# A Framework for Generative AI-Driven Assessment in Higher Education

**Preprint** · May 2025

DOI: 10.20944/preprints202505.0800.v1



**4 authors:**

Galina Ilieva
Plovdiv University
**99** PUBLICATIONS **798** CITATIONS
SEE PROFILE

Tania Yankova
Plovdiv University
**51** PUBLICATIONS **670** CITATIONS
SEE PROFILE

Margarita Ruseva
Plovdiv University
**36** PUBLICATIONS **187** CITATIONS
SEE PROFILE

Stanimir Kabaivanov
Plovdiv University
**75** PUBLICATIONS **265** CITATIONS
SEE PROFILE

# A Framework for Generative AI-Driven Assessment in Higher Education

Galina Ilieva *, Tania Yankova, Margarita Ruseva, Stanimir Kabaivanov

Posted Date: 12 May 2025

doi: 10.20944/preprints202505.0800.v1

Keywords: generative artificial intelligence; higher education; digital pedagogy; educational technology, assessment; AI-supported assessment; framework for assessment; academic integrity; responsible AI use

*Article*

# A Framework for Generative AI-Driven Assessment in Higher Education

**Galina Ilieva[1,*], Tania Yankova [1], Margarita Ruseva [1] and Stanimir Kabaivanov [2]**

[1] Department of Management and Quantitative Methods in Economics, University of Plovdiv Paisii Hilendarski, 4000 Plovdiv, Bulgaria
[2] Department of Finance and Accounting, University of Plovdiv Paisii Hilendarski, 4000 Plovdiv, Bulgaria
* Correspondence: galili@uni-plovdiv.bg

**Abstract:** The rapid integration of generative artificial intelligence (AI) into educational environments raises both opportunities and concerns regarding assessment design, academic integrity, and quality assurance. While new generation AI tools offer new modes of interactivity, feedback, and content generation, their use in assessment remains insufficiently pedagogically framed and regulated. In this study, we propose a new framework for generative AI-supported assessment in higher education, structured around the needs and responsibilities of three key stakeholders (branches): instructors, students, and control authorities. The framework outlines how teaching staff can design adaptive and AI-informed tasks and provide feedback, how learners can engage with these tools transparently and ethically, and how institutional bodies can ensure accountability through compliance standards, policies, and audits. This three-branch multi-level model contributes to the emerging discourse on responsible AI adoption in higher education by offering a holistic approach for integrating AI-based systems into assessment practices while safeguarding academic values and quality.

**Keywords:** generative artificial intelligence; higher education; digital pedagogy; educational technology; assessment; AI-supported assessment; framework for assessment; academic integrity; responsible AI use

## 1. Introduction

The digital transformation of education has significantly reshaped both the delivery of knowledge and the assessment of learning outcomes. In higher education, the rapid adoption of online platforms for teaching, assessment and evaluation has faced growing challenges related to access, affordability, academic integrity, and quality assurance [1,2]. These challenges are further intensified by the emergence of generative artificial intelligence (AI), which is transforming educational practices at a fundamental level and reshaping the entire learning and evaluation ecosystem [3]. Recent advancements in large language models (LLMs) have enabled the development of powerful content generating AI tools that can interpret user intent, engage in natural language dialogue, and offer interactive, personalised, and scalable support. In higher education, these technologies are no longer limited to tutoring functions as they are increasingly being embedded into assessment design, feedback generation, and academic support.

While generative AI introduces significant opportunities for personalised learning experience and adaptive assessment process [4], it also raises concerns around the assessment accuracy [5]. Such developments have prompted renewed demands for updated quality standards from organizations such as the European Association for Quality Assurance in Higher Education (ENQA) [6], particularly in response to the growing integration of AI technologies in assessment practices.

At the same time, higher education accreditation procedures place strong emphasis on the quality of student assessment as a key indicator of educational effectiveness, requiring modern, transparent, and multi-component evaluation methods that align with intended learning outcomes. For example, under certain criteria [7], lecturers are expected to demonstrate the use of varied assessment

 doi:10.20944/preprints202505.0800.v1



formats, including interactive and digital tools, while ensuring ~~fairness~~, academic integrity, and alignment with course objectives.

Given the rapid adoption of AI-based tools into educational practice, along with increasing regulatory demands, there is a growing imperative to rethink and redesign assessment models to ensure they remain relevant, reliable, and equitable within digitally mediated, AI-enhanced learning environments [8]. Despite the proliferation of AI-powered instruments in higher education, existing assessment frameworks often lack structured guidance on how to incorporate such technologies in a pedagogically appropriate and ethically responsible manner.

Our study addresses this gap by proposing and validating a structured framework for integrating generative AI into assessment processes in higher education. It takes into account the roles, responsibilities, and expectations of three key stakeholder groups: instructors, students, and higher education quality assurance bodies. To verify the framework, a case study was conducted within the context of a university-level course on business data analysis. A comparative evaluation was performed between assessments graded by a team of academic lecturers and those generated by ChatGPT. This comparison between human- and AI-based assessments reveals key challenges and opportunities, contributing to the refinement of the framework application in line with both quality assurance standards and pedagogical objectives.

The main objectives of this study are as follows:

- To identify the needs, expectations, challenges, and perspectives of key actors regarding the integration of AI-based tools for assessment purposes;
- To propose quality assurance measures and practical guidelines for the responsible use of generative AI in formal assessment settings;
- To design a new conceptual framework that enables a systematic and stakeholder-sensitive integration of generative AI into assessment practices in higher education;
- To support the development of transparent, scalable, and learner-centred evaluations that maintain academic integrity while embracing technological innovation;
- To examine the technological, pedagogical and ethical implications of deploying AI-powered systems for designing, administering, and evaluating academic assessments.

The primary contribution of this paper is the development of a conceptual framework for innovative assessment in higher education, offering a holistic approach for embedding AI-based instruments across different stages of the evaluation cycle. The proposed model supports both formative and summative assessments, emphasising ethical considerations, transparency, and fairness. By proactively addressing potential challenges, including overdependence on automation and improper AI tool usage, the framework facilitates a more robust, accountable, and inclusive assessment environment aligned with contemporary educational objectives.

The structure of the paper is as follows: Section 2 presents an overview of recent advancements in generative AI technologies, focusing particularly on LLM-based platforms and their educational applications. Section 3 reviews existing research on the integration of generative AI in academic assessment, covering its benefits, limitations, and institutional responses. Section 4 introduces the proposed conceptual assessment framework, describing its three branches (teaching staff, learners, and quality assurance authorities), their interactions, and practical implementation aspects. The paper concludes by summarising key findings, outlining implications for academic policy and practice, and providing recommendations for future research directions.

## 2. State-of-the-Art Review of Generative AI Tools and Their Assessment Applications

Generative AI tools, particularly those built upon advanced LLMs and next-generation intelligent chatbots, originated as software applications to facilitate human-computer interaction through natural language. Initially used in customer service and business communication to automate routine tasks, such as answering frequently asked questions, these tools have quickly grown in complexity and capability, now offering sophisticated solutions in multiple domains, including education.

 doi:10.20944/preprints202505.0800.v1



Within higher education contexts, generative AI has evolved to become integral not only to instructional support but also to the design, delivery, and evaluation of academic assessments. Generative assessment tools specifically refer to software platforms that leverage AI to create educational content, streamline grading, and assist learners in achieving targeted learning outcomes. By integrating capabilities such as natural language processing, machine vision, and multimodal data interpretation, these tools enhance instructional efficiency and allow for greater personalisation in educational practice [9,10]. Nevertheless, despite the expanding role of generative AI in education, many lecturers, administrators, and policymakers continue to face challenges in fully understanding the capabilities and implications of these emerging technologies. As generative AI becomes more deeply embedded within academic environments, exploring its transformative potential for assessment practices becomes essential.

This section examines the current state of generative AI tools and their relevance to academic assessment. The first subsection begins with a classification of assessment types in higher education based on key criteria. Next, we explore the emerging use of generative AI to support both instructors and students, with a focus on instructional efficiency and enhancing personalisation. The second subsection reviews established metrics for quality evaluation in higher education, followed in the third subsection by an overview of the technical capabilities of leading generative AI tools. Finally, we present a comparative analysis of these tools in terms of their functionalities within educational assessment settings.

### *2.1. Classification of Assessment in Higher Education and the Role of Generative AI in This Process*

Assessment in higher education can be categorised across multiple dimensions. Some of the most widely used classification criteria include the stakeholder involvement (instructor-led, peer-reviewed, self-assessed, or externally moderated), setting (in-person, online, hybrid), purpose (formative or summative), personalization (uniform or adaptive), format (written, oral, practical, digital), timing (continuous or final), and feedback mode (quantitative, qualitative, automated, personalised). Furthermore, assessments differ in authenticity, ranging from purely academic exercises to real-world tasks, and in frequency, from frequent low-stakes activities (micro-assessments) to infrequent, high-stakes (final) evaluations [11].

