Journal of General Education and Humanities
Vol. 5, No. 6, December 2026, pp. 7245 – 7275, https://doi.org/10.58421/gehu.v5i6.2254
ISSN 2963-7147

Design, Implementation, and Empirical Evaluation of a Human–AI Collaborative Adaptive Assessment Model in Higher Education

Indra Saputra[1], Alkadri Masnur[2], Resti Elma Sari[3], Ani Nurhayati[4], Elviza Yeni Putri[5], Melda Mahniza[6], Feriantano Sundang Pranata[7], Retno Triswandari[8]
[1]Department of Beauty and Cosmetology Education, Universitas Negeri Padang, Padang, Indonesia
[2,4]Department of Educational Technology, Universitas Negeri Padang, Padang, Indonesia
[3]Department of Early Childhood Teacher Education, Universitas Negeri Padang, Padang, Indonesia
[5,6]Department of Family Welfare Sciences, Universitas Negeri Padang, Padang, Indonesia
[7]Department of Tourism, Universitas Negeri Padang, Padang, Indonesia
[8]Department of Special Education, Universitas Negeri Padang, Padang, Indonesia

Article Info

Article history:

Received 2026-10-10
Revised 2026-11-08
Accepted 2026-11-11

Keywords:

Adaptive Assessment
Artificial Intelligence
Formative Feedback
Human–AI Collaboration
Natural Language Processing

ABSTRACT

AI-supported assessment increasingly enables adaptive tasks, automated analysis, and personalized feedback, yet their integration into coherent, human-supervised formative assessment remains insufficiently established. This study developed and evaluated the Human–AI Collaborative Adaptive Assessment Model (HA-CAAM) using Design and Development Research structured through the ADDIE framework. Implemented in a higher-education Media Education course, the model employed performance-responsive task adaptation using predefined thresholds, rubric-guided AI-assisted scoring, formative feedback, and instructor verification. Five experts validated the model, 148 students provided quantitative and qualitative implementation feedback, and 50 student responses were independently assessed by two human raters and AI. Expert validation provided initial evidence of content validity (mean Aiken's V = 0.818). Students reported generally favorable experiences, particularly regarding clarity and feedback usefulness, although Learning Support and Overall User Experience received comparatively lower ratings. AI-generated scores demonstrated strong agreement with human-reference scores (ICC = 0.898), despite individual discrepancies. These findings provide initial evidence of content validity, positive implementation experiences, and scoring agreement, rather than demonstrated improvements in learning outcomes or equivalence between AI and human assessment decisions. HA-CAAM contributes an empirically refined, seven-stage assessment framework that coordinates performance-responsive adaptation, AI-assisted analysis, formative feedback, and instructor-led quality assurance within a human-supervised assessment process.

This is an open-access article under the CC BY-SA license.

Corresponding Author:

Indra Saputra
Department of Beauty and Cosmetology Education, Universitas Negeri Padang, Padang, Indonesia
Email: indrasaputra@fpp.unp.ac.id

1. INTRODUCTION

Artificial intelligence (AI) is increasingly reshaping higher education by extending the capacity of digital environments to support personalized learning, assessment, feedback, and instructional decision-making. Recent advances in natural language processing (NLP), machine learning, and generative AI have accelerated this transformation, enabling educational systems to analyze complex learner-generated data and provide increasingly responsive support [1], [2], [3]. Assessment represents a particularly important area of this transformation. The emergence of generative AI has not only expanded opportunities for rapid evaluation, personalized feedback, and self-assessment. However, it has also challenged conventional assessment practices, which are often static, labor-intensive, and insufficiently responsive to differences in learners' performance and needs [4], [5]. These limitations become particularly salient for open-ended assessment, where constructed responses can reveal deeper conceptual understanding but require substantial time and effort to evaluate at scale [6], [7]. These challenges are especially relevant to the Media Education course examined in this study, where assessment must accommodate differences in student performance while supporting timely feedback and subsequent instructional decisions. Consequently, the growing integration of AI into higher education creates a need to move beyond simply digitizing conventional assessment toward approaches that are more adaptive, formative, responsive, and explicitly aligned with learning and instructional decision-making.

Existing research has established several complementary capabilities that can support such transformation. First, adaptive learning technologies can use information about learners' performance and needs to provide personalized tasks, support, and feedback; recent work has also demonstrated the potential of combining adaptive technologies with NLP to improve contextual personalization [8], [9]. Second, advances in NLP and large language models have substantially expanded the capacity to analyze constructed responses. A systematic review of 93 studies showed that AI- and NLP-based approaches are increasingly being used for automatic assessment of text-based responses in post-secondary education [6]. At the same time, empirical systems have shown that NLP pipelines can analyze essay content and translate the results into content-specific formative feedback [10], [11]. Third, generative AI further extends assessment by producing rapid and individualized feedback. GPT-based models, for example, have shown considerable potential to generate readable, pedagogically relevant feedback on open-ended higher education tasks [12], [13]. At the same time, emerging evidence cautions against treating automated assessment as a standalone replacement for academic judgment. Human–AI collaborative grading approaches show that instructor involvement in refining criteria, reviewing AI-supported scoring outputs, and validating assessment decisions can improve transparency and alignment with human grading [14], [15]. More recent evidence similarly suggests that collaborative human–AI marking can improve scoring consistency when AI-generated evaluations remain anchored to explicit criteria and human evaluative judgment [16], [17]. Taken together, the literature indicates that the technological building blocks for adaptive,

NLP-supported, feedback-oriented, and human-supervised assessment are increasingly well established.

However, these developments do not yet fully resolve how such capabilities should operate together as a coherent pedagogical assessment process. Previous studies have integrated selected assessment functions, including adaptive learning with NLP [8], automated response analysis with formative feedback [10], and LLM-supported grading with instructor validation [14]. These contributions demonstrate the feasibility of combining AI-supported assessment capabilities but differ in their primary emphasis. Adaptive approaches focus on performance-responsive learning pathways, NLP-supported assessment emphasizes response analysis and feedback, and collaborative grading foregrounds human verification of AI-generated scores. Broader reviews likewise identify assessment, personalized feedback, and intelligent tutoring as prominent but not always pedagogically integrated AI applications in higher education [1], [18]. More recent human-in-the-loop frameworks emphasize instructor responsibility, transparency, and institutional governance [19], [20]. Building on these contributions, HA-CAAM addresses the need to coordinate adaptive task progression, rubric-guided AI assessment, human quality assurance, formative feedback, and reassessment within a single learning-oriented assessment cycle. Thus, the unresolved issue is less whether AI can adapt learning, analyse constructed responses, generate feedback, or support grading, and more how these complementary capabilities can be pedagogically orchestrated within an end-to-end assessment architecture. Comparatively less attention has been given to a closed assessment cycle that systematically connects performance-responsive task adaptation, rubric-guided AI/NLP analysis, personalized and actionable formative feedback, human oversight, revision and reassessment, instructional follow-up, and cross-cutting safeguards. Moreover, many empirical implementations remain anchored to particular assessment tasks or disciplinary settings, leaving the configurability of such an integrated architecture across course contexts comparatively underexplored. Accordingly, it remains unclear how these capabilities can be organized into a coherent, trustworthy, and configurable adaptive assessment model for AI-enhanced higher education.

To address this gap, the present study develops and empirically evaluates the Human–AI Collaborative Adaptive Assessment Model (HA-CAAM) for higher education. HA-CAAM conceptualizes assessment as an integrated, iterative, and human-supervised process in which assessment tasks respond to learner performance, constructed responses are analyzed through AI/NLP-assisted processes anchored to explicit scoring criteria, personalized formative feedback is generated from assessment evidence, and human oversight supports quality assurance and accountable assessment decisions. Rather than introducing a new NLP algorithm or replacing human assessors, HA-CAAM contributes an integrated pedagogical process in which performance-responsive task progression determines subsequent assessment challenges, rubric-guided AI analysis supports scoring and formative feedback, and instructor oversight maintains assessment quality and accountability. These functions connect through iterative reassessment so evidence from one assessment activity can inform subsequent learning and assessment decisions. The model is configurable because course-specific learning outcomes, assessment tasks, difficulty specifications, and scoring rubrics can be adjusted while retaining the same underlying

assessment cycle; however, this design flexibility does not establish its empirical effectiveness or transferability across different disciplines.

Accordingly, this study aimed to develop, validate, implement, evaluate, and empirically refine HA-CAAM as a configurable human–AI collaborative assessment model for higher education. Three research questions guided the study:

RQ1. To what extent does the initial HA-CAAM demonstrate adequate content-validity evidence based on expert evaluation, and what conceptual or operational refinements are identified through expert feedback?

RQ2. How do students experience the implementation of HA-CAAM in terms of clarity and usability, perceived adaptivity, assessment relevance, AI feedback, fairness and trust, learning support, and overall experience, and what complementary insights emerge from their qualitative feedback?

RQ3. To what extent do AI-assisted scores generated within HA-CAAM correspond with human-reference scores under shared rubric-based assessment conditions?

2. THEORETICAL BACKGROUND

2.1. Assessment as a Pedagogical Process in AI-Enhanced Higher Education

Assessment in higher education increasingly extends beyond measuring learning outcomes to support learning itself. Contemporary formative assessment perspectives emphasize that assessment evidence becomes educationally meaningful when it is used to identify learners' current understanding, inform feedback, and guide subsequent learning actions rather than merely document achievement [21], [22]. This orientation is closely aligned with Assessment for Learning (AfL), in which assessment activities, evaluation, and feedback are deliberately embedded within teaching and learning processes to support ongoing improvement [23]. In this perspective, feedback is not an endpoint after assessment; its value depends on whether learners can interpret and use the information to improve subsequent performance [24], [25]. This learning-oriented conception is particularly relevant to AI-supported assessment because technological responsiveness has pedagogical value only when assessment evidence contributes to subsequent learning decisions.

The growing use of artificial intelligence further expands assessment's pedagogical role. Generative AI can support immediate feedback, self-assessment, automated evaluation, and more responsive assessment practices, but its emergence also requires higher education institutions to reconsider how they maintain assessment quality, learning integrity, and human judgment [4]. Reviews of emerging technologies in higher education similarly caution against evaluating AI-supported assessment primarily by technological performance. Instead, formative purposes, human-centered implementation, and ethical considerations should remain central to assessment design [18]. More specifically, the intersection of AI and AfL suggests that AI can enhance the generation and use of assessment evidence while also introducing challenges related to feedback quality, learner agency, and the pedagogical interpretation of AI-generated information [26], [27]. Accordingly, AI-enhanced assessment should be understood not merely as automated scoring but as a pedagogical process in which evidence of student learning is generated, interpreted, and translated into feedback and instructional action. This learning-oriented assessment principle

provides a foundational rationale for HA-CAAM: assessment evidence should continuously inform feedback, revision, and subsequent learning rather than function as a terminal score.

2.2. Adaptive Assessment and Performance-Responsive Pathways

A central limitation of uniform assessment is that learners with different levels of knowledge and performance are exposed to the same assessment conditions and levels of challenge. Adaptive approaches address this limitation by modifying learning or assessment pathways based on learner information. A scoping review of 69 studies in higher education found that personalized adaptive learning systems commonly tailor learning pathways to individual student needs and performance indicators, with prior-knowledge assessments frequently triggering adaptive content delivery. The review also reported improvements in academic performance in 59% of the included studies and increased engagement in 36% [28].

