn on throughout are taken principally from Australian regulatory practice, where the two-lane secure and open model has been implemented at sector scale and is therefore observable in a way that more diffuse national responses are not. These illustrations demonstrate the analytic purchase of the framework; they do not test it. Testing is the task set out in Section 4.6.

4. Results and Discussion

4.1 A four-dimensional framework of assessment trustworthiness

The framework proposed here reconstructs the object of governance for a setting in which artificial intelligence participates on both sides. Assessment trustworthiness is defined as the warranted confidence of students, staff and external stakeholders that a mark or qualification represents what it purports to represent, under conditions where generative artificial intelligence may have contributed both to the work and to the judgement of it. It comprises four dimensions, presented in Figure 1.

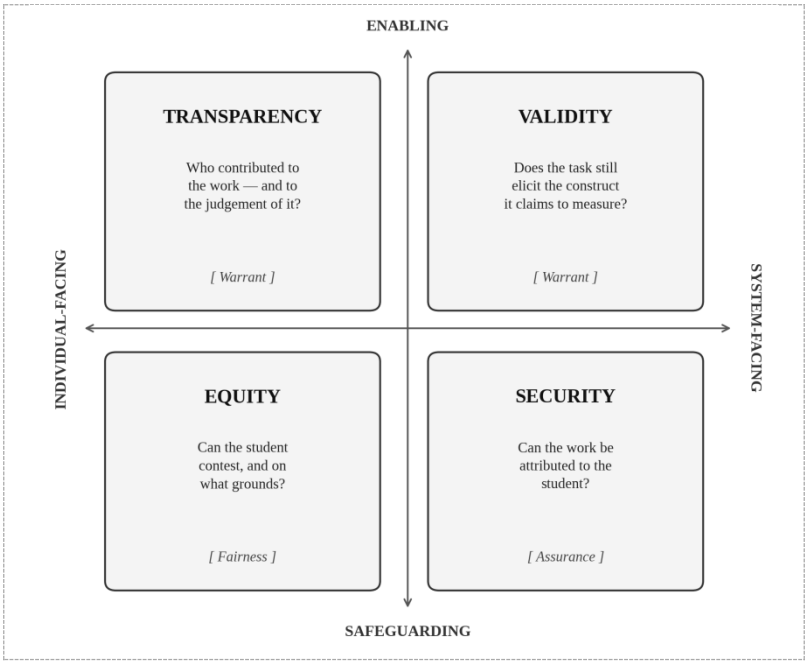


Figure 1. Four-dimensional framework of assessment trustworthiness under artificial intelligence convergence

Figure 1 arranges the dimensions on two axes. The vertical axis distinguishes enabling from safeguarding conditions; the horizontal distinguishes those that are individual-facing, bearing on the relationship between a student and a marker, from those that are system-facing, bearing on the institution and its regulator. Validity and transparency constitute the warrant of trustworthiness; security constitutes its assurance; equity constitutes its fairness.

Three properties warrant emphasis. The first is hierarchy without sufficiency. The dimensions are ordered but not substitutable. Security without validity yields judgements that are defensible but no longer measure the outcome claimed for them; validity without transparency yields performance that neither party can interpret, since neither knows what the other contributed; transparency without equity yields disclosure regimes that penalise most heavily the students least able to meet them. Evaluation reporting only assurance will systematically overstate achievement, which is the pattern currently visible in institutional reporting on assessment reform, where progress is counted in secured tasks.

The second is role-sensitivity. Each dimension corresponds to a different position of the technology in the assessment relationship. Validity answers to the producer position, transparency to the producer and arbiter positions jointly, security to the institution's assurance obligation under threshold standards, and equity to the student as claimant against an institutional determination.

The third is measurability. Each dimension admits indicators. Validity admits rubric-to-outcome mapping audits and evidence of construct-irrelevant variance introduced by task conditions; transparency admits declaration rates on the student side and disclosure and override rates on the marker side; security admits the proportion of programme learning outcomes covered by at least one verified assurance point; equity admits allegation and appeal volumes, sustained-allegation rates disaggregated by student characteristic, and the existence of a stated ground of appeal against an automation-assisted determination. Institutional dashboards presently measure the third and, occasionally and partially, the second.

4.2 Validity: the question displaced by security

Validity is the dimension with the deepest theoretical foundation and the least presence in current institutional practice. The classical position is that validity is a property not of an instrument but of an inference: the question is whether the interpretation placed on a performance is warranted, and the principal threats are construct under-representation and construct-irrelevant variance (Kane, 2013; Messick, 1995). Recent work has argued directly that the sector's preoccupation with cheating has displaced this prior question, and that a task which cannot support a valid inference is not rescued by being made secure (Dawson et al., 2024).