Another widely recognised assessment framework is Bloom's taxonomy. Combined with active learning approaches, this framework offers a structured method for evaluating not only what students know but also how they engage with and apply that knowledge. This taxonomy identifies six hierarchical cognitive levels (remembering, understanding, applying, analysing, evaluating, and creating), linked to specific instructional methods and assessment formats [12].

Generative AI aligns with and enhances multiple dimensions of these assessment classifications. It enables digital and adaptive assessment formats, particularly within formative and criterion-referenced models. AI-based tools can provide personalised assessment tasks aligned with students' cognitive levels, provide instant feedback to support iterative learning, and offer dynamic content that adjusts to learner needs – thereby enriching the formative assessment process. In summative contexts, AI-driven assessment can streamline the creation of diverse test versions and support instructors with rubric-based grading, helping maintain consistency and reduce workload.

Moreover, AI-powered tools also enhance inclusivity by adapting content for learners with varying needs – through language simplification, alternative formats (e.g., audio), and multimodal delivery. Their integration further expands possibilities for authentic assessment, enabling the creation of simulations, scenario-based tasks, and problem-solving environments reflective of professional contexts.

However, the adoption of AI-based applications necessitates a re-evaluation and adaptation of traditional assessment taxonomies and standards. Institutions must carefully consider how generative AI influences academic integrity and the quality of feedback. As new AI-based tools become increasingly embedded in assessment design and evaluation, updating quality assurance measures to reflect these changes is essential.

 doi:10.20944/preprints202505.0800.v1



*2.2. Standards, Policies and Procedures for Assessment Quality in Higher Education*

Quality assessment in higher education requires that assessment practices be valid, reliable, fair, and aligned with the intended learning outcomes. High-quality assessment accurately reflects what students are expected to learn and demonstrates whether those outcomes have been achieved. Validity refers to the degree to which an assessment measures what it is intended to measure, while reliability relates to the consistency of results across different assessors, contexts, and student cohorts. Fairness entails that all students are assessed under comparable conditions, with appropriate accommodation to support equity and inclusion. Additional indicators of assessment quality include alignment with curriculum objectives, transparency of assessment criteria, and the provision of constructive feedback [13].

To uphold these standards, institutions typically use internal moderation processes, external examiner reviews, student feedback, benchmarking against national qualification frameworks, and systematic analyses of grade distributions. Increasingly, digital tools and analytics enhance institutions' capacity to monitor assessment effectiveness and maintain academic integrity.

Internationally, frameworks such as ISO 21001:2018 [14] provide structured guidance for educational organisations, emphasising learner satisfaction, outcome alignment, and continuous improvement in assessment practices. Although its predecessor, ISO 29990:2010, specifically targeted non-formal learning, its key principles, such as needs analysis and outcome-based evaluation, remain integral within current guidelines [15].

Within Europe, the Standards and Guidelines for Quality Assurance (ESG), established by the ENQA, serve as a key regional benchmark. ESG Standard 1.3 focuses specifically on student-centred learning and assessment, underscoring the importance of fairness, transparency, alignment with learning objectives, and timely feedback [16]. These principles have shaped quality assurance systems across Europe and influence institutional practice at national and institutional levels.

National agencies such as the UK's Quality Assurance Agency (QAA) provide detailed guidance on assessment practices [17]. Their frameworks cover areas such as the reliability of marking, moderation protocols, assessment-objective alignment, and the balanced integration of formative and summative methods. These guidelines promote transparency, support diverse assessment types, and ensure that feedback is timely and meaningful.

At the institutional level, internal quality assurance policies typically define assessment expectations in terms of validity, fairness, and consistency. Tools such as grading rubrics, standardised criteria, and peer moderation are used to ensure alignment with programme-level outcomes and to enhance the reliability of evaluation. Institutions often conduct regular reviews, audits, and data-driven evaluations to track assessment performance and drive continuous improvement in teaching and curriculum design [18].

Collectively, these international standards, national regulations, and institutional policies provide a comprehensive structure for ensuring robust assessment practices. They ensure that assessment not only serves as a tool for evaluation but also supports meaningful student learning, accountability, and educational enhancement within contemporary higher education systems.

*2.3. Generative AI Tools for Assessment in Higher Education*

Assessment tools in higher education have evolved significantly, transitioning from traditional paper-based examinations and manual grading toward digitally enhanced, personalised evaluation methods. Early developments, such as computer-based testing, increased grading efficiency and objectivity, paving the way for Learning Management Systems (LMSs) that supported online quizzes, automated scoring, and instant feedback [19]. Subsequently, the integration of learning analytics and adaptive assessment platforms enabled educators to track student progress and tailor instruction dynamically. Massive Open Online Courses (MOOCs) often incorporate peer review systems, automated grading, and interactive tasks, enabling them to accommodate thousands of learners simultaneously [20]. More recently, AI-driven platforms have emerged, facilitating automated question generation, rubric-based grading, and scalable formative assessment [21].

 doi:10.20944/preprints202505.0800.v1



The rise of generative AI technologies has further transformed the assessment landscape, offering capabilities for real-time, interactive, and context-aware evaluations. Institutions exploring these advanced tools encounter new opportunities for personalised assessment while addressing critical challenges concerning academic integrity, transparency, and equity. Below is an overview of prominent generative AI-based platforms, each tailored to specific assessment needs within higher education.

ChatGPT (OpenAI, San Francisco, CA, USA; https://openai.com/chatgpt, accessed on 10 April 2025) is one of the most widely used generative AI tools in education. Based on the GPT-4.5 architecture, ChatGPT enables instructors to generate quiz questions, test prompts, scoring rubrics, and feedback explanations. With a user-friendly interface and conversational capabilities, it supports academic writing, peer simulation, and self-assessment for students. Its ability to provide consistent, context-sensitive responses makes it suitable for both formative and summative assessment support. While the basic version is free, advanced features are available through the ChatGPT Plus subscription.

Eduaide.ai (Eduaide, Annapolis, MD, USA; https://eduaide.ai, accessed on 10 April 2025) is a dedicated platform for educators that leverages generative AI to support teaching and assessment design. It provides over 100 instructional resource formats, including quiz generators, essay prompts, rubrics, and feedback templates. Instructors can generate differentiated tasks aligned with Bloom's taxonomy or customize assessments by difficulty, format, or objective. Eduaide.ai also includes alignment tools to ensure curriculum coherence and academic standards, making it suitable for both classroom and online environments.

Quizgecko (Quizgecko, London, United Kingdom; https://quizgecko.com, accessed on 10 April 2025) is an AI-powered quiz generator that enables instructors to create assessment items from any text input, such as lecture notes, articles, or web content. It supports multiple-choice, true/false, short answer, and fill-in-the-blank formats. Quizgecko also allows export to LMS platforms and offers analytics on question difficulty and answer patterns. This tool is useful for formative assessments, quick reviews, and automated test creation aligned with instructional content.

GrammarlyGO (Grammarly, San Francisco, CA, USA; https://www.grammarly.com, accessed on 10 April 2025) is an AI-based writing assistant that offers contextual feedback on grammar, clarity, and tone. It supports students in developing academic writing skills and can serve as a real-time feedback mechanism during essay composition. For instructors, GrammarlyGO helps in providing formative feedback, especially in large-scale or asynchronous courses. The tool offers suggestions for revision while maintaining a focus on the original intent, making it suitable for self-assessment and peer-review training. GrammarlyGO assists instructors in delivering formative feedback efficiently, especially in large-scale or online courses.

Gradescope (Turnitin, Oakland, CA, USA; https://www.gradescope.com, accessed on 10 April 2025) is a grading platform enhanced with AI features such as automatic grouping of similar answers, rubric-based scoring, and feedback standardization. It supports handwritten, typed, and code-based responses and is widely used in STEM courses. While not generative per se, Gradescope's AI-assisted evaluation streamlines the grading process, improves consistency, and facilitates large-class assessments with reduced instructor workload.

Copilot (Microsoft, Redmond, WA, USA; https://copilot.microsoft.com, accessed on 10 April 2025) is integrated into Microsoft 365 suite and enables instructors and students to generate content, summarize documents, create study guides, and automate tasks such as rubric generation and feedback provision. Its integration into Word, Excel, and Teams enhances workflow efficiency in assessment processes.

