

Article

Beyond AI Detection: A Pilot Study of Integrevise™ and Viva-Based Verification of Student Understanding in AI-Mediated Assessment

James Hutson ^{1,*}, Kyle Poyer ², Ebenezer Ogoe ² and Kelvin Adeshola Atologun ²

¹ Art History, AI, and Visual Culture, Lindenwood University, Saint Charles, MO 63301, USA

² Integrevise™, Kingsway, Newport NP20 1HG, UK; ebenezer@integrevise.com (E.O.); kelvin@integrevise.com (K.A.A.)

* Correspondence: jhutson@lindenwood.edu

Abstract

This article examines the Integrevise™ platform through a repeated cross-sectional, multi-cycle pilot case study of viva-based verification in AI-mediated assessment environments. Integrevise™ pairs a submitted written artifact with a short adaptive viva in which students explain their work, reasoning, and application in their own words. Rather than functioning as an AI detector or automated grading system, the platform operates as a diagnostic assessment layer intended to surface comprehension, authorship confidence, and disengagement risk before final grades become the only available signal. The pilot was conducted across Fall 2025 and Spring 2026 at a private liberal arts college in the Midwest; these phases involved different student groups and are therefore treated as iterative implementation cycles rather than a longitudinal cohort. Results should be interpreted as preliminary pilot evidence. In Spring 2026, 52 vivas were completed, but formal student survey data were limited to seven respondents and showed mixed perceptions: only 14.3% agreed that the oral assessment helped them think more deeply about the assignment, whereas 57.1% disagreed or strongly disagreed. Platform feedback was also incomplete, with 20 of 52 vivas (38.5%) producing no student feedback record. Qualitative feedback, tutor observations, and implementation notes nevertheless suggest that viva-based verification may help identify some comprehension gaps and implementation barriers that written artifacts alone may not reveal. The findings, therefore, support continued investigation of Integrevise™ as a process-rich assessment intervention, but not broad claims of efficacy or scalability without larger, more systematic validation.

Keywords: Integrevise™; AI assessment; viva-based assessment; academic integrity; authentic assessment



Academic Editor: Hani Morgan

Received: 5 June 2026

Revised: 2 July 2026

Accepted: 3 July 2026

Published: 6 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

In May 2026, Princeton University's faculty voted to require proctoring for all in-person examinations beginning in summer 2026, ending a 133-year tradition of unproctored exams under its Honor Code in response to concerns that generative AI had made cheating easier to commit and more difficult to observe [1]. The significance of this decision is not limited to Princeton; it functions as a highly visible marker of a wider institutional reckoning in which even long-standing trust-based assessment cultures are being reconsidered under the pressure of AI-enabled academic misconduct. Reporting on the decision noted faculty concerns that students could use AI or other digital supports during laptop-based exams,

while survey data cited in the same coverage suggested that 30% of Princeton seniors had admitted cheating on an assignment or examination and nearly half knew of an Honor Code violation, although fewer than 1% reported one. This story captures the core problem facing contemporary assessment: generative AI has not simply added a new cheating mechanism; it has destabilized the assumptions by which educational systems infer learning from submitted work.

The problem is especially urgent because generative AI has made knowledge-like production fast, inexpensive, fluent, and difficult to attribute. Traditional assessment depends on a chain of inference in which an essay, examination response, code file, project report, or reflection is treated as evidence of a student's knowledge, reasoning, judgment, and effort. That chain weakens when a submitted artifact may result from independent cognition, AI-supported drafting, external editing, prompt iteration, paraphrasing tools, or some combination of these practices. Cotton, Cotton, and Shipway [2] argue that ChatGPT creates new pressures for academic integrity because AI-generated text does not fit traditional plagiarism models. Scarfe et al. [3] demonstrate the severity of the validity problem empirically: 94% of AI-generated submissions inserted into a real university assessment system went undetected, and the AI submissions often received stronger marks than genuine student submissions. HEPI's 2026 student survey further shows that use has become normalized rather than exceptional, with 94% of students reporting that they used such tools to help with assessed work and 95% reporting use of AI in at least one way [4]. These findings indicate that assessment can no longer presume unaided production as the default condition of student work.

Institutional responses have often moved in three directions, none of which is sufficient by itself. The first response is surveillance, including proctoring, lockdown browsers, AI-detection software, and behavioral monitoring. These tools may deter some forms of misconduct, but they can produce privacy concerns, procedural distrust, and inequitable false positives. Liang et al. [5] found that GPT detectors disproportionately misclassified writing by non-native English writers as AI-generated, while Weber-Wulff et al. [6] concluded that detection systems lacked the reliability required for definitive academic integrity judgments. The second response is a retreat to closed-book, in-person, or hand-written assessments. Such approaches may verify some forms of independent performance, but they risk narrowing the evidence of learning to what can be reproduced under artificial time pressure. The third response is AI-inclusive assessment, which recognizes that generative tools are now part of academic and professional practice and therefore asks instructors to specify allowable use, assess process, and evaluate judgment. Perkins et al. [7] and Furze et al. [8] show the value of structured AI-use frameworks, yet such frameworks still require mechanisms for verifying that students understand and can defend the work they submit.