The adaptive principle becomes more powerful when systems can capture richer information about learner needs. Mejeh and Rehm [8], for example, demonstrated the potential of combining adaptive learning technology with natural language processing to identify learner information that can support more contextualized and personalized adaptation. Their work indicates that adaptation need not depend solely on fixed indicators or objective-response data; textual evidence from learners can also inform personalized support. Recent work in adaptive assessment similarly demonstrates the feasibility of connecting learner-performance information with AI-supported assessment processes and iterative personalized responses within a scalable assessment architecture [29], [30].

In the context of assessment, adaptivity can therefore be conceptualized as a performance-responsive mechanism in which evidence obtained from a learner's current response informs the level or nature of subsequent assessment tasks. This logic aligns with the broader principle of adaptive testing, in which assessment delivery is dynamically adjusted based on information obtained during the assessment process rather than following an identical fixed sequence for all learners. However, HA-CAAM does not adopt a psychometrically calibrated computerized adaptive testing procedure; instead, it operationalizes adaptivity pedagogically through predefined performance-based decision rules that determine subsequent task difficulty. This distinction matters because the purpose of adaptation in the present model is not to produce a latent ability estimate with maximum measurement efficiency, but to maintain an appropriate level of assessment challenge while generating evidence that supports feedback and further learning. In HA-CAAM, performance-responsive adaptation therefore connects current assessment evidence directly to the subsequent assessment pathway.

2.3. NLP-Supported Analysis of Constructed Responses

Adaptive assessment is only as meaningful as the evidence that drives adaptation. Although fixed-response items are relatively easy to score automatically, they may provide limited information about learners' conceptual reasoning, explanation, or argumentation. Open-ended text-based responses can capture richer evidence of understanding, yet their evaluation becomes increasingly time-consuming as class size grows. This tension has

stimulated growing interest in using artificial intelligence (AI) and natural language processing (NLP) to support the analysis of constructed responses.

A systematic review of 93 studies in post-secondary education showed that AI- and NLP-based approaches have increasingly been used to support the automatic assessment of text-based responses, including short answers and extended written responses [6]. The review indicates that developments in NLP and large language models (LLMs) have expanded the possibilities for analyzing open-ended responses at scale while potentially enabling faster assessment and feedback. Empirical work further demonstrates how automated textual analysis can be connected to formative purposes. Gombert et al. [10], for example, developed an NLP-supported approach that analyzed student essay content and translated the resulting evidence into informative feedback, showing how automated response analysis can contribute not only to scoring but also to identifying learning-relevant information.

Recent advances in LLM-based assessment further demonstrate the potential of AI-supported analysis of constructed responses. Morris et al. [31], for example, developed LLM-based models to score constructed-response mathematics items automatically and reported close correspondence with human scoring for most evaluated tasks. Importantly, however, their approach involved item-specific preprocessing, input modification, and model configuration, indicating that scoring performance depends not only on the underlying AI model but also on how assessment information and task characteristics are represented within the scoring process. Similarly, Xue et al. [32] demonstrated that prompt design can materially influence automatic-scoring performance. Using constructed responses from 930 students across 11 items, they found that systematically optimized prompts improved scoring accuracy relative to prompts based only on the original human-scoring instructions. Their findings also raise validity considerations when LLM-generated prompt modifications depart from established scoring specifications.

Taken together, these findings suggest that the quality of AI-supported constructed-response assessment cannot be attributed to the language model alone. It depends on the alignment among assessment tasks, scoring criteria, response evidence, prompt specifications, and the procedures used to interpret AI-generated outputs. Accordingly, HA-CAAM treats AI/NLP analysis as rubric-guided assessment support rather than unconstrained automated judgment. Constructed responses provide evidence that can support scoring, diagnosis, and personalized feedback, while explicit assessment criteria and human oversight establish the interpretive boundaries within which AI-generated analysis is used. This principle enables richer learner evidence to inform subsequent adaptive decisions without transferring final assessment authority from human assessors to the AI system.

2.4. AI-Driven Formative Feedback

The educational value of AI-supported assessment depends not only on the capacity to analyze student responses but also on how the resulting evidence is translated into feedback that supports subsequent learning. Contemporary feedback perspectives emphasize that feedback should be learner-centered rather than simply information transmitted from an assessor to a learner. A rapid review and qualitative synthesis of 95 publications by Ryan et

al. [33] identified multiple design attributes of learner-centered text-based feedback, highlighting the importance of considering its context, characteristics, and components. From this perspective, effective feedback should provide information that learners can meaningfully interpret and use to support subsequent learning action. This principle is particularly relevant to AI-supported assessment because the capacity to generate feedback rapidly and at scale does not, by itself, ensure that such feedback is pedagogically useful.

Artificial intelligence extends the possibilities for providing timely, individualized, and scalable formative feedback. Dai et al. [12] and Wan & Chen [34] demonstrated large language models' capacity to generate formative feedback on student work, while Gombert et al. [10] and Pozdniakov et al. [35] showed how automated analysis of student responses can be translated into content-specific, informative feedback. More broadly, Ba et al. [36], in a systematic review of 129 studies, found that AI-assisted feedback can operate across multiple feedback domains—including task, process, self-regulation, and self-level feedback—and can support learner perceptions, actions, and learning outcomes. Their synthesis nevertheless emphasizes that meaningful AI-assisted feedback requires integrating pedagogical and technological considerations rather than a predominantly technology-driven approach.

Recent developments in generative AI further demonstrate how assessment evidence can be transformed into individualized feedback. AlGhamdi et al. [37], for example, examined prompt-based LLMs for automated scoring and feedback generation in higher education, demonstrating the potential of structured prompting to support personalized feedback alongside assessment processes. Such findings reinforce the importance of prompt and assessment design in determining the pedagogical quality of AI-generated feedback. The usefulness of AI feedback therefore depends not simply on generative capability, but on how the AI-supported process represents assessment evidence, scoring criteria, feedback specifications, and learner needs [38], [39].

At the same time, immediacy and personalization do not necessarily guarantee pedagogically effective feedback. Recent systematic evidence on LLM-generated formative feedback identifies recurring concerns including hallucinations, over-positivity, and misclassification, while indicating that grounding AI feedback in student work and maintaining teacher-in-the-loop oversight can mitigate some of these limitations [40], [41]. Accordingly, AI-generated feedback should be evaluated not merely by how rapidly or individually it is produced, but by whether it is sufficiently accurate, specific, interpretable, and actionable to support subsequent learning.

Within HA-CAAM, AI-generated feedback is therefore positioned as part of a broader formative sequence rather than as a terminal output of automated assessment. Evidence from student responses informs the diagnosis of strengths and learning gaps, which in turn supports personalized, actionable feedback, revision, and subsequent reassessment. This design connects AI-supported feedback to the Assessment for Learning principle established earlier: assessment evidence acquires formative value when learners can use it to improve subsequent performance. Consequently, AI-driven feedback in HA-CAAM does not simply provide faster comments; it helps transform assessment evidence into learning-oriented action within an iterative assessment cycle.

2.5. Human–AI Partnership, Trustworthiness, and Assessment Quality

Assessment decisions carry pedagogical and ethical consequences, making reliability, validity, fairness, transparency, and contextual judgment essential considerations in AI-supported assessment. Transforming assessment through generative AI therefore requires careful consideration not only of what AI systems can do, but also of how assessment authority and responsibility are distributed between technological systems and human assessors. Xia et al. [4] and Bower et al. [42] emphasize that effective assessment transformation depends not only on technological capability but also on teachers' understanding of AI, beliefs about human–AI assessment, and capacity to adopt innovative and holistic assessment practices.

Human-centered concerns are similarly prominent in research on technology-supported assessment. Sembey et al. [18] and Alfredo et al. [43] identified the need for greater attention to human-centric implementation and ethical considerations when emerging technologies are used for assessment and feedback. More recent evidence reinforces this concern. A systematic review of LLM-generated formative feedback identified hallucinations, over-positivity, and misclassification as recurring limitations and indicated that teacher-in-the-loop oversight can help mitigate such risks [40]. Human-in-the-loop assessment research likewise emphasizes that AI's increasing generativity, adaptivity, and algorithmic feedback capabilities create a corresponding need for transparent assessment criteria, informed teacher oversight, and ethical safeguards [19].

Empirical evidence further supports collaborative rather than replacement-oriented approaches to AI-supported assessment. Win Myint et al. [14] showed that an AI–instructor collaborative grading approach, in which instructors refined the marking scheme and validated assessment logic, produced results that more closely reflected human grading than standalone prompting approaches. Furthermore, Cohn et al. [44] suggested that the value of AI in assessment does not depend on transferring assessment authority to automated systems, but on structuring complementary roles in which AI contributes computational and analytic capabilities. In contrast, human assessors contribute contextual interpretation, pedagogical judgment, and professional accountability.

Accordingly, AI capability and assessment authority should remain conceptually distinct. Within HA-CAAM, AI can support response analysis, rubric-guided scoring, diagnosis, personalized feedback, and analytic reporting. In contrast, human assessors retain responsibility for reviewing selected or flagged outputs, verifying or adjusting scores when necessary, interpreting assessment evidence within the learning context, and making pedagogically consequential decisions. Human oversight is therefore not treated as an external safeguard applied only after AI processing; it is a structural component of the assessment process itself. In this arrangement, AI functions as an analytic and feedback partner, while human assessors retain professional discretion and final accountability for assessment decisions.

2.6. Toward a Configurable Closed-Loop Adaptive Assessment Architecture

The preceding literature streams demonstrate that the major capabilities required for AI-enhanced assessment are increasingly available. Adaptive systems can modify learning

or assessment pathways according to learner information; NLP and LLMs can support the analysis of constructed responses; generative AI can produce personalized formative feedback; and human involvement can strengthen the contextual appropriateness, transparency, and accountability of AI-supported assessment. However, researchers often examine these capabilities as individual functions or in partially connected combinations rather than as components of a unified pedagogical assessment cycle [45]. This observation is consistent with broader reviews showing that AI-based learning tools in higher education commonly perform assessment and evaluation, personalized feedback and recommendation, and intelligent tutoring functions, while effective strategies for pedagogically integrating these capabilities remain insufficiently established [1].

Existing research nevertheless provides important foundations for greater integration. Mejeh and Rehm [8] demonstrate how NLP can enrich adaptive personalization; Gombert et al. [10] show how automated textual analysis can be translated into informative feedback; and Win Myint et al. [14] demonstrate the value of explicit human–AI collaboration in assessment. Emerging frameworks extend this integration further. Ilieva et al. [20], for example, proposed a generative AI-driven assessment framework structured around the complementary responsibilities of instructors, students, and institutional control authorities, emphasizing transparency, quality assurance, and responsible implementation. Human-in-the-loop perspectives similarly emphasize the importance of connecting AI-enabled generativity, adaptivity, and algorithmic feedback with teacher competencies and ethical safeguards [19]. Collectively, these developments indicate movement toward more integrated AI-supported assessment, while also leaving scope for an architecture that more explicitly connects performance-responsive task progression, rubric-guided AI/NLP analysis, human quality assurance, personalized formative feedback, iterative reassessment, and cross-cutting safeguards within a single learning-oriented assessment cycle.

On this basis, the literature supports a closed-loop assessment logic in which evidence generated at one stage is used to shape subsequent assessment and learning actions [46]. Conceptually, this process connects assessment preparation and adaptive task delivery with student response, AI/NLP-assisted analysis, human quality assurance, diagnosis and personalized feedback, and subsequent revision and reassessment. Assessment outcomes are therefore not treated as terminal outputs. Instead, evidence produced during assessment informs learner feedback, human judgment, instructional follow-up, and the configuration of subsequent assessment activities. This cyclical relationship matters because it distinguishes an adaptive formative assessment process from approaches that use AI primarily as an isolated tool for scoring, feedback generation, or task production.