Table 1 presents a comparative overview of key characteristics, functionalities, supported platforms, input limits, and access types of the discussed generative AI-based assessment tools.

 doi:10.20944/preprints202505.0800.v1



**Table 1.** Comparison of the most widely used generative AI tools for assessment in higher education.

| Tool name | Based on LLM | Main educational functionality | Supported platforms | Input Limit (Max Tokens) | Access type |
|---|---|---|---|---|---|
| ChatGPT | GPT-4.5 | Assessment creation, rubric assistance, feedback generation | Web, mobile, API | Up to 32,000 tokens | Freemium and subscription |
| Eduaide.ai | GPT- and Anthropic-based | Instructional/assessment resource generation, Bloom's alignment | Web | N/A | Freemium and subscription |
| Quizgecko | GPT-based | Quiz generation from text, export to LMS | Web | N/A | Freemium and subscription |
| GrammarlyGO | Proprietary GPT-based | Real-time writing support, revision suggestions | Desktop, browser, mobile | N/A | Freemium and subscription |
| Gradescope | Custom AI models | Rubric-based AI-assisted grading and feedback | Web, LMS integration | N/A | Institutional license |
| Copilot | GPT-4.5 (via Azure) | Automation of content generation, rubric writing, document summarization | Microsoft 365 apps | Up to 32,000 tokens | Subscription (Microsoft 365) |

Depending on their primary function and technological scope, generative AI tools for assessment can be grouped into several distinct categories:

- Content-generation tools – these platforms (e.g., ChatGPT, Eduaide.ai, Copilot) focus on the automatic creation of assessment elements such as quiz questions, prompts, rubrics, and feedback. They support instructors in designing assessments aligned with learning taxonomies and learning outcomes.
- Interactive quiz generators – tools like Quizgecko use generative algorithms to convert any text into structured assessment formats, such as multiple-choice, fill-in-the-blank, or true/false questions. Their focus is on streamlining formative assessment, particularly in asynchronous or online contexts.
- AI-assisted feedback systems – platforms such as GrammarlyGO provide real-time, formative feedback on student writing, helping learners improve clarity, coherence, and argumentation. These tools are particularly useful in large classes and for supporting autonomous learning.
- Grading automation systems – tools like Gradescope apply AI models to evaluate student submissions using rubrics, group similar responses, and standardise feedback. They are widely adopted in STEM education and are integrated with major LMSs.
- Learning assistants and peer support tools – ChatGPT and similar apps serve as AI tutors that guide students through problem-solving steps and help reinforce foundational concepts, indirectly supporting assessment preparation through personalised coaching.

These AI tools also vary according to their intended users. Eduaide.ai and Copilot primarily serve educators, assisting with assessment design and workflow automation. GrammarlyGO targets students directly, providing immediate, actionable feedback. Tools like ChatGPT and Copilot offer flexible usage scenarios suitable for both instructors and students, depending on instructional contexts and needs.

Deployment options further differentiate these tools. Most operate via web interfaces, ensuring broad accessibility. GrammarlyGO's mobile deployment accommodates informal or flexible study environments. Gradescope and Quizgecko offer compatibility with institutional LMS platforms, while Microsoft Copilot integrates AI functionalities seamlessly within widely used productivity tools such as Word and Excel.

Different access models also influence the adoption and usage of these tools. Platforms such as ChatGPT, Eduaide.ai, Quizgecko, and GrammarlyGO typically follow freemium models with

 doi:10.20944/preprints202505.0800.v1



subscription-based advanced features. Gradescope primarily operates under institutional licenses, managed centrally by universities or departments, and Copilot requires a Microsoft 365 subscription, aligning with enterprise workflows.

Underlying AI models further distinguish these tools. ChatGPT and Copilot leverage advanced LLMs like GPT-4.5, capable of handling complex, contextually nuanced tasks, thus enhancing their effectiveness in supporting comprehensive assessments and feedback. Gradescope's approach, which focuses on clustering and rubric-based evaluation rather than generative capabilities, prioritises grading efficiency and transparency, making it particularly suited for structured and high-stakes assessments.

Each AI tool also aligns differently with assessment types. GrammarlyGO and Quizgecko are particularly well-suited for formative purposes, providing iterative, low-stakes learning reinforcement. ChatGPT, Eduaide.ai, and Gradescope support both formative and summative assessment contexts, facilitating assessment creation, scoring, and rubric-based feedback. Copilot bridges both categories effectively by aiding material preparation, content summarisation, and feedback provision within institutional workflows.

Alignment with pedagogical frameworks further differentiates these tools. Eduaide.ai explicitly integrates Bloom's taxonomy, enabling targeted assessments tailored to distinct cognitive levels. ChatGPT, through guided prompts, can foster reflective and metacognitive learning processes. A shared advantage of LLM-based platforms is their inherent adaptability and personalisation, tailoring outputs to specific instructional contexts and individual learner needs.

Despite these benefits, there are important limitations to consider. Generative tools such as ChatGPT and GrammarlyGO can produce outputs that are difficult to verify, raising concerns regarding source traceability and explainability—critical aspects in high-stakes assessments. In contrast, tools like Gradescope emphasise transparent evaluation criteria, structured feedback, and clear grading standards. Additionally, risks such as AI-generated inaccuracies, known as hallucinations, necessitate careful oversight by educators to ensure assessment validity and reliability.

In this subsection, we have explored the evolving role of generative AI tools in reshaping higher education assessment procedured. These tools vary considerably in their functionality, degree of automation, adaptability, and LMS integration. They enable a broad spectrum of assessment activities, from automated content creation and real-time formative feedback to personalised learning analytics, opening new possibilities for designing inclusive, scalable, and pedagogically rich assessment experiences.

The next section discusses related work, outlining existing assessment frameworks within higher education and summarising current research on user attitudes towards AI-driven assessment tools.

## 3. Related Work

### *3.1. Theoretical Frameworks for Application of Generative AI in Assessment Processes*

Several studies have developed models addressing different facets of modern AI integration, particularly the use of tools like ChatGPT in frameworks for academic assessment design and delivery (Table 2).

Smolansky et al. [22] introduced a comprehensive six-dimension framework for evaluating online assessment quality, incorporating academic integrity, student experience, authenticity, information integrity, quality feedback, and equity of access. This framework is grounded in empirical data from both students and educators across two universities.

Thanh et al. [12] created a framework to evaluate generative AI's effectiveness in solving authentic economics tasks. Their findings indicated that generative AI performs exceptionally well at lower levels of Bloom's taxonomy and maintains acceptable performance in higher-order tasks, particularly in quantitative contexts.

 doi:10.20944/preprints202505.0800.v1



Agostini and Picasso [23] proposed a pedagogical and technological framework focused on sustainable assessment using LLMs. Their study highlights the ability of generative AI tools to automate feedback and grading for large student groups without compromising educational quality.

Kolade et al. [24] conducted a quasi-experimental study on ChatGPT's application in essay-based assessments. They found that AI could generate original, high-quality essays, but faced limitations when asked to create multiple unique outputs from the same account and struggled with accurate citation practices. Based on these findings, they proposed a framework for AI-assisted assessment that supports lifelong learning by emphasizing higher-order cognitive skills.

Salinas-Navarro et al. [25] employed a methodological approach called "thing ethnography" to promote authentic assessment in higher education. By focusing on the relationship between material artefacts, learners, and AI tools, their work underscores the value of embedding assessment practices in real-world and context-rich learning environments.

Khlaif et al. [26] explored faculty attitudes and redesign strategies in response to generative AI adoption. They introduced the "Against, Avoid, Adopt, and Explore" conceptual framework, offering structured guidance for navigating assessment transformations in the AI era. The authors call for further validation and refinement of this framework across broader educational contexts.

Several of these frameworks contribute valuable insights, yet they also reveal limitations. For instance, Thanh et al. [12] and Kolade et al. [24] focus on specific assessment types – quantitative economics tasks and essay-based assignments, respectively – without addressing the broader curriculum or skill development across disciplines. Frameworks like those proposed by Smolansky et al. [22] and Khlaif et al. [26] target particular stakeholder groups, such as students and faculty, but do not offer fully integrated institutional strategies that encompass administrators, curriculum designers, and policymakers. Moreover, evaluation approach in the study of Smolansky et al. [22] primarily rely on user perceptions and satisfaction surveys, rather than incorporating functional performance metrics or systematic usability modelling. These limitations highlight the need for a comprehensive and adaptable framework that can guide the scalable, ethical, and pedagogically aligned integration of generative AI into higher education assessment.

Despite growing interest and innovation in this area, a universally accepted and operational framework for the design, development, and implementation of generative AI tools in academic assessment remains absent. As generative AI continues to evolve, there is a pressing need for holistic, transdisciplinary framework that address pedagogical, technological, and ethical dimensions of assessment in higher education.

### *3.2. Measuring the Attitudes Towards Generative AI Assessment and its Quality*

This section explores the perceptions and attitudes of both students and educators towards the use of generative AI tools in academic assessment, along with the perceived quality and impact of these tools on learning and academic integrity. As generative AI becomes increasingly embedded in higher education, understanding stakeholder responses is essential to ensuring effective, ethical, and sustainable implementation.

User experience and perceived value play a central role in shaping attitudes toward generative AI-supported assessment. As with educational chatbots, student satisfaction and perceived usefulness are key indicators of how well these tools support learning goals. Several studies have investigated these dynamics, identifying motivations for adoption, concerns about misuse, and implications for assessment quality and design (Table 2).