This article positions IntegreviseTM as a response to that verification problem. The platform is not an AI detector, plagiarism tool, or autonomous grading system (Figure 1). Rather, it operates as a viva-based assessment layer that pairs a submitted written artifact with a short adaptive oral exchange in which students explain their reasoning, clarify their choices, and demonstrate understanding in their own words. The platform addresses the weakness of artifact-only assessment by shifting part of the evidentiary burden from polished product to demonstrated comprehension. It also avoids the central limitation of detection-based approaches because the primary question is not whether AI was used, but whether the student can explain, apply, and defend the knowledge represented in the submission. In this respect, IntegreviseTM aligns with emerging work on conversational and oral assessment in the AI era. Ipeirotis and Rizakos [9] describe scalable voice-AI oral assessment as a way to test genuine understanding, while Barba and Stegner [10] argue

that conversational examinations can restore evidentiary validity by requiring students to explain thinking in real time.

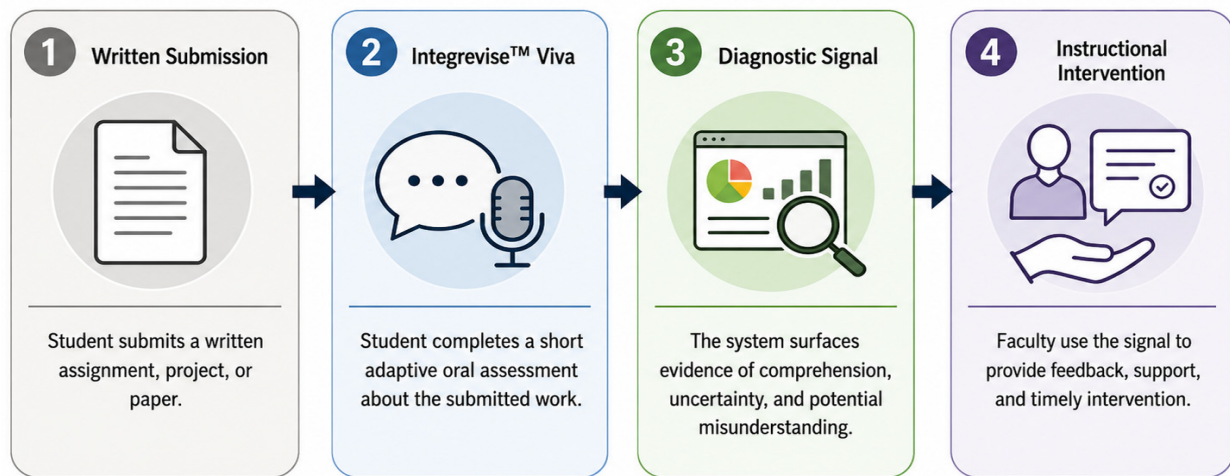


Figure 1. Integrevise™ as a Four-Stage Assessment Process.

The Integrevise™ pilot was implemented as a repeated cross-sectional, multi-cycle pilot case study across Fall 2025 and Spring 2026. The Fall 2025 phase functioned as an exploratory implementation, while the Spring 2026 phase served as a revised deployment informed by postmortem findings. Because the two phases involved different groups of students, the study does not track individual-level change over time and should not be understood as longitudinal in the strict methodological sense. Fall findings revealed both the pedagogical rationale for viva-based verification and the barriers that limit adoption when a platform is introduced late or explained unclearly. Postmortem findings showed that students experienced confusion about whether the tool was grading them, whether it was evaluating them through AI, and how the viva related to their coursework. Additional barriers included pacing issues, inconsistent faculty setup, limited student-facing feedback, and difficulty integrating the platform into Canvas workflows. These findings were not treated as evidence that viva-based assessment lacked value; instead, they clarified the implementation conditions required for trust, validity, and usability. The Spring 2026 phase therefore introduced design and communication revisions, including clearer onboarding, push-to-talk control, visible transcription, refined question flow, improved feedback delivery, and a more explicit formative-to-summative assessment sequence.

