International Journal of Education and Humanities (IJEH), 6(3) 2026:647-660
http://i-jeh.com/index.php/ijeh/index
E-ISSN: 2798-5768

What counts as evidence of learning when students use generative AI? A critical integrative review of assessment validity in higher education

Wenjie Peng[1]

Abstract

Generative artificial intelligence (GenAI) can support feedback, practice, and drafting, but it weakens a familiar assumption of take-home assessment: that the submitted product reliably records students' own learning. Higher education now offers many AI-aware assessment proposals, yet their recommendations often overlap, and the evidential purpose of each safeguard remains unclear. This critical integrative review examined 31 verified works published from 2021 to August 2026, comprising 26 primary empirical, technical, implementation, or document-analysis studies and five review or conceptual papers. Sources were coded for the learning claim at stake, the role allowed for GenAI, the evidence collected, the inference supported, the implementation burden, and the equity risk. The review found that detector scores cannot establish learning, authentic tasks alone remain vulnerable, and no single process record proves independent competence. Stronger designs combine evidence channels: a final product, selected process traces, a short oral or live check, and, where retention matters, delayed unaided performance. Building on this finding, the article presents an evidence-calibration model that asks instructors to define the learning claim, locate the AI-sensitive link in the inference, choose the smallest useful evidence bundle, and match verification intensity to stakes, workload, access, and student rights. The model is a design aid rather than a validated instrument. It offers a practical basis for assessment decisions while clarifying where further empirical testing is needed. The review contributes by synthesizing overlapping recommendations, making each safeguard's evidential purpose explicit, and foregrounding equity and student rights in AI-aware assessment design.

Keywords: *Academic Integrity, Assessment Validity, Generative Artificial Intelligence, Higher Education, Process Evidence*

A. Introduction

Generative artificial intelligence (GenAI) has made drafting, coding, translation, and explanation available on demand. For universities, the difficult question is not simply whether students use these tools. It is whether an assessment still justifies the claim made from the resulting mark. A well-written report may show sound judgment, but it may also conceal how much of the reasoning, source selection, and revision the student carried out. The problem is therefore one of assessment validity as well as academic integrity: what evidence allows an instructor to say that a learner understands, can perform, or can transfer the knowledge being assessed?

The empirical picture is more unsettling than a simple narrative of benefit or harm. Students describe faster access to explanations and help with difficult tasks, alongside concerns about accuracy, dependence, privacy, and unequal access (Chan & Hu, 2023; Johnston, Wells, Shanks, Boey, & Parsons, 2024; Ma, 2026; Ravšelj et al., 2025). Controlled studies point in a similarly

[1]College of Physical Education, Southwest University, China. 2536644952@qq.com

mixed direction. A carefully designed university tutor improved immediate learning and engagement, whereas unrestricted assistance in secondary mathematics improved practice performance but reduced later unaided performance; pedagogical guardrails narrowed that gap (Bastani et al., 2025; Kestin, Miller, Klales, Milbourne, & Ponti, 2025). How the tool is used and what students must still do themselves matter more than the label attached to the technology. This finding is surprising because it undercuts the assumption that either permitting or prohibiting GenAI will, by itself, preserve the meaning of an assessment.

Post-submission detection does not resolve this difficulty. Independent evaluations have found unstable performance across detectors and disproportionate false positives for non-native English writing (Liang, Yuksekgonul, Mao, Wu, & Zou, 2023; Weber-Wulff et al., 2023). Human judgment is also uncertain. In experiments conducted at two British business schools, markers generally could not distinguish human, AI-modified, and AI-generated work, and greater task authenticity did not, by itself, make the distinction reliable (Kofinas, Tsay, & Pike, 2025). A detector score or a suspicion about style may prompt a conversation, but neither shows what a student learned. The empirical phenomenon thus creates a double problem: GenAI can improve performance in the short term while weakening the evidential link between that performance and independent learning, and the most familiar post hoc remedies do not restore that link.

Recent scholarship has consequently moved toward redesign. The AI Assessment Scale (AIAS) links permission to assessment purpose (Perkins, Furze, Roe, & MacVaugh, 2024). DRIVE examines the quality of students' interactions with a chatbot (Oliveira, Zednik, Bombaerts, Sadowski, & Conijn, 2025). GENESIS and the Assessment Design Matrix (ADM) connect institutional conditions, graduate competencies, strategies, and methods (Ahangama, 2026). Assessment twins combine two modes of evidence for the same outcome (Roe, Perkins, & Giray, 2026). Black Box Assessment foregrounds drafts, errors, and revision (Winstone, Gravett, & Elkington, 2026), while a four-pillar model brings together process documentation, oral defense, authentic tasks, and clear policy (Nikolić & Basta Nikolić, 2026). An earlier scoping review of 32 empirical studies documented the wider shift toward assessment literacy, self-regulation, and policy change (Xia, Weng, Ouyang, Lin, & Chiu, 2024).