Nikolic et al. [9] conducted a large-scale study involving seven Australian universities, where ChatGPT was tested on real engineering assessments. Faculty graded AI-generated answers using standard procedures, finding that ChatGPT often produced passable results with only minimal prompt engineering. While students and staff acknowledged the tool's usefulness for solving structured tasks, concerns emerged about originality and the potential erosion of academic standards. The study emphasizes the need for revised assessment strategies that account for the presence and capabilities of generative AI.

 doi:10.20944/preprints202505.0800.v1



Williams [27] presented a conceptual model that frames generative AI as one of the central forces disrupting traditional assessment practices. Through literature synthesis and SWOT analysis, the author advocates a shift away from rigid summative models toward more formative, student-centred assessment approaches. This model suggests that AI tools can support continuous feedback and authentic learning experiences, provided institutions implement pedagogically sound frameworks and safeguard against misuse.

Chiu [28] builds on this perspective by emphasizing the transformative potential of generative AI while also identifying necessary conditions for its responsible use. His analysis stresses the importance of fostering AI literacy among both students and educators. Students must be equipped to use AI tools ethically and effectively, while educators should develop the ability to design AI-resilient assessments and critically evaluate AI-assisted student outputs. Chiu recommends institutional policies that promote coexistence with AI and encourages leveraging generative AI for personalized feedback and efficiency, while maintaining educational quality.

Gruenhagen et al. [10] surveyed over 300 university students regarding their use of ChatGPT in coursework and their views on academic integrity. A substantial number of respondents admitted to using AI tools in assignments and did not consider it cheating. The study found that psychosocial variables such as motivation and academic stress influence AI adoption. These findings point to the need for updated academic integrity policies and redesigned assessments that are both AI-aware and learning-focused.

Ogunleye et al. [29] provided empirical evidence on the quality of AI-generated outputs in real university coursework. Comparing ChatGPT and Google Bard on postgraduate-level tasks, the researchers found that both systems demonstrated subject knowledge and problem-solving capacity, achieving passing scores. However, challenges such as incomplete referencing and varying depth in responses were noted. The study concludes that while generative AI tools are capable of supporting academic tasks, their uncritical use may hinder learning. Recommendations include integrating AI literacy, adopting mixed assessment formats (e.g., oral exams or reflective tasks), and developing detection and monitoring strategies to preserve educational value.

Several of the reviewed studies offer important contributions to understanding stakeholder attitudes and perceived quality of generative AI assessment tools, yet they also present notable limitations. For example, Nikolic et al. [9] and Ogunleye et al. [29] provide empirical insights into the performance of AI tools on specific types of academic tasks, but focus primarily on technical evaluation without fully addressing the pedagogical or psychological dimensions of user experience. Conceptual contributions such as those by Williams [27] and Chiu [28] propose strategic shifts in assessment design and institutional policy but remain largely theoretical, lacking empirical validation across diverse educational contexts. Studies like Gruenhagen et al. [10] explore student perceptions and psychosocial factors influencing AI use but do not integrate these findings into a broader framework for institutional implementation or educational design. Moreover, most investigations prioritize either student or educator perspectives in isolation, without offering a holistic view that includes administrative, curricular, and policy-level considerations. Finally, much of the existing research relies heavily on self-reported perceptions or satisfaction measures, rather than employing systematic performance metrics, behavioural data, or usability modelling.

**Table 2.** Comparison of recent studies on the application of generative AI in higher education assessment.

| Reference | Study Type | Target Assessment | Primary Focus | Methodological Approach |
|---|---|---|---|---|
| Nikolic et al. [9] | Empirical (experiment) | Engineering assignments/exams | ChatGPT performance on engineering tasks – implications for authenticity and design | Empirical (performance analysis) |
| Smolansky et al. [22] | Empirical (survey study) | Assessments (general) | Educator vs. student perspectives on GAI impact (integrity, usage, attitudes) | Empirical (survey research) |

 doi:10.20944/preprints202505.0800.v1



| | | | | |
|---|---|---|---|---|
| Thanh et al. [12] | Empirical (experiment) | Authentic assessments | GAI's ability to solve real-world tasks – benefits and limitations | Empirical (comparative) |
| Williams [27] | Conceptual (analysis) | Summative HE assessment (general) | Post-COVID assessment model redesign incorporating AI influences | Conceptual (theoretical model) |
| Agostini & Picasso [23] | Conceptual (analysis) | Assessment processes (general) | Sustainable, alternative assessment practices in response to LLMs | Conceptual (strategy proposal) |
| Chiu [28] | Survey (qualitative) | General (multiple assessment forms) | Policy and strategy | Empirical (focus groups) |
| Gruenhagen et al. [10] | Conceptual (commentary) | Written exams & assignments | Academic integrity challenges posed by GAI in assessments | Conceptual (issue commentary) |
| Kolade et al. [24] | Empirical (experiment + framework) | Multiple course assessments | ChatGPT's effect on learning/assessment – framework for AI-era assessment | Empirical (with conceptual elements) |
| Ogunleye et al. [29] | Empirical (analysis) | Various HE assessment tasks | Evaluating GAI tools' capabilities on student assessments – impact on practice | Empirical (multi-task analysis) |
| Perkins et al. [30] | Conceptual (framework proposal) | Assessment practices (general) | AIAS framework for ethical GAI integration in assessment policies | Conceptual (theoretical) |
| Salinas-Navarro et al. [25] | Empirical (exploratory study) | Experiential/authentic tasks | Using GAI to enhance experiential learning and authentic assessments | Empirical (AI tool exploration) |
| Khlaif [26] | Framework proposal (qualitative study) | General (all course assessments) | Assessment redesign | Empirical (interviews & focus groups) |

Taken together, these studies highlight the multifaceted nature of attitudes towards the use of AI-based assessment. While there is growing recognition of its potential to enhance learning, streamline feedback, and improve efficiency, persistent concerns remain regarding academic integrity, originality, and the risk of over-reliance. Importantly, perceptions are shaped not only by the technical performance of AI tools but also by broader pedagogical, psychological, and institutional dynamics. This complexity underscores the need for a holistic assessment framework that integrates generative AI thoughtfully while safeguarding fairness, authenticity, and learning quality. Such a framework must address diverse stakeholder perspectives and align technological innovation with academic standards to ensure that AI-supported assessments remain both meaningful and equitable.

## 4. Framework for Generative AI-Supported Assessment in Higher Education

In this section, we introduce a new conceptual framework that supports the integration of generative AI tools, particularly intelligent chatbots, into university-level assessment practices. It promotes a more personalized, scalable, and explainable assessment ecosystem through the responsible use of AI.

This framework is designed to address the emerging challenges and opportunities in higher education assessment by structuring the interaction between three key participants groups. In the proposed framework (Figure 1), the stakeholders (represented by the left, right, and upper branches of the flowchart) are supported by generative AI through their distinct functions. The framework has a hierarchical structure and it is organized into three levels, corresponding to the standard phases of the academic assessment process: (1) course-level assessment; (2) midterm testing, and (3) unit tasks.

 doi:10.20944/preprints202505.0800.v1



Level 1. Course Assessment

At the course assessment level, generative AI supports the final, cumulative evaluation of students' performance. For lecturers, AI tools assist in designing comprehensive exam formats, generating final exam questions aligned with course outcomes, and producing consistent, rubric-based grading and feedback at scale. For students, intelligent chatbots offer support in reviewing course content, simulating exam environments, and answering last-minute queries with tailored feedback and revision prompts. For control authorities, generative AI enables the systematic monitoring of assessment quality, fairness, and alignment with accreditation standards. AI dashboards can summarize grading patterns and highlight anomalies in final results, ensuring transparency and integrity in the summative assessment process.

Level 2. Midterm Assessment

At the midterm exam level, generative AI enhances the quality and responsiveness of ongoing evaluations conducted mid-course. For lecturers, AI can generate customized midterm questions, assist in balancing question difficulty, and evaluate structured responses using predefined rubrics. Chatbots can also assist in clarifying exam instructions and interpreting student performance trends. For students, AI tools provide personalized practice tasks, simulate exam-style questions, and offer constructive feedback on performance to guide future learning. For control authorities, this stage offers an opportunity to monitor continuous assessment integrity. AI tools can ensure consistency in exam administration, track progress data, and assess the degree to which midterm evaluations align with curriculum benchmarks.

Level 3. Unit Assessments

At the unit assessment level, generative AI facilitates frequent, formative evaluations embedded within the teaching-learning cycle. Lecturers benefit from AI-enabled automation of low-stakes quizzes, knowledge checks, and quick feedback mechanisms, allowing real-time insight into students' understanding of specific topics. Students interact with AI chatbots to clarify confusing concepts, test their knowledge, and receive immediate feedback to support mastery learning. These interactions also promote critical thinking and self-regulated learning strategies. For control authorities, this level provides granular data on assessment engagement and learning progression. AI tools can help identify early signs of disengagement or learning gaps, enabling timely interventions and supporting evidence-based quality assurance at the micro-curricular level.