The Spring 2026 evidence was more limited and mixed than a simple efficacy claim would suggest. Platform records showed 52 completed vivas, with 22 positive feedback records, 10 negative feedback records, and 20 sessions without feedback. The formal student survey included only seven respondents; one respondent (14.3%) agreed that the viva helped them think more deeply about the assignment, while four respondents (57.1%) disagreed or strongly disagreed. These data do not support broad claims that most students experienced stronger awareness or that the platform demonstrably prevents disengagement. Instead, they support a narrower pilot-stage interpretation: Integrevise™ may produce diagnostically useful evidence in some cases, while student perceptions, missing response data, technical usability, and implementation context substantially shape interpretation. Tutor observations suggested that some viva outputs surfaced amber and red indicators of weak comprehension, incorrect application, or disengagement not fully visible in written submissions alone. One case revealed that a student had been using the wrong supplemental material for weeks, a problem that conventional final grading likely would have identified too late for effective intervention. The study therefore proceeds

from a broader crisis in AI-era assessment to a focused pilot analysis of Integrevise™ as a candidate verification intervention. Its methodological contribution lies in examining whether adaptive viva-based verification can preserve additional evidence of learning while acknowledging that generative AI has permanently altered the conditions of academic production.

2. Literature Review: AI and Assessment in Education, 2023 to 2026

2.1. Key Trends in AI and Assessment

Scholarship from the past three years, starting with the rapid proliferation of AI use, such as ChatGPT 3.5 Turbo launched in February 2023, has shifted educational assessment from a primarily procedural concern about cheating toward a deeper validity problem concerning what student work can now be said to evidence (Table 1). Earlier assessment models assumed that the production of essays, solutions, summaries, code, reflections, and research syntheses required a sustained relationship between student cognition and submitted artifact, yet generative AI destabilizes that relationship by making fluent knowledge-like production fast, inexpensive, and widely accessible. Swiecki et al. [11] anticipated this transition by arguing that assessment in the age of AI would require a rethinking of what evidence is collected, how that evidence is interpreted, and how human judgment remains meaningfully involved. Chan [12] subsequently framed the technology as a system-level policy issue rather than a classroom-level anomaly, identifying academic integrity, privacy, governance, staff readiness, and pedagogical adaptation as interdependent concerns. The pace of student adoption underscores the magnitude of this shift: Freeman [13] reported that 88% of surveyed students used these generative tools for assessed work, while Stephenson and Armstrong [4] reported that 94% used them to help with assessed work and 95% used AI in at least one way. These figures do not indicate that all use is misconduct, but they do show that assessment can no longer operate on the assumption that unaided production is the default condition of student work. The most important change, therefore, is conceptual: institutions must move from asking only whether AI was used to asking what form of learning, judgment, authorship, transfer, and disciplinary competence the assessment is intended to verify. This literature consequently converges around a central claim: assessment systems built for scarcity of information and effortful knowledge production must be redesigned for an environment of abundant, fluent, and partially automated production. The Higher Education Policy Institute’s (HEPI) 2026 report describes AI use among undergraduates as “near universal,” and its 2025 report explicitly recommends stress-testing assessments to determine whether they can be completed easily using, especially, generative writing tools.

Table 1. Core assessment challenges in the age of generative AI.

Challenge	Consideration	Assessment Implication	Sources
Artifact-only evidence is no longer sufficient	AI can generate or substantially revise essays, code, summaries, reflections, and reports with limited student cognition.	Final submissions must be supplemented with evidence of reasoning, process, and transfer.	[2,3]
AI use is now normalized	Student use of generative AI for assessed work is widespread rather than exceptional.	Assessment design should assume AI access and specify acceptable, prohibited, or required uses.	[4,13]

Table 1. Cont.

Challenge	Consideration	Assessment Implication	Sources
Detection is unreliable and inequitable	AI detectors produce false positives, struggle with edited or hybrid writing, and may misclassify multilingual writers.	Detection should not serve as the primary basis for academic integrity decisions.	[5,6]
Performance can mask weak learning	AI can improve fluency and organization without improving understanding, judgment, or retention.	Assessments must verify whether students can explain, apply, and defend their work.	[14,15]
Integrity policies remain ambiguous	Students and faculty often lack shared definitions of appropriate AI assistance.	Clear task-level AI-use guidance is needed for fairness, consistency, and validity.	[7,12]
Secure exams alone are insufficient	Proctored or closed-book formats may verify independent performance but can narrow authenticity.	AI-era assessment requires balance between secure checkpoints and authentic disciplinary work.	[1,3]
Authentic assessment must be redesigned	Real-world tasks increasingly involve AI, which complicates traditional notions of independent authorship.	Students should document process, justify choices, and demonstrate competent AI-mediated judgment.	[7,8]
Scalable verification is the central challenge	Vivas, oral defenses, portfolios, and process reviews improve validity but can increase workload.	Institutions need hybrid systems that make understanding visible without overwhelming faculty.	[9,10]