These approaches are useful, but the literature remains limited because they do not all answer the same question. A permission scale clarifies what use is allowed. A process record can reveal decisions made during the task. An oral defense samples explanation under live conditions. A second, unaided task tests whether performance transfers when assistance is removed. Because these measures serve different evidential purposes, treating them as interchangeable can create unnecessary work while leaving the central learning claim weakly supported. Existing proposals also overlap, and the evidential purpose of each safeguard is not always clear. The practical gap is a way to select evidence according to the inference at risk, rather than adding every available safeguard to every task.

This review responds to that gap by comparing recent approaches according to the evidential purpose they serve and by developing a brief calibration process for selecting a proportionate combination of evidence for the intended learning claim. Specifically, this critical integrative review examines 31 verified works published from 2021 to August 2026, including 26 primary empirical, technical, implementation, or document-analysis studies and five review or conceptual papers. Sources are coded for the learning claim at stake, the role allowed for GenAI, the evidence collected, the inference supported, implementation burden, and equity risk. Three questions guide the review: RQ1. What does the evidence show about the limits of final products, detection, and single safeguards in GenAI-enabled assessment?. RQ2. What learning inference can each commonly proposed evidence channel support, and where does it remain weak? RQ3.

How can instructors choose a proportionate combination of evidence for different outcomes and levels of consequence?.

This focus is important because assessment decisions must be defensible in terms of both validity and integrity. If detector scores cannot establish learning, authentic tasks alone remain vulnerable, and no single process record proves independent competence, then instructors need a principled way to combine evidence channels: a final product, selected process traces, a short oral or live check, and, where retention matters, delayed unaided performance. The proposed evidence-calibration model asks instructors to define the learning claim, identify the AI-sensitive link in the inference, select the smallest useful evidence bundle, and match verification intensity to the stakes, workload, access, and student rights. The contribution is therefore practical and conceptual: it synthesizes overlapping recommendations, makes each safeguard's evidential purpose explicit, and offers a proportionate basis for assessment decisions while clarifying where further empirical testing is needed.

B. Methods

A critical integrative review was selected because the question spans several forms of evidence: student and faculty studies, AI-literacy measures, classroom interventions, detector tests, assessment experiments, implementation studies, and conceptual models. These sources do not estimate a common effect, so statistical pooling would have obscured important differences in setting and outcome. The purpose was instead to compare what each study can support and to identify where different lines of evidence converge.

The search was completed on 31 August 2026. Crossref metadata, publisher websites, and journal archives were searched using combinations of "generative AI", "ChatGPT", "higher education", "assessment validity", "academic integrity", "AI literacy", "process assessment", "oral defense", "authentic assessment", and "AI detection". Backward and forward citation checks were used to locate empirical studies behind major claims and to identify frameworks published during 2025 and 2026.

Works were eligible when they were published in English from 2020 to August 2026; addressed higher education, AI literacy, assessment, or a directly transferable learning or detector-validity question; and could be verified through a DOI or an official publisher record. Primary evidence included surveys, interviews, psychometric studies, classroom interventions, experiments, technical evaluations, implementation studies, and analyses of original documents. Reviews and conceptual papers were retained when they supplied an assessment model that could be compared with the empirical evidence. Promotional material, unverifiable citations, duplicate reports, and claims without a clear educational link were excluded.

One randomized secondary-school study was retained because its comparison of unrestricted and guardrailed assistance directly tests a learning mechanism relevant to university assessment. It is not treated as direct evidence about higher-education populations. Conversely, the recent framework papers were included because they represent the design choices now available to university teachers, even when those choices have not yet been validated. This distinction between mechanism evidence, population evidence, and design proposals was maintained throughout the synthesis.

The final corpus contained 31 works. Twenty-six were primary empirical, technical, implementation, or original document-analysis studies; five were reviews or conceptual framework papers. Bibliographic details and DOI destinations were checked against publisher records. The search was purposive rather than systematic, so the synthesis does not claim exhaustive coverage. The corpus categories are summarized in

Table **1**.

Each source was coded for setting, participants or test objects, design, outcome, permitted or observed AI role, evidence channel, principal finding, and limit on interpretation. For assessment proposals, two additional questions were recorded: which learning inference the proposal was intended to strengthen, and what burden or equity risk it might introduce. Self-report findings were treated as perceptions rather than measured learning. Stronger statements about learning were reserved for controlled or performance-based studies, and conceptual claims were identified as proposals rather than validated effects.

