

From Principles to Practice: A Systematic Review of Generative AI-Responsive Assessment Interventions in Higher Education

Dr. Zafer YUCA

Eastern Mediterranean University, School of Computing and Technology, Famagusta, Turkey.

zafer.yuca@emu.edu.tr

ORCID ID: 0000-0002-4898-3881

Assoc. Prof. Dr. Fahme DABAJ

Eastern Mediterranean University, Faculty of Education, Famagusta, Turkey.

fahme.dabaj@emu.edu.tr

ORCID ID: 0000-0002-6233-7931

Abstract

Generative artificial intelligence (GenAI) challenges higher education assessment because a polished submitted artefact may no longer provide transparent evidence of the learner's knowledge, reasoning, or competence. This systematic review examines empirical studies in which assessment practices were deliberately configured in response to GenAI and implemented with higher education students. It synthesizes evidence concerning assessment design, the role assigned to GenAI, evidence of student learning, educational outcomes, and implementation conditions. Five recurring functional responses are identified: regulating AI use, structuring human-AI collaboration, requiring critical evaluation of AI-generated material, making the human learning process visible, and verifying competence through oral or performance-based evidence. Across these approaches, the central design issue is not simply whether GenAI is permitted, but how cognitive responsibility is distributed between learner and technology and how the learner's contribution is evidenced. The review therefore proposes an AI-Human Assessment Responsibility Model linking AI delegation, human cognitive responsibility, and evidentiary mechanisms. Sustainable GenAI-responsive assessment should be designed from intended learning outcomes and evaluated in relation to learning, validity, integrity, equity, workload, accessibility, and scalability rather than seeking a universally 'AI-proof' format.

Keywords: generative artificial intelligence; GenAI; higher education; assessment; assessment redesign; academic integrity; assessment validity; AI literacy

1. Introduction

The rapid development and widespread accessibility of generative artificial intelligence (GenAI) have introduced a significant transition in higher education. Since the public release of ChatGPT in late 2022, generative AI systems have become increasingly capable of producing sophisticated text, code, images, analyses, and other academic artefacts in response to natural-language prompts. Their growing integration into students' academic practices has created opportunities for personalized learning, feedback, idea generation, writing support, and problem solving, while simultaneously challenging established assumptions about how student learning is demonstrated and assessed (Kasneci et al., 2023; Xia et al., 2024). Consequently, the educational debate surrounding GenAI has progressively shifted from whether students will use these technologies toward how higher education should design learning and assessment in environments where AI assistance is readily available.

Assessment represents one of the areas most directly affected by this transformation. Higher education assessment serves several interconnected purposes: it supports learning, provides feedback, evaluates achievement, and enables institutions to make judgments about whether students have attained intended learning outcomes. These functions depend on the capacity of assessment evidence to support defensible interpretations of students' knowledge, skills, reasoning, and competence. Generative AI complicates this relationship because the quality of an assessed artefact may no longer provide sufficient evidence of the cognitive processes or capabilities of the student who submits it. A student may use GenAI to explore alternative explanations, obtain feedback, refine an argument, or critically evaluate a draft; alternatively, the same technology may perform substantial portions of the reasoning, synthesis, writing, or problem solving that the assessment was intended to elicit. Similar-looking final products may therefore represent markedly different levels of student engagement and learning.

Emerging empirical evidence illustrates the significance of this problem. Recent research has demonstrated that AI-generated academic work can satisfy conventional assessment criteria without necessarily providing reliable evidence of the student's underlying competence.

McGroarty et al. (2026), for example, submitted AI-generated psychology coursework through an authentic university marking process and reported that the submissions received passing grades without being identified by markers as unusual. Similarly, Brattli et al. (2026), using assessment data from 1,066 undergraduate students across five years, found substantial changes in grade distributions when a course moved from an AI-permissive take-home examination to an AI-restricted supervised examination, raising questions about whether performance under different AI-access conditions represents equivalent evidence of student competence. Such findings suggest that GenAI does not merely create a new mechanism for academic misconduct; rather, it challenges the relationship between observable assessment performance and the learning that assessment is intended to represent.

Initial institutional responses to this challenge frequently emphasized prohibition, detection, and academic-integrity regulation (Cotton et al., 2024). However, reliance on these approaches alone is increasingly difficult to sustain as generative AI capabilities develop and AI functionality becomes embedded within widely used digital platforms. Moreover, treating GenAI exclusively as a threat risks overlooking its emerging role in professional and disciplinary practice. Graduates are increasingly likely to encounter environments in which the ability to work critically and responsibly with AI constitutes part of professional competence. The educational challenge is therefore not simply to create assessments that prevent AI use, but to determine when AI use is educationally appropriate, what capabilities students should demonstrate with and without AI assistance, and how assessment can generate credible evidence of those capabilities.

This challenge has stimulated growing interest in assessment redesign. Scholars have proposed a range of responses, including authentic assessment, oral examinations and viva voce, staged or multi-component tasks, process-based assessment, reflective components, portfolios, supervised performance, AI-use declarations, prompt or interaction records, and assessment designs in which students are explicitly required to use, critique, verify, or improve AI-generated outputs. These approaches reflect a broader movement away from relying exclusively on polished final products and toward assessment architectures that make aspects of students' reasoning, decision making, verification, and learning processes more visible.

Recent conceptual work has described this transition as a movement from authentic products toward "authenticated processes," emphasizing the need for assessment to provide interpretable evidence of human judgment and responsibility when AI contributes to the production of assessed work (Tsiligkiris, 2026).

The literature has consequently advanced rapidly from identifying the disruption caused by GenAI toward proposing principles and frameworks for assessment redesign. Recent scholarship has examined assessment validity, authentic assessment, academic integrity, transparency mechanisms, AI susceptibility, and approaches for distinguishing students' own understanding from AI-supported performance. For example, Ali and Maroulis (2026) critically reviewed assessment validity in AI-rich higher education, while Tsiligkiris (2026) developed a multidimensional framework emphasizing process transparency, cognitive demand, student agency, inclusion, and AI-aware validity. Other emerging approaches include assessment twins, process-focused assessment, AI-supported oral verification, and mechanisms intended to establish students' epistemic ownership of AI-mediated work.

Although the emerging literature offers increasingly sophisticated guidance concerning how assessment should respond to generative artificial intelligence (GenAI), an important distinction remains between principles for assessment redesign and evidence concerning their implementation. Conceptual recommendations may be pedagogically compelling without demonstrating how effectively they operate when translated into assessment practice. An assessment designed to increase confidence in students' authorship or understanding may, for example, increase staff workload, introduce accessibility challenges, generate student anxiety, or prove difficult to implement with large cohorts. Conversely, assessments that deliberately incorporate GenAI may strengthen AI literacy, evaluative judgement, and professional authenticity while simultaneously creating new questions concerning how individual learning should be evidenced. The educational value of GenAI-responsive assessment therefore cannot be established through conceptual plausibility alone; it also requires systematic examination of what occurs when redesigned assessment is implemented with students.

A growing body of empirical research indicates that this transition from principle to practice has begun. Researchers and educators have implemented approaches including structured AI integration, critical evaluation of AI-generated material, reflective documentation of AI use, process-oriented assessment, and oral or performance-based verification. These interventions represent substantively different responses to GenAI. Some seek to preserve opportunities for students to demonstrate unaided competence; others permit AI while making its use transparent; still others deliberately incorporate human-AI interaction into the capabilities being assessed.

Consequently, the emerging question is no longer simply whether assessment should change in response to GenAI, but which assessment responses have actually been implemented, how they have been designed, and what empirical evidence exists concerning their educational and practical consequences.

Existing research also suggests that implementation outcomes cannot be reduced to academic performance alone. GenAI-responsive assessment may influence students' critical engagement, AI literacy, evaluative judgement, perceptions of authenticity, and approaches to learning. At the same time, redesigned assessment may affect academic integrity, confidence in authorship, accessibility, educator workload, scalability, and institutional feasibility. These outcomes are particularly important because an assessment intervention that appears pedagogically desirable in a small class may not necessarily remain viable when applied across disciplines or large student cohorts. Understanding GenAI-responsive assessment therefore requires attention not only to what is designed, but also to how it is implemented and what evidence emerges from its implementation.

Despite the increasing number of empirical studies, this evidence remains distributed across disciplines, assessment formats, methodological traditions, and educational contexts. Individual studies frequently focus on a particular intervention within a specific programme or course, making it difficult to determine broader patterns across the developing evidence base. Moreover, empirical studies of implemented assessment interventions coexist with a much larger literature comprising conceptual frameworks, institutional recommendations, perception surveys, academic-integrity discussions, and studies testing GenAI performance against conventional assessments. Although each contributes to understanding the broader transformation of assessment, they answer different questions. A focused synthesis of interventions implemented with higher education students is therefore necessary to distinguish recommendations about what educators could do from evidence concerning what educators have done and what happened as a result.