Configurability constitutes a complementary design principle. AI-based learning and assessment tools are increasingly implemented across diverse disciplinary contexts, learning objectives, assessment formats, and technological environments [1]. A generic pedagogical architecture should therefore preserve its underlying assessment logic while allowing course-specific components—including learning outcomes, disciplinary content, structured item/task banks, difficulty specifications, scoring rubrics, and AI prompts—to be configured according to the intended learning context [47]. Configurability in this sense refers to design flexibility rather than demonstrated cross-disciplinary generalizability. The architecture may

be configured for different course contexts, but its transferability across disciplines and institutions remains an empirical question requiring subsequent validation.

Taken together, these theoretical and empirical strands provide the conceptual foundation for the Human–AI Collaborative Adaptive Assessment Model (HA-CAAM) developed in this study. At the conceptual level, HA-CAAM integrates learning-centered assessment, performance-responsive adaptation, rubric-guided AI/NLP analysis, human quality assurance, personalized and actionable formative feedback, revision and reassessment, and responsible-AI safeguards within an iterative assessment process. Its contribution lies not in introducing these capabilities individually, but in pedagogically orchestrating them so that AI-supported responsiveness remains connected to learning improvement, professional human judgment, and accountable assessment practice. The literature synthesis therefore establishes the design requirements from which the initial HA-CAAM was developed and subsequently validated empirically, implemented, evaluated, and refined.

3. METHOD

3.1. Research Design and Development Procedure

This study used Design and Development Research (DDR) as the overarching methodology to systematically develop, validate, implement, evaluate, and refine the Human–AI Collaborative Adaptive Assessment Model (HA-CAAM). DDR is appropriate for research concerned with the systematic development and empirical investigation of educational products, tools, and models [48]. Within this methodological framework, the ADDIE framework—Analysis, Design, Development, Implementation, and Evaluation—structured the development process [49]. Accordingly, DDR served as the overarching research methodology for developing and empirically evaluating HA-CAAM, while ADDIE functioned as the procedural framework organizing its five development stages. The two frameworks therefore served complementary rather than overlapping methodological roles. An overview of the research design and development procedure is presented in Figure 1.

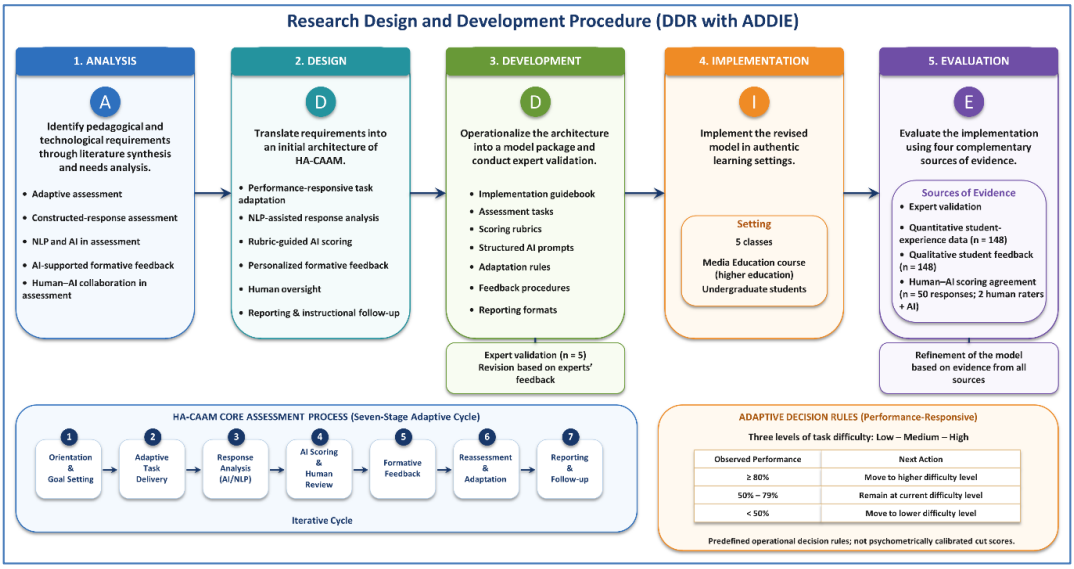


Figure 1. Research Design and Development Procedure of HA-CAAM Using the DDR and ADDIE Framework

During the Analysis stage, pedagogical and technological requirements were identified through a synthesis of literature concerning adaptive assessment, constructed-response assessment, natural language processing (NLP), AI-supported formative feedback, and human–AI collaboration in assessment. The Design stage translated these requirements into an initial architecture connecting performance-responsive task adaptation, NLP-assisted response analysis, rubric-guided AI scoring, personalized formative feedback, human oversight, reporting, and instructional follow-up. During Development, the architecture was operationalized into a model package comprising an implementation guidebook, assessment tasks, scoring rubrics, structured AI prompts, adaptation rules, feedback procedures, and reporting formats. The resulting model was subsequently subjected to expert validation and revised before student implementation.

During the Implementation stage, the revised model was applied in five classes of a Media Education course. The Evaluation stage integrated four complementary sources of evidence: expert validation, quantitative student-experience data, qualitative student feedback, and human–AI scoring agreement. These sources of evidence were used to evaluate implementation and identify components needing further refinement. The resulting refinements informed the final specification of the model's adaptive decision procedures, AI/NLP-assisted analysis and scoring, human oversight, formative feedback and reassessment, and governance safeguards while retaining its underlying assessment architecture.

HA-CAAM was designed as a configurable assessment architecture rather than a course-specific assessment system. The implementation maintained its core assessment process, while learning outcomes, disciplinary content, assessment tasks, difficulty specifications, and scoring rubrics were configured to match the intended course competencies. The adaptive mechanism used three levels of task difficulty developed by the instructor to align with the intended course learning outcomes. Students began at the intermediate level. Performance scores of ≥80% indicated progression to a greater difficulty level, scores of 50%–79% indicated maintenance of the current level, and scores below 50% indicated progression to a lower difficulty level. These thresholds were based on instructional mastery criteria and served as operational pedagogical decision rules rather than psychometrically calibrated cut scores. AI-generated recommendations were reviewed by the instructor, who determined the subsequent difficulty level based on the student's performance. Thus, the mechanism supported instructor-mediated, performance-responsive assessment rather than psychometric estimation of latent ability.

3.2. Participants and Research Context

The study involved two participant groups corresponding to the validation and implementation phases. The first group comprised five expert validators purposively selected for their academic and research expertise relevant to the components evaluated in the model. Collectively, the experts represented complementary expertise in educational assessment and evaluation, AI/NLP-supported learning technologies, psychometrics, instructional design, and ethics, privacy, and fairness in AI-supported assessment. The number of validators was determined to ensure representation across these five

complementary areas of expertise, and experts were purposively selected based on their relevance to the model's validation dimensions. This expertise-based selection enabled the model to be examined from complementary assessment, technological, pedagogical, psychometric, and ethical perspectives. The experts independently evaluated the conceptual coherence, content relevance, operational feasibility, and implementation characteristics of the initial HA-CAAM and its supporting guidebook.

The implementation phase involved undergraduate students enrolled in a Media Education course at a public university in Indonesia. The implementation was conducted during the July–December 2025 academic semester over six class meetings. Assessment activities used a blended learning approach combining face-to-face and online activities, with instructors using ChatGPT to support assessment and formative feedback. HA-CAAM was implemented across five classes, each with approximately 30 third-semester students, for an initial implementation group of 150 students. Of these, 148 students completed the student-experience questionnaire and provided usable responses, yielding a response rate of 98.7%. The Media Education course provided the context for implementing and evaluating the model; however, the model was designed as a configurable assessment architecture in which course-specific learning outcomes, content, assessment tasks, difficulty specifications, and scoring rubrics could be adapted to the intended learning context.

For the human–AI scoring evaluation, we selected a class-balanced subsample of 50 student responses, comprising 10 responses from each of the five participating classes. This sampling procedure ensured equal representation of the five implementation classes in the scoring-agreement analysis. The sample-size rationale and selection procedure for the human–AI scoring evaluation are described in Section 3.4.

3.3. Research Instruments

Three instruments were used to collect evidence during the validation and implementation phases. First, expert evaluation was conducted using a 25-item validation instrument covering 10 aspects of the proposed HA-CAAM: model foundation and rationale; constructs and model structure; adaptive assessment mechanism; NLP-assisted analysis mechanism; AI-supported feedback; assessment validity and reliability considerations; implementation syntax and procedures; reporting and follow-up; ethics, privacy, and responsible AI use; and guidebook quality. Each item was rated on a five-point relevance scale. The instrument was designed to obtain expert judgments on the relevance and adequacy of the model components and to identify areas requiring refinement before implementation.

Second, students' experiences with the implementation of HA-CAAM were examined using a 33-item questionnaire comprising eight dimensions: Clarity and Ease of Use (CEU; 4 items), Perceived Adaptivity (PA; 4 items), Perceived Assessment Relevance (PAR; 4 items), AI Feedback Quality (AIFQ; 5 items), Feedback Usefulness and Improvement (FUI; 4 items), Fairness and Trust (FFT; 4 items), Learning Support (LS; 4 items), and Overall User Experience (OUE; 4 items). All items were rated on a five-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The questionnaire served as an implementation-evaluation instrument to capture students' perceptions and

experiences of the assessment process, rather than to measure causal changes in academic performance or learning outcomes.

Third, three open-ended questions were included to obtain complementary qualitative evidence about students' experiences with the model. The questions invited students to describe aspects of the assessment and feedback process that they found useful, difficulties or limitations they experienced, and suggestions for improving the model. All 148 students who completed the quantitative questionnaire also responded to the open-ended questions. These responses were used to complement the quantitative findings by providing contextual information about students' experiences of adaptivity, feedback, learning support, trust, human oversight, and areas for improvement. The assessment tasks used during implementation comprised essay-based and multiple-choice questions aligned with the course learning outcomes. Essay responses were evaluated using predefined scoring rubrics, while multiple-choice responses were evaluated using answer keys and corresponding scoring criteria. The same assessment criteria were provided to GPT-4o and the human assessors to ensure a common scoring basis.

3.4. Validation and Implementation Procedure

The validation and implementation procedure was conducted sequentially. First, the five expert validators independently evaluated the initial HA-CAAM and its supporting guidebook using the 25-item validation instrument. Experts also provided written comments and recommendations concerning the conceptual structure, adaptive procedures, AI/NLP-assisted assessment processes, scoring and feedback mechanisms, implementation procedures, and ethical safeguards. Item-level content-validity evidence and qualitative expert feedback were subsequently used to identify components requiring refinement before classroom implementation. Revisions focused on clarifying the alignment among learning outcomes, assessment tasks, scoring criteria, adaptive decision rules, AI-supported analysis and feedback, human oversight, and follow-up procedures.

Following expert validation and refinement, HA-CAAM was configured for implementation in the Media Education course. The instructor developed essay-based and multiple-choice assessment tasks at three difficulty levels aligned with the intended course learning outcomes, supported by corresponding scoring rubrics and answer keys. Student performance was assessed against these predefined criteria, and GPT-4o generated recommendations concerning the difficulty level of subsequent tasks. The instructor reviewed these recommendations and determined whether students progressed to a higher level, remained at the same level, or received a lower-level task according to the operational rules described in Section 3.1. Students subsequently completed new assessment tasks based on their preceding performance. This procedure provided instructor-mediated, performance-responsive assessment rather than psychometric estimation of latent ability.