Analysis proceeded in two passes. The first grouped sources into five evidence functions: use and perceptions; AI-literacy measurement and instruction; learning and assessment implementation; detector validity; and assessment frameworks. The second compared six evidence channels found across those groups: final products, process traces, oral or live performance, delayed unaided tasks, contextual tasks, and linked evidence from more than one mode. The calibration model was derived only after these channels had been compared against the learning claims they could and could not support. No human participants or private data were involved, so institutional ethics approval was not required.

Table 1. Composition of the review corpus

Evidence function	n	Typical design	Contribution to the synthesis
Use and perceptions	8	Surveys, interviews, mixed methods	Adoption, benefits, concerns, and policy needs
AI literacy	8	Scale/test studies and course evaluations	Competencies, measurement, and instruction
Learning and assessment	7	Experiments, implementation, interaction analysis	Conditions that support or obscure learning
Detection validity	2	Technical and benchmark evaluations	Accuracy, instability, and differential error
Frameworks, reviews, and document analysis	6	Conceptual work, scoping review, document analysis	Design options and their intended purpose

C. Results and Discussion

Learning effects depend on the instructional arrangement

Most higher-education studies in the corpus describe use, expectations, or confidence. They show that students use GenAI for information-seeking, writing, translation, brainstorming, and problem-solving, and that attitudes vary by discipline, prior experience, and local policy (Mansoor, Bawazir, Alsabri, Alharbi, & Okela, 2024; Ravšelj et al., 2025). Interviews and surveys also document a recurring tension: students value speed and individualized help but remain concerned about misinformation, dependence, privacy, and the ongoing need for teachers (Bissessar, 2026; Chan & Hu, 2023; Johnston et al., 2024; Ma, 2026). These findings help explain behavior, but they cannot establish that learning improved.

The distinction matters because convenience and learning are not the same outcome. Abbas, Jam, and Khan (2024) found that workload and time pressure predicted greater use of ChatGPT, which was associated with increased procrastination and memory loss, and lower academic performance. Because the study was observational, the results identify associations rather than causal effects; they nevertheless point to circumstances in which students may delegate work rather than learn from support. Zou and Huang (2023) similarly found that perceived usefulness

and ease of use shape doctoral students' intentions to use ChatGPT for writing. Acceptance data explain uptake; they do not validate the quality of the resulting knowledge.

AI-literacy research provides a better description of what responsible use requires. Validated measures separate technical knowledge, practical use, critical appraisal, self-efficacy, ethics, and self-management (Carolus, Koch, Straka, Latoschik, & Wienrich, 2023; Hornberger, Bewersdorff, & Nerdel, 2023; Laupichler, Aster, & Raupach, 2023; Laupichler, Aster, Haverkamp, & Raupach, 2023). Course and module studies suggest that these capabilities can be taught rather than assumed (Kong, Cheung, & Zhang, 2021; Kong, Cheung, & Zhang, 2022; Persada & Heryandi, 2026). A student who can provide a fluent answer is therefore not necessarily AI-literate; stronger evidence lies in how the answer is checked, revised, and used.

The smaller group of performance studies reinforces that point. Guardrails and task design changed whether AI assistance supported later performance in the studies by Bastani et al. (2025) and Kestin et al. (2025). Oliveira et al. (2025) examined 1,450 annotated chatbot interactions associated with 70 graded essays in science, technology, engineering, and mathematics (STEM) writing-intensive courses. Higher-quality interactions, marked by visible subject knowledge and purposeful direction, were positively related to essay scores. The association does not establish causality, but the process data added information not available from the final essay alone.

Neither detection nor authenticity is sufficient on its own

Detector-led decisions fail for both technical and educational reasons. Tool performance changes with the model, text length, editing, threshold, and test corpus. Weber-Wulff et al. (2023) found that available detectors were not reliable enough to support confident high-stakes judgments under realistic text conditions. Liang et al. (2023) found that several detectors classified non-native English writing as AI-generated at disproportionate rates. In a multilingual university, this is not a minor measurement error; it can shift the burden of suspicion toward students because of linguistic style.

Even perfect authorship classification would answer the wrong educational question. A detector cannot show whether a cited source was read, whether an argument was understood, or whether a student could reproduce the reasoning without assistance. It also cannot separate acceptable editing from prohibited substitution unless the task has already defined those categories. Detector output may be used as one signal that warrants ordinary academic review, but it should not replace task evidence, a student's explanation, or established due process.

Authentic assessment is valuable but is not a complete safeguard. Kofinas et al. (2025) compared human-authored, AI-modified, and AI-generated submissions from authentic business assessments. Eight markers, working in pairs, were generally unable to identify AI involvement, and authenticity did not protect the written tasks on its own. The finding leaves the educational case for authentic work intact while showing that a realistic topic or professional scenario still needs evidence of how the student interpreted the context and made decisions.