2.2. From Academic Integrity Panic to Evidentiary Validity

The earliest institutional response to OpenAI's ChatGPT and related tools often treated generative AI as a novel academic integrity threat, but the scholarship increasingly shows that the more durable challenge concerns the evidentiary logic of assessment. Cotton, Cotton, and Shipway [2] argued that ChatGPT, in particular, forced higher education to reconsider academic integrity practices because generative text can produce original-seeming work that evades traditional plagiarism detection. Rudolph, Tan, and Tan [16] similarly positioned ChatGPT as a direct challenge to traditional written assessments, especially when institutions continue to privilege short essays and take-home assignments as proxies for student understanding. Sullivan, Kelly, and McLaughlan [17] emphasized that integrity concerns cannot be separated from student learning, because overly punitive responses may suppress legitimate experimentation while permissive ambiguity may encourage substitution of AI output for student cognition. Scarfe et al. [3] provided one of the clearest empirical demonstrations of this problem by injecting AI-generated submissions into an actual university assessment system and finding that 94% of submissions went undetected while often achieving grades above real student work. Yeadon et al. [18] reached a similar conclusion in physics, showing that AI-generated short essays could achieve first-class marks in an existing assessment format. Nikolic et al. [19] extended the problem into engineering education, demonstrating that the performance of ChatGPT across assessment contexts required multidisciplinary reconsideration of assessment integrity. Together, these studies indicate that the submitted artifact has become a weaker proxy for student mastery unless the task design requires process evidence, situated judgment, oral verification, iterative production, or supervised demonstration. The question is no longer whether students can use AI to complete assessments; the evidence indicates that they can. The more difficult question is whether assessment can be redesigned to distinguish polished output

from learned competence. For instance, Scarfe et al. [3] explicitly framed AI infiltration as a fundamental problem for unsupervised coursework and take-home examinations.

2.3. Detection, False Positives, and the Failure of Policing as a Primary Strategy

The literature on AI detection shows that detection tools may have limited investigative or formative value, but they cannot provide a stable foundation for high-stakes academic integrity enforcement. Ardito [20] argued that generative detection is misaligned with contemporary higher education because the technology is unreliable, susceptible to evasion, and inconsistent with the growing legitimacy of AI-supported learning. OpenAI's own experience with AI text detection reinforces this limitation. In January 2023, OpenAI released a classifier intended to indicate whether text was AI-written, but the organization later made the tool unavailable because of its low rate of accuracy [21]. Even before withdrawal, OpenAI cautioned that the classifier was "not fully reliable," correctly identifying only 26% of AI-written text in its challenge set while incorrectly labeling 9% of human-written text as AI-written [21]. The same source warned that the classifier should not be used as a primary decision-making tool and noted that AI-written text could be edited to evade detection [21]. This institutional example strengthens the broader assessment argument: if even a leading AI developer could not provide a sufficiently reliable detector for high-stakes educational decisions, then assessment strategy should move from identifying whether AI was used toward verifying whether students can understand, explain, and defend the work they submit.

Liang et al. [5] demonstrated an equity problem that is especially important for diverse institutions: GPT detectors disproportionately misclassified writing by non-native English writers as AI-generated, creating a procedural risk for multilingual students. Weber-Wulff et al. [6] tested multiple AI-generated text detection systems and concluded that they were neither sufficiently accurate nor sufficiently reliable, particularly when texts were paraphrased, translated, or otherwise obfuscated. Elkhataat, Elsaid, and Almeer [22], likewise, found uneven performance across AI detection systems, reinforcing the conclusion that no detector should be treated as a definitive adjudicator of misconduct. Perkins et al. [23] showed that adversarial techniques could reduce already weak detector accuracy, which further undermines the use of detectors as punitive instruments. Sadasivan et al. [24] made the broader technical argument that reliable generative text detection is unstable because model outputs can be revised, paraphrased, human-edited, or generated by models outside the training assumptions of a detector. Public reporting has followed the same pattern: Goodier [25] reported rising proven AI-cheating cases in UK universities but also noted that many institutions did not separately track such misuse and that experts viewed recorded cases as a partial measure of a much larger phenomenon. A detection-centered policy, therefore, risks both under-enforcement and over-enforcement: sophisticated misuse may pass unnoticed, while innocent students, especially multilingual writers, may face false accusations. A more defensible approach treats detection as one weak signal among many and prioritizes assessment designs that make learning processes visible. Liang et al. [5] summarize the equity concern directly by noting that detectors frequently misclassify non-native English writing as AI-generated.

