Beyond AI Detection: Transforming Educational Assessment in Higher Education through Generative AI

Ayesha Nawaz[1], Danish Kamal[1]*

[1] Department of Education, The University of Lahore, Pakistan

**Corresponding author: Danish Kamal, Email: deshecho60@gmail.com*

Abstract

Generative artificial intelligence (GenAI) has destabilized the evidentiary basis of higher-education assessment. The dominant institutional reflex has been detection: software that attempts to separate machine-written from human-written work. This review argues that detection is not merely imperfect but structurally self-defeating. It commits a category error, authenticating an artifact rather than assuring a person, and it locks institutions into a losing arms race in which detectors and generators co-evolve while a severe error asymmetry falls hardest on already-marginalized students. Drawing on assessment theory and the fast-growing GenAI-in-education literature, we reframe the problem around validity: the useful question is not whether a machine touched a submission, but whether a task still licenses a trustworthy inference about what a student can do. We introduce the Assessment Trilemma, a heuristic in which validity, security, and authenticity form a three-way tension that no single task can fully resolve, and use it to explain why prevailing responses, from prohibition and proctoring to the AI Assessment Scale, the two-lane model, authentic redesign, and postplagiarism, each succeed only by sacrificing one corner. The article then argues that the trilemma eases at the level of the program rather than the task: assurance should be distributed across secured checkpoints, authentic AI-integrated tasks, and process evidence that together triangulate a warranted judgment of capability. We close with the ethical stakes and a research agenda for authentic, ethical, and future-ready assessment.

Keywords: generative artificial intelligence; assessment validity; academic integrity; authentic assessment; programmatic assessment; higher education

1. Introduction

When large language models became freely available to students at the end of 2022, they did not introduce a new form of misconduct so much as expose an old assumption. Higher education had long treated the unsupervised written artifact, the essay, the take-home problem set, the literature review, as a reliable proxy for what a student knew and could do. That proxy worked because producing a competent artifact was assumed to require the competence being certified. Generative artificial intelligence (GenAI) severs this link. A fluent, structurally sound, superficially well-referenced text can now be produced in seconds by someone who has learned little, and the resulting output is frequently indistinguishable from capable human work [1,2].

Faced with this rupture, the sector's first and still most common response has been to reach for detection: commercial classifiers that assign a probability that a submission was machine-generated. The appeal is obvious, because detection promises to restore the old proxy without redesigning anything. This review contends that the appeal is a trap. Detection cannot deliver what institutions

want from it, not because the tools are immature, but because the strategy misidentifies the problem. The peril that GenAI presents is not chiefly to creativity; rather, it concerns validity, which pertains to how well an evaluation substantiates accurate conclusions on student competence [3,4]. When seen from this perspective, the objective shifts from apprehending machines to crafting tasks and, importantly, programs whose outcomes remain reliable.

This article transforms the reframing into a reasoned, organized narrative. We initially demonstrate why detection is fundamentally self-sabotaging rather than simply erroneous. We subsequently shift the issue from the artifact to the individual and from the task to the program, using a heuristic termed the Assessment Trilemma to elucidate the conflicts that render every singular task resolution incomplete. In that context, we analyze and evaluate the foremost reactions within the field, and we propose a systematic paradigm where assurance is disseminated rather than centralized. Ultimately, we regard the ethical and forward-looking implications with gravity, as a justifiable evaluation framework must be fair and equip graduates for professional environments where AI techniques are commonplace rather than prohibited. Our contribution is synthetic: we do not present new empirical findings, but rather we consolidate a disjointed and rapidly evolving body of literature into a cohesive argument and a limited set of decision-making tools for educators and institutions to utilize.

This evaluation is essentially constructive and argument-driven, concentrating on integrating and appraising different perspectives rather than carefully recording every scholarly work. We strategically focus on three research subjects: empirically examining AI's strengths and constraints in test environments, recognized theories related to examination accuracy and realistic testing, and the fast-evolving design study on GenAI and ethics, focusing on research published in credible test journals and insights from quality control associations. Considering the accelerated developments in both model efficiency and organizational approaches, we see certain findings as representative of patterns instead of absolute truths, underscoring fundamental notions that are likely to remain relevant regardless of the emergence of certain instruments.

2. The Detection Trap

Four characteristics collectively render detection an inadequate basis for the validity of evaluation. The initial aspect is empirical inconsistency. Independent evaluations have consistently demonstrated that AI-writing detectors inaccurately categorize both artificial and human-generated text at levels unsuitable for critical choices, with vendors explicitly warning that their evaluations are suggestive rather than definitive and should not be regarded as evidence of wrongdoing [5]. A seemingly minimal document-level false-positive frequency can have significant implications when used at scale, since a university handling hundreds of thousands of submissions will erroneously identify a significant percentage of legitimate people.