Task susceptibility also varies by discipline and format. In a pilot involving digital health and health-information assessments, GenAI performed strongly on objective, well-structured tasks but was much weaker on Structured Query Language (SQL), clinical coding, and context-dependent judgment; more detailed prompting improved some outputs (Wani et al., 2026). Although the small quasi-experimental study cannot establish a general ranking of assessment types, it illustrates why labels such as essay, quiz, reflective task, or authentic task are too broad. Teachers need to examine the specific reasoning a task elicits and whether a model can produce an acceptable product without that reasoning.

Recent approaches address different parts of the problem

The recent literature contains several practical responses, but a comparison shows that their design units differ. AIAS is primarily a communication and permission tool. DRIVE treats student-AI interaction as assessable evidence. GENESIS and ADM operate across institutional, program, and task levels. Assessment twins use two linked performances to strengthen the same validity claim. Black Box Assessment examines the trajectory of work, while the four-pillar model combines several safeguards into a single task.

Table **2** summarizes these differences.

Table 2. Comparison of recent GenAI assessment approaches

Approach	Primary purpose	Main evidence or mechanism	Current limitation
AI Assessment Scale	Clarify the permitted role of AI in a task	Task-level permission and disclosure	Useful for communication; permission alone does not verify learning
DRIVE	Assess how students direct and contribute knowledge to AI interaction	Annotated interaction records	Empirical support from 70 essays; transfer beyond the study context is untested
GENESIS and ADM	Align institutional support, competencies, strategies, and methods	Document-based institutional and task design	Broad and focused on information and communication technology; learning effects were not tested
Assessment twins	Cross-check one outcome through two linked modes	Complementary tasks completed close together	Clear validity rationale; workload and equity require testing
Black Box Assessment	Make learning, error, and revision visible	Drafts, logs, mistakes, and reflective accounts	Process records can be curated and need selective verification
Four-pillar model	Combine process, oral defense, authentic work, and policy	Multiple evidence channels and a vulnerability audit	Operational indicators are provisional rather than validated cutoffs

There is substantial agreement across these approaches. Students need clear rules; final products should not be the only source of evidence when independent competence matters; reasoning and revision should carry marks; and high-stakes claims benefit from another mode of performance. The disagreement is mainly about emphasis and scale. Ahangama (2026) places institutional readiness before task redesign. Roe et al. (2026) focus on cross-verification. Winstone et al. (2026) treat the history of learning as the object of assessment. Nikolić and Basta Nikolić (2026) provide explicit design indicators but state that their thresholds remain heuristic. The institutional pilot reported by Furze, Perkins, Roe, and MacVaugh (2024) also shows that a permission scale requires local interpretation and staff support in practice.

Process evidence, oral defense, disclosure, and authentic tasks are already established proposals. The less settled question is how an instructor should choose among them. A short oral check may be valuable when independent explanation is central, but excessive when the outcome concerns collaborative tool use. A complete prompt history may be informative for a study of

interaction, but wasteful for a low-stakes draft. Selection requires attention to the learning claim, not the popularity of a particular framework.

Evidence channels support different inferences

Six evidence channels recurred across the corpus. Each can strengthen an assessment, but none establishes every aspect of learning. The final product shows the quality of what was submitted. Process traces show selected decisions and changes. Oral or live performance samples explanation and adaptation. A delayed unaided task addresses retention or transfer. A contextual task elicits situated judgment. Linked evidence tests whether signals converge across modes. Their purposes and limitations are set out in Table 3.

Table 3. What common evidence channels can and cannot support

Evidence channel	Inference supported	Main weakness	Proportionate use
Final product	Quality of the completed disciplinary work	Independent contribution and reasoning may be hidden	Retain, but add another channel when individual competence matters
Selected process trace	Planning, source use, checking, and revision	Can be curated; volume does not equal quality	Sample diagnostic artifacts and grade their relevance
Oral or live check	Explanation, adaptation, and ownership under questioning	Anxiety, accessibility, time, and marker consistency	Use short standardized prompts and provide equivalent adjustments
Delayed unaided task	Retention and transfer without tool support	Narrow sampling and additional assessment time	Reserve for foundational or safety-critical capability
Contextual task	Application of knowledge to local constraints	A generic scenario may still be simulated by AI	Use specific data and require reasons for consequential choices
Linked evidence pair	Consistency of performance across two modes	Greater scheduling and marking burden	Use for capstones and other high-consequence decisions

Process evidence is most useful when it is chosen for a reason. A source map can show why particular literature was used; it cannot show independent recall. A revision note can show how feedback changed an argument; it cannot prove that every earlier draft was student-written. Chat records may reveal subject knowledge, as Oliveira et al. (2025) demonstrated, but a student can also rehearse or selectively submit them. Requiring more records does not automatically strengthen the inference.