ChatGPT GPT-4o, developed by OpenAI, supported AI-assisted assessment and formative feedback. The instructor used ChatGPT with a standardized prompt template, adjusting task content, scoring rubrics, or answer keys as needed. For each student response, GPT-4o was provided with the assessment task, relevant scoring criteria, and the student's pseudonymized response. The AI was instructed to evaluate the response against the

supplied criteria and generate a score, formative feedback, and a recommendation for the difficulty level of the subsequent task. Essay responses were assessed using predefined rubrics, whereas multiple-choice responses were assessed against the supplied answer keys and scoring criteria. The same criteria were used by human assessors to maintain consistency in the assessment standards.

AI-generated outputs were treated as assessment support rather than autonomous assessment decisions. All AI-generated scores, feedback, and task-difficulty recommendations were reviewed by the instructor against the student's response, the predefined scoring rubric or answer key, and the instructional context. Greater attention was given to flagged outputs requiring more detailed verification against the student's response and the predefined assessment criteria. The instructor retained responsibility for determining final scores and subsequent assessment tasks and could revise AI-generated outputs when inconsistencies with the assessment criteria or response evidence were identified. Thus, AI-assisted assessment operated within a human-supervised process, with final assessment authority remaining with the instructor.

A separate scoring-agreement evaluation was conducted to examine the correspondence between AI-generated scores and human-reference scores. The required number of student responses was determined using the reliability-study sample-size approach proposed by Walter et al. [50]. With two ratings per response, a two-sided significance level of .05, statistical power of .80, a minimum acceptable intraclass correlation of .60, and an anticipated intraclass correlation of .80, the estimated minimum sample size was 49 responses. This was rounded to 50 responses, with 10 responses randomly selected from each of the five participating classes to maintain balanced class representation.

Two faculty members on the Media Education teaching team who were not on the research team independently scored the same 50 pseudonymized student responses evaluated by GPT-4o. The responses included essay-based and multiple-choice assessments. Both human raters and GPT-4o used identical scoring rubrics and answer keys, while the human raters were blinded to the AI-generated scores during their independent evaluations. Human–human agreement was examined first to establish the consistency of the human-reference ratings. The two human ratings for each response were subsequently averaged to produce the Human Mean, which served as the reference score for statistical comparison with the independently generated AI score.

After completing the model implementation, the 148 participating students completed the 33-item student-experience questionnaire and responded to the three open-ended questions described in Section 3.3. Participation in the research evaluation was voluntary. The study was conducted with institutional research permission. Written or electronic informed consent was obtained from all participants prior to data collection. Participants were informed about the study's purpose, the voluntary nature of participation, the use of AI in the assessment process, and the confidentiality of their data. Student identities were replaced with pseudonymous codes during AI-assisted assessment.

3.5. Data Analysis

Expert-validation data were analyzed using Aiken's V to quantify the degree of agreement among the five experts regarding the relevance of each model component. Item-level V coefficients were calculated for the 25 validation items, and descriptive means of the item-level V coefficients were subsequently calculated for each of the 10 validation aspects to summarize the pattern of expert judgments. These aspect means were treated as descriptive summaries rather than as separately pooled Aiken's V coefficients. Quantitative expert judgments were interpreted together with the experts' qualitative comments to identify components requiring refinement prior to implementation. Aiken's V coefficients were interpreted as indicators of expert-judged content relevance, with higher coefficients reflecting stronger agreement regarding the relevance of the evaluated components; the coefficients were not treated as evidence of overall model effectiveness.

The internal consistency of the 33-item student-experience questionnaire was examined separately for each of its eight dimensions using Cronbach's alpha (α) and corrected item–total correlations. These analyses provided evidence of the instrument's internal consistency and were not interpreted as comprehensive evidence of construct validity. Students' experiences of HA-CAAM implementation were then summarized descriptively using means, standard deviations, and 95% confidence intervals for each dimension. Because this analysis aimed to characterize students' experiences of model implementation rather than test causal effects or between-group hypotheses, no inferential comparisons among student groups were conducted.

Responses to the three open-ended questions were analyzed using descriptive thematic content analysis. The lead researcher conducted the analysis manually, with assistance from the research team, using a structured coding matrix. Responses were read repeatedly to identify meaningful units, which were assigned initial descriptive codes. The team then grouped conceptually related codes into broader categories and consolidated them into recurring themes. The research team helped review the coding structure and discuss the interpretation and organization of the resulting themes. Theme frequencies were calculated as the number and percentage of participants whose responses reflected each theme. Because a single response could reflect more than one theme, the themes were not treated as mutually exclusive. The qualitative findings were used to complement and contextualize the quantitative student-experience results rather than to develop an independent explanatory theory.

Human–human and human–AI scoring agreement was examined using intraclass correlation coefficients (ICCs) based on a two-way mixed-effects model with an absolute-agreement definition, with 95% confidence intervals reported for the estimates. Consistent with recommendations for transparent ICC reporting, we explicitly specified the model, type, and definition of the ICC [51]. Human–human agreement was examined first using the independent scores assigned by the two human raters. After establishing the consistency of the human ratings, the two scores were averaged for each response to create the Human Mean, which served as the human-reference score for comparison with the AI-generated score. For the AI–human comparison, the single-measure absolute-agreement ICC was treated as the primary estimate of correspondence between the AI score and the human-

reference score, while the average-measure ICC was reported as complementary information. Mean Difference (MD), calculated as AI Score − Human Mean, was used to describe the direction of score differences, whereas Mean Absolute Difference (MAD) summarized the average magnitude of disagreement irrespective of direction. The scoring-agreement analysis compared AI-generated and human-reference scores for the same student responses under identical assessment criteria. Because the sample included both essay-based and multiple-choice assessments, the resulting agreement estimates were interpreted as overall estimates for the combined scoring sample rather than separate estimates for each question format. All quantitative analyses were conducted using IBM SPSS Statistics.

4. RESULTS AND DISCUSSION

4.1. Expert Validation of the Developed Model

Expert validation provided generally positive content-validity evidence for the proposed HA-CAAM. Across the 25 validation items, Aiken's V coefficients ranged from 0.70 to 0.90, with an overall mean of 0.818. Eighteen items (72%) achieved coefficients of 0.80 or higher, while the remaining seven items (28%) ranged from 0.70 to 0.75; no item yielded a coefficient below 0.70. At the aspect level, the descriptive means of item-level Aiken's V coefficients ranged from 0.750 to 0.900. Guidebook Quality had the highest descriptive aspect mean (0.900), followed by Ethics, Privacy, and AI Use (0.850), while Model Foundation and Rationale and Reporting and Follow-up each had the lowest descriptive aspect mean (0.750). Overall, the pattern indicated generally positive expert judgments while also identifying specific components requiring further operational refinement.

Table 1. Expert Validation Results of the Developed Model

No.	Validation Model Aspect and Indicator	Aiken's V	Aspect Mean
Model Foundation and Rationale			
1	Clarity of model purpose, rationale, and relevance to AI-based learning assessment	0.70	0.750
2	Conceptual support for integrating adaptive assessment, NLP, and AI-driven feedback	0.80	
Model Constructs and Structure			
3	Clarity and functional appropriateness of model components	0.75	0.783
4	Logical, systematic, and integrated relationships among model components	0.90	
5	Flexibility of model structure across AI-supported courses	0.70	
Adaptive Mechanism			
6	Clarity of the mechanism for determining task difficulty	0.90	0.833
7	Appropriateness of difficulty-level transitions based on learner performance	0.90	
8	Appropriateness of adaptive challenge according to learner performance	0.70	
NLP Mechanism			
9	Appropriateness of NLP for analyzing students' constructed responses	0.85	0.833
10	Clarity of rubric- and prompt-guided NLP analysis	0.85	
11	Flexibility of NLP criteria/rubrics across course competencies	0.80	
AI-Driven Feedback			
12	Relevance of AI-generated feedback to students' responses and assessment results	0.90	0.800
13	Personalization of strengths, weaknesses, and improvement suggestions	0.75	

No.	Validation Model Aspect and Indicator	Aiken's V	Aspect Mean
14	Appropriate positioning of AI feedback within formative assessment and instructor oversight	0.75	
Assessment Validity and Reliability			
15	Alignment of assessment results with the competencies being measured	0.90	0.850
16	Mechanism for maintaining consistency and reliability of AI-assisted scoring	0.80	
17	Incorporation of fairness and human verification in assessment quality assurance	0.85	
Syntax and Implementation			
18	Clarity, logic, and usability of model implementation stages	0.80	0.833
19	Clarity of instructor and AI roles throughout implementation	0.80	
20	Feasibility and flexibility of implementation across different courses	0.90	
Reporting and Follow-up			
21	Informativeness of individual- and class-level assessment reporting	0.80	0.750
22	Usefulness of assessment results and feedback for instructional follow-up	0.70	
Ethics, Privacy, and AI Use			
23	Adequacy of AI ethics, data protection, transparency, and instructor oversight	0.85	0.850
Guidebook Quality			
24	Systematic and comprehensible presentation of model concepts and procedures	0.90	0.900
25	Adequacy of the guidebook for implementing the model across course contexts	0.90	
Overall Mean Aiken's V			0.818

The quantitative validation results indicated generally positive expert judgments regarding the relevance of HA-CAAM components. The comparatively lower descriptive aspect means for Model Foundation and Rationale and Reporting and Follow-up (both 0.750) highlighted areas requiring further attention. The final model specification emphasizes explicit adaptive decision rules, structured AI prompting, rubric-guided scoring, instructor oversight, formative feedback, and instructional follow-up. It also incorporates provisions for privacy, fairness, transparency, and accountable AI use. Accordingly, the validation findings are interpreted as evidence of expert-judged content relevance, while broader procedures such as item/task-bank calibration and cross-group measurement evaluation remain priorities for future validation.

4.2. Internal Consistency of the Student Experience Instrument

Before examining students' experiences with the implementation of HA-CAAM, we examined the internal consistency of the student-experience instrument separately for each of its eight dimensions. Cronbach's alpha coefficients and corrected item–total correlations were used to evaluate the consistency of the items within their respective dimensions. The results are presented in Table 2.

Table 2. Internal Consistency of the Student Experience Instrument

Dimension	No. of Items	Cronbach's α	Corrected Item–Total Correlation Range
Clarity and Ease of Use	4	0.864	0.626–0.777
Perceived Adaptivity	4	0.791	0.520–0.686
Perceived Assessment Relevance	4	0.850	0.622–0.726
AI Feedback Quality	5	0.883	0.581–0.799
Feedback Usefulness and Improvement	4	0.782	0.494–0.647
Fairness and Trust	4	0.846	0.618–0.722
Learning Support	4	0.837	0.591–0.730
Overall User Experience	4	0.831	0.634–0.682

As shown in Table 2, Cronbach's alpha coefficients ranged from 0.782 to 0.883, indicating satisfactory internal consistency across all eight dimensions. AI Feedback Quality demonstrated the highest internal consistency (α = 0.883), followed by Clarity and Ease of Use (α = 0.864) and Perceived Assessment Relevance (α = 0.850). Corrected item–total correlations ranged from 0.494 to 0.799, with all items showing adequate relationships with their respective dimension scores. Accordingly, all 33 items were retained for the subsequent analysis of students' experiences with HA-CAAM.

4.3. Student Experience of Model Implementation

Students' experiences with HA-CAAM implementation were examined across eight dimensions using mean composite scores derived from the corresponding questionnaire items. Descriptive statistics, including means, standard deviations, and 95% confidence intervals, were calculated for each dimension. The results are presented in Table 3.