The scope of this difficulty is frequently misinterpreted. At the beginning of AI-writing identification deployment, the vendors reported a document-level false-positive percentage near one percent, but objective analyses depicted a less optimistic scenario: small-scale investigations erroneously classified genuine, completely human-written student texts as AI-generated, with sentence-level error rates considerably surpassing the whole record statistic. Providers have since integrated clauses, underscoring low-confidence findings and alerting institutions that a score does not denote wrongdoing

[5]. The openness is esteemed, although it detracts from the initial promise. An apparatus that demands a cautionary statement suggesting it should not serve as verification cannot sustain an academic structure, and an institution that perceives it as such is forming essential conclusions on a premise that its own producer denies.

The second trait is a significant and morally impactful error imbalance. The consequences of a detector's malfunction are inherently unequal. A false negative permits an occurrence of misconduct to remain undetected, which is regrettable yet amendable. A false positive designates a student as devoid of misconduct, resulting in an accusation that is anxiety-provoking, detrimental to their credibility, and difficult to contest, as the suspect finds it hard to contradict a statistical claim about their personal work. This disparity erodes the perception of legitimacy that a fair process depends upon and shifts the burden of proof onto the student [6].

Detectors focus on the quantitative attributes within writings, rendering material that seems conventional, organized, or insufficiently unique prone to identification. This hinders non-native English authors, who often employ a limited range of vocabulary and syntax, and it endangers neurodivergent and disabled children when their distinctive writing approaches depart from conventional educational benchmarks [7]. A method that consistently transforms linguistic diversity into distrust is not an impartial protection; it is an issue of fairness hiding as precision.

The fourth attribute is adaptability and, over time, crucial. Detection and production develop in parallel. Each advancement in creative readability undermines the statistical signals upon which detectors relied, yet easily obtainable paraphrasing and humanizing techniques allow a determined student to bypass any classification system using machine language. The steadfast fraudster is therefore the one least susceptible to being apprehended, whereas the truthful but unconventional scribe endures misleading implications. This illustrates a Red Queen dynamic: entities are compelled to enhance recognition capabilities incessantly to remain stationary, as the target perpetually retreats with every model introduction [8].

Underneath these four attributes resides a more profound conceptual error. Detection endeavors to verify an artifact, ensuring that a specific text was generated without the assistance of machinery. However, evaluation should not, or ought not to, regard the artifact for its intrinsic value. It considers the individual: their comprehension, their abilities, and the conclusions about their potential that the artifact permits. Once GenAI is capable of generating the artifact, verifying its authenticity reveals little about the individual. The type of mistake is continuing to monitor the agent once it has ceased its function as a proxy. Evading a recognition mechanism necessitates shifting the focus of validation from the material to the underlying competence [4,9].

3. From Artifact to Capability: Putting Validity First

Validity constitutes the essential tenet mandated by the industry. In modern contexts, validity relates not to the assessment itself but to the inferences drawn from its results: a judgment is considered accurate if results support the interpretations and subsequent measures taken [3]. GenAI endangers legitimacy in a specific way. It allows students to complete an activity without possessing the necessary skills that the task is designed to assess, hence weakening the link between educational success and competence. Dawson and colleagues expressed unequivocally: the foremost issue for educators is not

identifying when a student committed academic fraud, but instead asking whether we can still have confidence that graduates retain the skills their credentials assert they do [4].

It is beneficial to recognize that GenAI does not undermine validity in a singular, straightforward manner but rather assaults multiple aspects simultaneously. According to Messick, validity encompasses various dimensions, including the extent to which a task reflects the intended content, requires the requisite mental operations, aligns with the construct's internal makeup, predicts results across different contexts, correlates suitably with external criteria, and ensures defensible outcomes from its application [3]. An AI-generated unauthorized writing may exhibit superficial material while lacking depth and structural integrity, as the cognitive activity it was intended to demonstrate did not take place; its applicability diminishes, as the identical student might not replicate the results in different contexts; and its consequential validity is compromised, as decisions made based on it could be unreliable. Identifying the components is crucial in practice, since it provides a designer with clear directives on what a revamped task must reinstate, rather than allowing validity to remain an ambiguous goal.

This transition is emancipating since it redirects focus from monitoring to creation. It additionally establishes an essential differentiation between construct-relevant and construct-irrelevant applications of AI. The issue of AI utilization is contingent upon the specific metrics being assessed. If the objective measure is a student's independent capacity to deduce a proof, then the utilization of AI is construct-irrelevant contamination and should be disregarded. If the goal is to assess the capacity to address a genuine professional issue, and that issue is typically resolved using AI techniques, then prohibiting AI would jeopardize validity by evaluating a contrived, diminished representation of the skill [8,10]. There is no definitive response regarding the usage of AI by students; rather, the pertinent inquiry revolves around the specific competencies being validated and the circumstances under which a performance might be deemed reliable.