2.4. AI-Inclusive Policy Frameworks and Assessment Transparency

By 2024, a significant strand of scholarship had shifted from prohibition toward structured inclusion, especially through frameworks that specify what kinds of use are allowed for particular learning outcomes. Perkins et al. [7] proposed the AI Assessment Scale as a practical framework for aligning permitted use with the purpose of assessment, thereby transforming AI from an invisible integrity threat into an explicit design variable. Furze

et al. [8] then piloted the AI Assessment Scale and reported institutional benefits, including reduced GenAI-related academic misconduct cases, improved student attainment, and more multimodal assessment practices, although those findings require replication across institutional contexts and disciplines. Perkins, Roe, and Furze [26] later revised the scale to clarify its relationship to assessment validity, learning design, and social constructivist principles. Chan [12] complements these assessment-specific frameworks by proposing an ecological policy approach organized around pedagogy, governance, and operations, which is useful because assessment reform cannot succeed without staff development, student training, privacy protocols, tool access, and clear communication. Moreover, Wang et al. [27] found that many highly ranked universities adopted open but cautious approaches, emphasizing ethical use, privacy, accuracy, syllabus guidance, and multifaceted evaluation rather than simple bans. Ghimire and Edwards [28] similarly identified a policy gap in educational institutions, arguing that many policies remain underdeveloped relative to the speed of adoption. As such, UNESCO [29] offered international guidance that emphasizes human agency, teacher support, transparency, inclusion, and policy-level governance for generative AI in education and research. The central assessment implication is that students need precise, task-level guidance: whether AI is prohibited, permitted for brainstorming, allowed for drafting, required for comparison, or assessed as part of professional workflow. Without such specificity, students may experience inconsistent norms across courses, while faculty may unintentionally produce contradictory expectations within the same institution. The policy literature therefore supports assessment transparency as a validity condition, not merely as a compliance practice. Chan's framework identifies pedagogical, governance, and operational dimensions as interdependent elements of AI policy in university teaching and learning.

2.5. Learning, Cognitive Offloading, and False Performance

A second major body of literature concerns the difference between enhanced performance and actual learning. These tools, which expanded to include Anthropic's Claude and Google's Gemini, can improve the surface quality of student work by strengthening organization, fluency, grammar, formatting, idea generation, and citation-like presentation, but those improvements do not necessarily demonstrate durable knowledge, independent reasoning, or transfer. Fan et al. [14] framed this as a problem of metacognitive laziness, finding that the tools can alter learning motivation, process engagement, and performance in ways that complicate the interpretation of assessment outcomes. Georgiou [30] similarly argued that ChatGPT use in particular may reduce cognitive engagement when learners outsource demanding stages of argument formation, although such claims require continued empirical testing across disciplines and assessment types. Lo [31], in a rapid review of ChatGPT's educational implications, emphasized both pedagogical potential and risks related to accuracy, overreliance, academic integrity, and the need for new assessment approaches. Kasneci et al. [15], likewise, presented large language models (LLMs) as educationally promising but risky, especially when students use them without sufficient metacognitive control or verification habits. Not surprisingly then, Li et al. [32] identified academic integrity, assessment, and data protection as recurring ethical concerns in the early ChatGPT literature, noting that the field still needed more empirical research. Student surveys by HEPI make this tension visible in student language: some students describe AI as enabling deeper analysis by saving time, while others describe it as making them feel that they are not using their own minds [4]. The assessment challenge is that both experiences may produce similar-looking final submissions. Therefore, a valid AI-era assessment must collect evidence not only of product quality, but also of problem formulation, verification, revision, explanation, and reflective judgment. This is why process-rich assessment is not

a cosmetic redesign; it is a necessary response to false performance, where the artifact improves while the underlying learning remains uncertain.

2.6. Automated Assessment, Feedback, and the Risks of Efficiency

These tools, nevertheless, are not only discussed in terms of academic dishonesty. AI also appears in the literature as a possible assessment support system, especially through automated scoring, formative feedback, question generation, and learning analytics. Gao et al. [33] reviewed automatic assessment of text-based responses in post-secondary education and found that natural language processing (NLP) systems can support feedback and scoring for open-ended responses, particularly in contexts where instructor workload limits the feasibility of frequent formative assessment. Yan et al. [34] identified a wide range of LLM use cases in education, including feedback, grading, detection, content generation, teaching support, and recommendation, while also highlighting challenges around technological readiness, transparency, replicability, privacy, and beneficence. Dempere et al. [35] discussed the impact on higher education more broadly, emphasizing that its use in learning environments produces institutional opportunities as well as risks. Gill et al. [36] similarly argued that ChatGPT and related systems require updated academic regulations and evaluation practices because the tools can support instruction while also generating inaccurate information and circumventing conventional plagiarism detection. The efficiency promise is, therefore, real but ambiguous. Automated feedback may make large-scale formative assessment more feasible, yet automated scoring can narrow constructs if institutions begin assessing only what machines can reliably parse. In writing-intensive, humanities, design, clinical, and professional fields, assessment often depends on audience awareness, ethical positioning, disciplinary judgment, creativity, and contextual interpretation, which are difficult to reduce to automated measures. Human oversight remains essential because generated feedback can be plausible but shallow, fluent but incorrect, and confident without sufficient grounding. The literature consequently supports using AI as an assessment aid rather than an autonomous assessment authority. In this model, AI can help generate formative feedback, compare drafts, identify gaps, and scaffold revision, but the summative judgment of learning remains tied to transparent criteria, human expertise, and evidence of student agency.