Live and delayed evidence answer different questions. A brief oral defense can test whether the student can explain and extend submitted work. A delayed unaided item is better suited to retention and transfer. Both may disadvantage students if they are introduced without clear criteria, practice, reasonable adjustments, and consistent marking. These concerns are not arguments against verification. They are reasons to make verification proportionate and to provide an equivalent route when disability, language, connectivity, or privacy conditions affect participation.

An evidence-calibration model

The synthesis supports a short sequence of five decisions for selecting evidence. It makes the assessment claim, the vulnerable link in the inference, and the reason for each added requirement explicit. Figure 1 Summarizes the sequence.

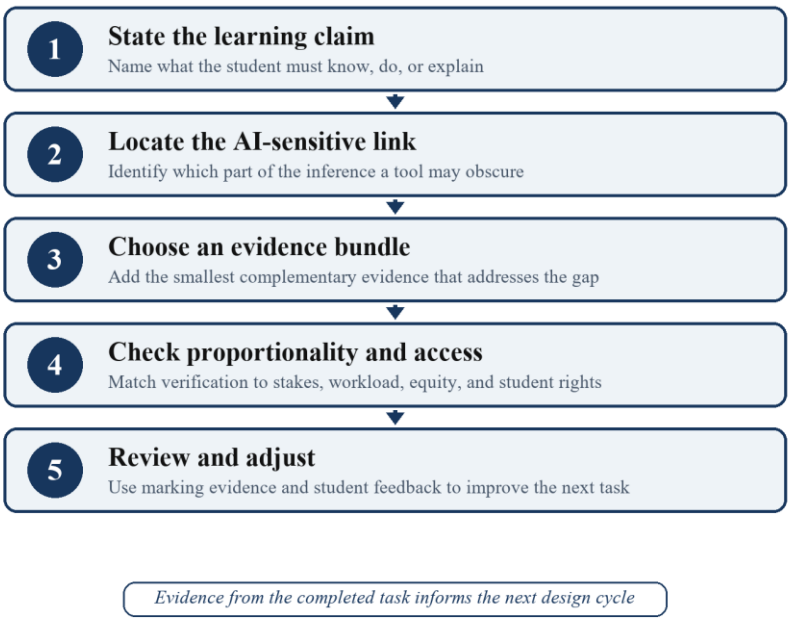


Figure 1. Evidence calibration for GenAI-enabled assessment

First, state the learning claim in terms that can be observed. "Understands research" is too broad. A usable claim might be that the student can select credible sources, compare conflicting findings, and justify a conclusion for a defined audience. The instructor should also decide which part must be demonstrated independently and which part may legitimately include professional tool use.

Second, locate the AI-sensitive link in the inference. The risk is not identical for every task. In an essay, GenAI may obscure the selection of sources and the development of arguments. In programming, it may obscure algorithm design, testing, or code modification. In a lesson plan, it may supply plausible activities without showing whether the student can adapt them to learners, facilities, or safety requirements. Naming the weak link avoids blanket restrictions.

Third, choose a complementary evidence bundle. The final product typically remains useful, but it is paired with the smallest additional source to address the weak link. One annotated source decision may be enough for a low-stakes literature task. A capstone may need staged work, a decision log, and a sampled oral defense. Where transfer without a tool is part of the outcome, an unaided checkpoint should be included rather than inferred from a polished submission.

Fourth, match verification intensity to consequence, access, and due process. The amount of evidence should rise with the stakes of the decision, the importance of independent performance, and the ease with which the final product can be produced without the target learning. It should fall when the same information can be obtained through sampling. A five-minute oral check, for example, may be more defensible than collecting an entire semester of prompt logs. Before the task is released, students should know what use is permitted, what records are required, how those records affect the grade, and what equivalent route is available if a commercial tool is inaccessible or unsuitable. If concerns arise, any response should consider

the declared rules, task evidence, the student's account, and the institution's standard procedures; a detector score should not serve as the standard of proof.

Fifth, review and adjust after marking. The instructor considers whether the added evidence actually informed judgment. An artifact that was never consulted should probably be removed. An oral prompt that produced inconsistent decisions needs clearer criteria or moderation. Student feedback can reveal whether instructions supported learning or merely created paperwork. This step keeps assessment design open to correction rather than turning a temporary response to current tools into a permanent rule.

Worked applications

For a research essay, the learning claim may concern source selection, synthesis, and argument. Brainstorming or language feedback can be allowed, while the evidence bundle includes the final essay, a short source rationale, one documented claim check, and a sampled oral question. The source rationale addresses selection; the claim check addresses verification; the oral question addresses explanation. None needs to reproduce the student's complete writing history.

For a programming or data-analysis task, the final code can be paired with an initial design sketch, repository history, student-written tests, and a brief live modification. The design sketch shows the intended approach; tests show which failures the student anticipated; and the live change samples transfer. GenAI may still be used for debugging if that reflects professional practice. The grade should concentrate on correctness, test quality, explanation, and the student's response when the initial solution fails. A short classroom experiment found higher computational thinking, self-efficacy, and motivation scores in a ChatGPT-supported programming activity, although the small sample size and brief duration limit generalization (Yilmaz & Karaoglan Yilmaz, 2023).