Three qualifications recur. First, the supervised tasks to which assurance has migrated introduce their own construct-irrelevant variance, including time pressure, handwriting, recall under observation, and unfamiliarity with a mode of performance that no longer resembles professional practice, so the shift purchases assurance at a validity cost that is rarely counted (Kane, 2013; Messick, 1995), and one that falls unevenly across the student body (Nieminen, 2024). Second, the tension between authenticity and security is genuine and not fully resolvable: authentic tasks resemble professional practice, professional practice now incorporates generative tools, and the more closely a task resembles that practice the less it can be secured (Ajjawi et al., 2024). Third, validity attaches to programme-level claims about graduate capability rather than to individual tasks, which is the strongest argument for the programmatic assurance pathway set out by TEQSA (2025), and also the reason that pathway is beyond the coordination capacity of many providers.

The governance implication is that assurance decisions should be recorded as validity decisions. Securing a task is a trade: it strengthens one inference and weakens another. An institution that documents which outcomes it has secured, without documenting which it has thereby ceased to assess authentically, has recorded only half of what it has done.

4.3 Transparency: declaration without symmetry

Transparency is the dimension on which institutional activity has been greatest and institutional consistency least. The student-facing apparatus is now well developed: declaration statements, permitted-use scales specifying levels of assistance task by task (Perkins et al., 2024a), and institutional policy frameworks distinguishing prohibited, permitted and required use (Chan, 2023). International guidance has converged on disclosure as a first-order principle (United Nations Educational, Scientific and Cultural Organization [UNESCO], 2023), and the ethical frameworks developed within the field make the same commitment (Holmes et al., 2022; Nguyen et al., 2023).

Three qualifications recur. First, declaration depends on self-report and is unverifiable, which is tolerable if its purpose is pedagogical, prompting reflection on how the work was made (Tai et al., 2018), and untenable if it is treated as an evidentiary basis for allegation, which is increasingly how it functions, given that no independent corroboration is available (Fleckenstein et al., 2024; Weber-Wulff et al., 2023). Second, permitted-use scales were designed as instruments of task design and are being implemented as instruments of compliance (Perkins et al., 2024a), a transposition that strips them of the disciplinary judgement they were intended to carry and that policy analyses of institutional adoption have repeatedly observed (Chan, 2023; Yan et al., 2024). Third, and most consequentially, the apparatus is unilateral. No comparable instrument exists on the marker side: institutions that require students to specify the level of assistance used in producing an assignment do not require markers to specify the level of assistance used in judging it, and do not report such use to students in aggregate or in the individual case.

The governance implication is symmetry. A declaration regime that binds only one party to an assessment relationship is not a transparency mechanism but a surveillance mechanism, and is likely to be understood as such by students, with predictable consequences for their trust in the resulting judgements.

4.4 Security: defensibility without validity

Security is the dimension on which the sector has moved fastest and with the greatest coordination. The two-lane model distinguishes secure tasks, conducted under conditions that permit confident attribution of the work to the student, from open tasks in which generative tools may be used deliberately and productively (Liu & Bridgeman, 2023; TEQSA, 2023). Subsequent regulatory guidance has elaborated implementation pathways, including programme-wide reform, unit-level assurance, and hybrid arrangements combining the two, and has been explicit that detection cannot substitute for structural change (Corbin et al., 2025; TEQSA, 2025).

Three qualifications recur. First, the model's coherence depends on secure tasks assuring programme-level rather than unit-level outcomes; where every unit adds a supervised task to demonstrate compliance, the sector acquires the cost of the reform without its logic, and the fragmentation this produces is acknowledged in the guidance itself (TEQSA, 2025) and anticipated in the sustainable assessment tradition that preceded it (Boud & Soler, 2016; Boud and Associates, 2010). Second, the resource intensity of supervision is unevenly distributed, and the pathway a provider selects is likely to reflect its capability rather than its pedagogical

position, a hypothesis that has not been tested and that the sector currently has no mechanism to observe, notwithstanding that comparable capability disparities have been documented in the adoption of educational technology more broadly (Yan et al., 2024; Zawacki-Richter et al., 2019). Third, the model is a response to a technology that continues to change; assurance conditions calibrated to text generation are already being tested by multimodal and agentic systems (Corbin et al., 2025), and a governance instrument that requires recalibration with each capability release is a costly one to depend upon.