2.7. Authentic Assessment, Secure Verification, and Process Evidence

Because of these developments, the post-2023 literature increasingly recommends authentic assessment, but the concept requires refinement because workplace authenticity now often includes AI use. Traditional calls for authentic assessment emphasized real-world relevance, complex problem-solving, audience awareness, and situated performance, yet generative AI complicates this model because many real-world tasks can now be partially automated. Mahmoud [37] points toward one possible response by examining open-book assessment with ChatGPT in which interaction traces can become part of the evidence base. Far from merely permitting use, such approaches ask students to show how they prompt, test, critique, revise, verify, and justify AI-mediated work. Perkins et al. [7] similarly position AI-inclusive assessment as a design problem, where educators determine which cognitive processes must be protected from machine assistance and which AI-mediated capabilities should be assessed as emerging professional skills. Weale [38] reported that UK universities were warned to “stress-test” assessments because use had become widespread, and Freeman [13] made that stress-testing recommendation explicit in HEPI’s report. Goodier [25] added a complementary public accountability dimension by reporting that nearly 7000 proven AI-related cheating cases were recorded in UK higher

education in 2023 to 2024, while also noting that many institutions did not yet consistently track AI misuse.

The University of Sydney's two-lane approach provides a useful program-level frame for this shift from prohibition toward structured assurance. Bridgeman, Liu, and Weeks [39] distinguish between Lane 1 assessments, which validate attainment of learning outcomes through secured assessment, and Lane 2 assessments, which support learning in more open environments where generative AI may be scaffolded and used responsibly. This distinction is valuable because it avoids the false binary between banning AI and permitting it without verification. Instead, it positions secure demonstration of capability as a deliberate program-level design choice while allowing AI-supported learning activities to develop students' capacity to work productively and ethically with emerging tools. IntegreviseTM aligns most closely with the Lane 1 assurance function when it is used to verify that students can explain and apply the knowledge represented in their submissions, yet it can also support Lane 2 learning when deployed formatively to help students recognize gaps in understanding before final assessment. The two-lane model therefore strengthens the conceptual basis for the present study by situating viva-based verification within a broader assessment architecture for AI-enabled education.

These developments suggest that assessment reform should combine authentic tasks with secure verification, rather than treating authenticity and security as opposites. Secure components may include in-class writing, oral defense, supervised problem-solving, practical demonstrations, viva-style questioning, version histories, annotated drafts, and reflective process logs. Authentic components may include AI-supported workplace simulations, client briefs, disciplinary case analyses, design sprints, research memos, and multimodal deliverables. The strongest model is therefore hybrid: students complete meaningful tasks, use AI according to clearly stated rules, and provide evidence of learning through checkpoints that are difficult to outsource. Guardian reporting in 2025 connected the rise in student AI use to calls for assessment stress-testing and later documented increasing confirmed AI-misconduct cases across UK universities [25,38].

2.8. Gaps in the Literature and Limits of Current Approaches

Despite rapid growth in the literature, the evidence base remains uneven, reactive, and insufficiently cumulative. Many influential studies are conceptual, exploratory, single-institution, discipline-specific, preprint-based, or focused on immediate detection and integrity concerns rather than durable learning outcomes. Scarfe et al. [3], Yeaton et al. [18], and Nikolic et al. [19] provide important demonstrations of vulnerability, but the field needs more studies comparing redesigned assessments against legacy assessments using measures of learning, equity, reliability, workload, and transfer. Furze et al. [8] offer promising evidence for structured AI-inclusive assessment, yet that work requires broader replication across institution types, class sizes, disciplines, and student populations. The literature also lacks a mature validity theory for AI-mediated assessment. Existing frameworks rarely specify with sufficient precision when use of the tools invalidates evidence, when it supports learning, when it should be considered part of the construct, and when unaided performance must be protected because the learning outcome requires independent mastery. Equity gaps remain significant: Liang et al. [5] exposed bias against non-native English writers in detectors, but less research addresses neurodivergent learners, first-generation students, multilingual students, economically disadvantaged students, students with paid AI access versus free-tier access, and learners in performance-based disciplines.

Another unresolved limitation is scalability. Oral defenses, process portfolios, iterative checkpoints, and individualized feedback can increase validity, but they can also increase faculty workload unless institutions develop shared rubrics, calibrated sampling, peer

review structures, AI-supported formative workflows, and program-level assessment architectures. Current approaches, therefore, remain caught between two insufficient poles: detection-heavy enforcement that cannot reliably prove authorship and exam-heavy retreat that may verify recall while weakening authenticity. The needed methodological turn is toward scalable authentic assessment models that preserve credible evidence of learning under conditions of ubiquitous generative AI. The present pilot addresses this gap by examining Integrevise™ as a repeated cross-sectional, multi-cycle implementation of viva-based verification, with attention to both its diagnostic promise and the methodological constraints that shape interpretation.