Table 3. Descriptive Statistics of Students' Experience with the Implemented Model

Dimension	M	SD	95% CI
Clarity and Ease of Use	4.132	0.611	[4.033, 4.231]
Perceived Adaptivity	3.905	0.600	[3.808, 4.003]
Perceived Assessment Relevance	3.965	0.621	[3.864, 4.065]
AI Feedback Quality	3.943	0.599	[3.846, 4.041]
Feedback Usefulness and Improvement	4.068	0.565	[3.976, 4.159]
Fairness and Trust	4.025	0.618	[3.925, 4.126]
Learning Support	3.765	0.634	[3.662, 3.868]
Overall User Experience	3.789	0.634	[3.686, 3.892]

As shown in Table 3, mean scores across the eight dimensions ranged from 3.765 to 4.132 on the five-point response scale. Clarity and Ease of Use received the highest mean score (M = 4.132, SD = 0.611, 95% CI [4.033, 4.231]), followed by Feedback Usefulness and Improvement (M = 4.068, SD = 0.565, 95% CI [3.976, 4.159]) and Fairness and Trust (M = 4.025, SD = 0.618, 95% CI [3.925, 4.126]). This descriptive pattern indicates that clarity and usability, feedback usefulness, and fairness and trust were among the most positively perceived aspects of HA-CAAM implementation.

Comparatively lower mean scores were observed for Learning Support (M = 3.765, SD = 0.634, 95% CI [3.662, 3.868]) and Overall User Experience (M = 3.789, SD = 0.634, 95% CI [3.686, 3.892]), although both remained above the midpoint of the response scale. Perceived Adaptivity (M = 3.905, SD = 0.600), AI Feedback Quality (M = 3.943, SD = 0.599), and Perceived Assessment Relevance (M = 3.965, SD = 0.621) occupied intermediate positions. Taken together, the findings indicate generally positive student experiences across the eight dimensions. However, the comparatively lower ratings for Learning Support and Overall User Experience suggest that these aspects warrant further attention in subsequent implementation and refinement. These descriptive differences do not establish the underlying reasons for the lower ratings.

4.4. Students' Qualitative Experiences

Analysis of the open-ended responses identified five recurring themes concerning students' experiences with the implemented assessment model. The most frequently identified theme was immediate and actionable feedback (60.0%), indicating that students particularly valued receiving feedback shortly after completing the assessment. One student noted, “I was helped because I received feedback immediately after answering. Usually, I have to wait a long time to know my mistakes” (M33). Learning and self-regulation support (50.0%) was also prominent, with students describing the assessment as supporting learning beyond earning a score. As one participant explained, “I feel this model is more useful for learning than merely for obtaining a grade” (M103).

Other responses revealed more nuanced experiences. Trust, human oversight, and AI limitations (33.3%) reflected concerns that AI might misinterpret responses, particularly when students used less formal language, alongside expectations that instructors should remain involved in reviewing AI-supported assessment. Personalized and adaptive assessment experiences (30.0%) indicated that students generally recognized task adaptation, although some were uncertain about why particular questions were assigned to them. Finally, usability and improvement needs (30.0%) highlighted requests for more specific feedback and clearer recommendations for subsequent learning activities. Overall, the qualitative findings suggest that students valued the model primarily for its immediacy and learning-oriented feedback, while also identifying areas for further refinement concerning feedback specificity, transparency of adaptive decisions, and human oversight of AI-supported assessment.

4.5. Human–AI Scoring Agreement

To establish a human reference for evaluating AI-generated assessment scores, two human raters independently assessed the same 50 student responses evaluated by GPT-4o, comprising both essay-based and multiple-choice assessments. All assessments were scored on a common 1–100 scale, using identical rubrics and answer keys for AI and human scoring. The human raters were blinded to the AI-generated scores. Human–human agreement was first examined using a two-way mixed-effects intraclass correlation coefficient (ICC) with an absolute-agreement definition. The single-measure ICC was 0.892 (95% CI [0.817, 0.937]), while the average-measures ICC was 0.943 (95% CI [0.899, 0.968]), indicating strong agreement between the two human raters. Given the strong agreement between the human raters, the mean score of the two raters was subsequently used as the Human Mean, which served as the human-reference score for each response.

Agreement between the AI-generated scores and the Human Mean was subsequently examined using the same two-way mixed-effects, absolute-agreement ICC approach. The single-measure ICC was 0.898 (95% CI [0.828, 0.941]), while the average-measures ICC was 0.947 (95% CI [0.906, 0.970]). The single-measure coefficient was treated as the primary indicator of AI–human scoring correspondence because the analysis concerned agreement between an individual AI-generated score and the corresponding human-reference score, while the average-measures coefficient was reported as complementary information. The resulting single-measure coefficient indicates a high degree of absolute

agreement between AI-generated and human-reference scores across the 50 evaluated responses. Table 4 summarizes the human–human and AI–human agreement estimates, along with descriptive indicators of score differences.

Table 4. Human–Human and AI–Human Scoring Agreement

Comparison / Agreement Indicator	Estimate	95% CI
Human–Human ICC, single measures	0.892	[0.817, 0.937]
Human–Human ICC, average measures	0.943	[0.899, 0.968]
AI–Human ICC, single measures	0.898	[0.828, 0.941]
AI–Human ICC, average measures	0.947	[0.906, 0.970]
Mean difference (AI − Human)	−0.150	—
Mean absolute difference	1.850	—
Observed difference range	−5.00 to 4.00	—
Maximum absolute difference	5.00	—

As shown in Table 4, the Human Mean was calculated as the mean of the scores independently assigned by the two human raters and served as the human-reference score. ICCs were estimated using a two-way mixed-effects model with an absolute-agreement definition. The mean difference was calculated as AI score minus Human Mean; therefore, negative values indicate lower AI-generated scores relative to the human-reference scores.

To further characterize score discrepancies beyond the ICC, we examined directional and absolute differences between AI-generated and human-reference scores. The mean difference was −0.150 points (SD = 2.207), with observed differences ranging from −5.00 to 4.00 points. Thus, the AI scores showed only a small average directional difference relative to the human reference, with the negative sign indicating slightly lower AI-generated scores on average. The mean absolute difference was 1.850 points (SD = 1.183), and the largest observed absolute difference was 5.00 points.

Taken together, the high absolute-agreement ICC and the relatively small average directional difference indicate close scoring correspondence between AI-generated scores and the human reference within the evaluated response subset, although individual score discrepancies remained. These findings are interpreted as evidence of scoring agreement under the specific rubric, prompting procedure, assessment tasks, and implementation conditions used in this study, not as evidence that AI-generated scoring is equivalent to, or can replace, human judgment. Within HA-CAAM, these findings support using AI-generated scores as supplementary assessment evidence under instructor supervision rather than as replacements for human judgment. Instructors reviewed all AI-generated assessment outputs for potential errors or inconsistencies, with more detailed verification directed toward flagged cases, while retaining final responsibility for assessment decisions.

4.6. Final Refined AI-Supported Adaptive Assessment Model

The final model was refined by integrating evidence from expert validation, student implementation, qualitative feedback, and human–AI scoring evaluation. The refinement process primarily strengthened the model's operational specification rather than altering its underlying assessment logic. Particular attention was given to performance-responsive adaptive decision procedures, structured AI/NLP-assisted analysis and scoring, human

oversight, personalized and actionable formative feedback, iterative reassessment, and ethical and governance safeguards. The resulting Human–AI Collaborative Adaptive Assessment Model (HA-CAAM) comprises a seven-stage adaptive assessment cycle supported by an AI–human collaboration layer, key implementation enablers, cross-cutting ethical, legal, and trust safeguards, and a continuous improvement mechanism, as illustrated in Figure 2. The architecture is guided by six overarching principles: learning-centered assessment, fairness and equity, transparency and explainability, human oversight and accountability, privacy and data protection, and continuous improvement. The final architecture includes procedures examined during classroom implementation alongside broader design provisions intended to guide subsequent application and validation.

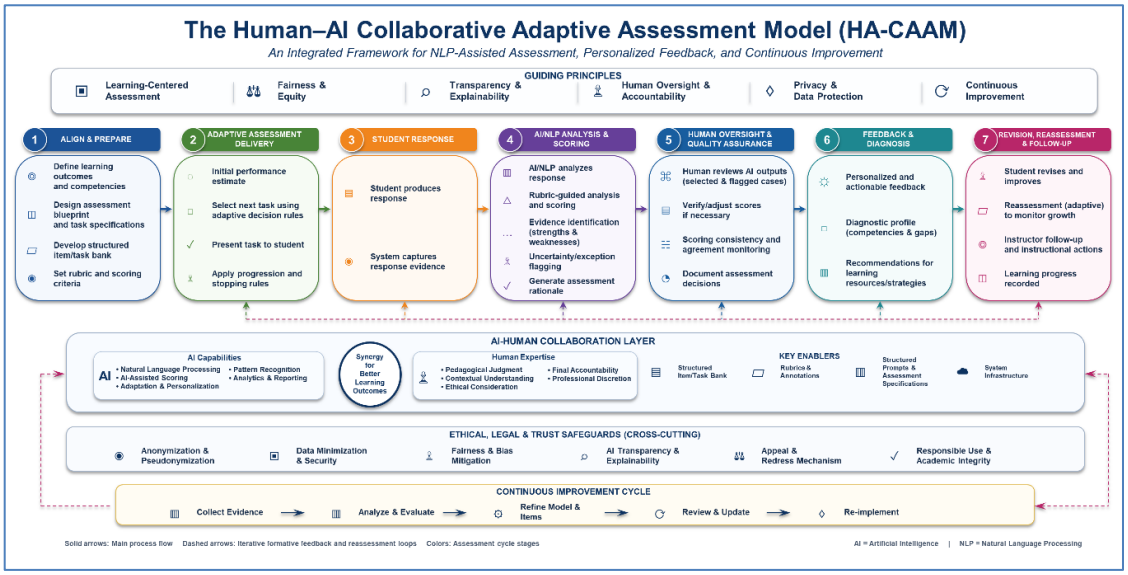


Figure 2. Final Refined Human–AI Collaborative Adaptive Assessment Model (HA-CAAM)

As shown in Figure 2, the seven-stage adaptive assessment cycle begins with (1) Align and Prepare, in which learning outcomes and competencies are defined and aligned with the assessment blueprint, item/task specifications, a structured item/task bank, and scoring criteria and rubrics. (2) Adaptive Assessment Delivery begins with an initial performance estimate and uses predefined adaptive decision rules to select and present subsequent tasks and determine progression or stopping based on observed performance. In (3) Student Response, students produce responses that are captured by the system as assessment evidence. (4) AI/NLP Analysis and Scoring uses rubric-guided AI-assisted analysis to evaluate student responses, generate scores, identify strengths and weaknesses, provide formative feedback, and flag cases requiring closer verification. (5) Human Oversight and Quality Assurance involves reviewing all AI-generated assessment outputs for potential errors or inconsistencies, with more detailed verification of flagged cases. Instructors may verify or adjust scores when necessary and retain final responsibility for assessment decisions. (6) Feedback and Diagnosis translates assessment evidence into personalized and actionable feedback, diagnostic profiles of competencies and learning gaps, and recommendations for learning resources or strategies. Finally, (7) Revision, Reassessment, and Follow-up enables students to revise and improve their work, engage in

adaptive reassessment to monitor growth, receive instructor follow-up and instructional actions, and have their learning progress recorded.