The framework is also based on a five-step assessment procedure that reflects the typical phases of the academic assessment process. At each step, the roles of generative AI are illustrated through the supportive functions it provides to key stakeholders.

Step 1. Assessment Planning

This stage begins before course delivery and focuses on assessment design. Lecturers define intended learning outcomes (ILOs), assessment types (e.g., essays, quizzes, projects), and evaluation criteria. Generative AI tools support this process by offering alternative rubric templates, generating sample tasks aligned with Bloom's taxonomy, and mapping course outcomes to competency standards. Students are informed about assessment formats and expectations early on, with chatbots offering clarification and examples on request. Accreditation bodies or quality assurance units may access AI-generated syllabi and task banks to evaluate coherence with institutional policies.

Step 2. Pre-Assessment Preparation

At this stage, students engage in self-assessment or formative exercises to prepare for the main assessments. Intelligent chatbots provide quiz questions, prompt writing exercises, and offer adaptive feedback to address knowledge gaps. Lecturers can use AI to automatically generate review materials and practice exams tailored to class performance. Control bodies benefit from AI-generated metadata describing assessment difficulty, scope, and alignment with curricular standards. This fosters transparency and consistency in academic oversight.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: Posted: 12 May 2025 doi:10.20944/preprints202505.0800.v1



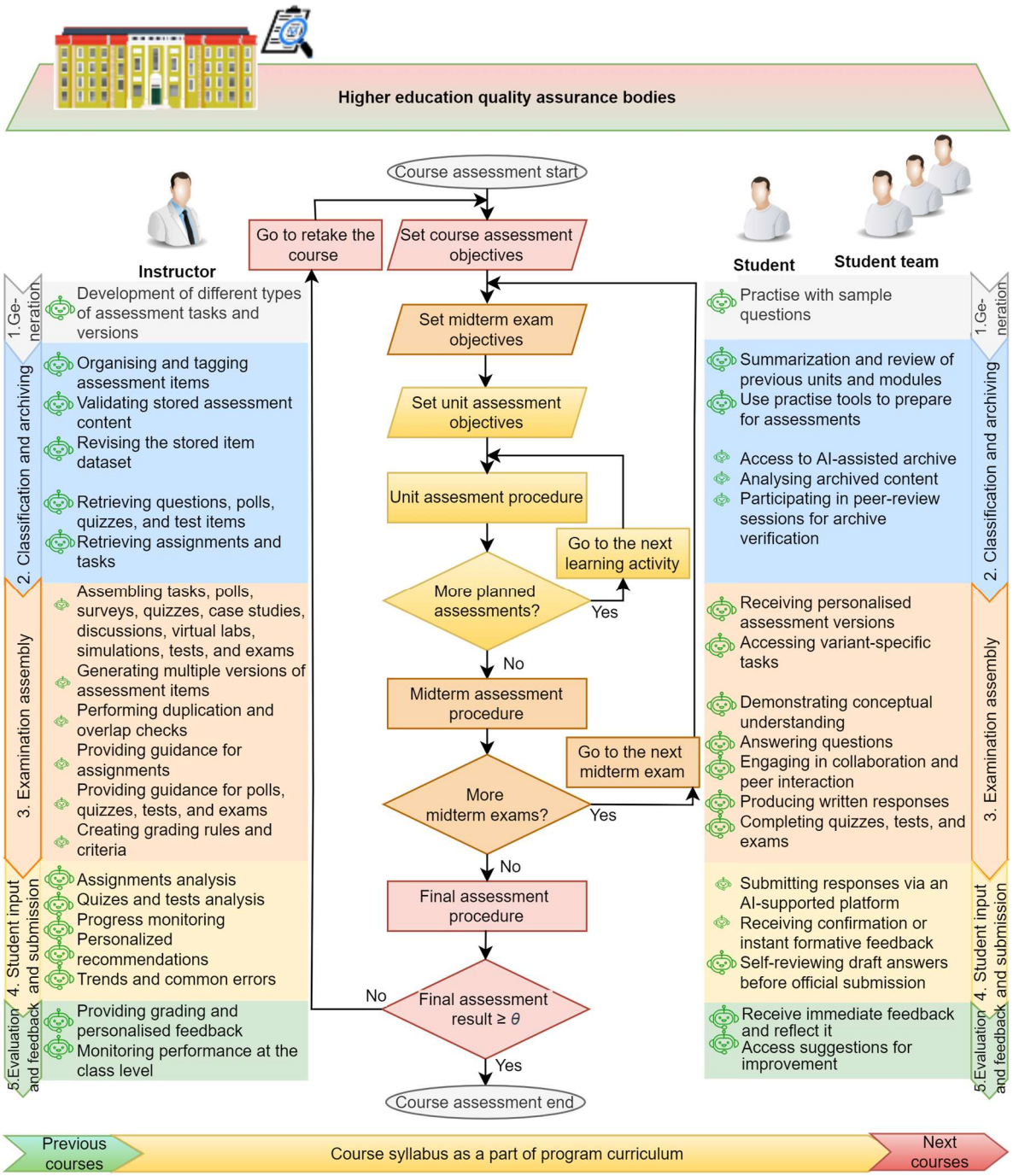


**Figure 1.** Framework for generative AI application in higher educational assessment.

Notes: The symbol $\theta$ denotes the minimum achievement threshold required to successfully pass the final exam, as defined in the course syllabus. This threshold typically falls within the range of 60% to 70% of the total available exam points. The size of the green icons depicting generative AI tools (left and right) corresponds to the extent of their applicability in assessment activities.

Step 3. Assessment Delivery

During actual assessments, students interact with assessment tasks in various formats. For open-ended or interactive exams, AI chatbots may simulate oral examination scenarios or clarify task instructions. For objective assessments, they can proctor online sessions, check integrity, and ensure compliance. Lecturers rely on intelligent AI for administering assessments through LMS-integrated chatbot interfaces and may enable AI-supported answer validation. Although AI involvement is highest in pre/post stages, in real-time assessments, its role is more assistive than central.

 doi:10.20944/preprints202505.0800.v1



Step 4. Feedback and Evaluation

This stage focuses on grading, feedback generation, and progress analysis. Lecturers benefit from AI-powered tools such as Gradescope, ChatGPT, or custom LMS bots that offer feedback drafts, detect inconsistencies, or suggest rubric-aligned scoring. Students receive structured, transparent, and timely feedback that is traceable to their submissions and rubric standards. Accreditation bodies can audit aggregated feedback quality and check AI-driven assessment traceability via digital trails or SHAP/LIME explanations, especially for high-stakes assessments.

Step 5. Post-Assessment Reflection and Improvement

After grades are finalized, all stakeholder groups engage in review and improvement. Students may receive AI-curated learning paths based on their performance. They may also use chatbots to reflect on assessment outcomes and plan future studies. Lecturers analyze assessment reports and student engagement analytics to refine future task design. Control bodies access de-identified summaries and metadata to evaluate the fairness, consistency, and effectiveness of the assessment strategies applied across departments or cohorts.

This framework supports alignment between technological advancements and academic integrity, aiding all actors within the university ecosystem. It is flexible enough to be adapted for different educational contexts and disciplines, and it encourages a balance between automation and human judgment in the assessment process. In the next section, we validate the effectiveness of the framework through a practical experiment involving final exam evaluation.

According to the proposed framework, generative AI tools play a supportive role throughout the five-step assessment procedure, implemented across the three primary levels of academic evaluation: unit assessments, midterm exams, and final exams. At the unit assessment level, they can assist students during the early stages of the process by offering access to curated learning resources, clarifying task requirements, and providing formative feedback on drafts and quiz attempts. For lecturers, this stage allows chatbots to handle routine queries, automate initial grading of objective tasks, and flag common misconceptions, enabling more targeted instructional interventions.

During the midterm exam phase, intelligent tools can be used to support students through personalized revision tools, practice tasks, and simulated test environments. Lecturers can rely on chatbot-generated analytics to identify learning gaps and adjust mid-semester teaching strategies accordingly. At this level, AI tools can also assist in rubric generation and consistent grading of structured responses, saving time while maintaining assessment quality.

In the final exam stage, generative AI tools continue to provide real-time clarification of exam-related questions and post-assessment feedback. They help students reflect on their performance, guide them through self-assessment prompts, and offer suggestions for improvement. For instructors, intelligent instruments can streamline final grading procedures and generate performance reports that support curriculum review and course evaluation.

While generative AI tools offer significant benefits, such as immediate feedback, reduced administrative workload, and individualized student support, it is essential to emphasize that all critical decisions regarding instructional design, assessment criteria, and course outcomes remain the responsibility of the instructor. The tool function is to complement, not replace, the academic judgment and pedagogical expertise of educators.