A physical education lesson plan presents a different case. GenAI can generate a polished sequence of activities, but the important claim may be whether the student can adapt the plan to age, ability, equipment, space, inclusion, and safety. A proportionate bundle could include the written plan, an annotation explaining three adaptations, a short micro-teaching segment, and questions based on an unexpected constraint. The live element tests situated judgment rather than writing style.

For concept learning, a guardrailed tutor may provide hints and examples during practice. If later independent use of the concept is important, the assessment should include a delayed, unaided problem or explanation. This arrangement is consistent with the contrast between supported performance and later performance observed by Bastani et al. (2025), while retaining the potential benefits of structured tutoring reported by Kestin et al. (2025). The student is not penalized for receiving help during practice; the design simply checks whether the help became usable knowledge.

A professional project may permit extensive use of GenAI because responsible delegation is part of the outcome. In that case, the strongest evidence may be a decision record that shows what was delegated, what risks were considered, how important outputs were verified, and why the final choice was accepted. The assessment then measures professional judgment rather than pretending that independent production is the goal. Clear disclosure also makes feedback more useful because the marker can respond to the actual workflow.

Implementation and equity

Implementation should begin with one consequential task, not a complete course rewrite. The instructor identifies the learning claim, tests the task with currently available tools, and adds

one or two evidence channels that address the observed weakness. A worked example should show students what acceptable disclosure and verification look like. The pilot can then be judged by whether the additional evidence changed marking decisions, improved feedback, or revealed confusion.

At program level, staff should map where independent and AI-supported performance are sampled. If every task permits unrestricted assistance, a program may lose evidence of foundational capability. If every task prohibits it, students may graduate without practice in evaluating tools used in their field. A mixed portfolio can include protected foundations, scaffolded experimentation, critical evaluation, and integrated professional use. Shared language across courses reduces the need for students to interpret a new set of rules in every module.

Staff preparation remains important. Faculty in Jamaica reported a need for training, clearer policy, and assessment adaptation (Madden, McKenzie, & Daley, 2025), while studies from China and a global student sample show that experience and expectations differ across contexts (Ma, 2026; Ravšelj et al., 2025). Ahangama's (2026) document analysis likewise places governance and support before sustainable task redesign. Institutions should therefore provide time, examples, moderation support, accessible tool options, and advice on privacy rather than transferring all responsibility to individual teachers.

Workload and equity should be treated as design constraints. Oral checks can be sampled rather than attached to every submission. Process records can use a common one-page form instead of unbounded chat histories. Students who cannot or do not wish to use a commercial model need an equivalent route unless tool use is itself an explicit outcome and access is provided. Adjustments for disability and language should preserve the learning claim while changing the mode or timing of verification. Equal treatment does not require identical evidence collection.

Limitations and research priorities

This review has four main limitations. It was structured and transparent but not exhaustive, and one reviewer made the final inclusion and coding decisions. The literature uses evolving terminology, and several relevant models emerged in 2026, so a later search may identify additional approaches. The corpus is English-language and uneven across regions and disciplines. Finally, much of the higher-education evidence relies on self-reports or small pilot studies. These limits mean that the synthesis should guide design questions rather than be read as a ranked effectiveness review.

The evidence-calibration model is also unvalidated. Its steps are derived from problems and design responses in the literature, but the review does not demonstrate that following them improves reliability, fairness, or learning. Existing models face the same gap. Nikolić and Basta Nikolić (2026), Roe et al. (2026), and Winstone et al. (2026) explicitly identify the need for empirical testing. The present model should therefore be evaluated against ordinary product-only assessment and against simpler redesigns.

Useful studies would compare evidence bundles rather than asking only whether AI was allowed. Outcomes should include unaided retention, transfer, inter-rater agreement, marking time, students' understanding of the rules, and the frequency with which decisions are changed by process or oral evidence. Factorial designs could test whether a source rationale, a revision note, or an oral check provides information beyond the final product. Longitudinal work is needed to determine whether repeated verification practices become part of students' independent disciplinary judgment.

Distributional effects require equal attention. Paid access, connectivity, language, prior digital experience, disability, and privacy constraints can alter who benefits from an AI-permissive task and who bears the burden of documentation. Detector studies show why an acceptable average error rate may still be unfair when mistakes cluster in one group (Liang et al., 2023). Evaluations of redesigned assessment should therefore report access conditions, accommodations, subgroup outcomes, and student appeals rather than relying on a single overall performance measure.