Accordingly, the present study systematically reviews empirical research on GenAI-responsive assessment interventions implemented in higher education. A GenAI-responsive assessment intervention is understood as an assessment practice whose design, procedures, conditions of AI use, verification mechanisms, or evaluative activities have been deliberately configured in response to the availability or educational integration of generative AI. The review focuses specifically on interventions implemented with higher education students for which empirical evidence concerning implementation or outcomes was collected. In doing so, it distinguishes assessment interventions from studies concerned exclusively with attitudes toward GenAI, AI-detection technologies, institutional policies, or the ability of GenAI systems themselves to complete conventional assessments.

The review is guided by the following research questions:

- RQ1. What GenAI-responsive assessment strategies have been empirically implemented in higher education?
- RQ2. What pedagogical purposes, design characteristics, and approaches to GenAI use characterize these assessment strategies?
- RQ3. What educational, assessment-related, and implementation outcomes have been investigated, and what does the empirical evidence indicate about these outcomes?
- RQ4. What challenges, limitations, and contextual factors influence the implementation and effectiveness of GenAI-responsive assessment strategies?
- RQ5. What gaps in the existing empirical evidence should guide future research and the development of GenAI-responsive assessment practice in higher education?

By addressing these questions, this review seeks to move the assessment debate from principles to practice. Rather than attempting to identify a universally "AI-proof" assessment format, the study examines how higher education is beginning to negotiate different relationships between human and artificial intelligence within assessment. Through systematic comparison of assessment design, implementation, and reported outcomes, the review aims to identify patterns in the emerging evidence, clarify where empirical support remains limited, and provide an evidence-informed foundation for future assessment design and research in AI-rich higher education.

2. Method

2.1 Review Design

This study employed a systematic literature review to identify, evaluate, and synthesize empirical research on generative artificial intelligence (GenAI)-responsive assessment strategies implemented in higher education. The review was designed in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines (Page et al., 2021), with search reporting guided by PRISMA-S (Rethlefsen et al., 2021). PRISMA 2020 was used to support transparent reporting of the identification, screening, eligibility assessment, and inclusion of studies.

The review differed from broader reviews of artificial intelligence in educational assessment by focusing specifically on assessment practices that had been implemented and empirically investigated in higher education settings. Accordingly, conceptual papers proposing assessment redesign without implementation, policy commentaries, general discussions of academic integrity, and studies examining perceptions of GenAI without an implemented assessment component were outside the principal scope of the review.

2.2 Eligibility Criteria

Eligibility criteria were established before the final screening process to maintain consistency between the research questions, literature search, and study-selection decisions.

Studies were considered eligible when they met all of the following criteria: (a) they were conducted in a higher education context; (b) they addressed generative artificial intelligence, including systems such as ChatGPT or comparable large-language-model-based generative tools; (c) they described an assessment strategy, assessment task, assessment intervention, or assessment-related educational practice that had been implemented with students; (d) they reported empirical evidence concerning the implementation or outcomes of that practice; (e) they were published as peer-reviewed journal articles or full peer-reviewed conference papers; and (f) the full text was available in English.

An assessment strategy was considered GenAI-responsive when the design of the assessment was introduced, modified, or investigated in explicit response to the availability, use, restriction, integration, or educational implications of generative AI. This definition permitted the inclusion of strategies that intentionally integrated GenAI as well as strategies designed to verify student learning in contexts where GenAI was restricted or could otherwise influence assessment performance.

Studies were excluded when they: (a) focused exclusively on primary, secondary, or non-formal education; (b) examined artificial intelligence generally without a clear GenAI component; (c) evaluated automated grading, conventional intelligent tutoring systems, learning analytics, or predictive AI without relevance to generative AI; (d) reported only students' or educators' attitudes, intentions, acceptance, or perceptions without implementing an assessment practice; (e) focused solely on AI-detection accuracy or the detection of AI-generated text; (f) discussed academic misconduct or institutional policy without an implemented assessment strategy; (g) were conceptual articles, editorials, commentaries, opinion pieces, protocols, dissertations, book chapters, abstracts, or literature reviews; or (h) did not report original empirical data.

Existing systematic and scoping reviews were excluded from the final empirical synthesis but were retained separately for background, identification of potentially relevant primary studies, and comparison of the contribution of the present review with previous syntheses.

2.3 Information Sources and Search Strategy

The literature search was designed to capture research at the intersection of three concepts: generative artificial intelligence, higher education, and assessment. Searches were conducted in Scopus, Web of Science Core Collection, ERIC, and Education Source. These databases were selected to provide complementary coverage of educational research, multidisciplinary scholarship, and technology-enhanced learning research.

Because the contemporary educational impact of GenAI accelerated following the public release of ChatGPT in November 2022, the search period began in November 2022. The upper boundary corresponded to the date on which the final database searches were conducted.

The syntax was adapted to the indexing conventions and search fields of each database. Searches were conducted primarily in titles, abstracts, and keywords where database functionality permitted.

A deliberately broad assessment block was retained during the initial search because emerging GenAI-responsive practices may not consistently be described using terms such as “assessment redesign.” Restricting retrieval to this terminology could therefore omit relevant empirical interventions described using discipline-specific language.

Backward reference searching of included articles and relevant previous reviews was also undertaken to identify eligible studies not captured through database searching. Forward citation searching was used for particularly relevant foundational or early empirical studies where appropriate.

The complete database-specific search strings and the dates on which each search was conducted should be reported in an appendix to ensure reproducibility, consistent with PRISMA-S (Rethlefsen et al., 2021).

2.4 Study Selection

All records identified through the database searches were exported to a reference-management system and duplicates were removed before screening. Study selection then proceeded in two stages.

During the first stage, titles and abstracts were screened against the predefined eligibility criteria. Records clearly unrelated to GenAI, higher education, or assessment implementation were excluded. During the second stage, the full texts of potentially eligible studies were retrieved and evaluated against all inclusion and exclusion criteria.

Particular attention during full-text screening was given to distinguishing an implemented assessment intervention from a proposed assessment strategy. A study was not included solely because its authors recommended oral examinations, authentic assessment, reflective tasks, or other redesign strategies. Evidence that the assessment practice had actually been implemented with higher education students and investigated empirically was required.

Reasons for full-text exclusion were recorded systematically. The study-selection process and the numbers of records identified, screened, excluded, assessed for eligibility, and included in the synthesis were documented using a PRISMA 2020 flow diagram.

2.5 Data Extraction

A structured data-extraction form was developed to capture information relevant to the five research questions. For each included study, the following information was extracted:

1. bibliographic information, including author(s), publication year, and publication outlet;
2. country and institutional context;
3. academic discipline or subject area;
4. educational level and participant characteristics;
5. sample size;
6. research design and data-collection methods;
7. GenAI technology or platform used;
8. type of assessment and assessment context;
9. description of the GenAI-responsive assessment strategy;
10. whether GenAI use was prohibited, restricted, optional, permitted, required, or deliberately integrated;
11. pedagogical rationale and theoretical or conceptual framework, where reported;
12. student learning and performance outcomes;
13. assessment-validity or academic-integrity outcomes;
14. student and educator experiences or perceptions of the implemented assessment;
15. implementation challenges, workload, scalability, accessibility, and feasibility considerations; and
16. authors' reported limitations and recommendations.

The extraction framework was designed to capture both the characteristics of the interventions and the evidence reported regarding their implementation. This distinction enabled the review to move beyond identifying assessment formats toward examining what is currently known about how these formats operate in practice.

2.6 Methodological Quality Appraisal

Because empirical educational research on GenAI includes qualitative, quantitative, and mixed-methods designs, methodological quality was assessed using the Mixed Methods Appraisal Tool (MMAT; Hong et al., 2018). The MMAT provides design-specific criteria for evaluating qualitative studies, randomized controlled trials, non-randomized quantitative studies, quantitative descriptive studies, and mixed-methods research.

Quality appraisal was used to inform interpretation of the strength and limitations of the evidence rather than to calculate an overall numerical quality score. Studies were not automatically excluded solely because they exhibited methodological limitations; instead, appraisal findings were considered when evaluating the confidence that could be placed in particular conclusions and when identifying methodological weaknesses in the emerging evidence base.

2.7 Data Synthesis

Given the expected heterogeneity of assessment interventions, academic disciplines, research designs, and reported outcomes, a narrative and thematic synthesis was planned rather than a conventional meta-analysis.

The first stage of synthesis involved descriptive mapping of the included studies according to publication year, geographical location, discipline, research design, participant group, GenAI platform, and assessment context.

The second stage focused on developing a taxonomy of empirically implemented GenAI-responsive assessment strategies. Categories were generated inductively from the included studies rather than imposed in advance. Potentially related interventions were compared according to their underlying assessment purpose and design characteristics rather than merely their labels.