The seven-stage cycle is supported by an AI–Human Collaboration Layer that combines complementary technological capabilities and human expertise. AI contributes natural language processing, AI-assisted scoring, adaptation and personalization, pattern recognition, and analytics and reporting, whereas human expertise contributes pedagogical judgment, contextual understanding, ethical consideration, professional discretion, and final accountability. This collaboration is supported by four Key Enablers: a structured item/task bank, rubrics and annotations, structured prompts and assessment specifications, and system infrastructure. Within this architecture, AI functions as an analytic, scoring, adaptation, and feedback-support mechanism rather than as an autonomous assessor, while human judgment remains central to contextual interpretation and final assessment decisions.

The Ethical, Legal, and Trust Safeguards operate across the entire assessment process rather than as a separate sequential stage. These cross-cutting safeguards comprise anonymization and pseudonymization, data minimization and security, fairness and bias mitigation, AI transparency and explainability, appeal and redress mechanisms, and responsible use and academic integrity. The model also incorporates a Continuous Improvement Cycle in which evidence from implementation is collected, analyzed, and evaluated to inform refinement of the model and assessment items, followed by review and updating before subsequent reimplementation. These safeguards and continuous-improvement mechanisms represent design provisions incorporated through the refinement process and should not be interpreted as evidence that each individual safeguard or mechanism was independently validated during the present implementation.

Taken together, the final HA-CAAM positions adaptive assessment as an iterative, learning-oriented, and human-supervised process in which assessment evidence informs subsequent task adaptation, diagnosis, feedback, revision, reassessment, and instructional decision-making. The model integrates AI-supported analytical and adaptive capabilities with human assessment authority while maintaining explicit implementation enablers, cross-cutting safeguards, and mechanisms for continuous refinement.

4.7. Discussion

The findings provide generally positive evidence for the conceptual and operational quality of the developed HA-CAAM, while also showing that its initial architecture required further operational specification. Expert validation yielded an overall mean Aiken's V of 0.818, with most items reaching 0.80 or higher, whereas expert feedback primarily called for clearer adaptive decision rules, more structured AI-assisted scoring, explicit human verification, iterative reassessment, and stronger governance safeguards. These findings indicate that the central challenge was not simply whether AI could be incorporated into assessment, but how AI, adaptive decision-making, pedagogical design, and human judgment could be coordinated within an accountable assessment process [4], [19], [52], [53].

Previous research provides complementary but distinct foundations for AI-supported assessment. Xia et al. [4] and broader systematic evidence [52] highlight the opportunities

and challenges of generative AI for assessment, while Ilieva et al. [20] emphasize how assessment responsibilities are distributed among instructors, students, and institutional authorities. Fajardo-Ramos et al. [19] foreground human verification, transparency, data stewardship, and accountability within human-in-the-loop assessment. These approaches establish important principles for responsible AI integration, but their primary emphases differ from HA-CAAM's operational coordination of performance-responsive task progression, rubric-guided AI assessment, instructor verification, formative feedback, and subsequent reassessment. HA-CAAM therefore extends these contributions by connecting assessment functions and human responsibilities within a single iterative pedagogical process, rather than treating technological capabilities or governance requirements as separate considerations. This interpretation is consistent with broader evidence emphasizing the importance of pedagogical and contextual design in higher-education AI implementation [18], [53]. The expert-driven refinements in the present study can therefore be understood as a transition from a technologically feasible model toward a more pedagogically explicit and accountable assessment architecture, in which assessment alignment, adaptive decisions, AI functionality, and human responsibility are specified before implementation.

Students' experiences further suggest that the educational value of HA-CAAM lies primarily in connecting assessment evidence with timely and usable learning support. Clarity and Ease of Use received the highest rating (M = 4.132), followed by Feedback Usefulness and Improvement (M = 4.068) and Fairness and Trust (M = 4.025); qualitatively, immediate and actionable feedback was the most frequently identified theme (60.0%), while learning and self-regulation support appeared in 50.0% of responses. Together, these findings indicate that AI-supported adaptivity becomes pedagogically meaningful when evidence of student performance is translated rapidly into feedback that can inform subsequent learning action [12], [36], [54], [55]. This interpretation is consistent with systematic evidence showing that AI-assisted feedback can support task-, process-, and self-regulation-oriented learning processes [36], [55], while empirical studies indicate that LLM-generated feedback can provide timely, personalized, and potentially useful support for subsequent learning and revision [12], [54]. Qian [56] similarly identified immediate assessment and feedback as prominent pedagogical applications of GenAI in higher education, while Nazaretsky et al. [57] found AI-generated and human-authored formative feedback comparable in overall pedagogical quality, although limitations remained in metacognitive support.

However, Learning Support (M = 3.765) and Overall User Experience (M = 3.789) received comparatively lower ratings, despite the relatively positive evaluation of Feedback Usefulness and Improvement (M = 4.068). This contrast suggests that students may distinguish between receiving useful feedback and getting enough guidance to act on it during subsequent learning activities. Their requests for more specific explanations, clearer recommendations for subsequent learning, and greater transparency about adaptive task selection may explain this pattern, although the available descriptive and qualitative evidence does not establish its underlying causes. This interpretation is consistent with research indicating that students value detailed, constructive, and context-sensitive AI feedback, while unmet expectations may limit their engagement with AI-supported learning processes [58], [59]. Importantly, these favorable student experiences should not be

interpreted as evidence that HA-CAAM improved learning achievement or other educational outcomes, because such effects were not directly evaluated in the present study. Thus, immediate feedback should not be equated with pedagogically complete feedback; its value depends on whether students can understand, evaluate, and act on it in subsequent learning [24], [60], [61]. This explains why HA-CAAM explicitly links diagnosis and personalized feedback with revision, reassessment, and instructor follow-up rather than treating feedback as the endpoint of assessment.

The scoring results further clarify AI's appropriate role within HA-CAAM. The two independent human raters first showed strong agreement (ICC = 0.892), and AI-generated scores showed similarly strong agreement with the Human Mean reference (ICC = 0.898). The small mean difference (MD = −0.150) indicates little directional scoring difference, whereas the mean absolute difference (MAD = 1.850) confirms that discrepancies remained at the individual-response level. The findings therefore support close human–AI scoring correspondence under the implemented rubric-guided conditions, but not equivalence between AI and human judgment. The distinction has practical implications for assessment decisions with different levels of consequence. In low-stakes formative assessment, relatively small scoring discrepancies may be manageable when AI-generated evaluations are used primarily to guide feedback and subsequent learning activities under instructor supervision. In high-stakes assessment, however, even limited individual discrepancies may affect consequential decisions, particularly when scores fall near grading or progression boundaries. Therefore, the observed agreement should not be interpreted as sufficient evidence for autonomous AI scoring in high-stakes contexts. The present findings instead support the continued verification of AI-generated scores and the retention of final assessment authority by human instructors.

This distinction matters because evidence increasingly shows that LLM scoring performance varies by assessment tasks, scoring criteria, rubric design, prompting procedures, and model characteristics [31], [37], [62], [63]. Flodén [62], for example, found that although ChatGPT-generated university examination scores often closely matched teacher scores, substantial discrepancies remained for individual responses. At the same time, Wang et al. [63] showed that human–LLM discrepancies can vary across scoring contexts and that agreement in relative ranking does not necessarily imply reliability for absolute judgments. Similarly, Usher [64] and Usher and Faraon [65] demonstrated variability in AI–instructor grading according to assessment conditions and student-work quality, whereas rubric-guided and structured approaches have produced stronger human–AI correspondence in other studies [14], [31], [37]. The comparatively close correspondence observed in the present study may therefore be partly attributable to a common scoring rubric, structured AI prompts, and a human reference derived from two independent raters. Importantly, 33.3% of students' qualitative responses also raised concerns related to trust, human oversight, or AI limitations. Taken together with evidence emphasizing the need for human verification, transparency, and accountability in AI-assisted assessment [18], [19], this convergence suggests treating human oversight as a structural quality-assurance mechanism within HA-CAAM, not a temporary measure until AI becomes sufficiently accurate to replace human assessors.

Taken together, evidence from expert validation, student experience, qualitative feedback, and human–AI scoring informed the final specification of HA-CAAM as a seven-stage adaptive assessment cycle supported by an AI–human collaboration layer, key implementation enablers, cross-cutting ethical, legal, and trust safeguards, and a continuous improvement mechanism. HA-CAAM contributes not by introducing adaptive assessment, AI/NLP-assisted analysis, personalized feedback, or human oversight individually, but by pedagogically orchestrating these functions as an interconnected, iterative assessment architecture. This responds to a literature in which important components of AI-supported assessment have often been advanced through complementary but partially separated strands, including GenAI-enabled assessment functions and challenges [4], [52], adaptive and performance-responsive assessment processes [29], AI-supported scoring and feedback [20], [37], and human-in-the-loop assessment and accountability [19]. Recent frameworks similarly emphasize that assessment in the GenAI era should be understood as a socio-technical and human–AI collaborative process in which technological affordances are coordinated with pedagogical judgment, student learning, and governance rather than treated as substitutes for them [66], [67], [68].

HA-CAAM operationalizes this integration through a closed formative loop connecting performance-responsive assessment, student response, rubric-guided AI/NLP analysis, human quality assurance, diagnosis and personalized feedback, and revision and reassessment. In contrast, AI–human collaboration, implementation enablers, responsible-AI safeguards, and continuous improvement support the process overall. The adaptive component of this cycle was operationalized through three predefined difficulty levels and pedagogical performance thresholds: scores of 80% or higher prompted progression to a more challenging task, scores between 50% and 79% maintained the current difficulty level, and scores below 50% prompted movement to an easier task. Unlike psychometrically calibrated computerized adaptive testing, these thresholds were intended to support instructional decisions rather than estimate latent ability or optimize measurement precision. During implementation, students received tasks at different difficulty levels based on their performance, providing evidence that performance-responsive task selection was enacted in the classroom.

However, because systematic individual adaptation-path records were not retained, the study cannot establish the frequency, consistency, or longer-term educational consequences of particular adaptive pathways. Thus, the findings support the feasibility of implementing predefined adaptive rules but do not demonstrate the psychometric effectiveness or learning impact of the adaptation mechanism. Accordingly, HA-CAAM positions AI not as an autonomous assessor but as one component within a configurable, human-supervised adaptive assessment ecosystem in which technological responsiveness is deliberately coordinated with formative learning processes, pedagogical judgment, and accountable assessment practice.

From an Assessment for Learning perspective, HA-CAAM contributes a structured process for translating assessment evidence into feedback, instructional decisions, and subsequent assessment opportunities rather than treating scores as terminal outcomes. Its human-centered contribution lies in positioning instructors as active interpreters and

decision-makers throughout the assessment cycle, while AI supports analysis, feedback generation, and performance-responsive task recommendations. These contributions are primarily conceptual and design-oriented, supported by initial implementation evidence; their effects on learning outcomes and broader assessment quality require further empirical evaluation.

4.8. Implications and Contributions

This study contributes conceptually by positioning AI-supported adaptive assessment as a human-supervised assessment ecosystem rather than a technology-centered process of automated task selection, scoring, or feedback. More specifically, HA-CAAM integrates a seven-stage adaptive assessment cycle with an AI–human collaboration layer, key implementation enablers, cross-cutting ethical, legal, and trust safeguards, and a continuous improvement mechanism. Empirically, converging evidence from expert validation, quantitative and qualitative student experiences, and human–AI scoring agreement supports this conceptualization. Across these evidence sources, a consistent pattern emerged: AI can support performance-responsive assessment, rubric-guided analysis, and timely formative feedback, but its educational value depends on explicit assessment criteria, human oversight, actionable feedback, and opportunities for revision and reassessment. HA-CAAM's contribution therefore lies not in introducing these capabilities individually, but in pedagogically orchestrating them within an empirically refined, iterative, human-supervised assessment architecture.