3. Methodology

This study used a repeated cross-sectional, multi-cycle pilot case study design to evaluate Integrevise™ as a viva-based assessment layer for verifying student understanding in AI-mediated assessment environments. The pilot was conducted at a private liberal arts college in the Midwest across Fall 2025 and Spring 2026, with the first phase functioning as an exploratory implementation and the second phase serving as a revised deployment informed by postmortem findings. The Fall 2025 and Spring 2026 phases involved different groups of students; consequently, the study did not follow the same participants over time and does not support individual-level longitudinal claims. The study included courses delivered across on-ground, hybrid, and fully online modalities to examine whether the platform could operate across varied instructional conditions. Integrevise™ was implemented after written submissions, prompting students to complete a short adaptive viva in which they explained, clarified, and applied ideas from their own work. The platform was not evaluated as an AI detector, plagiarism checker, or autonomous grading system; rather, it was studied as a diagnostic verification mechanism designed to make student comprehension, reasoning, and possible disengagement risk more visible than written artifacts alone (Figure 2).

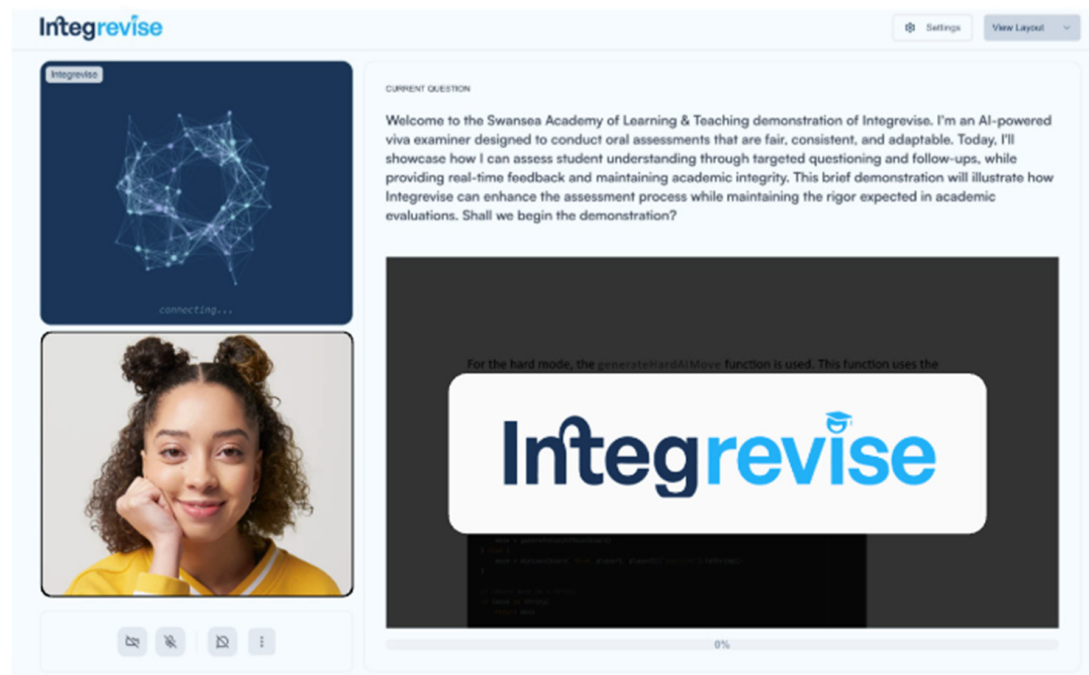


Figure 2. Main viva screen, included for illustrative purposes only and not treated as research evidence.

Data collection included student surveys, platform-generated usage and feedback records, implementation notes, postmortem analysis, tutor observations, and targeted student interviews (Table 2). Faculty-facing items were piloted to support implementation reflection, but faculty survey responses were not analyzed as outcome data and are therefore not reported in the Results. Student surveys captured usability, clarity of instructions, perceived fairness, anxiety or trust concerns, feedback use, and whether the viva helped students recognize what they understood or did not understand. Targeted student interviews were used to clarify patterns that could not be fully interpreted from survey or platform data, especially student confusion about AI-mediated evaluation, resistance to the tool, perceived usefulness, and shifts in self-awareness after explaining work aloud. This triangulated approach allowed the study to examine measurable implementation patterns alongside qualitative experiences across different learning contexts, while preserving the distinction between formal survey results, voluntary feedback records, interview comments, and tutor observations (Figure 3).

Table 2. Spring 2026 student survey instrument used in the Integrevise™ pilot.