The third stage synthesized evidence concerning the reported consequences of implementation. Outcomes were organized around recurring domains emerging from the literature, with particular attention to student learning and performance, cognitive engagement, assessment validity, academic integrity, student experience, AI literacy, educator workload, scalability, accessibility, and feasibility.

Finally, relationships among assessment design, the role assigned to GenAI, implementation conditions, and reported outcomes were examined across studies. This stage was intended to identify not only which strategies had been implemented but also the conditions under which particular approaches appeared promising, problematic, or insufficiently evaluated.

2.8 Development of an Evidence-Informed Framework

Following the descriptive and thematic synthesis, findings were integrated to develop an evidence-informed framework for GenAI-responsive assessment practice. The framework was derived from recurring patterns observed across the included empirical studies rather than established a priori.

Its purpose was to connect three dimensions of the evidence: assessment design, the role of GenAI, and evidence of student learning. Particular attention was given to how assessment designs make students' reasoning, judgment, decision making, verification, reflection, or performance observable when GenAI contributes to academic work.

The resulting framework was intended as a synthesis of the available empirical evidence rather than a prescriptive model applicable to all educational contexts. Areas supported by limited or methodologically weak evidence were identified explicitly as priorities for further research.

2.9 Operational Definition of a GenAI-Responsive Assessment Intervention

For the purposes of this review, a GenAI-responsive assessment intervention was defined as an assessment practice in which the task design, assessment procedure, conditions governing AI use, student–AI interaction, verification mechanism, or evidentiary requirements were deliberately configured in response to the availability or educational use of generative artificial intelligence.

This operational definition distinguished GenAI-responsive assessment interventions from studies in which GenAI was merely present in the educational environment. Student use of GenAI during an otherwise unchanged assessment was not, by itself, considered sufficient for inclusion unless the assessment design or associated requirements explicitly addressed that use. Similarly, studies investigating students' or educators' perceptions of how GenAI might affect assessment were excluded when no assessment intervention had been implemented.

To support consistent eligibility decisions, potentially relevant studies were evaluated using four sequential questions:

1. Was the study conducted in a higher education context?
2. Did the study concern generative AI in relation to assessment?
3. Was an assessment practice, requirement, procedure, or condition deliberately designed or modified in response to GenAI?
4. Was that practice implemented with students and investigated using empirical data?

Studies meeting all four conditions were eligible for the principal synthesis. Studies failing the third condition but providing empirical evidence concerning students' naturally occurring GenAI use during assessment were retained as contextual literature but excluded from the principal intervention synthesis. Conceptual proposals and assessment frameworks that had not been implemented with students were similarly retained for theoretical comparison but excluded from the empirical corpus.

This distinction was intended to preserve the review's focus on the translation of GenAI-responsive assessment principles into educational practice.

3. Results

3.1 Emerging Characteristics of GenAI-Responsive Assessment Interventions

The emerging empirical literature indicates that higher education institutions have responded to generative artificial intelligence through substantively different forms of assessment intervention.

These interventions cannot be adequately distinguished solely according to conventional assessment formats such as essays, examinations, presentations, or oral assessments. Rather, a more meaningful distinction concerns the function that the intervention assigns to GenAI and the mechanisms through which students remain responsible for demonstrating learning.

Across the identified studies, GenAI-responsive interventions ranged from assessments that regulated or documented students' AI use to tasks in which interaction with GenAI constituted an intentional component of the assessed activity. In some interventions, students were permitted to use GenAI while being required to disclose or reflect upon their use. Other interventions structured the purposes or stages for which GenAI could be employed. A further group repositioned AI-generated content as an object of disciplinary evaluation, requiring students to identify limitations, verify claims, correct errors, or critically evaluate the quality of AI-generated work. Finally, some assessment redesigns introduced additional process, oral, or performance evidence intended to make students' own understanding more directly observable.

An important distinction therefore emerged between AI-permissive and pedagogically structured AI use. Merely allowing students to use GenAI did not necessarily result in cognitively sophisticated engagement. Evidence from postgraduate management students, for example, indicated that GenAI use was concentrated largely on writing, paraphrasing, and rephrasing activities, whereas comparatively few students demonstrated more advanced forms of critical engagement. In contrast, assessment interventions that explicitly required students to interrogate or evaluate AI-generated content repositioned critical judgement as part of the assessed performance.

This distinction suggests that the educational significance of GenAI-responsive assessment depends not simply on whether artificial intelligence is permitted, but on the cognitive responsibilities assigned to the learner when AI is present. Where GenAI performs part of the productive work associated with an assessment, redesigned tasks may require students to demonstrate responsibility through verification, critique, modification, justification, reflection, or direct performance.

The identified interventions also differed in the evidence used to represent student learning. Conventional assessment typically locates evidence predominantly in the submitted product. GenAI-responsive interventions increasingly supplement or partially replace product evidence with process records, reflective accounts, documented human–AI interactions, critical evaluations, oral explanations, or observable performance. This movement indicates an emerging concern with making the human contribution to AI-mediated work more visible to assessors.

Taken together, the developing evidence suggests that GenAI-responsive assessment may be understood through three interconnected dimensions: the functions delegated to AI, the cognitive responsibilities retained by the student, and the evidentiary mechanisms through which those responsibilities are demonstrated. These dimensions provide an initial analytical structure for comparing otherwise heterogeneous assessment interventions across disciplines and educational contexts.

3.2 Types of GenAI-Responsive Assessment Intervention

The verified empirical literature indicates at least five recurring functional responses to GenAI. First, some assessments regulate AI use by defining purpose-specific, open, restricted, or disclosure-based conditions. Quince and Faulconer (2026), for example, examined three redesigned undergraduate physics assessments operating under purpose-specific and open-use policies and found that assessment context shaped the functions for which students reported using GenAI.

Second, some interventions structure human–AI collaboration rather than merely permitting it. Martin et al. (2025) co-designed an AI-integrated postgraduate assessment in which students in a compliant condition used AI to support a blog and critically reflected on the outputs, while an unrestricted condition permitted broader AI use. Cochrane et al. (2026) similarly examined a scaffolded first-year assessment requiring students to generate, critically evaluate, revise, and reflect on GenAI-assisted outputs.

Third, a group of interventions makes AI-generated content an explicit object of disciplinary critique. Tam et al. (2026) required students to critically evaluate ChatGPT-generated scientific output. Casey (2026) required public-policy students to use ChatGPT to draft a ministerial policy brief and then reflect critically on the process and limitations of the resulting output. These designs reposition verification and evaluative judgement as part of assessed performance rather than treating AI solely as a production aid.

Fourth, some approaches attempt to evidence the human learning process through reflections, interaction records, or other process data. Fischer et al. (2024) embedded reflection on GenAI use into a postgraduate management assessment and found that students' reported uses were concentrated on writing, paraphrasing, and rephrasing, with comparatively limited evidence of advanced sensemaking. Oliveira et al. (2025) went further by evaluating 1,450 annotated GenAI interactions linked to 70 graded STEM essays; interaction quality was positively associated with essay performance, and different interaction profiles were associated with idea co-development versus text refinement.

Fifth, some redesigns emphasize direct verification of human competence through oral or performance evidence. Ogle and Jarrett (2026) replaced a written nursing case study with an oral ISBAR assessment for 740 students and reported broadly stable cohort performance, no identified integrity breaches under the oral format, greater educator confidence in authorship, and reductions in marking time and cost. Hutson et al. (2026) examined an adaptive viva layer paired with written submissions; the pilot suggests that oral explanation can reveal comprehension and authorship-related information not visible in the written artefact alone, while also underscoring the need for larger and more systematic evaluation.

These functional strategies are not mutually exclusive. A single assessment may simultaneously regulate AI use, structure interaction, require critique, collect process evidence, and include oral verification.