The governance implication is that security is necessary and is not an achievement in itself. An institution can secure every unit in a course and answer no question about whether its graduates can do what its course claims they can do.

4.5 Equity: the dimension without a mechanism

Equity is the least developed dimension in both scholarship and practice, and the one where the consequences of the current settlement fall most unevenly.

Three features of the present arrangement bear on it. The first is differential access. Capability differences between free and subscription tiers of generative systems are substantial and widening, so a regime that permits or requires tool use in open tasks distributes advantage by capacity to pay, and one that prohibits it distributes advantage by capacity to conceal (Selwyn, 2019; Yan et al., 2024). Concern that unequal access to these systems will deepen existing educational disparities is widely expressed in the field (Holmes et al., 2022; Nguyen et al., 2023; UNESCO, 2023), but it has been examined empirically at scale in neither direction, and the assessment literature has largely inherited the concern without testing it. The second is the differential burden of supervision. Timed, invigilated, handwritten performance is not neutral with respect to disability, chronic illness, caring responsibility or English as an additional language, and a sector-wide migration of assurance into that mode is a sector-wide redistribution of disadvantage, precisely the concern that scholarship on assessment for inclusion has articulated in relation to conventional examinations (Nieminen, 2024). The third is contestation. Integrity allegations arising from or supported by detection tools rest on instruments of unknown and unstable accuracy (Weber-Wulff et al., 2023), and the bias in those instruments is not random: controlled evaluation has found that widely used detectors misclassify a majority of writing by non-native English authors as machine-generated while classifying comparable native-speaker writing correctly, an effect attributed to the lower lexical variability of the former (Liang et al., 2023). Automated marking systems raise a parallel concern, since scoring divergences in these systems are systematic rather than random and are concentrated in the context-dependent dimensions of a rubric (Flodén, 2025; Pack et al., 2024). Yet the student's capacity to contest either an allegation or a mark depends on institutional procedures designed for determinations grounded in documentary evidence, and the ethical frameworks the sector has adopted specify transparency and fairness as principles without specifying the machinery through which either is enforced (Holmes et al., 2022; Nguyen et al., 2023).

This is where the analytic contribution of administrative accountability scholarship is clearest. Peeters and Widlak (2018), analysing exclusion produced by information architecture, and Busuioc (2021), on holding algorithmic determinations to account, supply the relevant apparatus: where determinations are automated or automation-assisted, accountability requires a functioning individual route to challenge, not merely a policy assertion that a human remains responsible. A student who cannot know that a generative system contributed to their mark, or who cannot obtain the basis on which a detection tool flagged their work, holds a right of appeal that they have no means of exercising. A right without a remedy is, in this setting, a procedure recorded in a policy that the student cannot reach.

4.6 Propositions and research agenda

The framework yields six propositions, offered as testable statements for subsequent empirical work.

P1. Where an institution's assessment reform is oriented primarily to security, gains in the defensibility of its judgements will exceed gains in their validity, and the divergence will be greatest in disciplines whose learning outcomes are least amenable to supervised demonstration.

P2. Institutions that impose disclosure obligations on students without imposing symmetrical obligations on markers will exhibit lower student trust in assessment outcomes than institutions imposing both, controlling for institutional type and cohort composition.

P3. The proportion of academic integrity allegations that are subsequently withdrawn or not sustained will be higher where the allegation originated in a detection tool than where it originated in unaided academic judgement.

P4. Non-self-accrediting providers will adopt unit-level assurance in preference to programme-wide reform at a higher rate than self-accrediting providers, controlling for discipline mix, reflecting differences in coordination capability rather than in pedagogical position.

P5. Where automated marking or feedback is used without recorded override, the resulting distribution of marks will exhibit lower variance than under human-only marking, and the reduction will be concentrated at the tails of the distribution.

P6. Assessment designs that assess process alongside product will exhibit smaller measured attainment differences between students with and without subscription access to frontier generative tools than product-only designs.

Four methodological priorities follow. First, comparative designs across provider types, since the evidence base is drawn almost entirely from large self-accrediting institutions and the framework's central equity claims concern providers that are absent from it. Second, longitudinal work capable of distinguishing the effects of assessment reform from cohort effects and from the continuing capability trajectory of the underlying technology, which changes the treatment condition mid-study in a way few educational interventions do. Third, administrative-data studies of allegation, appeal and override, covering volumes, outcomes, resolution times and disaggregation by student characteristic, which are almost entirely absent from current scholarship and which would test P3 and P5 directly. Fourth, participatory and co-design research with students on declaration and disclosure regimes, which would address the transparency asymmetry both as an object of study and as a method.