Reconceptualizing validity alleviates the ethical hysteria associated with hybrid human-AI collaboration. Eaton's concept of a postplagiarism epoch illustrates the trend: collaboration using intelligent technologies is becoming commonplace, and distinguishing between human and machine contributions is increasingly impractical [9]. What remains intact is accountability. An author may assign the technical aspects of creation to a tool, yet cannot transfer responsibility for precision, discernment, and the assertions presented. Validity-focused evaluation can simultaneously encompass two concepts: that AI-aided creation is permissible in many situations, and that the learner is accountable for the caliber and authenticity of the outcome.

4. The Assessment Trilemma

If validity is the objective, two additional requirements impose themselves on any genuine evaluation. It must be sufficiently secure to link success to the student's educational progress, and it is expected to be sufficiently true that the skills it validates are applicable outside the classroom. We assert that the trio of requirements: validity, security, and authenticity, constitutes a trilemma: enhancing any two within a single task generally compromises the third (**Figure 1**). The trilemma serves as a pragmatic rather than an empirical principle, yet it possesses genuine explanatory strength, as it forecasts the compromises that have hindered the industry's quest for a singular resolution.

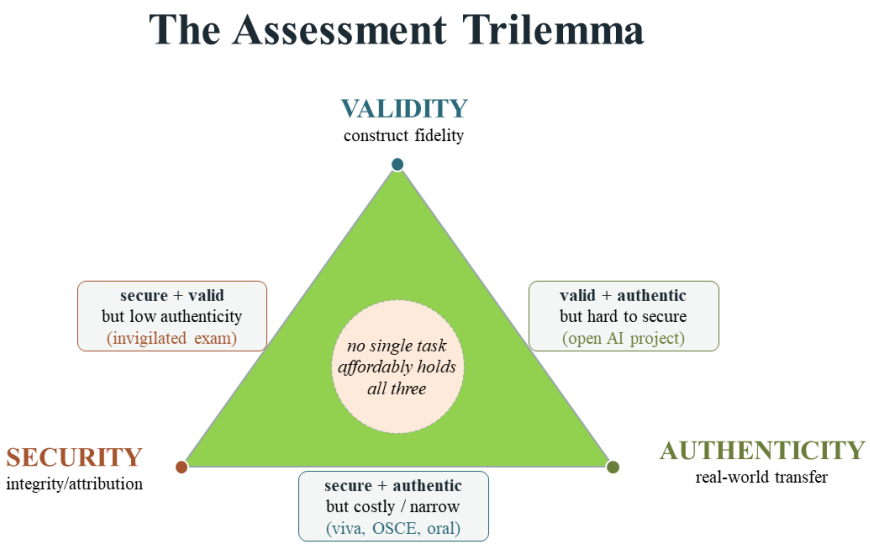

Figure 1. The Assessment Trilemma. Validity (construct fidelity), security (attribution/integrity), and authenticity (real-world transfer) conflict with one another in each given job. Dominant reactions vary primarily in the specific aspect they compromise; no singular task can economically dominate the center, necessitating that confidence be derived from an extensive curriculum rather than being focused on a solitary evaluation.

Analyze these three corners; a traditional monitored investigation straddles the boundary of security and credibility: we may delegate the content to the student and, with well-structured items, have confidence in the inferences about knowledge, yet the sterile prerequisites bear scant analogy to actual life situations, and for various fields, this test represents a more restricted premise than its syllabus claims. An unrestricted, home-based project enabling students to leverage AI exists on the threshold of validity and authenticity: it can reflect true professional activities and, if well-designed, enhance higher-order skills, but its security is inadequate as confirming individual contributions is not straightforward. An interactive oral, viva, or clinical examination, exemplified by an OSCE, straddles the boundary of security and authenticity: it is both accountable and realistic, but it incurs high costs, is complex to scale, and often represents only a tiny segment of the intended construct within the allotted time.

Viewing the trilemma as a recommendation of despair would constitute a misjudgment. It does not imply that valuable appraisal is unattainable; it emphasizes that its excellence must be found in the suitable arena. A pertinent example is the formulation of a financial portfolio, where no single asset can achieve the trifecta of yield, safety, and liquidity; hence, a judicious investor curates a well-rounded strategy instead of chasing an unachievable flawless product. Those who design assessments are in a comparable position. Accepting that the quintessential single task is a myth alleviates the concern that has propelled the industry towards more pervasive recognition, allowing for a shift in focus to the more solvable challenge of how a well-curated selection of jobs can collectively fulfill what any single activity fails to accomplish.