D. Conclusion

This critical integrative review aimed to examine why GenAI has made assessment validity and academic integrity harder to ensure, to compare the evidential purpose of commonly proposed safeguards, and to develop a proportionate approach to selecting evidence for the learning claim at stake. Drawing on 31 verified works published from 2021 to August 2026, the review addressed the limits of final products, detection, and single safeguards; the learning inference each evidence channel can support; and how instructors can choose a proportionate combination of evidence across different outcomes and levels of consequence. Its most important finding is that GenAI has not made assessment impossible, but it has made an old weakness harder to ignore: a polished final product is not always enough to support a claim about individual learning. Detector scores cannot repair that weakness, and neither authenticity nor process documentation is sufficient in every case. The evidence instead favors a conditional view: structured assistance can support learning, yet it can also separate successful completion from retained capability. Stronger designs, therefore, combine evidence channels: a final product, selected process traces, a short oral or live check, and, where retention matters, delayed unaided performance, because no single channel proves independent competence.

The review contributes conceptually and practically by synthesizing overlapping recommendations and making the evidential purpose of each safeguard explicit. It offers an evidence-calibration model that asks instructors to define the learning claim, locate the AI-sensitive link in the inference, choose the smallest useful evidence bundle, and match verification intensity to stakes, workload, access, and student rights. The model is a design aid rather than a validated instrument. Its implication is that assessment decisions should be justified by the inferences they support, not by the accumulation of safeguards. This shifts attention from detection toward defensible evidence design and keeps equity, due process, and student rights within the assessment conversation.

The review has limitations. The evidence-calibration model has not yet been empirically validated, and the conclusions rest on a heterogeneous body of empirical, technical, implementation, document analysis, review, and conceptual work. Because the review is integrative rather than meta-analytic, it does not estimate effect sizes or resolve inconsistencies across contexts; publication bias, variation in disciplinary assessment traditions, and limited evidence on long-term retention and equity effects also constrain the conclusions. Future research should test whether calibrated evidence bundles improve judgment sufficiently to justify the effort they require, compare bundles across disciplines and levels of stakes, examine effects on equity and access, and develop validated instruments for implementation. Longitudinal and experimental studies are especially needed to establish when particular combinations of evidence support valid inferences about understanding, performance, and transfer.

References

657

Abbas, M., Jam, F. A., & Khan, T. I. (2024). Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. International Journal of Educational Technology in Higher Education, 21(1), Article 10. https://doi.org/10.1186/s41239-024-00444-7

Ahangama, N. (2026). Designing assessments in the generative AI era: A tailored assessment framework for ICT tertiary education. International Journal of Educational Technology in Higher Education, 23, Article 9. https://doi.org/10.1186/s41239-026-00582-0

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. Proceedings of the National Academy of Sciences, 122(26), e2422633122. https://doi.org/10.1073/pnas.2422633122

Bissessar, C. (2026). To use or not to use ChatGPT and assistive artificial intelligence tools in higher education institutions? The modern-day conundrum - students' and faculty's perspectives. Equity in Education & Society, 5(2), 132-144. https://doi.org/10.1177/27526461231215083

Carolus, A., Koch, M. J., Straka, S., Latoschik, M. E., & Wienrich, C. (2023). MAILS - Meta AI literacy scale: Development and testing of an AI literacy questionnaire based on well-founded competency models and psychological change- and meta-competencies. Computers in Human Behavior: Artificial Humans, 1(2), 100014. https://doi.org/10.1016/j.chbah.2023.100014

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), Article 43. https://doi.org/10.1186/s41239-023-00411-8

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. https://doi.org/10.14742/ajet.9434

Hornberger, M., Bewersdorff, A., & Nerdel, C. (2023). What do university students know about artificial intelligence? Development and validation of an AI literacy test. Computers and Education: Artificial Intelligence, 5, 100165. https://doi.org/10.1016/j.caeai.2023.100165

Johnston, H., Wells, R. F., Shanks, E. M., Boey, T., & Parsons, B. N. (2024). Student perspectives on the use of generative artificial intelligence technologies in higher education. International Journal for Educational Integrity, 20(1), Article 2. https://doi.org/10.1007/s40979-024-00149-4

Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. Scientific Reports, 15(1), Article 17458. https://doi.org/10.1038/s41598-025-97652-6

Kofinas, A. K., Tsay, C.-H., & Pike, D. (2025). The impact of generative AI on academic integrity of authentic assessments within a higher education context. British Journal of Educational Technology, 56(6), 2522-2549. https://doi.org/10.1111/bjet.13585

Kong, S.-C., Cheung, W. M.-Y., & Zhang, G. (2021). Evaluation of an artificial intelligence literacy course for university students with diverse study backgrounds. Computers and Education: Artificial Intelligence, 2, 100026. https://doi.org/10.1016/j.caeai.2021.100026