For practice, the framework implies a revision of institutional assessment reporting. A report stating the proportion of units containing a secured task measures assurance alone. Adding three indicator families, namely evidence that secured outcomes were selected on validity grounds, marker-side disclosure and override rates alongside student declaration rates, and disaggregated allegation and appeal outcomes, would render warrant and fairness visible alongside assurance. For course designers, the immediate implication is narrower and more actionable: the decision to secure a task should be recorded together with the outcome that decision has made it harder to assess authentically, so that the trade is documented at the point it is made rather than discovered at the point of course review.

4.7 Limitations

Three limitations qualify the argument. The paper is conceptual: the framework is derived from the literature rather than tested against it, and while the four dimensions are analytically distinct, their empirical separability, particularly the boundary between validity and security, which interact rather than merely coexist, requires validation. The empirical illustrations are drawn disproportionately from Australian regulatory practice, reflecting the unusual degree of coordination in that setting and the consequent observability of institutional response; the framework's applicability to jurisdictions without a comparable national instrument, and to systems in which assessment governance is devolved to the discipline, requires separate examination. And the review underpinning the argument is narrative rather than systematic, selected for theoretical relevance rather than exhaustive coverage; a systematic review testing the coverage claims made in Section 4, particularly the claim that marker-side use is under-examined relative to student-side use, would be a valuable complement and is the natural next study.

5. Conclusion

Assessment in the generative artificial intelligence era has been theorised as a problem of student conduct. That theorisation was adequate while the technology sat on one side of the desk. It is no longer adequate now that the same systems draft the work and, increasingly, judge it. Where both are true, the question is not whether a particular student used a tool but whether the institution can still warrant what its marks and qualifications claim, and that warrant depends on decisions about task design, disclosure, assurance and contestation that are made within the institution, by identifiable staff, and are therefore susceptible to governance in a way that student conduct is not.

The four-dimensional framework advanced here is offered as a corrective to an evaluative discourse in which success is measured by tasks secured. Security is necessary but not sufficient; validity is the question that security displaced rather than answered; transparency binding one party only is surveillance; and an institution that judges with tools it does not disclose, on grounds a student cannot reach, has a procedure rather than a remedy. Asking only how many tasks have been secured leaves unasked the questions that determine what a qualification is worth: what it now measures, who contributed to the judgement, and what a student can do when that judgement is wrong.

For the field, the invitation is straightforward. The technology that has destabilised assessment is not only in the hands of students. Its governance on the institutional side is an assessment governance problem, and the conceptual apparatus the field already possesses, namely validity, transparency, fairness and contestability, is precisely what the study of artificial intelligence in assessment currently lacks.

Ethics Statement

Ethics approval was not required, as the paper is a conceptual analysis of published literature.

References

Ajjawi, R., Tai, J., Dollinger, M., Dawson, P., Boud, D., & Bearman, M. (2024). From authentic assessment to authenticity in assessment: Broadening perspectives. *Assessment & Evaluation in Higher Education, 49*(4), 499–510. https://doi.org/10.1080/02602938.2023.2271193

Bearman, M., & Ajjawi, R. (2023). Learning to work with the black box: Pedagogy for a world with artificial intelligence. *British Journal of Educational Technology, 54*(5), 1160–1173. https://doi.org/10.1111/bjet.13337

Bearman, M., Ryan, J., & Ajjawi, R. (2023). Discourses of artificial intelligence in higher education: A critical literature review. *Higher Education, 86*(2), 369–385. https://doi.org/10.1007/s10734-022-00937-2

Boud, D. and Associates. (2010). *Assessment 2020: Seven propositions for assessment reform in higher education*. Australian Learning and Teaching Council.

Boud, D., & Soler, R. (2016). Sustainable assessment revisited. *Assessment & Evaluation in Higher Education, 41*(3), 400–413. https://doi.org/10.1080/02602938.2015.1018133

Busuioc, M. (2021). Accountable artificial intelligence: Holding algorithms to account. *Public Administration Review, 81*(5), 825–836. https://doi.org/10.1111/puar.13293

Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. *International Journal of Educational Technology in Higher Education, 20*(1), 38. https://doi.org/10.1186/s41239-023-00408-3

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education, 20*(1), 43. https://doi.org/10.1186/s41239-023-00411-8

Corbin, T., Dawson, P., & Liu, D. (2025). Talk is cheap: Why structural assessment changes are needed for a time of GenAI. *Assessment & Evaluation in Higher Education, 50*(7), 1087–1097. https://doi.org/10.1080/02602938.2025.2503964