The trilemma alters the whole argument. The reason the contention remains unresolved is not that the

field has been unsuccessful in developing an innovative design that addresses all three needs simultaneously; it is that no such comprehensive design is feasible. Each argument serves, essentially, as a resolution over which element to abandon, either explicitly spoken or implicitly understood. Seen this way, the sterile opposition between prohibition and permission is revealed as a false choice between two edges of the same triangle, and the productive question becomes how to combine tasks so that a program covers all three corners even though no individual task can.

5. Mapping the Field's Responses

The leading responses to GenAI in assessment can now be read as different bets on the trilemma. **Table 1** summarizes them; the discussion that follows draws out what each achieves and what it forfeits.

Table 1. Prevailing responses to generative AI in assessment mapped against the Assessment Trilemma

Response	Core mechanism	Corner prioritized	What it sacrifices	Representative sources
Detection and proctoring	Classify text or surveil the exam to authenticate the artifact	Security	Equity, trust, and validity; loses the arms race over time	[5,7]
Prohibition	Forbid AI use across the board	Security (nominal)	Authenticity and enforceability; unverifiable and often ignored	[6,11]
AI Assessment Scale (AIAS)	Named levels of permitted AI use per task, communicated to students	Validity via clarity	Relies on disclosure and integrity; not itself a security mechanism	[5,12]
Two-lane / secured-vs-open	Split tasks into secured (no AI) and open (AI-rich) lanes at program level	Security + Authenticity, separately	Treats a continuum as a dichotomy; a bare "all-or-none" is insupportable	[6,13]
Authentic redesign	Realistic, higher-order tasks tied to professional practice	Authenticity + Validity	Lower security unless combined with attribution mechanisms	[14–16]
Postplagiarism / integrity reframing	Accept hybrid human-AI work; foreground disclosure and responsibility	Validity via norms	Offers ethics and culture, not task-level assurance on its own	[9]

Note: Each response is coherent, but each purchases its strength by conceding a corner; none is sufficient alone.

5.1 Prohibition and Detection

Prohibition and detection are the reflex pairing, and they fail together. A blanket prohibition is unenforceable in unsupervised settings and, as Curtis argued, an "all-or-none" logic is insupportable once one notices that the sector already tolerates other undetectable risks, such as contract cheating, without abandoning assessment altogether [6]. Prohibition presents a tangible price: it obstructs the valid and pertinent applications of AI that are integral to numerous occupational skill sets. Detection, as stated in Section 2, is unable to verify the prohibition it aims to maintain.

5.2 The AI Assessment Scale

The AI Assessment Scale (AIAS) represents a significant progression by substituting a binary system with a nuanced, articulate array of grades, ranging from no AI utilization to complete AI interaction, tailored to align with the educational objectives of each activity [13]. A preliminary application indicated more defined objectives and less uncertainty for personnel and learners [11], and the framework has subsequently been enhanced to bolster its connection to validity and to shift away from a disciplinary perspective [12]. Its fundamental constraint is that it serves as an interaction and creative instrument, rather than a security apparatus: it informs students of what is allowed; however, it does not guarantee that the authorization was respected on its own. That is less a defect and more an allocation of responsibilities, highlighting the necessity for supplementary assurance.

5.3 The Two-Lane Model and Its Critics

The two-pronged strategy, established at the University of Sydney, spreads the obstacles within an activity by assigning some evaluations as verified, with AI prohibited and credit confirmed, and others as open, where AI is employed to promote genuine, future-oriented education [12,17]. The main insight is correct and prefigures our argument: assurance is a program-level attribute, and the secured lane need only endorse the principal outputs it is capable of certifying. The assessments are likewise informative. Curtis revealed that a hard binary viewpoint is unsustainable and suggested that a controlled compromise, where constrained AI utilization is allowed with stipulations, is necessary. Fundamentally, safety is a spectrum as opposed to an immediate operation; defining a route as only secure inflates the assurances that any situation may deliver. The two-lane paradigm is best comprehended not as a binary distinction but as a mechanism for controlling a dimension of the trilemma, and it is effective solely when the secured lane is authentically safeguarded and the open lane is honestly oriented towards training [6].

5.4 Authentic Redesign and Postplagiarism

Genuine evaluation, long promoted for its autonomous educational rationale, acquires heightened significance in this context. Its characteristic aspects, authenticity, intellectual rigor, and the cultivation of critical assessment, delineate activities that defy cursory execution and validate valuable competencies [14,18]. However, validity by itself does not address safety; a practical take-home assignment remains a take-home assignment. Postplagiarism, on the other hand, provides an academic and moral framework, legitimizing transparent, accountable human-AI partnerships and transforming organizations from a regulatory stance to an educational approach [9]. Neither genuine reform nor postplagiarism offers a comprehensive solution; nonetheless, each provides a perspective that the other lacks, which is why they should be integrated rather than prioritized.