Kong, S.-C., Cheung, W. M.-Y., & Zhang, G. (2022). Evaluating artificial intelligence literacy courses for fostering conceptual learning, literacy and empowerment in university students:

Refocusing to conceptual building. Computers in Human Behavior Reports, 7, 100223. https://doi.org/10.1016/j.chbr.2022.100223

Laupichler, M. C., Aster, A., Haverkamp, N., & Raupach, T. (2023). Development of the Scale for the Assessment of Non-Experts' AI Literacy (SNAIL) - An exploratory factor analysis. Computers in Human Behavior Reports, 12, 100338. https://doi.org/10.1016/j.chbr.2023.100338

Laupichler, M. C., Aster, A., & Raupach, T. (2023). Delphi study for the development and preliminary validation of an item set for the assessment of non-experts' AI literacy. Computers and Education: Artificial Intelligence, 4, 100126. https://doi.org/10.1016/j.caeai.2023.100126

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

Ma, Y. (2026). Experiences with artificial intelligence among students at the University of Linyi. International Journal of Education and Humanities, 6(1), 233-251. https://doi.org/10.58557/(ijeh).v6i1.439

Madden, O. N., McKenzie, N., & Daley, J.-L. (2025). Effects of ChatGPT and generative artificial intelligence in higher education: Voices of Jamaican academic faculty. International Journal of Education and Humanities, 5(2), 297-308. https://doi.org/10.58557/(ijeh).v5i2.314

Mansoor, H. M. H., Bawazir, A., Alsabri, M. A., Alharbi, A., & Okela, A. H. (2024). Artificial intelligence literacy among university students - A comparative transnational survey. Frontiers in Communication, 9, 1478476. https://doi.org/10.3389/fcomm.2024.1478476

Nikolić, D., & Basta Nikolić, M. (2026). Designing AI-resilient assessment in higher education: A four-pillar conceptual framework. Frontiers in Artificial Intelligence, 9, 1841682. https://doi.org/10.3389/frai.2026.1841682

Oliveira, M., Zednik, C., Bombaerts, G., Sadowski, B., & Conijn, R. (2025). Assessing students' DRIVE: A framework to evaluate learning through interactions with generative AI. Computers and Education: Artificial Intelligence, 9, 100497. https://doi.org/10.1016/j.caeai.2025.100497

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 & Learning Practice, 21(6). https://doi.org/10.53761/q3azde36

Persada, A. R., & Heryandi, Y. (2026). Instructional design of e-modules and e-courses for AI literacy: Enhancing conceptual, critical, and ethical understanding. International Journal of Education and Humanities, 6(1), 180-193. https://doi.org/10.58557/(ijeh).v6i1.416

Ravšelj, D., Keržič, D., Tomaževič, N., Umek, L., Brezovar, N., Iahad, N. A., ... Aristovnik, A. (2025). Higher education students' perceptions of ChatGPT: A global study of early reactions. PLOS ONE, 20(2), e0315011. https://doi.org/10.1371/journal.pone.0315011

Roe, J., Perkins, M., & Giray, L. (2026). Assessment twins: An approach for strengthening assessment validity in the age of generative AI. Journal of Applied Learning & Teaching, 9(2). https://doi.org/10.37074/jalt.2026.9.2.3

Wani, T. A., Liem, M., Prasad, N., Robinson, K., Nexhip, A., Tassos, M., ... Riley, M. (2026). Susceptibility of assessment types to AI-generated content in digital health and health

information management education: Quasi-experimental pilot study. JMIR Medical Education, 12, e82988. https://doi.org/10.2196/82988

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., ... Waddington, L. (2023). Testing of detection tools for AI-generated text. International Journal for Educational Integrity, 19(1), Article 26. https://doi.org/10.1007/s40979-023-00146-z

Winstone, N. E., Gravett, K., & Elkington, S. (2026). Black Box Assessment: Rethinking integrity and learning for a time of generative AI. Assessment & Evaluation in Higher Education. Advance online publication. https://doi.org/10.1080/02602938.2026.2661365

Xia, Q., Weng, X., Ouyang, F., Lin, T. J., & Chiu, T. K. F. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. International Journal of Educational Technology in Higher Education, 21(1), Article 40. https://doi.org/10.1186/s41239-024-00468-z

Yilmaz, R., & Karaoglan Yilmaz, F. G. (2023). The effect of generative artificial intelligence (AI)-based tool use on students' computational thinking skills, programming self-efficacy and motivation. Computers and Education: Artificial Intelligence, 4, 100147. https://doi.org/10.1016/j.caeai.2023.100147

Zou, M., & Huang, L. (2023). To use or not to use? Understanding doctoral students' acceptance of ChatGPT in writing through the technology acceptance model. Frontiers in Psychology, 14, 1259531. https://doi.org/10.3389/fpsyg.2023.1259531.