Table 1. Core empirical GenAI-responsive assessment studies synthesized in the manuscript

Study	Context	Participants	Assessment response	Evidence	Primary contribution
Casey (2026)	Public policy	25 students	Required ChatGPT policy brief + reflection	Reflections; focus group	Critical evaluation; AI literacy
Foung et al. (2024)	University English	74 reflections; 28 focus-group participants	Redesigned assessments using GenAI/digital tools	Reflections; focus groups	Tool choice; equity/access
Fischer et al. (2024)	Postgraduate management	118 students	Assessment plus reflection on GenAI use	Student claims; assessor evaluation	Permissive use often remained surface-level
Martin et al. (2025)	Postgraduate coursework	Implemented co-designed assessment	Structured/compliant vs unrestricted AI-use conditions	Mixed evaluation	Feasibility; integrity; structured use
Oliveira et al. (2025)	STEM writing	70 essays; 1,450 interactions	GenAI interaction as analyzable learning process	Annotated interaction traces; grades	Interaction quality and learning process
Tam et al. (2026)	Undergraduate science	Coursework participants	Critical evaluation of ChatGPT-generated scientific content	Student-perspective evidence	Critical evaluation and verification
Cochrane et al. (2026)	First-year HE	164 students	Scaffolded generate-evaluate-revise-reflect assessment	Cross-sectional survey	Evaluative engagement; perceived learning
Quince & Faulconer (2026)	Undergraduate physics	Three redesigned assessments	Purpose-specific/open AI-use conditions + disclosure	Mixed methods; performance	Policy conditions shape AI use
Ogle & Jarrett (2026)	Nursing	740 students	Written case study replaced by oral ISBAR assessment	Cohort outcomes; staff experience	Verification; scalability; workload/cost
Hutson et al. (2026)	Higher education pilot	Small pilot; student survey n=7	Adaptive viva following written submission	Pilot evaluation	Oral verification; feasibility caveat
Stratton-Maher & Kelly (2026)	First-year nursing	Large cohort; focus-group subsample	Structured ChatGPT-integrated assessment	Qualitative/student experience	Ethics; guidance; digital inequity

Table 2. Functional taxonomy of GenAI-responsive assessment strategies

Strategy	Central design question	Typical mechanisms	Illustrative studies
Regulating AI	What may AI do?	Permission levels, restrictions, purpose-specific conditions, disclosure	Quince & Faulconer, 2026
Structuring AI collaboration	How should students work with AI?	Scaffolds, staged use, specified functions, revision and reflection	Martin et al., 2025; Cochrane et al., 2026
Critiquing AI	Can students judge AI output?	Verification, correction, comparison, error identification, justification	Casey, 2026; Tam et al., 2026
Evidencing the human process	What did the student contribute?	Interaction records, prompts, drafts, annotations, reflections, decision traces	Fischer et al., 2024; Oliveira et al., 2025
Verifying human competence	Can the student demonstrate understanding?	Viva, oral defence, practical or supervised performance	Ogle & Jarrett, 2026; Hutson et al., 2026

3.3 Positioning of Generative AI Within Assessment

The studies show that the educational role of GenAI varies substantially across interventions. In some cases, GenAI is a permitted support tool; in others, it is deliberately integrated into task completion, becomes an object of critique, or contributes process data that are themselves assessed. Fount et al. (2024), for example, redesigned university English assessments to allow students to select among GenAI and other digital tools and found that students could make purpose-sensitive choices, while also identifying concerns about inequities between free and premium tools. Martin et al. (2025), Casey (2026), Tam et al. (2026), and Cochrane et al. (2026) illustrate more structured forms of integration in which students are required to evaluate, reflect on, or revise AI-supported work.

This variation supports a distinction between AI access and pedagogically structured AI use. Access alone does not establish that students are engaging in higher-order reasoning. Fischer et al. (2024) found that only a minority of students demonstrated advanced critical engagement with GenAI, whereas interventions explicitly designed around critique, verification, and revision place those activities within the assessed construct.

3.4 Evidencing Human Learning in AI-Mediated Assessment

A recurring feature of the empirical literature is the use of evidence beyond the final submitted product. Process-oriented approaches use reflective accounts, AI-interaction histories, annotated revisions, or documented decision making to make aspects of the learner's contribution visible (Casey, 2026; Fischer et al., 2024; Oliveira et al., 2025). Oral and performance-oriented approaches instead seek direct evidence that students can explain, justify, or apply the knowledge represented in submitted work (Hutson et al., 2026; Ogle & Jarrett, 2026).

The need for such evidence is reinforced by studies outside the intervention corpus. Kofinas et al. (2025) found that markers could not reliably distinguish human-authored, GenAI-modified, and GenAI-generated work and that greater assessment authenticity did not, by itself, safeguard against GenAI involvement. McGroarty et al. (2026) likewise reported that two AI-generated psychology reports received passing grades under routine university marking conditions. Brattli et al. (2026), using five years of assessment data from 1,066 students, observed a substantial change in grade distributions when a course moved from an AI-permissive take-home examination to an AI-restricted supervised examination. Although these studies do not all constitute implemented GenAI-responsive interventions, they provide contextual evidence for the concern that final-product performance and independent student competence may not be measurement-equivalent under different AI-access conditions.

3.5 Educational and Assessment-Related Outcomes

The empirical studies report outcomes at different evidentiary levels. Several studies primarily capture experience and perception, including acceptability, perceived learning, confidence, and students' views of AI integration (Casey, 2026; Fount et al., 2024; Stratton-Maher & Kelly, 2026; Tam et al., 2026). Such evidence is important for implementation but should not be interpreted as equivalent to demonstrated learning gains.

Other studies provide behavioural or process evidence. Fischer et al. (2024) compared students' claims about GenAI use with assessors' evaluations, Oliveira et al. (2025) analysed interaction traces, and Quince and Faulconer (2026) coded student declarations of GenAI use across redesigned assessments.

A smaller group provides performance or comparative outcome evidence. Ogle and Jarrett (2026) compared cohort-level outcomes before and after an oral assessment redesign, while Martin et al. (2025) compared structured and unrestricted AI-use conditions within an implemented assessment. Across this heterogeneous evidence base, the strongest conclusion is therefore not that GenAI-responsive assessment uniformly improves learning, but that outcomes depend on how AI use is structured, what students are required to do cognitively, and what evidence is collected.

3.6 Implementation, Feasibility, Equity, and Scalability

Implementation outcomes are central to judging whether promising assessment designs are sustainable. Martin et al. (2025) explicitly examined feasibility, acceptability, and perceived integrity. Ogle and Jarrett (2026) provide evidence that oral assessment can be implemented at large cohort scale under particular conditions and may reduce marking time and cost, challenging assumptions that oral verification is necessarily impractical. Conversely, Hutson et al. (2026) describe a smaller pilot and appropriately call for broader validation.

Equity also emerges as an implementation concern. Fount et al. (2024) reported student concerns about differences between free and premium AI tools, while Stratton-Maher and Kelly (2026) identified digital inequities and inconsistent institutional guidance as issues in a structured nursing assessment. These findings suggest that requiring GenAI within assessment may create new access and support obligations for institutions.

3.7 Methodological Quality and Limitations of the Evidence Base

The verified evidence base is methodologically heterogeneous. It includes qualitative case studies, mixed-methods evaluations, cross-sectional surveys, observational studies, process analyses, pilots, and cohort-level comparisons. This diversity supports the use of the MMAT for design-sensitive appraisal (Hong et al., 2018), but it also limits direct comparison across studies. Final MMAT judgments for each included study should be reported in the study-characteristics/appraisal table once the definitive corpus is locked; no overall numerical MMAT score should be calculated.

A recurring limitation is reliance on self-reported experience or perceived learning. Cochrane et al. (2026), for example, provide useful evidence concerning perceived learning, evaluative engagement, confidence, and competence, but the cross-sectional survey design does not establish causal learning effects. Casey (2026), Fount et al. (2024), Stratton-Maher and Kelly (2026), and Tam et al. (2026) similarly provide valuable evidence about student experience and critical engagement while remaining context-specific. Stronger causal claims require comparative, longitudinal, or experimental designs with objective measures of learning.

3.8 Evidence Gaps

Several gaps remain evident. First, more comparative studies are needed to establish whether structured GenAI-responsive interventions improve learning, evaluative judgement, or transfer relative to alternative assessment designs. Second, the literature needs stronger longitudinal evidence concerning whether students retain the knowledge and judgement demonstrated during AI-supported assessment. Third, workload, cost, accessibility, reliability, and equity require systematic evaluation across larger and more diverse cohorts.

Fourth, research remains concentrated in individual courses and institutions, limiting cross-disciplinary and cross-institutional generalization. Finally, the field requires more explicit validation of process evidence. Interaction logs, reflections, declarations, and oral explanations may make student contribution more visible, but future research should establish when these sources provide reliable and valid evidence of the intended learning outcomes.

4. Discussion

The rapid integration of generative artificial intelligence (GenAI) into higher education has generated extensive debate about academic integrity, assessment validity, and the continuing suitability of established assessment practices. While much of the early discussion concentrated on preventing unauthorized AI use or detecting AI-generated work, the emerging empirical literature examined in this review indicates a broader transition. Higher education assessment is increasingly moving from attempts to control the presence of GenAI toward deliberate decisions about what role AI should play, what intellectual responsibility should remain with the student, and what evidence is required to demonstrate student learning. This transition represents an important shift from reactive responses to GenAI toward the intentional redesign of assessment for environments in which AI assistance is increasingly difficult, and in some contexts undesirable, to exclude.