Practically, HA-CAAM gives educators a configurable pathway to integrate generative AI into assessment while retaining human assessment authority. Implementation begins with explicit learning outcomes and competencies, assessment specifications, a structured item/task bank, and scoring rubrics; these foundations support performance-responsive task progression, rubric-guided AI/NLP analysis and scoring, diagnosis and personalized feedback, human quality assurance, and subsequent revision and reassessment. At the institutional level, the findings imply that AI policies should address not only students' use of generative AI but also educators' and institutions' use of AI in evaluating student work. Such implementation requires explicit provisions for privacy and data protection, fairness and bias mitigation, transparency and explainability, human accountability, responsible use, and student review or appeal mechanisms. These requirements align with emerging evidence that responsible AI-assisted assessment requires governance to develop alongside pedagogical design and human oversight, rather than being treated as a separate compliance concern [19], [66], [69]. Accordingly, responsible adoption of HA-CAAM requires technological responsiveness and scalability to remain subordinate to pedagogical judgment, learning-oriented assessment, and institutional accountability.

4.9. Limitations and Future Research

Several limitations should be considered when interpreting the findings and provide directions for further research. First, HA-CAAM was implemented within a single higher-education course and institutional context; therefore, although the model was designed to be configurable, its transferability across disciplines and institutions requires further empirical

validation. Future studies should implement HA-CAAM in diverse disciplinary and institutional settings to examine whether its adaptive procedures, feedback mechanisms, human-oversight processes, and governance safeguards remain applicable across contexts. Second, the present evaluation focused on content-validity evidence, student experience, qualitative feedback, and human–AI scoring agreement rather than causal improvements in learning outcomes. Experimental or quasi-experimental studies are therefore needed to determine whether HA-CAAM improves learning performance, feedback uptake, self-regulated learning, or other educational outcomes compared with conventional assessment approaches. Third, the human–AI agreement findings were obtained under specific assessment tasks, rubrics, prompting procedures, and implementation conditions and should not be generalized to other AI systems or assessment contexts without further testing. Future research should examine the robustness of scoring and feedback across AI models, prompt configurations, task types, and disciplinary contexts. Finally, subsequent validation should strengthen the psychometric and responsible-AI dimensions of HA-CAAM through empirical item/task calibration, cross-group measurement evaluation, longitudinal monitoring of adaptive progression, and systematic examination of fairness and potential bias in AI-supported assessment decisions.

5. CONCLUSION

This study developed, validated, implemented, evaluated, and refined the Human–AI Collaborative Adaptive Assessment Model (HA-CAAM) as a configurable approach to AI-supported adaptive assessment in higher education. Addressing the three research questions, the findings demonstrate that: (1) expert validation indicated positive content relevance (mean Aiken's V = 0.818), with validator recommendations informing refinements to the model's conceptual and operational components; (2) students reported generally positive implementation experiences, particularly regarding feedback usefulness, although Learning Support and Overall User Experience received comparatively lower ratings; and (3) rubric-guided AI-generated scores demonstrated strong agreement with human ratings (ICC = 0.898), supporting their potential use as supplementary assessment evidence under instructor supervision. These findings provide initial evidence of content validity, positive implementation experiences, and human–AI scoring agreement, but do not establish improvements in learning outcomes or equivalence between AI-generated and human assessment decisions.

The resulting HA-CAAM integrates performance-responsive task progression, rubric-guided AI/NLP analysis and scoring, instructor-led quality assurance, personalized formative feedback, and iterative reassessment within a seven-stage adaptive assessment cycle. Its principal contribution is coordinating these functions within a unified, learning-oriented assessment architecture that connects assessment evidence to subsequent instructional decisions. Rather than replacing human assessors, HA-CAAM positions AI as a supporting component of adaptive and formative assessment while maintaining human accountability, instructor authority over final assessment decisions, and responsible AI governance. Given its implementation within a single Media Education course and institutional context, further experimental evaluation, psychometric refinement of adaptive

decision rules, and cross-disciplinary validation are needed to establish the model's broader educational effectiveness and applicability. Overall, HA-CAAM offers a promising framework for advancing transparent, formative, and accountable AI-supported assessment practices in higher education.

6. ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support provided by the Institute for Research and Community Service (LPPM), Universitas Negeri Padang, Indonesia, under Research Contract No. 2022/UN35.15/LT/2025.

REFERENCES

[1] J. Luo, C. Zheng, J. Yin, and H. H. Teo, "Design and assessment of AI-based learning tools in higher education: a systematic review," *Int. J. Educ. Technol. High. Educ.*, vol. 22, no. 1, p. 42, 2025, doi: 10.1186/s41239-025-00540-2.

[2] H. Crompton and D. Burke, "Artificial intelligence in higher education: the state of the field," *Int. J. Educ. Technol. High. Educ.*, vol. 20, no. 1, p. 22, 2023, doi: 10.1186/s41239-023-00392-8.

[3] M. Bond *et al.*, "A meta systematic review of artificial intelligence in higher education: a call for increased ethics, collaboration, and rigour," *Int. J. Educ. Technol. High. Educ.*, vol. 21, no. 1, p. 4, 2024, doi: 10.1186/s41239-023-00436-z.

[4] Q. Xia, X. Weng, F. Ouyang, T. J. Lin, and T. K. F. Chiu, "A scoping review on how generative artificial intelligence transforms assessment in higher education," *Int. J. Educ. Technol. High. Educ.*, vol. 21, no. 1, p. 40, 2024, doi: 10.1186/s41239-024-00468-z.

[5] E. D. Lindsay, M. Zhang, A. Johri, and J. Bjerva, "The Responsible Development of Automated Student Feedback with Generative AI," in *2025 IEEE Global Engineering Education Conference (EDUCON)*, 2025, pp. 1–10. doi: 10.1109/EDUCON62633.2025.11016572.

[6] R. Gao, H. E. Merzdorf, S. Anwar, M. C. Hipwell, and A. R. Srinivasa, "Automatic assessment of text-based responses in post-secondary education: A systematic review," *Comput. Educ. Artif. Intell.*, vol. 6, p. 100206, 2024, doi: 10.1016/j.caeai.2024.100206.

[7] A. Botelho, S. Baral, J. A. Erickson, P. Benachamardi, and N. T. Heffernan, "Leveraging natural language processing to support automated assessment and feedback for student open responses in mathematics," *J. Comput. Assist. Learn.*, vol. 39, no. 3, pp. 823–840, Jun. 2023, doi: 10.1111/jcal.12793.

[8] M. Mejeh and M. Rehm, "Taking adaptive learning in educational settings to the next level: leveraging natural language processing for improved personalization," *Educ. Technol. Res. Dev.*, vol. 72, no. 3, pp. 1597–1621, 2024, doi: 10.1007/s11423-024-10345-1.

[9] U. Maier and C. Klotz, "Personalized feedback in digital learning environments: Classification framework and literature review," *Comput. Educ. Artif. Intell.*, vol. 3, p. 100080, 2022, doi: 10.1016/j.caeai.2022.100080.

[10] S. Gombert *et al.*, "From the Automated Assessment of Student Essay Content to Highly Informative Feedback: a Case Study," *Int. J. Artif. Intell. Educ.*, vol. 34, no. 4, pp. 1378–1416, 2024, doi: 10.1007/s40593-023-00387-6.

[11] H. Zhang *et al.*, "eRevise: Using natural language processing to provide formative feedback on text evidence usage in student writing," in *Proceedings of the AAAI conference on artificial intelligence*, 2019, pp. 9619–9625. doi: 10.1609/aaai.v33i01.33019619.

[12] W. Dai *et al.*, "Assessing the proficiency of large language models in automatic feedback generation: An evaluation study," *Comput. Educ. Artif. Intell.*, vol. 7, p. 100299, 2024, doi: 10.1016/j.caeai.2024.100299.

[13] J. Escalante, A. Pack, and A. Barrett, "AI-generated feedback on writing: Insights into efficacy and ENL student preference," *Int. J. Educ. Technol. High. Educ.*, vol. 20, no. 1, p. 57, 2023, doi: 10.1186/s41239-023-00425-2.

[14] P. Y. Win Myint, S. L. Lo, and Y. Zhang, "Harnessing the power of AI-instructor collaborative grading approach: Topic-based effective grading for semi open-ended multipart questions," *Comput. Educ. Artif. Intell.*, vol. 7, p. 100339, 2024, doi: 10.1016/j.caeai.2024.100339.

[15] T. Heinrich, S. Baily, K.-W. Chen, J. DeOliveira, S. Park, and N. C.-H. Wang, "AI-assisted grading and personalized feedback in large political science classes: Results from randomized controlled

trials," *PLoS One*, vol. 20, no. 8, p. e0328041, 2025, doi: 10.1371/journal.pone.0328041.

[16] D. Wu, T.-T. Goh, D. Chen, and H. Yin, "Enhancing graduate students' evaluative skills through AI-supported collaborative marking," *Front. Psychol.*, vol. 17, no. March, pp. 1–24, 2026, doi: 10.3389/fpsyg.2026.1739222.

[17] D.-W. Zhang, M. Boey, Y. Y. Tan, and A. H. S. Jia, "Evaluating large language models for criterion-based grading from agreement to consistency," *npj Sci. Learn.*, vol. 9, no. 1, p. 79, 2024, doi: 10.1038/s41539-024-00291-1.

[18] R. Sembey, R. Hoda, and J. Grundy, "Emerging technologies in higher education assessment and feedback practices: A systematic literature review," *J. Syst. Softw.*, vol. 211, p. 111988, 2024, doi: 10.1016/j.jss.2024.111988.

[19] D. C. Fajardo-Ramos, A. Chiappe, and J. Mella-Norambuena, "Human-in-the-loop assessment with AI: implications for teacher education in Ibero-American universities," *Front. Educ.*, vol. 10, no. October, pp. 1–15, 2025, doi: 10.3389/feduc.2025.1710992.

[20] G. Ilieva, T. Yankova, M. Ruseva, and S. Kabaivanov, "A Framework for Generative AI-Driven Assessment in Higher Education," 2025. doi: 10.3390/info16060472.

[21] P. Black and D. Wiliam, "Developing the theory of formative assessment," *Educ. Assessment, Eval. Account. (formerly J. Pers. Eval. Educ.*, vol. 21, pp. 5–31, 2009, doi: 10.1007/s11092-008-9068-5.

[22] E. G. Offerdahl, M. McConnell, and J. Boyer, "Can I have your recipe? Using a fidelity of implementation (FOI) framework to identify the key ingredients of formative assessment for learning," *CBE—Life Sci. Educ.*, vol. 17, no. 4, p. es16, 2018, doi: 10.1187/cbe.18-02-0029.

[23] C. J. Watling and S. Ginsburg, "Assessment, feedback and the alchemy of learning," *Med. Educ.*, vol. 53, no. 1, pp. 76–85, 2019, doi: 10.1111/medu.13645.

[24] D. Carless and D. Boud, "The development of student feedback literacy: enabling uptake of feedback," *Assess. Eval. High. Educ.*, vol. 43, no. 8, pp. 1315–1325, Nov. 2018, doi: 10.1080/02602938.2018.1463354.

[25] E. Molloy, D. Boud, and M. Henderson, "Developing a learning-centred framework for feedback literacy," *Assess. Eval. High. Educ.*, vol. 45, no. 4, pp. 527–540, May 2020, doi: 10.1080/02602938.2019.1667955.