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International, 61*(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148

Dawson, P. (2021). *Defending assessment security in a digital world: Preventing e-cheating and supporting academic integrity in higher education*. Routledge. https://doi.org/10.4324/9780429324178

Dawson, P., Bearman, M., Dollinger, M., & Boud, D. (2024). Validity matters more than cheating. *Assessment & Evaluation in Higher Education, 49*(7), 1005–1016. https://doi.org/10.1080/02602938.2024.2386662

Fleckenstein, J., Meyer, J., Jansen, T., Keller, S. D., Köller, O., & Möller, J. (2024). Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays. *Computers and Education: Artificial Intelligence, 6*, 100209. https://doi.org/10.1016/j.caeai.2024.100209

Flodén, J. (2025). Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. *British Educational Research Journal, 51*(1), 201–224. https://doi.org/10.1002/berj.4069

Gardner, J., O'Leary, M., & Yuan, L. (2021). Artificial intelligence in educational assessment: 'Breakthrough? Or buncombe and ballyhoo?' *Journal of Computer Assisted Learning, 37*(5), 1207–1216. https://doi.org/10.1111/jcal.12577

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education, 32*, 504–526. https://doi.org/10.1007/s40593-021-00239-1

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

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*(7), 100779. https://doi.org/10.1016/j.patter.2023.100779

Liu, D., & Bridgeman, A. (2023, July 12). What to do about assessments if we can't out-design or out-run AI? *Teaching@Sydney*. The University of Sydney. https://educational-innovation.sydney.edu.au/teaching@sydney/what-to-do-about-assessments-if-we-cant-out-design-or-out-run-ai/

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Nguyen, A., Ngo, H. N., Hong, Y., Dang, B., & Nguyen, B.-P. T. (2023). Ethical principles for artificial intelligence in education. *Education and Information Technologies, 28*, 4221–4241. https://doi.org/10.1007/s10639-022-11316-w

Nieminen, J. H. (2024). Assessment for inclusion: Rethinking inclusive assessment in higher education. *Teaching in Higher Education, 29*(4), 841–859. https://doi.org/10.1080/13562517.2021.2021395

Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence, 6*, 100234. https://doi.org/10.1016/j.caeai.2024.100234

Peeters, R., & Widlak, A. (2018). The digital cage: Administrative exclusion through information architecture — The case of the Dutch civil registry's master data management system. *Government Information Quarterly, 35*(2), 175–183. https://doi.org/10.1016/j.giq.2018.02.003

Perkins, M. (2023). Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching and Learning Practice, 20*(2), 07. https://doi.org/10.53761/1.20.02.07

Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2024a). The Artificial Intelligence Assessment Scale (AIAS): A framework for ethical integration of generative AI in educational assessment. *Journal of University Teaching and Learning Practice, 21*(6), 49–66. https://doi.org/10.53761/q3azde36

Perkins, M., Roe, J., Postma, D., McGaughran, J., & Hickerson, D. (2024b). Detection of GPT-4 generated text in higher education: Combining academic judgement and software to identify generative AI tool misuse. *Journal of Academic Ethics, 22*, 89–113. https://doi.org/10.1007/s10805-023-09492-6

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119–144. https://doi.org/10.1007/BF00117714

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

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research, 104*, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Stephenson, R., & Armstrong, C. (2026). *Student generative artificial intelligence survey 2026 (HEPI Report 199)*. Higher Education Policy Institute. https://www.hepi.ac.uk/reports/student-generative-ai-survey-2026/

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence, 3*, 100075. https://doi.org/10.1016/j.caeai.2022.100075

Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement: Enabling students to make decisions about the quality of work. *Higher Education, 76*, 467–481. https://doi.org/10.1007/s10734-017-0220-3

Tertiary Education Quality and Standards Agency. (2023). *Assessment reform for the age of artificial intelligence*. TEQSA.

Tertiary Education Quality and Standards Agency. (2025). *Enacting assessment reform in a time of artificial intelligence*. TEQSA.

United Nations Educational, Scientific and Cultural Organization. (2023). *Guidance for generative AI in education and research*. UNESCO. https://doi.org/10.54675/EWZM9535

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity, 19*, 26. https://doi.org/10.1007/s40979-023-00146-z

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology, 55*(1), 90–112. https://doi.org/10.1111/bjet.13370

Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology, 56*(1), 150–166. https://doi.org/10.1111/bjet.13494

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*, 39. https://doi.org/10.1186/s41239-019-0171-0