For students, these tools enhance the learning experience by reducing wait times, enabling self-paced exploration of concepts, and promoting deeper engagement through interactive dialogue. By posing open-ended questions and encouraging evidence-based responses, they can stimulate critical thinking, support argument development, and present alternative perspectives.

Furthermore, integrating generative AI into this structured assessment model supports key didactic principles such as flexibility, accessibility, consistency, and systematic progression. AI tools help educators identify knowledge gaps, personalize learning trajectories, and reinforce skill development through continuous, tailored practice. Their interactive and adaptive nature allows them to be applied effectively across various assessment contexts—for systematization, revision, reflection,

 doi:10.20944/preprints202505.0800.v1



and mastery—resulting in a more inclusive, engaging, and pedagogically sound assessment environment.

The third branch (top section of Figure 1) of the proposed framework addresses the role of control authorities, such as quality assurance units, academic oversight committees, and accreditation bodies, in upholding assessment standards and institutional integrity. Generative AI can support these stakeholders through advanced analytics and automated reporting tools that enable real-time monitoring of assessment practices across departments or academic programmes. For example, AI-powered dashboards can consolidate data on grading trends, feedback consistency, and student performance, helping control authorities identify anomalies, evaluate alignment with intended learning outcomes, and verify compliance with institutional and regulatory standards.

Additionally, generative AI can assist in auditing assessment materials for originality, fairness, and alignment with curriculum objectives. Tools based on natural language processing can evaluate the cognitive level of test items (e.g., using Bloom's taxonomy), check for redundancy or bias in exam questions, and flag potential academic integrity issues. AI-driven analysis of student feedback and peer evaluations can further inform quality reviews. By integrating these capabilities, control authorities can enhance transparency, streamline accreditation processes, and uphold academic standards more efficiently and consistently across the institution.

The proposed framework offers a holistic and multi-level approach to integrating generative AI into assessment processes, balancing innovation with pedagogical integrity, and supporting all key stakeholders in achieving a more transparent and adaptive higher education ecosystem. While the framework emphasizes intelligent chatbot-enabled interventions, it also accommodates broader generative AI functionalities, including rubric generation, automated feedback, and analytics dashboards.

The educators can employ the methodology described in the following section to assess the efficiency of generative AI in a specific university course.

## 5. Verification of the Proposed Generative AI-Based Assessment Framework

To evaluate the effectiveness of the proposed generative-AI-enhanced assessment framework, we conducted an experimental study in a data-analysis course during the winter semester of 2024–2025. The course's final examination was marked in parallel: first by a team of three subject-expert instructors and then by ChatGPT 4.5, following the same rubric. This pilot comparison allowed us to measure the reliability and validity of the AI-based approach against traditional human assessment.

The case study covered the scripts of fifteen students (S1–S15). The exam comprised sixteen questions designed to test analytical, interpretative, and applied competencies: six open-ended items (Q1–Q6), each worth 4 or 6 points, and ten closed-response items (Q7–Q16), each worth 1 point, for a maximum total of 40 points. The teaching team independently (without AI) evaluates the students' written submissions and assigns a final grade to each student based on a predefined scale linking total points to corresponding grades. The AI performs the same evaluation, having been provided with the criteria and scale used by the lecturers, and grades the submissions in accordance with these instructions.

For the purposes of this experimental validation, a single open-ended question (Q1) was selected. This question involved a specialised task requiring interpretation of GARCH model output and was intended to assess students' statistical reasoning and applied econometric skills. It was evaluated on a scale of up to 4 points. This task was independently assessed by both a team of lecturers and ChatGPT using a structured evaluation rubric specifically developed for this purpose (Table 3). The rubric comprised seven criteria: understanding of the output table, explanation of the model's purpose, interpretation of key parameters, use of statistical significance, identification of convergence warnings, structure and clarity of the response, and authenticity or originality. Its design reflects the cognitive levels outlined in Bloom's taxonomy, incorporating both lower-order skills (e.g. understanding and applying statistical output) and higher-order thinking (e.g. analysing parameter relationships and demonstrating original interpretation), thereby supporting a comprehensive

 doi:10.20944/preprints202505.0800.v1



evaluation of students' cognitive performance. Each criterion was scored on a scale from 0 to 4, resulting in a maximum possible score of 20 points per submission. To align with the final exam format, these rubric scores were proportionally scaled down to a 4-point grading system—the maximum allocated for this specific open-ended question.

**Table 3.** Assessment rubric used to evaluate the GARCH model question.

| Criterion | Max points | Scoring guidance |
|---|---|---|
| 1. Understanding of the Output Table | 4 | 4: Fully explains structure and roles of model components (mean, volatility, distribution);<br>3: Minor inaccuracies;<br>2: Partial understanding;<br>1: Minimal;<br>0: None |
| 2. Explanation of Model Purpose | 3 | 3: Clearly explains GARCH use in modelling volatility;<br>2: Some context;<br>1: Basic idea but unclear;<br>0: Incorrect or missing |
| 3. Interpretation of Key Parameters | 4 | 4: Accurate and insightful on all key terms (mu, omega, alpha, beta, nu);<br>3: Mostly accurate;<br>2: Some errors;<br>1: Misinterpretation;<br>0: Not attempted |
| 4. Use of Statistical Significance | 2 | 2: Discusses p-values/confidence intervals correctly;<br>1: Some understanding;<br>0: No mention or incorrect |
| 5. Discussion of Convergence Warning | 2 | 2: Identifies warning and explains implications;<br>1: Recognised but misinterpreted;<br>0: Ignored or incorrect |
| 6. Structure and Clarity | 2 | 2: Logically structured, easy to follow;<br>1: Some clarity issues;<br>0: Disorganised or unclear |
| 7. Authenticity and Originality (human vs. AI only) | 3 | 3: Demonstrates unique insight or personal approach;<br>2: Generic but plausible;<br>1: AI-like phrasing;<br>0: Likely copied or lacks original reasoning |
| Total | 20 | |

The results from the comparison of the obtained points showed a reasonable closeness between the two assessments, with ChatGPT successfully replicating human judgement in many instances, particularly when student responses followed structured patterns and demonstrated clear technical accuracy. However, discrepancies emerged in responses that required contextual understanding or critical insight, where lecturer evaluation more effectively captured the depth and nuance of reasoning.

For example, in several cases (Students 1, 4, 6, and 10), both ChatGPT and the lecturers assigned identical or closely aligned scores, reflecting a high degree of agreement in well-structured and technically sound answers. In a further eight cases, the difference between the two assessments was minimal—just one point. Only one case showed a two-point difference. Notably, Student 3 received a score of 1 from ChatGPT but 4 from the lecturers, indicating that the AI may have underestimated the student's depth of interpretation or originality of reasoning. В този случай, преподавателите са стимулирали с допълнителни точки нетрадиционно хрумване, въпреки неточното и непълно изпълнение на конкретното задание. От страна на хората е проявен субективизъм, но той е с позитивен знак, докато чатботът не може да се отклони от зададените правила.

 doi:10.20944/preprints202505.0800.v1



Overall, these validation results demonstrate that the proposed generative AI-based assessment framework can provide consistent and scalable evaluation for clearly defined open-ended tasks. Nevertheless, the findings also highlight the value of a hybrid assessment approach, whereby AI tools support preliminary scoring and formative feedback, while final grading decisions remain under academic supervision. Such a model ensures both operational efficiency and pedagogical integrity, particularly in high-stakes assessment scenarios.

The remaining open-ended questions were evaluated using the same procedure. The closed-type questions were excluded from the ChatGPT-based assessment, as their scoring relies on fixed, unambiguous answers that are best evaluated using standard automated or manual marking procedures.

To evaluate the consistency between the lecturer-based and GAI-based assessments at the level of the final exam, we compared the total scores and corresponding grades assigned to each of the 15 students (Table 4). The results indicate a high degree of alignment. In 12 out of 15 cases, the final grades assigned by ChatGPT matched those given by the lecturer team. In one case (Student 1), ChatGPT assigned a grade one level higher than the human assessors did, while in another two cases (Student 3 and Student 15), it assigned a grade one level lower. These minor discrepancies suggest that the GAI-based assessment approach is largely consistent with expert human judgement in grading summative assessments.

**Table 4.** Comparison of lecturer and ChatGPT-based final exam evaluations (points and grades).

| Student: | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 | S9 | S10 | S11 | S12 | S13 | S14 | S15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| L(points) | 32 | 17 | 31 | 30 | 34 | 37 | 28 | 33 | 18 | 30 | 40 | 33 | 23 | 23 | 37 |
| L (grade) | 5 | 3 | 5 | 5 | 5 | 6 | 4 | 5 | 3 | 5 | 6 | 5 | 4 | 4 | 6 |
| GAI (points) | 39 | 22 | 27 | 33 | 34 | 38 | 26 | 29 | 22 | 29 | 39 | 31 | 28 | 25 | 34 |
| GAI (grades) | 6 | 3 | 4 | 5 | 5 | 6 | 4 | 5 | 3 | 5 | 6 | 5 | 4 | 4 | 5 |
| Difference (grades) | -1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

Note: Points (P) were converted to grades in five-tier grading system: P ≤ 16 → 2 (Unsatisfactory); 17–22 → 3 (Fair); 23–28 → 4 (Good); 29–34 → 5 (Very Good); 35–40 → 6 (Excellent).