The findings of this review suggest that the central assessment problem created by GenAI cannot be reduced to whether students are permitted to use artificial intelligence. Rather, the more consequential issue concerns the relationship among AI delegation, human cognitive responsibility, and evidence of learning. GenAI can contribute to brainstorming, information synthesis, drafting, linguistic refinement, feedback, problem solving, and other activities traditionally performed by students. As these functions are partially delegated to AI, the quality of a final product may become a less transparent indicator of the knowledge, reasoning, judgement, and disciplinary competence possessed by the student.

GenAI-responsive assessment therefore requires educators to reconsider not only the format of assessment but

also the inferential relationship between the evidence produced by an assessment and the learning outcomes that the assessment is intended to represent.

4.1 Beyond the Binary of Prohibiting or Permitting GenAI

One of the clearest implications emerging from the reviewed literature is the inadequacy of treating GenAI use as a binary choice between prohibition and unrestricted permission. Such a distinction focuses primarily on access to technology rather than on the educational quality of students' interactions with it.

Evidence from studies in which students were permitted to use GenAI indicates that access alone does not necessarily generate sophisticated forms of learning. Fischer et al. (2024), for example, found that postgraduate students frequently used GenAI for writing-related functions such as paraphrasing and rephrasing, whereas substantially fewer demonstrated uses associated with more advanced critical thinking and sensemaking. This finding illustrates an important limitation of permissive approaches: allowing GenAI may normalize its use without ensuring that students engage with it critically, strategically, or in ways aligned with intended learning outcomes.

By contrast, interventions that deliberately structured students' interactions with GenAI appeared to reposition AI use as an object of pedagogical design. Students may be required to interrogate an AI response, identify inaccuracies, compare alternative outputs, justify whether AI-generated information should be accepted, revise inadequate responses, or reflect on the limitations of the technology. In these contexts, GenAI does not simply reduce the work required from the learner. Instead, the assessment redistributes cognitive activity by assigning some productive functions to AI while deliberately retaining evaluative and decision-making responsibilities for the student.

This distinction suggests that the educational question should move beyond whether GenAI is used toward how GenAI use is structured and what students remain responsible for doing.

Consequently, policies that classify assessment simply as "AI permitted" or "AI prohibited" may be insufficient for guiding pedagogically meaningful practice.

More useful assessment specifications may need to identify the functions that can legitimately be delegated to AI, the functions that students are expected to perform themselves, and the evidence through which those expectations will be assessed.

4.2 From Product Authenticity to Evidence of Human Learning

A second major implication concerns the status of the final assessment product. Conventional assessment frequently assumes that the submitted artefact provides meaningful evidence of the student's capabilities. Although this assumption has never been absolute, students have always had access to external resources, peers, editors, and digital tools, the generative capabilities of contemporary AI substantially increase the distance that can exist between the quality of a submitted product and the competence of the individual submitting it.

This issue explains why simply making assessment tasks more "authentic" may not be sufficient. Authentic assessment remains educationally valuable because it can connect academic learning with meaningful disciplinary and professional practice. However, authenticity does not automatically establish authorship, understanding, or intellectual ownership. Evidence showing that assessors may struggle to distinguish human-generated from AI-generated or AI-modified submissions reinforces the limitation of relying on product characteristics alone.

The reviewed interventions indicate an emerging response: shifting some assessment attention from the product of learning to evidence of the process and performance underlying that product. Several strategies reflect this movement. Students may document their interactions with GenAI, submit prompts or conversation histories, explain how outputs were evaluated and modified, reflect upon decisions made during task completion, or defend their conclusions orally. Other approaches require students to demonstrate disciplinary competence directly through oral, practical, dialogic, or supervised performance.

These approaches do not imply that final products have become irrelevant. In many disciplines, producing a high-quality artefact remains an important learning outcome. Rather, the findings suggest that AI-mediated products may need to be interpreted alongside complementary forms of evidence when educators need to make claims about individual student competence.

This distinction is especially important because a polished AI-assisted product and genuine student learning are not mutually exclusive. A student may use GenAI extensively while exercising substantial disciplinary judgement in selecting, verifying, rejecting, integrating, and improving its outputs. Conversely, a student may produce an equally polished artefact while exercising minimal understanding. Assessment therefore needs mechanisms capable of distinguishing these situations.

4.3 Human Cognitive Responsibility as a Principle for GenAI-Responsive Assessment

The synthesis supports the development of human cognitive responsibility as a useful organizing principle for GenAI-responsive assessment. Human cognitive responsibility refers to the knowledge, reasoning, judgement, verification, decision making, and disciplinary accountability that students are expected to retain when GenAI contributes to assessed work.

This concept differs from requiring students to complete all intellectual work without technological assistance. Such an expectation may be neither realistic nor educationally desirable in disciplines where responsible AI use is becoming part of professional practice. Instead, human cognitive responsibility requires assessment designers to specify what students must remain capable of understanding and doing even when some productive activities are delegated to AI.

Three questions follow from this principle:

What may be delegated to AI?

What intellectual responsibility must remain with the student?

What evidence demonstrates that the student has exercised that responsibility?

These questions provide a common analytical language for assessment strategies that otherwise appear very different. An assessment requiring students to critique AI-generated scientific explanations, for example, deliberately delegates initial content production to AI while retaining evaluation, verification, and disciplinary judgement as student responsibilities. An assessment involving an AI-generated policy brief may similarly shift content generation toward AI while requiring students to identify factual limitations, political assumptions, omissions, and inappropriate recommendations. An oral verification component takes a different approach by requiring students to demonstrate that they understand and can justify work submitted through an AI-mediated process.

Viewed in this way, GenAI-responsive assessment does not necessarily seek to minimize AI involvement. Its objective is to maintain alignment between intended learning outcomes, the responsibilities assigned to students, and the evidence used to evaluate achievement.

4.4 The AI–Human Assessment Responsibility Model

Based on the emerging synthesis, this review proposes the AI–Human Assessment Responsibility Model as a conceptual framework for interpreting and designing GenAI-responsive assessment. The model consists of three interconnected dimensions: AI delegation, human cognitive responsibility, and evidentiary mechanism.

AI delegation concerns the functions that an assessment permits or requires students to assign to GenAI. These may include idea generation, information organization, drafting, linguistic refinement, feedback, coding, problem solving, simulation, or content generation.

Human cognitive responsibility concerns the intellectual capabilities that students remain accountable for demonstrating. Depending on the learning outcomes, these may include disciplinary understanding, critical evaluation, verification, reasoning, synthesis, ethical judgement, decision making, creativity, or the ability to recognize limitations and errors in AI-generated material.

Evidentiary mechanism concerns how assessment makes those human capabilities observable. Evidence may be located in the final product but may also include process documentation, AI-interaction records, annotated revisions, reflective accounts, critical commentary, oral justification, supervised performance, or combinations of these approaches.

The central proposition of the model is that the validity of AI-mediated assessment depends less on minimizing the amount of AI used than on achieving appropriate alignment among these three dimensions. Increasing AI delegation without reconsidering human responsibility and assessment evidence may weaken the relationship between observed performance and student competence. In contrast, extensive AI use may remain compatible

with meaningful assessment when the learning outcomes legitimately involve human–AI collaboration and when students' distinctive responsibilities are clearly defined and evidenced.

This model also suggests that no single GenAI-responsive assessment strategy is likely to be universally appropriate. Appropriate delegation and evidence depend upon disciplinary expectations and intended learning outcomes. The level of unaided performance required from a nursing student demonstrating clinical judgement, for example, may differ substantially from the AI-supported capability expected of a student producing professional communication in a field where AI-assisted drafting has become routine. Assessment redesign therefore requires decisions about the educational construct being assessed rather than simply decisions about technology.