Two advancements merit attention as they effectively merge traditional classifications. The first aspect is the emergence of software that focuses on documenting processes instead of monitoring outcomes: systems that allow for writing and edit tracking enable evaluators to understand the evolution of a project, reinstating a degree of credit to otherwise collaborative jobs without the imposition of oversight. The second aspect is the increasing implementation of interactive oral assessments as a scalable and secure method, wherein a concise, organized dialogue confirms the actual competence

behind a submission. Neither option serves as a panacea, and both include expenses; yet, they are significant as they fit more seamlessly inside the trilemma's core compared to traditional examinations or essays, illustrating that the extremes are not static positions but areas that may be expanded by deliberate design.

6. Resolving the Trilemma: Programmatic Assurance

The ongoing learning is that the academic trilemma is incompatible within an individual job, but it may be orchestrated throughout a course of action. This is the basis of that recommendation, and it carries an esteemed pedigree. The concept of procedural evaluation within healthcare has historically renounced the assumption that one tool can authenticate ability, positing rather that effective choices stem from the methodical synthesis of multiple flawed data sources [19]. The same principle responds to the GenAI challenge. Rather than requiring synchronous validity, security, and authenticity for each work, we should develop a methodology where various duties represent distinct aspects, facilitating a reliable appraisal of competence through their confluence (**Figure 2**).

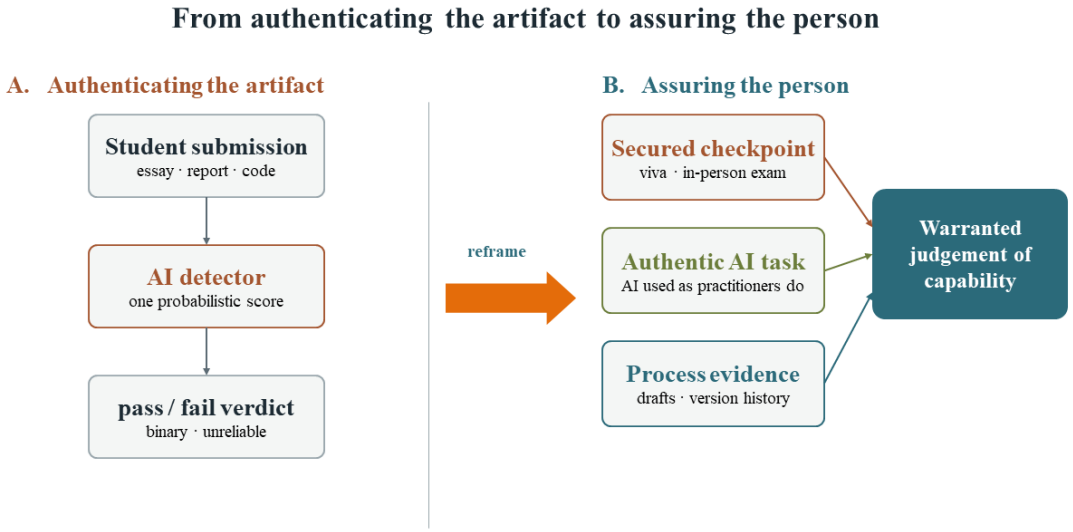

Figure 2. From authenticating the artifact to assuring the person

Three types of empirical proof, when integrated, address the trilemma. Established boundaries, positioned at significant milestones instead of at each activity, ensure privacy and authenticity for the results that need to be validated as the student's original work. Dynamic oral evaluations and vivas are particularly effective in this context, as a concise, organized dialogue can test comprehension in ways that are challenging to delegate and that disclose when the competence underlying a presented product genuinely exists. Genuine, AI-enhanced assignments establish credibility and, if the framework genuinely incorporates AI-supported duties, legitimacy as well; the educational objective is not to inhibit AI utilization but to render it apparent, intentional, and justifiable, frequently backed by a disclosure log documenting the usage of tools. The examination of proof substantiates identification requiring oversight: frameworks for writing, historical versions, drafts, and organized analysis render the progression of a work comprehensible, enabling evaluators to assess an evolving skill rather than a

solitary final product [8].

Table 2 transforms this into a functional design toolkit, correlating prevalent forms to their respective anchors, the AI functions they embody, and the security mechanisms that depend upon them. The crux of the matter does not require that all applications demand every sequence, but rather that a justifiable project must demonstrate that, throughout its evaluations, all three dimensions are addressed, and the fundamental results traverse at least one authentically protected barrier.