To further quantify the difference between the two approaches, we calculated several standard error metrics, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), for each open-ended question as well as for the total exam score (Table 5).

**Table 5.** Statistical error metrics for lecturer and ChatGPT scores by question and overall exam performance.

| Question | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Final exam score |
|---|---|---|---|---|---|---|---|
| MAE | 0.93 | 1.07 | 0.53 | 1.87 | 0.93 | 1.00 | 2.93 |
| MSE | 1.47 | 1.73 | 0.67 | 7.33 | 1.73 | 1.93 | 12.00 |
| RMSE | 1.21 | 1.32 | 0.82 | 2.71 | 1.32 | 1.39 | 3.46 |

The results show that ChatGPT achieved the closest agreement with lecturer scoring on Q3, with a MAE of just 0.53, suggesting very high consistency. Q1 and Q5 also demonstrated good alignment, each with a MAE of 0.93. These tasks likely involved well-structured, objective responses where ChatGPT performs reliably. Slightly higher error values were observed for Q2 and Q6, where MAEs were 1.07 and 1.00 respectively. This indicates acceptable, though somewhat less precise, agreement. In these cases, discrepancies may be due to borderline answers or interpretative elements that require human judgment. In contrast, Q4 presented the greatest challenge for ChatGPT, with the highest MAE (1.87) and RMSE (2.71). This question required students to submit an Excel file containing solver settings as part of a hands-on optimisation task, making it an example of assessment by doing. At the time of the experiment (December 2024), ChatGPT had limited capabilities for interpreting or

 doi:10.20944/preprints202505.0800.v1



evaluating attached spreadsheet files with embedded solver configurations, which likely contributed to the discrepancy in scoring.

When considering the final exam as a whole, the overall MAE was 2.93 points and the RMSE was 3.46 points (on a 40-point scale), indicating that ChatGPT scores generally differ from lecturer scores by approximately one grade level. While this reflects moderate accuracy, it also highlights the limitations of full automation in high-stakes summative assessment.

These findings suggest that generative AI is well suited for evaluating structured tasks and providing formative feedback but still requires human oversight for complex or high-impact assessments. Its integration into the assessment process is most effective when used to complement, rather than replace, academic judgment.

Each student received detailed feedback on their performance in the final exam, including both the assigned grades and targeted recommendations for addressing identified knowledge gaps. This individualised feedback supports the formative dimension of the assessment process and promotes self-directed learning. Based on the overall consistency of results between the generative AI-based and human-led evaluations, the proposed framework can be considered a reliable tool for assessing student performance in higher education settings. It offers a scalable solution for supporting both formative feedback and summative grading. For stakeholders such as academic staff, curriculum developers, and institutional quality assurance bodies, the framework provides a structured approach to integrating generative AI in assessment processes. Its use is especially recommended in contexts where efficiency, consistency, and timely feedback are essential, provided that it is complemented by appropriate human oversight in more interpretive or high-stakes tasks.

Discussion

The results of our pilot evaluation confirm that the proposed generative AI-based framework provides a reliable and scalable approach to assessment in higher education, particularly for structured analytical tasks. The strong alignment between intelligent tools and educator-assigned scores demonstrates its suitability for automating routine evaluation, while remaining limitations in complex tasks reinforce the need for a hybrid human-AI model.

Compared to existing frameworks, our approach offers several advantages. Smolansky et al. [22] propose a six-dimension model focused on perceptions of quality, but do not operationalise assessment at the task level; our rubric-based approach fills that gap. Similarly, Thanh et al. [12] show that generative AI performs well at lower levels of Bloom's taxonomy, an observation echoed in our findings, though we extend it by explicitly aligning rubric criteria with cognitive levels. Agostini and Picasso [23] emphasise sustainable and scalable feedback, an aspect also incorporated into our model. However, our framework includes full-cycle validation using real exam data. Kolade et al. [24] advocate rethinking assessment for lifelong learning but offer primarily conceptual insights; our work complements this by providing empirical evidence and practical implementation. Perkins et al. (2024) [30] stress ethical integration via their AI Assessment Scale; our framework similarly ensures academic oversight and responsible use. Finally, Khlaif et al. [26] present a model to guide educators in adapting assessments for the generative AI era. Our tested approach provides a ready-to-use structure for those in the "Adopt" or "Explore" phase of that model.

In summary, our framework builds on existing research by combining theoretical foundations with practical validation, supporting the effective and responsible integration of generative AI into assessment processes.

Different stakeholder groups can benefit from our framework for applying generative AI in higher education assessment. Lecturers can use the proposed framework to enhance assessment design by integrating generative AI tools for feedback, grading, and personalised task creation. The model also helps educators in managing workloads and documenting AI-supported innovations for teaching evaluations and institutional reporting. For students, the framework encourages the ethical and transparent use of generative AI in self-assessment, assignment revision, and project development. It fosters digital literacy and empowers learners to take an active and responsible role in AI-enhanced educational settings. Faculty quality assurance committees can adopt the framework as a benchmarking

 doi:10.20944/preprints202505.0800.v1



instrument to support the internal review of assessment practices. It facilitates structured evaluation of assessment formats, the appropriateness of AI tool integration, and alignment with intended learning outcomes. Furthermore, it provides a foundation for developing faculty-specific policies and training programmes to guide responsible AI use in educational evaluation.

At the institutional level, university-level governance bodies, such as academic boards, digital innovation units, or teaching and learning centres, can use the framework to ensure consistency across programmes and to support strategic planning in AI-supported assessment. It serves as a practical tool for monitoring compliance with institutional standards, promoting pedagogical innovation, and ensuring that AI adoption aligns with academic values and digital transformation goals. National accreditation agencies may also benefit from the framework by incorporating its principles into updated evaluation criteria that reflect the rise of digital and AI-enhanced assessment practices. It offers a structured reference for assessing transparency, fairness, and educational effectiveness at the programme and institutional levels. In this context, the framework supports policy development, informs audit procedures, and highlights best practices for maintaining academic quality in the era of generative AI.

## 6. Conclusions and Future Research

In this paper, we present a new conceptual framework for integrating generative AI into university-level assessment practices. The framework is organised into three branches and three levels, corresponding to key stakeholder groups and assessment purposes, respectively. It supports the systematic use of generative AI tools across the full assessment cycle within academic courses. By combining traditional methods with emerging technologies, the framework enables consistent evaluation, targeted feedback, and predictive insights into the effectiveness and objectivism of assessments in digitally enhanced higher education environment.

Building on earlier chatbot-based pedagogical models, the framework extends their functionality to include assessment-specific features such as automated task generation, rubric-based scoring, and real-time feedback delivery. It is designed to enhance transparency, personalise the evaluation process, and support institutional compliance and accreditation. Its structure is flexible and scalable, allowing application across diverse academic programmes and disciplines.

Initial experiments highlight the potential of new AI technology in supporting self-assessment, exam preparation, as well as formative and summative evaluation. Our case study revealed that current AI-based tools can successfully replicate human evaluation practices in accurately assessing student submissions.

We propose several recommendations for university stakeholders aiming to implement this framework: (1) institutional strategies for AI integration should align technological adoption with pedagogical objectives; (2) universities should invest in upskilling both instructors and learners in AI literacy and digital competence; and (3) regulatory and quality assurance bodies should establish clear guidelines for AI-supported assessment practices. Lifelong learning and micro-credentialing systems may also benefit from the framework by offering personalised learning pathways supported by AI analytics.

This study has several limitations. First, our empirical validation is currently restricted to a single academic institution, which restricts the generalisability of the findings. Second, real-time testing of AI-assisted assessments in live class settings has not yet been fully implemented. Third, our dataset is static and does not capture changes in students' assessment behaviour over time. Lastly, the correlation between the level of generative AI adoption and actual learning or teaching outcomes remains unexplored.

Future research will address these limitations by: (1) expanding the scope of the study to include multiple universities and diverse academic disciplines; (2) evaluating the effectiveness of AI-supported assessments through longitudinal analysis; and (3) benchmarking our framework against best practices in international settings. We also plan to investigate the ethical and psychological

 doi:10.20944/preprints202505.0800.v1



dimensions of AI-assisted grading and explore hybrid models where human expertise is combined with machine intelligence for optimal assessment design and delivery.