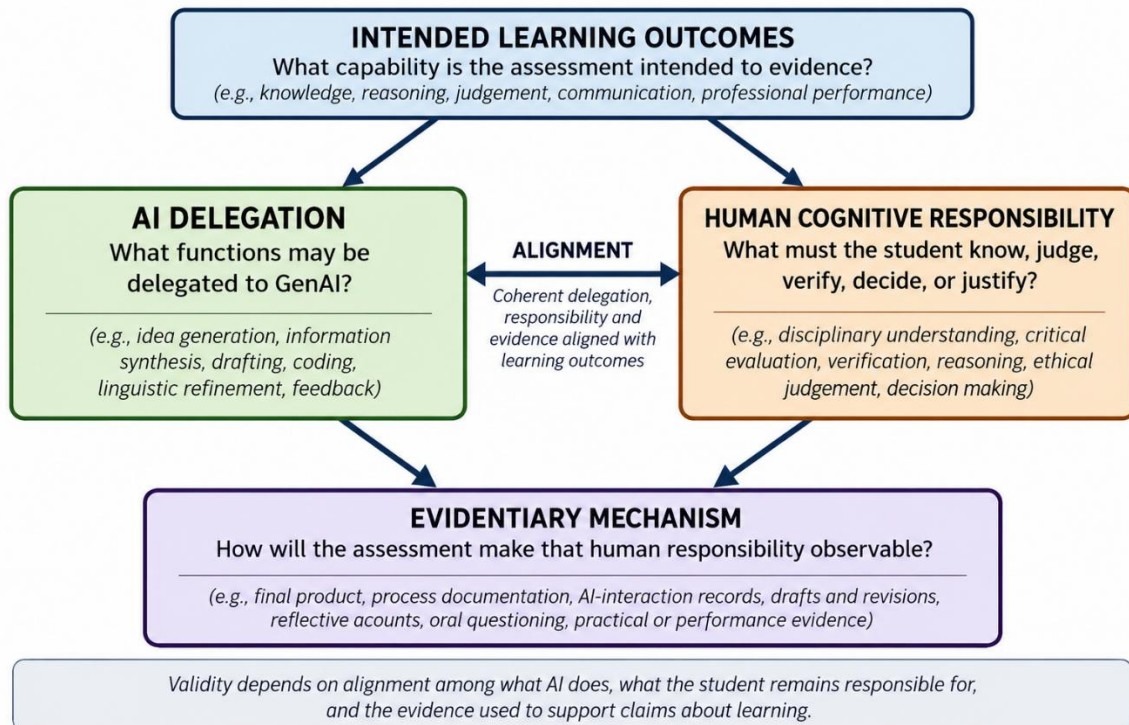


Figure 1. AI-Human Assessment Responsibility Model

4.5 Critical Evaluation May Be More Educationally Important Than Prompt Production Alone

Another implication concerns the competencies higher education chooses to develop around GenAI. Considerable attention has been given to prompt engineering and students' ability to obtain useful outputs from AI systems. Although effective prompting may be valuable, the reviewed interventions suggest that the more enduring educational capability may lie in students' ability to evaluate what AI produces.

GenAI systems can generate plausible but inaccurate, incomplete, biased, poorly reasoned, or contextually inappropriate responses.

Consequently, students who can obtain sophisticated outputs without possessing the knowledge required to evaluate them may become more dependent on AI while appearing increasingly competent through the quality of their products.

Assessment that rewards output generation without examining evaluative judgement may therefore inadvertently reinforce superficial AI dependence.

Interventions requiring students to identify errors, verify claims, compare AI-generated material against disciplinary knowledge, improve inadequate responses, or justify acceptance and rejection decisions represent a different pedagogical approach. In these designs, GenAI becomes a source of material upon which disciplinary judgement is exercised.

This distinction may be particularly important for higher education because expertise is not defined merely by the capacity to produce answers. It also involves the ability to recognize when an answer is inadequate and to explain

why. GenAI-responsive assessment may therefore provide an opportunity to place greater emphasis on evaluative judgement, verification, epistemic responsibility, and critical AI literacy.

4.6 Implementation and Scalability Matter

A pedagogically sophisticated assessment intervention has limited institutional value if it cannot be implemented sustainably. The reviewed evidence therefore indicates that implementation outcomes deserve attention alongside learning outcomes.

Oral verification provides a useful example. Oral assessment can potentially provide direct evidence of student understanding and reduce uncertainty concerning authorship. However, large-scale implementation may raise concerns about staff workload, scheduling, consistency of questioning, reliability of marking, student anxiety, accessibility, and cost.

Importantly, emerging evidence indicates that these consequences cannot simply be assumed. Large-cohort implementations suggest that carefully designed oral approaches may be more scalable than commonly expected, while smaller pilots also demonstrate that student experiences can be mixed.

Similarly, requiring students to submit complete AI interaction histories may improve process transparency but create substantial volumes of additional material for assessors. Reflective components may make reasoning visible but may also become formulaic if students learn to produce strategically desirable accounts of their AI use. AI-integrated assessment may increase professional authenticity while simultaneously creating new technological requirements for institutions and students.

These findings emphasize the need to evaluate GenAI-responsive assessment as an educational system rather than solely as a task design. Effectiveness should therefore encompass not only learning but also reliability, workload, scalability, accessibility, student experience, and institutional feasibility.

4.7 Equity and Access

The increasing integration of GenAI into assessment also raises questions of equity. Students do not necessarily have equivalent access to AI technologies, particularly where more capable models, larger context windows, advanced reasoning features, or specialist functions require paid subscriptions. Evidence from redesigned assessment contexts indicates that students themselves recognize differences between free and premium access as a potential source of inequality.

This creates an assessment-design problem. If effective completion of an assessment increasingly depends on the quality of AI assistance available to individual students, differences in technological access may become differences in assessment opportunity. Institutions that explicitly integrate GenAI into assessment therefore need to consider whether equivalent tools should be institutionally provided, whether assessment requirements can be completed adequately using universally accessible systems, and whether the competencies being assessed depend excessively on particular commercial technologies.

Equity also extends beyond financial access. Students differ in prior AI experience, digital competence, disciplinary knowledge, language background, disability, and confidence in experimenting with emerging technologies. GenAI-responsive assessment should therefore avoid assuming that all students enter AI-rich learning environments with equivalent capabilities.

4.8 Limitations of the Evidence Base

Despite rapid growth in empirical research, the evidence base remains methodologically immature. A substantial proportion of emerging studies rely on student perceptions, reflections, self-reported confidence, or evaluations of acceptability. Such evidence is valuable for understanding implementation and student experience, but positive perceptions should not be interpreted as equivalent to demonstrated improvements in learning.

Relatively fewer studies provide comparative evidence capable of establishing whether redesigned assessment improves learning, validity, evaluative judgement, or other educational outcomes relative to alternative assessment conditions. Short intervention periods, small samples, discipline-specific implementations, absence of comparison groups, and reliance on researcher-developed measures further limit the generalizability of some findings.

The rapidly changing capabilities of GenAI introduce an additional challenge. An assessment design evaluated with one generation of AI may operate differently when students gain access to substantially more capable

systems. Consequently, assessment strategies based primarily on what AI *cannot currently do* may have limited durability.

Future research should therefore place greater emphasis on the educational capabilities that students should retain irrespective of changes in AI performance. Longitudinal studies, cross-disciplinary comparisons, multi-institutional research, objective measures of learning, and evaluations of workload and cost are particularly important.

Research should also examine whether interventions that increase confidence in authorship actually improve assessment validity and whether process documentation accurately represents students' cognitive contributions rather than merely their ability to describe those contributions retrospectively.

4.9 Future Research

The emerging evidence base suggests several priorities for future research. First, studies should move beyond whether students *like* GenAI-responsive assessment and investigate whether specific interventions improve learning, reasoning, evaluative judgement, and transfer of knowledge.

Second, comparative research is needed to determine when process evidence, oral verification, critical AI evaluation, or structured AI collaboration provides meaningful advantages over conventional assessment.

Third, researchers should investigate combinations of strategies rather than treating individual assessment formats as isolated solutions. An AI-integrated project combined with process documentation and selective oral verification, for example, may provide stronger evidence than any of these mechanisms alone.

Fourth, implementation research should examine scalability, staff workload, accessibility, equity, reliability, and cost across large and diverse cohorts. Fifth, longitudinal research should investigate whether students exposed to GenAI-responsive assessment develop durable capabilities for responsible AI collaboration rather than task-specific strategies for satisfying assessment requirements.

Finally, future studies should investigate the concept of human cognitive responsibility directly. Research could examine which intellectual functions can be delegated to GenAI without undermining particular learning outcomes, which forms of evidence most effectively demonstrate students' retained capabilities, and how these relationships differ across disciplines and levels of study.

4.10 Concluding Interpretation

The emergence of GenAI does not signal the end of meaningful assessment, but it does challenge higher education to reconsider what assessment evidence represents. When AI can contribute substantially to the production of academic work, the quality of the final artefact alone may provide an incomplete account of student capability. The empirical literature examined in this review indicates that educators are beginning to respond by structuring AI use, requiring critical evaluation, making learning processes more visible, and introducing additional mechanisms for demonstrating human understanding.

The key challenge is therefore not to remove AI from assessment but to ensure that the presence of AI does not make human learning invisible. GenAI-responsive assessment is strongest when it makes explicit what AI may contribute, what intellectual responsibility remains with the learner, and how that responsibility will be demonstrated.

From this perspective, the future of assessment is unlikely to be defined by a simple choice between human and artificial intelligence. It will instead depend on the capacity of higher education to design credible evidence of human learning within increasingly AI-mediated forms of academic and professional work.

5. Implications for Practice

The findings of this review have several implications for educators, assessment designers, programme leaders, and higher education institutions responding to the increasing presence of generative artificial intelligence (GenAI). A central implication is that assessment redesign should move beyond the binary question of whether GenAI should be permitted or prohibited. Instead, assessment decisions should begin with the intended learning outcomes and explicitly consider what functions may appropriately be delegated to AI, what cognitive responsibilities should remain with students, and what evidence is necessary to demonstrate that students have exercised those responsibilities.