[26] B. Memarian and T. Doleck, "A review of assessment for learning with artificial intelligence," *Comput. Hum. Behav. Artif. Humans*, vol. 2, no. 1, p. 100040, 2024, doi: 10.1016/j.chbah.2023.100040.

[27] P. Williams, "AI, analytics and a new assessment model for Universities," 2023. doi: 10.3390/educsci13101040.

[28] E. du Plooy, D. Casteleijn, and D. Franzsen, "Personalized adaptive learning in higher education: A scoping review of key characteristics and impact on academic performance and engagement," *Heliyon*, vol. 10, no. 21, p. e39630, 2024, doi: 10.1016/j.heliyon.2024.e39630.

[29] A. Hadyaoui and L. Cheniti-Belcadhi, "Scalable adaptive assessment framework for collaborative eLearning: a service-oriented AI approach," *Smart Learn. Environ.*, vol. 13, no. 1, p. 8, 2026, doi: 10.1186/s40561-025-00426-w.

[30] Y. Zhuang, R. Zhao, Z. Xie, and P. L. H. Yu, "Enhancing language learning through generative AI feedback on picture-cued writing tasks," *Comput. Educ. Artif. Intell.*, vol. 9, no. December, p. 100450, 2025, doi: 10.1016/j.caeai.2025.100450.

[31] W. Morris, L. Holmes, J. S. Choi, and S. Crossley, "Automated Scoring of Constructed Response Items in Math Assessment Using Large Language Models," *Int. J. Artif. Intell. Educ.*, vol. 35, no. 2, pp. 559–586, 2025, doi: 10.1007/s40593-024-00418-w.

[32] M. Xue, Y. Liu, X. Xiao, and M. Wilson, "Automatic prompt engineering for automatic scoring," *J. Educ. Meas.*, vol. 62, no. 4, pp. 559–587, Dec. 2025, doi: 10.1111/jedm.70002.

[33] T. Ryan, M. Henderson, K. Ryan, and G. Kennedy, "Designing learner-centred text-based feedback: a rapid review and qualitative synthesis," *Assess. Eval. High. Educ.*, vol. 46, no. 6, pp. 894–912, Aug. 2021, doi: 10.1080/02602938.2020.1828819.

[34] T. Wan and Z. Chen, "Exploring generative AI assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning," *Phys. Rev. Phys. Educ. Res.*, vol. 20, no. 1, p. 10152, 2024, doi: 10.1103/PhysRevPhysEducRes.20.010152.

[35] S. Pozdniakov *et al.*, "Large language models meet user interfaces: The case of provisioning feedback," *Comput. Educ. Artif. Intell.*, vol. 7, no. December, p. 100289, 2024, doi: 10.1016/j.caeai.2024.100289.

[36] S. Ba, L. Yang, Z. Yan, C. K. Looi, and D. Gašević, "Unraveling the mechanisms and effectiveness of AI-assisted feedback in education: A systematic literature review," *Comput. Educ. Open*, vol. 9, p. 100284, 2025, doi: 10.1016/j.caeo.2025.100284.

[37] E. M. AlGhamdi, Y. Li, D. Gašević, and G. Chen, "Leveraging prompt-based LLMs for automated scoring and feedback generation in higher education," *Comput. Educ.*, vol. 243, p. 105511, 2026, doi: 10.1016/j.compedu.2025.105511.

[38] H. Seo *et al.*, "Large Language Models as Evaluators in Education: Verification of Feedback Consistency and Accuracy," 2025. doi: 10.3390/app15020671.

[39] J. Steiss *et al.*, "Comparing the quality of human and ChatGPT feedback of students' writing," *Learn. Instr.*, vol. 91, no. June, p. 101894, 2024, doi: 10.1016/j.learninstruc.2024.101894.

[40] U. Maier, M. Seibold, and C. Klotz, "LLM-generated formative feedback in education: A qualitative systematic literature review," *Comput. Educ. Open*, vol. 10, p. 100374, 2026, doi: 10.1016/j.caeo.2026.100374.

[41] S. Rüdian, J. Podelo, J. Kužílek, and N. Pinkwart, "Feedback on feedback: Student's perceptions for feedback from teachers and few-shot LLMs," in *Proceedings of the 15th international learning analytics and knowledge conference*, 2025, pp. 82–92. doi: 10.1145/3706468.3706479.

[42] M. Bower, J. Torrington, J. W. M. Lai, P. Petocz, and M. Alfano, "How should we change teaching and assessment in response to increasingly powerful generative Artificial Intelligence? Outcomes of the ChatGPT teacher survey," *Educ. Inf. Technol.*, vol. 29, no. 12, pp. 15403–15439, 2024, doi: 10.1007/s10639-023-12405-0.

[43] R. Alfredo *et al.*, "Human-centred learning analytics and AI in education: A systematic literature review," *Comput. Educ. Artif. Intell.*, vol. 6, no. June, p. 100215, 2024, doi: 10.1016/j.caeai.2024.100215.

[44] C. Cohn, N. Hutchins, T. Le, and G. Biswas, "A chain-of-thought prompting approach with llms for evaluating students' formative assessment responses in science," in *Proceedings of the AAAI conference on artificial intelligence*, 2024, pp. 23182–23190. doi: 10.1609/aaai.v38i21.30364.

[45] R. L. Moore and S. S. Lee, "Harnessing generative AI (GenAI) for automated feedback in higher education: A systematic review," *Online Learn. J.*, vol. 28, no. 3, pp. 82–104, 2024, doi: 10.24059/olj.v28i3.4593.

[46] D. J. Nicol and D. Macfarlane-Dick, "Formative assessment and self-regulated learning: a model and seven principles of good feedback practice," *Stud. High. Educ.*, vol. 31, no. 2, pp. 199–218, Apr. 2006, doi: 10.1080/03075070600572090.

[47] S. Oppl, F. Reisinger, A. Eckmaier, and C. Helm, "A flexible online platform for computerized adaptive testing," *Int. J. Educ. Technol. High. Educ.*, vol. 14, no. 1, p. 2, 2017, doi: 10.1186/s41239-017-0039-0.

[48] R. C. Richey, *Design and development research: Methods, strategies, and issues*. New York: Routledge, 2007.

[49] R. M. Branch and İ. Varank, *Instructional design: The ADDIE approach*, vol. 722. Springer, 2009. doi: 10.1007/978-0-387-09506-6.

[50] S. D. Walter, M. Eliasziw, and A. Donner, "Sample size and optimal designs for reliability studies," *Stat. Med.*, vol. 17, no. 1, pp. 101–110, 1998, doi: 10.1002/(SICI)1097-0258(19980115)17:1%3C101::AID-SIM727%3E3.0.CO;2-E.

[51] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *J. Chiropr. Med.*, vol. 15, no. 2, pp. 155–163, 2016, doi: 10.1016/j.jcm.2016.02.012.

[52] J. Zhao, E. Chapman, and P. G. P. Sabet, "Generative AI and educational assessments: A systematic review," *Educ. Res. Perspect.*, vol. 51, pp. 124–155, Jan. 2024, [Online]. Available: https://search.informit.org/doi/10.3316/informit.T202506180000500592641426

[53] M. Belkina *et al.*, "Implementing Generative AI (GenAI) in Higher Education: A Systematic Review of Case Studies," *Comput. Educ. Artif. Intell.*, vol. 8, no. June, p. 100407, Apr. 2025, doi: 10.1016/j.caeai.2025.100407.

[54] A. Kinder *et al.*, "Effects of adaptive feedback generated by a large language model: A case study in teacher education," *Comput. Educ. Artif. Intell.*, vol. 8, no. June, p. 100349, 2025, doi: 10.1016/j.caeai.2024.100349.

[55] M. Lan and X. Zhou, "A qualitative systematic review on AI empowered self-regulated learning in higher education," *npj Sci. Learn.*, vol. 10, no. 1, p. 21, 2025, doi: 10.1038/s41539-025-00319-0.

[56] Y. Qian, "Pedagogical Applications of Generative AI in Higher Education: A Systematic Review of the Field," *TechTrends*, vol. 69, no. 5, pp. 1105–1120, 2025, doi: 10.1007/s11528-025-01100-1.

[57] T. Nazaretsky, H. Gabbay, and T. Käser, "Can students judge like experts? A large-scale study on the pedagogical quality of AI and human personalized formative feedback," *Comput. Educ. Artif. Intell.*, vol. 10, no. June, p. 100533, 2026, doi: 10.1016/j.caeai.2025.100533.

[58] Y. Öztürk and A. Çebi, "The potential of AI-generated feedback from the students' perspective: a systematic review," *Assess. Eval. High. Educ.*, vol. 51, no. 1, pp. 41–62, Jan. 2026, doi: 10.1080/02602938.2025.2588385.

[59] F. J.-Y. Jin *et al.*, "Students' perceptions of generative AI–powered learning analytics in the feedback

process: A feedback literacy perspective," *J. Learn. Anal.*, vol. 12, no. 1, pp. 152–168, 2025, doi: 10.18608/jla.2025.8609.

[60] S. Chen and A. C. K. Cheung, "Effect of generative artificial intelligence on university students learning outcomes: A systematic review and meta-analysis," *Educ. Res. Rev.*, vol. 49, no. November, p. 100737, 2025, doi: 10.1016/j.edurev.2025.100737.

[61] B. P. Solis Trujillo, D. Velarde-Camaqui, C. A. Gonzales Nuñez, E. V. Castillo Silva, and M. del P. de la Oliva, "The current landscape of formative assessment and feedback in graduate studies: a systematic literature review," *Front. Educ.*, vol. 10, no. May, pp. 1–11, 2025, doi: 10.3389/feduc.2025.1509983.

[62] J. Flodén, "Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT," *Br. Educ. Res. J.*, vol. 51, no. 1, pp. 201–224, Feb. 2025, doi: 10.1002/berj.4069.

[63] Y. Wang, J. Huang, L. Du, Y. Guo, Y. Liu, and R. Wang, "Evaluating large language models as raters in large-scale writing assessments: A psychometric framework for reliability and validity," *Comput. Educ. Artif. Intell.*, vol. 9, no. December, p. 100481, 2025, doi: 10.1016/j.caeai.2025.100481.

[64] M. Usher, "Generative AI vs. instructor vs. peer assessments: a comparison of grading and feedback in higher education," *Assess. Eval. High. Educ.*, vol. 50, no. 6, pp. 912–927, Aug. 2025, doi: 10.1080/02602938.2025.2487495.

[65] M. Usher and M. Faraon, "Who grades best? Comparing ChatGPT, peer, and instructor evaluations across varying levels of student project quality," *Assess. Eval. High. Educ.*, pp. 1–20, Nov. 2025, doi: 10.1080/02602938.2025.2588682.

[66] Beth L Chapman and Barney Dalgarno, "Beyond the ban: A theoretical framework for integrating Generative AI in assessment," *Policy Futur. Educ.*, vol. 24, no. 5, pp. 670–682, Sep. 2026, doi: 10.1177/14782103251411725.

[67] M. Perkins, J. Roe, and L. Furze, "Reimagining the Artificial Intelligence Assessment Scale: A refined framework for educational assessment," *J. Univ. Teach. Learn. Pract.*, vol. 22, no. 7, pp. 1–26, 2025, doi: 10.53761/rrm4y757.

[68] H. Bilgin, "A Framework for Responsible AI in Higher Education," *High. Educ. Gov. Policy*, vol. 6, no. 2, pp. 155–174, 2025, doi: 10.55993/hegp.1788343.

[69] M. Alfaleh, "Sustainable AI-Driven assessment in Higher Education: A systematic review of fairness, transparency, pedagogical innovation, and governance," 2026. doi: 10.3390/su18020785.