Table 2. Assessment formats mapped to the Assessment Trilemma, intended constructs, AI roles, and assurance mechanisms

Format	Trilemma anchor	Intended construct	AI role	Assurance mechanism	References
Interactive oral / viva	Security + Validity	Depth of understanding; ability to reason live	None during; may follow AI-assisted prep	Real-time human probing; hard to outsource	[20]
In-person practical / OSCE / lab	Security + Authenticity	Applied, embodied professional skill	As used in real practice, if any	Supervised performance in situ	[21]
Supervised digital exam (secure browser)	Security	Core knowledge that must be certified as own	Excluded	Invigilation; locked environment	[22]
Process-tracked writing	Validity + attribution	Reasoning and composition over time	Permitted and disclosed	Draft/version history; write-in platform	[23,24]
Authentic AI-integrated project	Authenticity + Validity	Realistic problem-solving with tools	Actively integrated; disclosed	Disclosure log; oral defense checkpoint	[25]
Program portfolio + reflection	Validity across time	Growth, evaluative judgment, integration	Permitted; reflected upon	Longitudinal triangulation across tasks	[26]

This methodology does not claim to render deception unfeasible; none can. Its objective is better justifiable as well as transparent: to ensure that the whole assessment of competence remains resilient notwithstanding the outcome of any particular activity. Utilizing an example from protection technology, the software operates similarly to a pile of Swiss-cheese structures, filled with misaligned deficiencies, ensuring that an apprentice who excels in one activity must nevertheless exhibit proficiency in another area [4,27]. Reliability evolves into an inherent characteristic of the framework instead of an unattainable requirement imposed on each component.

The adoption of this model is an emergence rather than a decision, and it is useful to distinguish the phases commonly passed through by businesses. A lot of people start in a Detect phase, looking at GenAI as contraband to catch. Identification's limits push them to move to Restrict, to impose tighter

rules and to re-establish supervised settings, which purchases momentary safety at the cost of authenticity. The turning point is the switch to Redesign, where specific assignments are altered to reduce the likelihood of avoidance, and then Integrate, where the use of AI is explicitly integrated and assessed in the syllabus as appropriate for professional practice, with specific checkpoints set aside for validating the authenticity of the student's work. The progress is not automatic, and companies can quit at any time, but the important thing is that the identification of progress permits managers to appraise where they stand honestly, and to accept that the early phases, while reassuring, are simply way stations on the journey, not endpoints. Many start in a Detect mode, thinking of GenAI as contraband to be caught. With the failures of identification, this goes on to Restrict, introducing tighter rules and re-establishing surveilled spaces, providing brief protection at the cost of authenticity. The pivotal phase is a shift to Redesign, where some tasks are altered to mitigate the risk of avoidance, next to Integrate, in which the use of AI is explicitly integrated and assessed according to the syllabus for relevant academic or professional contexts, with defined benchmarks defined to validate the student's creative work [6,8,10,13]. **Figure 3** encapsulates the advancement of institutional maturity.

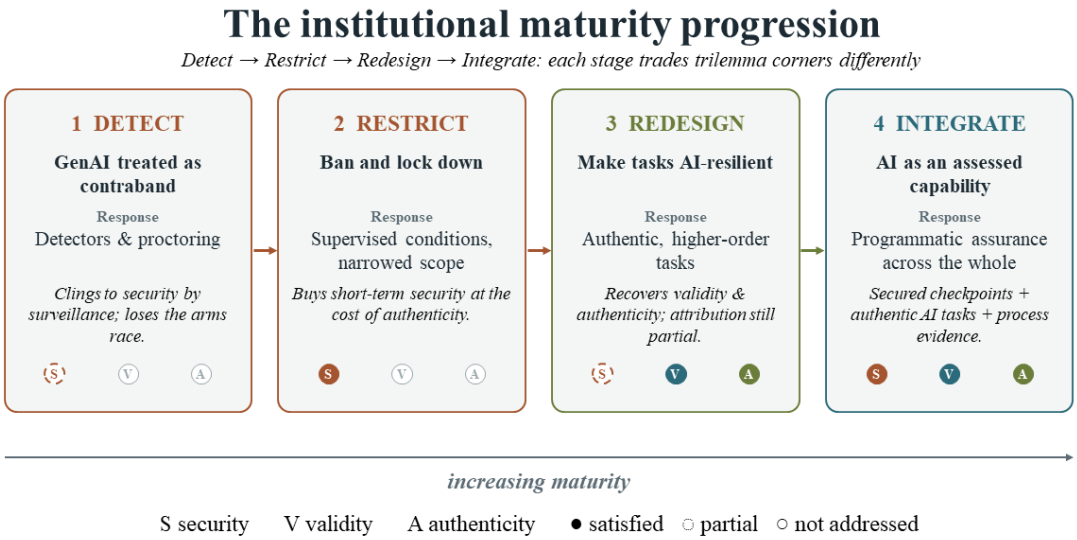

Figure 3. The institutional maturity progression for assessment in the age of GenAI. The framework proposes a progression from Detect and Restrict, through Redesign, to Integrate, reflecting increasing institutional maturity in responding to GenAI. Each phase addresses the tensions among security (S), validity (V), and authenticity (A) differently. The framework is author-developed as a synthesis of the assessment and GenAI literature.