5.1 Design from Learning Outcomes, Not from the Technology

Assessment redesign should begin with the capability that educators intend to assess rather than with the capabilities or limitations of a particular GenAI system. Designing assessment around what current AI tools cannot accomplish may provide only temporary protection because the capabilities of these systems continue to evolve. By contrast, identifying the knowledge, reasoning, judgement, communication, ethical decision making, or professional performance that students are expected to demonstrate provides a more durable basis for assessment design.

Educators should therefore distinguish among capabilities that students need to demonstrate independently, capabilities that may legitimately be augmented by AI, and capabilities involving responsible human–AI collaboration. The appropriate balance will differ across disciplines, educational levels, and learning outcomes. There is consequently unlikely to be a single assessment format that constitutes an optimal response to GenAI across higher education.

5.2 Specify the Role of GenAI Explicitly

Where GenAI is permitted, assessment instructions should specify its intended role rather than relying solely on broad statements that AI use is “allowed.” Students may need guidance concerning whether GenAI can be used for brainstorming, searching for alternatives, drafting, linguistic refinement, coding, feedback, data interpretation, or other functions.

Equally important is identifying activities for which students retain responsibility. These may include verifying factual claims, evaluating evidence, constructing arguments, applying disciplinary standards, identifying limitations, making decisions, or justifying conclusions.

Clear specifications of these boundaries can serve both pedagogical and integrity purposes. Students are given a clearer understanding of responsible AI use, while educators can align assessment criteria with the capabilities that students are actually expected to demonstrate.

5.3 Structure AI Use When It Supports the Learning Outcomes

Permitting GenAI without pedagogical structure should not be assumed to produce meaningful AI literacy or higher-order learning (Fischer et al., 2024). Where human–AI collaboration is relevant to the learning outcomes, educators may need to design explicit activities that require students to engage critically with AI-generated material.

Such activities might require students to compare alternative AI responses, identify factual or conceptual errors, verify claims against appropriate sources, revise inadequate outputs, explain why particular suggestions were accepted or rejected, or evaluate the assumptions underlying AI-generated recommendations.

In this way, GenAI becomes more than a productivity tool. It becomes an object through which students demonstrate disciplinary knowledge and evaluative judgement.

5.4 Consider Evidence Beyond the Final Product

Where GenAI can substantially contribute to the production of an assessed artefact, educators should consider whether the final product provides sufficient evidence of the intended learning outcomes. If not, complementary evidence may be necessary.

Depending on the discipline and assessment purpose, this evidence may include drafts, annotated revisions, selected AI-interaction records, reflective commentary, decision logs, critical evaluations, oral questioning, presentations, demonstrations, or practical performance. These mechanisms can help make students’ reasoning and decision making more visible.

However, increasing evidence requirements should not become an end in itself. Requiring students to submit extensive prompt histories or lengthy reflective accounts may increase workload without necessarily improving assessment validity. Additional evidence should therefore be collected selectively and only when it contributes meaningfully to judgements about student learning.

5.5 Use Oral and Performance Verification Strategically

Oral, dialogic, and performance-based assessment can provide valuable opportunities for students to demonstrate understanding independently of a polished AI-mediated product (Hutson et al., 2026; Ogle & Jarrett, 2026). Nevertheless, the evidence reviewed does not justify treating oral assessment as a universal solution to GenAI.

Implementation should consider cohort size, staff workload, consistency of questioning, marking reliability, student anxiety, accessibility, scheduling, and cost. Selective oral verification may in some contexts be more practical than converting an entire assessment regime to oral examination. For example, a brief defense could be combined with an AI-supported project when direct evidence of student reasoning is particularly important.

The principle is therefore not that every student should be examined orally, but that assessment designers should consider when direct or dialogic evidence is necessary to support credible judgements about competence.

5.6 Assess Critical AI Judgement, Not Merely AI Productivity

Higher education should be cautious about equating effective GenAI use with the ability to generate sophisticated outputs.

Students also require the capacity to recognize when AI output is inaccurate, unsupported, biased, incomplete, inappropriate, or inconsistent with disciplinary standards.

Assessment can contribute directly to this capability by requiring students to verify, critique, correct, and justify AI-generated material. Such activities position students as accountable decision makers rather than passive recipients of machine-generated answers.

This is particularly important in professional disciplines where inappropriate reliance on apparently plausible AI output may have significant consequences. Critical evaluation and verification should therefore be considered central components of AI literacy where they align with programme learning outcomes.

5.7 Address Equity Before Requiring GenAI

Institutions should also consider equity when GenAI becomes an explicit component of assessment. Students may have unequal access to paid AI systems, advanced models, specialist functions, reliable digital infrastructure, or previous experience with AI technologies (Foung et al., 2024; Stratton-Maher & Kelly, 2026).

Where the use of GenAI is required for assessment, institutions should seek to ensure that students have access to sufficiently comparable tools and appropriate guidance. Assessment performance should not inadvertently become a proxy for students' ability to purchase more capable AI services.

Accessibility, language background, digital competence, disability, and previous exposure to AI should similarly be considered when designing and evaluating GenAI-responsive assessment.

5.8 Evaluate Assessment Redesign as an Implementation Problem

Assessment innovation should be evaluated not only according to its pedagogical attractiveness but also according to its feasibility. Institutions introducing GenAI-responsive assessment should therefore monitor student learning, academic integrity, student experience, educator workload, marking reliability, accessibility, scalability, technological requirements, and cost.

Pilot implementation may be particularly valuable before large-scale adoption. Evidence collected during implementation can identify unintended consequences and enable assessment designs to be refined before they become embedded across programmes.

Ultimately, sustainable GenAI-responsive assessment requires alignment among learning outcomes, AI delegation, human cognitive responsibility, assessment evidence, and implementation conditions. The aim should not be to design assessment that is permanently resistant to technological change, but to develop assessment capable of producing credible evidence of student learning as the technological environment continues to evolve.

6. Limitations of the Present Review

The findings of this systematic review should be interpreted in light of several limitations.

First, GenAI in higher education is a rapidly developing field. The period between the widespread public availability of ChatGPT in late 2022 and the completion of this review is relatively short, while both technological capabilities and institutional responses have changed considerably during that period. Consequently, the evidence base represents an evolving rather than mature field of research. New assessment interventions and evaluations are likely to appear rapidly, and some conclusions of the present review may require reconsideration as stronger empirical evidence becomes available.

Second, terminology within the field remains inconsistent. Studies may describe similar practices using terms such as AI-integrated assessment, AI-enabled assessment, AI-supported assessment, assessment redesign, authentic assessment, process-oriented assessment, AI literacy assessment, or human–AI collaboration. Conversely, the same terminology may refer to substantively different interventions. Although the search strategy incorporated multiple terms associated with GenAI and assessment, relevant studies using unexpected terminology may have been missed.

Third, the review deliberately adopted a relatively focused definition of a GenAI-responsive assessment intervention. Studies were included in the principal synthesis when an assessment practice, procedure, AI-use condition, verification mechanism, or evaluative activity was deliberately configured in response to GenAI and implemented with higher education students. This decision strengthened the conceptual coherence of the review but excluded adjacent evidence concerning student perceptions, institutional policy, AI detection, automated grading, naturally occurring GenAI use, and the ability of AI systems to complete conventional assessments. Such literature remains relevant to the broader transformation of assessment but addresses a different research question.

Fourth, heterogeneity across the included studies limits direct comparison. Assessment interventions vary by discipline, educational level, cohort size, assessment purpose, role of GenAI, technological platform, research design, and outcome measurement. The studies also differ substantially in whether they investigate perceptions, behaviours, learning processes, academic performance, assessment validity, integrity, workload, or implementation feasibility. This heterogeneity makes statistical aggregation inappropriate for many outcomes and places greater emphasis on narrative and thematic synthesis.

Fifth, the strength of conclusions is constrained by the methodological characteristics of the available primary research. Emerging studies frequently rely on self-reported perceptions, reflections, convenience samples, single-course implementations, and relatively short intervention periods. Positive perceptions of an intervention cannot be assumed to represent improved learning, and associations between GenAI use and assessment performance do not necessarily demonstrate causal effects. The review therefore distinguishes different forms of evidence and interprets findings in relation to methodological quality rather than treating all reported outcomes as equivalent.

Sixth, the speed of technological development creates a particular problem for evidence synthesis. Studies conducted with earlier GenAI systems may not fully represent assessment conditions involving newer and more capable models. Findings concerning the limitations of specific AI systems may therefore become outdated more rapidly than findings concerning broader pedagogical principles. For this reason, the review places greater emphasis on assessment design, cognitive responsibility, and evidence of learning than on the performance characteristics of individual GenAI platforms.