**Author Contributions:** Conceptualization, G.I. and T.Y.; modelling, G.I., T.Y., M.R., and S.K.; validation, G.I. and T.Y.; formal analysis, T.Y.; resources, G.I., T.Y., M.B., and S.K.; writing—original draft preparation, G.I.; writing—review and editing, G.I. and T.Y.; visualization, G.I and T.Y.; supervision, S.K.; project administration, S.K.; funding acquisition, G.I. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Science Fund, co-founded by the European Regional Development Fund, Grant No. BG05M2OP001-1.002-0002 and BG16RFPR002-1.014-0013-M001, "Digitization of the Economy in Big Data Environment".

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** The authors thank the academic editor and anonymous reviewers for their insightful comments and suggestions.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Van Damme, D.; Zahner, D. (Eds.) Does Higher Education Teach Students to Think Critically? OECD Publishing: Paris, France, 2022. https://doi.org/10.1787/cc9fa6aa-en.
2. UNESCO. Reimagining Our Futures Together: A New Social Contract for Education. Available online: https://www.unesco.org/en/articles/reimagining-our-futures-together-new-social-contract-education (accessed on 20 May 2025).
3. Sevnarayan, K.; Potter, M.A. Generative Artificial Intelligence in Distance Education: Transformations, Challenges, and Impact on Academic Integrity and Student Voice. *J. Appl. Learn. Teach.* **2024**, *7*(1). https://doi.org/10.37074/jalt.2024.7.1.41.
4. World Economic Forum. Future of Jobs Report 2023. Available online: https://www.weforum.org/publications/the-future-of-jobs-report-2023 (accessed on 20 May 2025).
5. Gamage, K.A.; Dehideniya, S.C.; Xu, Z.; Tang, X. ChatGPT and Higher Education Assessments: More Opportunities Than Concerns? *J. Appl. Learn. Teach.* **2023**, *6*(2), 358–369. https://doi.org/10.37074/jalt.2023.6.2.32.
6. ENQA. Working Group Report on Academic Integrity. Available online: https://www.enqa.eu/wp-content/uploads/ENQA-WG-Report-on-Academic-Integrity-.pdf (accessed on 20 May 2025).
7. NEAA. Criteria for Programme Accreditation of Professional Field / Specialty from the Regulated Professions. Available online: https://www.neaa.government.bg/images/Criteria_EN/ENG_Kriterii_za_programna_akreditacija_na_PN-SRP.pdf (accessed on 20 May 2025).
8. European Commission. The European Higher Education Area in 2024: Bologna Process Implementation Report. 2024. Available online: https://eurydice.eacea.ec.europa.eu/publications/european-higher-education-area-2024-bologna-process-implementation-report (accessed on 20 May 2025).
9. Nikolic, S.; Daniel, S.; Haque, R.; Belkina, M.; Hassan, G.M.; Grundy, S.; Sandison, C. ChatGPT Versus Engineering Education Assessment: A Multidisciplinary and Multi-Institutional Benchmarking and Analysis. *Eur. J. Eng. Educ.* **2023**, *48*(4), 559–614. https://doi.org/10.1080/03043797.2023.2213169.
10. Gruenhagen, J.H.; Sinclair, P.M.; Carroll, J.A.; Baker, P.R.; Wilson, A.; Demant, D. The Rapid Rise of Generative AI and Its Implications for Academic Integrity: Students' Perceptions and Use of Chatbots. *Comput. Educ. Artif. Intell.* **2024**, *7*, 100273. https://doi.org/10.1016/j.caeai.2024.100273.
11. Larenas, C.D.; Díaz, A.J.; Orellana, Y.R.; Villalón, M.J.S. Exploring the Principles of English Assessment Instruments. *Ensaio: Aval. Polít. Públicas Educ.* **2021**, *29*, 461–483. https://doi.org/10.1590/S0104-403620210002902851.

12. Thanh, B.N.; Vo, D.T.H.; Nhat, M.N.; Pham, T.T.T.; Trung, H.T.; Xuan, S.H. Race with the Machines: Assessing the Capability of Generative AI in Solving Authentic Assessments. Australas. *J. Educ. Technol.* **2023**, *39*(5), 59–81. https://doi.org/10.14742/ajet.8902.
13. Kane, M. Validity and Fairness. *Lang. Test.* **2010**, *27*(2), 177–182. https://doi.org/10.1177/0265532209349467.
14. ISO 21001:2018. Educational Organizations—Management Systems for Educational Organizations—Requirements with Guidance for Use. Available online: https://www.iso.org/standard/66266.html (accessed on 20 May 2025).
15. Mai, F. Anforderungen an Lerndienstleister und Lerndienstleistungen. In Qualitätsmanagement in der Bildungsbranche, Springer Gabler: Wiesbaden, Germany, 2020. https://doi.org/10.1007/978-3-658-27004-9_8.
16. ENQA. Standards and Guidelines for Quality Assurance in the European Higher Education Area (ESG). 2015. Available online: https://www.enqa.eu/wp-content/uploads/2015/11/ESG_2015.pdf (accessed on 20 May 2025).
17. Ellis, R.; Hogard, E. (Eds.) Handbook of Quality Assurance for University Teaching; Routledge: London and New York, UK/USA, 2019. https://doi.org/10.4324/9781315187518.
18. Kalimullin, A.M.; Khodyreva, E.; Koinova-Zoellner, J. Development of Internal System of Education Quality Assessment at a University. *Int. J. Environ. Sci. Educ.* **2016**, *11*(13), 6002–6013. https://files.eric.ed.gov/fulltext/EJ1115530.pdf.
19. Xia, Q.; Weng, X.; Ouyang, F.; Lin, T.J.; Chiu, T.K. A Scoping Review on How Generative Artificial Intelligence Transforms Assessment in Higher Education. *Int. J. Educ. Technol. High. Educ.* **2024**, *21*(1), 40. https://doi.org/10.1186/s41239-024-00468-z.
20. Xiong, Y.; Suen, H.K. Assessment Approaches in Massive Open Online Courses: Possibilities, Challenges and Future Directions. *Int. Rev. Educ.* **2018**, *64*, 241–263. https://doi.org/10.1007/s11159-018-9710-5.
21. Vetrivel, S.C.; Vidhyapriya, P.; Arun, V.P. The Role of AI in Transforming Assessment Practices in Education. In AI Applications and Strategies in Teacher Education; IGI Global, 2025. https://doi.org/10.4018/979-8-3693-5443-8.ch003.
22. Smolansky, A.; Cram, A.; Raduescu, C.; Zeivots, S.; Huber, E.; Kizilcec, R.F. Educator and Student Perspectives on the Impact of Generative AI on Assessments in Higher Education. In Proceedings of the 10th ACM Conference on Learning@ Scale, Copenhagen, Denmark, 20–22 July 2023; pp. 378–382. https://doi.org/10.1145/3573051.3596191.
23. Agostini, D.; Picasso, F. Large Language Models for Sustainable Assessment and Feedback in Higher Education: Towards a Pedagogical and Technological Framework. *Intelligenza Artificiale* **2024**, *18*(1), 121–138. https://doi.org/10.3233/IA-240033.
24. Kolade, O.; Owoseni, A.; Egbetokun, A. Is AI Changing Learning and Assessment as We Know It? Evidence from a ChatGPT Experiment and a Conceptual Framework. *Heliyon* **2024**, *10*(4), e25953. https://doi.org/10.1016/j.heliyon.2024.e25953.
25. Salinas-Navarro, D.E.; Vilalta-Perdomo, E.; Michel-Villarreal, R.; Montesinos, L. Using Generative Artificial Intelligence Tools to Enhance Experiential Learning for Authentic Assessment. *Educ. Sci.* **2024**, *14*(1), 83. https://doi.org/10.3390/educsci14010083.
26. Khlaif, Z.N.; Alkouk, W.A.; Salama, N.; Abu Eideh, B. Redesigning Assessments for AI-Enhanced Learning: A Framework for Educators in the Generative AI Era. *Educ. Sci.* **2025**, *15*(2), 174. https://doi.org/10.3390/educsci15020174.
27. Williams, P. AI, Analytics and a New Assessment Model for Universities. *Educ. Sci.* **2023**, 13(10), 1040. https://doi.org/10.3390/educsci13101040.
28. Chiu, T.K. Future Research Recommendations for Transforming Higher Education with Generative AI. *Comput. Educ. Artif. Intell.* **2024**, *6*, 100197. https://doi.org/10.1016/j.caeai.2023.100197.
29. Ogunleye, B.; Zakariyyah, K.I.; Ajao, O.; Olayinka, O.; Sharma, H. Higher Education Assessment Practice in the Era of Generative AI Tools. *J. Appl. Learn. Teach.* **2024**, *7*(1). https://doi.org/10.37074/jalt.2024.7.1.28.
30. Perkins, M.; Furze, L.; Roe, J.; MacVaugh, J. The Artificial Intelligence Assessment Scale (AIAS): A Framework for Ethical Integration of Generative AI in Educational Assessment. *J. Univ. Teach. Learn. Pract.* **2024**, *21*(6), 49–66. https://doi.org/10.53761/q3azde36.

 doi:10.20944/preprints202505.0800.v1