A practical example elucidates the manner in which the edges meet. Considering a vocational qualification like medicine or architecture, essential information that should be validated as the students' own (such as accurate drug estimation, fundamental theories, and load-bearing fundamentals) may be guaranteed by only a handful of monitored assessments and brief active oral examinations, ensuring reliability and authenticity. The majority of the curriculum may comprise genuine, AI-enhanced assignments, style initiatives, care strategies, and data evaluations, where students are anticipated to employ tools like professionals do, with their methodology rendered transparent by sketches and a written record. A final verbal defense, when a student articulates and substantiates their

work in real time, completes the process by validating that the skills demonstrated throughout the unsecured assignments truly belong to the student. No individual component of such an approach is simultaneously reliable, valid, and entirely legitimate; nonetheless, the overall software assures competence with greater assurance than any set of detectors might.

7. Future-Ready Assessment: AI Literacy and Evaluative Judgment as Constructs

The last step is to acknowledge that proficient AI-augmented achievement is now core to the structure that schooling is meant to verify in several domains and to put down AI as a danger that needs to be restrained. Regardless of their level of expertise, a graduate designer who is unable to utilize creative resources or a researcher who is unable to seriously clear and assess an AI model is ill-prepared. An examination that prohibits the use of these instruments evaluates a retreating form of ability in cases when professional conduct has adopted them [8,10]. Thus, establishing AI proficiency as an obvious, evaluated result in the open lane rather than a covert talent learned in secret is necessary for future-ready evaluation.

Critical judgment is the meta-competence associated with these future-ready skills: the ability to compare the standard of an individual's and automated output to justifiable standards [28]. The capacity to distinguish between excellent and bad results, identify certain errors, improve prompts, and determine when a machine's response is insufficient is a rare and useful human talent if a device provides believable information on order. Bearman and associates contended that developing this judgment is exactly what evaluation for a GenAI era should focus on [29]. This offers the age-old concept of evaluation as training a fresh lease on life: assignments that compel students to evaluate, edit, and justify AI-assisted material help them acquire the precise evaluating skills that render their work credible and desirable [30]. The conceptual language for this change is derived from the conventional aspects of genuine evaluation, expanded to incorporate the virtual and the AI-mediated [31,32].

Daily responsibilities have a distinct feel when they are designed for critical judgment. In order to ensure that the measured skill is judgment rather than generation, a test could ask learners to create a proposal using AI, then evaluate it against curricular standards, point out any mistakes or implicit beliefs, and argue for their changes [29,33]. Although the machine lacks the evaluation that the learner is responsible for, such assignments are challenging to finish without comprehension. This is how true evaluation and predictive inspection come together: the identical approach that trains students to use AI ethically also reinstates the reliable reasoning that GenAI had undermined, transforming the risk into an opportunity for an improved review beyond what it replaced.

8. Conclusion

The instinct to detect AI misses the nature of the problem. GenAI has not made cheating uniquely easy so much as revealed that assessment was quietly relying on a proxy that no longer holds, and no classifier can rebuild that proxy without importing worse harms. The productive response is to change what we assure and where we assure it: to move from authenticating artifacts to assuring people, and from loading every demand onto a single task to distributing validity, security, and authenticity across a coherent program. This reframing turns a defensive scramble into a constructive project. It lets

institutions certify the outcomes that must be a student's own through a small number of genuinely secured checkpoints, embrace AI where professional practice already has, and cultivate the evaluative judgment that makes graduates both employable and trustworthy. Assessment reformed along these lines is not merely more resistant to AI; it is more authentic, more equitable, and better aligned with the world students will actually enter. That is a more demanding goal than catching cheats, and a far more worthwhile one.

Conflicts of Interest

The authors declare no conflicts of interest.

Author Contributions

Ayesha Nawaz: Concept of the Data, Literature Extraction, Manuscript Writing.
Danish Kamal: Concept of the Data, Literature Extraction, Manuscript Writing, Re-evaluation, Finalizing.
All authors have approved the final version of the manuscript.

Copyright and License

References

[1] Nikolic, S., Daniel, S., Haque, R., Belkina, M., Hassan, G. M., Grundy, S., Lyden, S., Neal, P., & Sandison, C. (2023). ChatGPT versus engineering education assessment: A multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity. *European Journal of Engineering Education, 48*(4), 559–614.

[2] Newton, P., & Xiromeriti, M. (2024). ChatGPT performance on multiple choice question examinations in higher education: A pragmatic scoping review. *Assessment & Evaluation in Higher Education, 49*(6), 781–798.

[3] Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education and Macmillan Publishing Company.

[4] Dawson, P., Bearman, M., Dollinger, M., & Boud, D. (2024). Validity matters more than cheating. *Assessment & Evaluation in Higher Education, 49*(7), 1005–1016.