Seventh, restricting the review to English-language publications may have reduced representation of assessment practices reported in other languages and educational systems. The geographical distribution of the available literature may also limit the transferability of findings across higher education contexts with different technological infrastructures, assessment cultures, regulatory environments, and institutional resources.

Finally, the analytical categories and the proposed AI–Human Assessment Responsibility Model are interpretive products of the synthesis. Although they are grounded in recurring characteristics of the empirical literature, they should not yet be regarded as a validated assessment theory. The model requires further empirical examination across disciplines, assessment formats, institutional contexts, and levels of study. Its value at this stage lies in providing a structured way to examine the alignment among AI delegation, human cognitive responsibility, and evidence of student learning.

These limitations do not diminish the importance of the emerging evidence but indicate the level of caution required when translating early findings into assessment policy and practice. The review should therefore be understood as a synthesis of a rapidly developing empirical field and as a foundation for further testing rather than as a definitive account of GenAI-responsive assessment.

7. Conclusion

Generative artificial intelligence is changing the conditions under which higher education assessment operates. As GenAI becomes increasingly capable of contributing to writing, reasoning, problem solving, coding, feedback, synthesis, and other activities traditionally associated with student work, assessment can no longer rely uncritically

on the assumption that the quality of a submitted product provides transparent evidence of the learner's capabilities.

This systematic review examined the emerging empirical literature on GenAI-responsive assessment interventions that have been implemented with higher education students. Rather than focusing on conceptual recommendations concerning how assessment *should* respond to AI, the review concentrated on the transition from principles to practice: how educators have actually redesigned or configured assessment in response to GenAI and what evidence has emerged from those implementations.

The synthesis indicates that GenAI-responsive assessment extends beyond a simple choice between prohibiting and permitting AI. Emerging interventions regulate the conditions of AI use, structure human–AI collaboration, require students to critically evaluate AI-generated material, make learning processes more visible, and introduce oral, dialogic, or performance-based mechanisms for verifying student competence. These approaches reflect different educational purposes and should not be treated as interchangeable solutions.

A central implication of the review is that the amount of AI used within an assessment may be less important than the distribution of cognitive responsibility between the student and the technology. Allowing GenAI does not necessarily result in meaningful engagement, while extensive AI involvement does not necessarily undermine assessment when students remain responsible for disciplinary judgement, verification, critical evaluation, decision making, and justification.

To capture this relationship, the review proposes the AI–Human Assessment Responsibility Model, organized around three interconnected dimensions: AI delegation, human cognitive responsibility, and evidentiary mechanism. The model suggests that GenAI-responsive assessment should explicitly determine what functions may be delegated to AI, what intellectual responsibilities must remain with students, and what evidence will demonstrate that students have exercised those responsibilities.

This perspective shifts attention away from the pursuit of universally “AI-proof” assessment. Assessment strategies built primarily around the current limitations of GenAI may become increasingly fragile as technological capabilities advance. More sustainable assessment design begins instead with intended learning outcomes and asks which capabilities students must demonstrate independently, which may legitimately be augmented by AI, and which involve responsible human–AI collaboration.

The emerging evidence also demonstrates that pedagogical promise alone is insufficient. GenAI-responsive assessment must be evaluated in relation to learning, validity, integrity, student experience, educator workload, scalability, accessibility, equity, and cost. An intervention that successfully makes student thinking visible but cannot be implemented reliably or equitably at scale may have limited practical value. Similarly, an assessment that students perceive positively should not be assumed to improve learning without appropriate outcome evidence.

The current evidence base remains young and methodologically heterogeneous. Stronger comparative, longitudinal, cross-disciplinary, and multi-institutional research is needed, particularly research employing objective measures of learning and examining the long-term consequences of different distributions of responsibility between humans and AI. As GenAI capabilities continue to change, research should prioritize durable educational constructs such as disciplinary understanding, evaluative judgement, verification, ethical responsibility, and the capacity to justify decisions.

Ultimately, the challenge posed by GenAI is not simply that artificial intelligence can produce academic work. The deeper challenge is determining what an assessment allows educators to conclude about the person who submitted that work. The future of assessment therefore depends less on separating human activity from artificial intelligence than on ensuring that human learning remains identifiable, demonstrable, and educationally meaningful within AI-mediated academic and professional practice.

GenAI-responsive assessment should consequently be designed around a simple but demanding principle: when AI contributes to assessed work, assessment must remain capable of showing what the student knows, understands, judges, and can do.

References

- Ali, R., & Maroulis, J. (2026). Assessment validity in the age of generative artificial intelligence: A critical review. *Assessment & Evaluation in Higher Education*.
<https://doi.org/10.1080/02602938.2026.2683046>
- Brattli, H., Utne, A., & Lynch, M. (2026). Assessment validity in the age of generative AI: A natural experiment. *Informatics*, 13(4), 56. <https://doi.org/10.3390/informatics13040056>
- Casey, D. (2026). ChatGPT in public policy teaching and assessment: An examination of opportunities and challenges. *Australian Journal of Public Administration*, 85(2), 362–376. <https://doi.org/10.1111/1467-8500.12647>
- Cochrane, D., Littschwager, S., Tsang, E., & Jin, Y. (2026). Beyond unstructured use: Scaffolded integration of generative AI in assessment for learning. *Cogent Education*, 13(1), 2731708.
<https://doi.org/10.1080/2331186X.2026.2731708>
- 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>
- Fischer, I., Sweeney, S., Lucas, M., & Gupta, N. (2024). Making sense of generative AI for assessments: Contrasting student claims and assessor evaluations. *The International Journal of Management Education*, 22(3), 101081. <https://doi.org/10.1016/j.ijme.2024.101081>
- Foung, D., Lin, L., & Chen, J. (2024). Reinventing assessments with ChatGPT and other online tools: Opportunities for GenAI-empowered assessment practices. *Computers and Education: Artificial Intelligence*, 6, 100250. <https://doi.org/10.1016/j.caeai.2024.100250>
- Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O’Cathain, A., Rousseau, M.-C., Vedel, I., & Pluye, P. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for Information*, 34(4), 285–291. <https://doi.org/10.3233/EFI-180221>
- Hutson, J., Poyer, K., Ogoe, E., & Atologun, K. A. (2026). Beyond AI detection: A pilot study of Integrevise™ and viva-based verification of student understanding in AI-mediated assessment. *Trends in Higher Education*, 5(3), 59. <https://doi.org/10.3390/higheredu5030059>
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kofinas, A. K., Tsay, C. H.-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>
- Martin, A. F., Tubaltseva, S., Harrison, A., & Rubin, G. J. (2025). Participatory co-design and evaluation of a novel approach to generative AI-integrated coursework assessment in higher education. *Behavioral Sciences*, 15(6), 808. <https://doi.org/10.3390/bs15060808>
- McGroarty, A., Elliott, M. A., Butler, S. H., Findlay, K., Hamilton, D., & March, J. (2026). Generative AI, written coursework, and assessment validity in higher education. *Teaching and Learning Inquiry*. Advance online publication.
- Ogle, T., & Jarrett, P. (2026). Replacing written assessment with an oral ISBAR assessment in a 700-student cohort: Cohort-level outcomes for performance, integrity, and staff experience. *Assessment & Evaluation in Higher Education*. <https://doi.org/10.1080/02602938.2026.2684284>
- 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>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71.
<https://doi.org/10.1136/bmj.n71>
- Quince, Z., & Faulconer, E. (2026). Student GenAI use under changing institutional policies: A mixed-methods case study in undergraduate physics. *International Journal of STEM Education*, 13, 61.
<https://doi.org/10.1186/s40594-026-00651-w>
- Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., Koffel, J. B., & PRISMA-S Group. (2021). PRISMA-S: An extension to the PRISMA statement for reporting literature searches

- in systematic reviews. *Journal of the Medical Library Association*, 109(2), 174–200.
<https://doi.org/10.5195/jmla.2021.962>
- Stratton-Maher, D., & Kelly, J. (2026). Learning with AI, not from AI: Ethical and pedagogical implications of ChatGPT in undergraduate education. *Nurse Education Today*, 158, 106961.
<https://doi.org/10.1016/j.nedt.2025.106961>
- Tam, A., Abraham, D., & Williams, A. (2026). Integrating generative AI into university coursework assessments – A student perspective. *Intersection: A Journal at the Intersection of Assessment and Learning*.
<https://doi.org/10.61669/001c.164293>
- Tsiligkiris, V. (2026). From authentic products to authenticated processes: A systematic conceptual review of authentic assessment in AI-rich higher education. *Assessment & Evaluation in Higher Education*.
<https://doi.org/10.1080/02602938.2026.2695376>
- 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, 40. <https://doi.org/10.1186/s41239-024-00468-z>