[5] Perkins, M., Furze, L., Roe, J., MacVaugh, J. (2024). 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.

[6] Curtis, G. J. (2025). The two-lane road to hell is paved with good intentions: Why an all-or-none approach to generative AI, integrity, and assessment is insupportable. *Higher Education Research &*

Development, 44(8), 2151–2158.

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

[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.

[9] Eaton, S. E. (2023). Postplagiarism: Transdisciplinary ethics and integrity in the age of artificial intelligence and neurotechnology. *International Journal for Educational Integrity, 19*(1), 23.

[10] Tertiary Education Quality and Standards Agency. (2024). *A new reality: The role of simulated learning activities in postgraduate psychology training programs*. Australian Government.

[11] Furze, L., Perkins, M., Roe, J., & MacVaugh, J. (2024). The AI Assessment Scale (AIAS) in action: A pilot implementation of GenAI-supported assessment. *Australasian Journal of Educational Technology, 40*(4), 38–55.

[12] Perkins, M., Roe, J., & Furze, L. (2025). Reimagining the artificial intelligence assessment scale: A refined framework for educational assessment. *Journal of University Teaching and Learning Practice, 22*(7), 1–26.

[13] Perkins, M., Roe, J., & Furze, L. (2024). The AI assessment scale revisited: A framework for educational assessment. *arXiv*. https://arxiv.org/abs/2412.09029

[14] Villarroel, V., Bruna, D., Bruna, C., Brown, G., & Boud, D. (2024). Authentic assessment training for university teachers. *Assessment in Education: Principles, Policy & Practice, 31*(2), 116–134.

[15] Ajjawi, R., Tai, J., Huu Nghia, T. L., Boud, D., Johnson, L., & Patrick, C. J. (2020). Aligning assessment with the needs of work-integrated learning: The challenges of authentic assessment in a complex context. *Assessment & Evaluation in Higher Education, 45*(2), 304–316.

[16] Villarroel, V., Bloxham, S., Bruna, D., Bruna, C., & Herrera-Seda, C. (2018). Authentic assessment: Creating a blueprint for course design. *Assessment & Evaluation in Higher Education, 43*(5), 840–854.

[17] Tran, P., & Dinneen, C. (2025). Reimagining EAP in the age of GenAI: Innovation, integrity, and the future of assessment. *University of Sydney Journal in TESOL, 4*.

[18] Ashford-Rowe, K., Herrington, J., & Brown, C. (2014). Establishing the critical elements that determine authentic assessment. *Assessment & Evaluation in Higher Education, 39*(2), 205–222.

[19] Van Der Vleuten, C. P., & Schuwirth, L. W. (2005). Assessing professional competence: From methods to programmes. *Medical Education, 39*(3), 309–317.

[20] Shivshankar, S. (2024). Assessment integrity and assessment security in the digital era. In *Teaching and learning in the digital era: Issues and studies* (pp. 137–163). World Scientific.

[21] Australian Postgraduate Psychology Simulation Education Working Group (APPESWG). (2021). A new reality: The role of simulated learning activities in postgraduate psychology training programs. In *Frontiers in Education*. Frontiers Media SA.

[22] Cronqvist, D., & Kortesaari, S. (2023). *Securing electronic exam environments*. Chalmers University of Technology.

[23] Fathi, J., & Rahimi, M. (2026). Utilising artificial intelligence-enhanced writing mediation to develop academic writing skills in EFL learners: A qualitative study. *Computer Assisted Language Learning, 39*(1–2), 263–302.

[24] Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication, 30*(3), 358–392.

[25] Salinas-Navarro, D. E., Vilalta-Perdomo, E., Michel-Villarreal, R., & Montesinos, L. (2024). Designing experiential learning activities with generative artificial intelligence tools for authentic assessment. *Interactive Technology and Smart Education, 21*(4), 708–734.

[26] Alkaabi, A. M., & Abdallah, A. K. (2024). Portfolio practices in the principal evaluation process: A qualitative case study. *Heliyon, 10*(21).

[27] Dawson, P. (2020). *Defending assessment security in a digital world: Preventing e-cheating and supporting academic integrity in higher education*. Routledge.

[28] 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*(3), 467–481.

[29] Bearman, M., Tai, J., Dawson, P., Boud, D., & Ajjawi, R. (2024). Developing evaluative judgement for a time of generative artificial intelligence. *Assessment & Evaluation in Higher Education, 49*(6), 893–905.

[30] Boud, D., & Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment & Evaluation in Higher Education, 31*(4), 399–413.

[31] Gulikers, J. T., Bastiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development, 52*(3), 67–86.

[32] Hu, A., Liu, Q., & Daniel, B. (2025). Digital technologies in authentic assessment in higher education: A systematic literature review and narrative synthesis. *SAGE Open, 15*(3), 21582440251357198.

[33] Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119–144.