Respondent Group	Item No.	Survey Item	Response Format/Options
Student	1	What is your major or degree program?	Open-ended
Student	2	What is your academic standing?	Freshman; Sophomore; Junior; Senior; Graduate
Student	3	What is the course title?	Open-ended
Student	4	Who is your instructor?	Open-ended
Student	5	Did the presence of the viva influence how you prepared or approached your assignment?	Strongly Agree; Agree; Neutral; Disagree; Strongly Disagree
Student	6	Did completing the oral assessment help you think more deeply about the assignment?	Strongly Agree; Agree; Neutral; Disagree; Strongly Disagree
Student	7	Did the oral assessment help you reflect on and deepen your understanding after submitting the assignment?	Strongly Agree; Agree; Neutral; Disagree; Strongly Disagree
Student	8	How often did you review the AI-generated feedback?	Always; Most of the time; Sometimes; Never
Student	9	Was the viva experience easy to navigate?	Strongly Agree; Agree; Neutral; Disagree; Strongly Disagree
Student	10	Was the oral assessment a fair way to verify understanding?	Strongly Agree; Agree; Neutral; Disagree; Strongly Disagree
Student	11	What did you find most helpful or most frustrating about the viva?	Open-ended

Survey items were developed by the research team for the pilot rather than drawn from a validated scale. Items were aligned with the implementation questions that guided the study: preparation effects, perceived depth of thinking, reflective value, feedback use, usability, fairness, trust, and open-ended frustration or benefit. No formal validity study, reliability analysis, or factor analysis was conducted; given the small Spring 2026 survey sample ($n = 7$) and heterogeneous item formats, such analyses would not have been meaningful. Analysis therefore remained descriptive and exploratory. Quantitative results were reported as counts and percentages for transparency, but percentages were

not interpreted inferentially. Qualitative comments, interviews, implementation notes, and tutor observations were coded thematically using categories derived from the pilot aims and recurring response patterns, including communication clarity, feedback adequacy, question relevance, platform usability, AI trust, perceived fairness, and diagnostic value. Fall 2025 findings identified several implementation barriers, including late introduction, unclear student messaging, uncertainty about whether AI was grading students, pacing issues, limited student-facing feedback, and inconsistent integration into Canvas workflows. These findings directly informed Spring 2026 refinements, including clearer onboarding, improved student support materials, push-to-talk control, visible live transcription, refined question flow, improved feedback delivery, and a stronger formative-to-summative sequence. The Spring 2026 deployment then provided the primary evidence for the revised model, including 52 completed vivas, 22 positive feedback records, 10 negative feedback records, 20 sessions without feedback, and seven formal student survey responses. The analysis focused on whether Integrevise™ offered preliminary diagnostic value by revealing student understanding, identifying weak comprehension earlier in selected cases, and clarifying the implementation conditions required for future assessment validation.

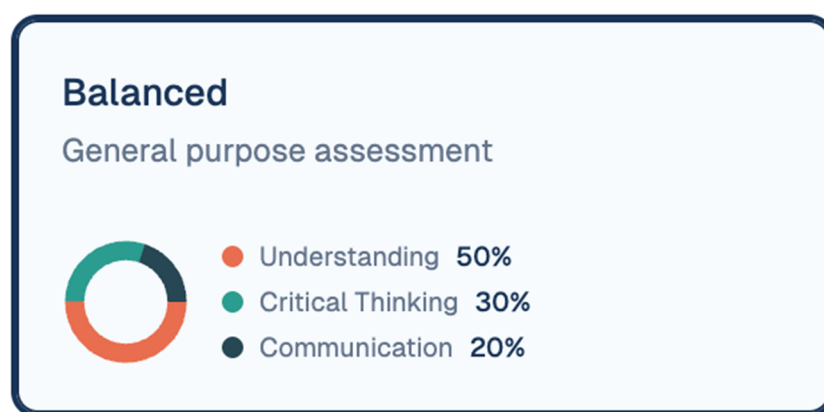


Figure 3. Viva settings, included for illustrative purposes only and not treated as research evidence.

4. Results

The results indicate that Integrevise™ functioned most plausibly as a diagnostic and reflective assessment layer, although student perceptions remained mixed and were highly dependent on communication, timing, and platform experience. Across the multi-cycle pilot, Fall 2025 revealed substantial implementation friction, while Spring 2026 showed a revised deployment after changes were made to onboarding, question flow, student-facing feedback, and the formative-to-summative sequence. The clearest pattern across the evidence is that students were more likely to recognize the value of the viva when they understood its purpose: it was not an AI detector, an automated grader, or a punitive integrity screen but a mechanism for verifying whether students could explain and apply the work they submitted (Figure 4). In the targeted Fall 2025 student interviews, three of four interviewees initially expressed frustration or confusion, yet all four acknowledged the potential value of the tool after the purpose was explained. Because this interview set was very small, these figures are treated as implementation observations rather than prevalence estimates. The same interview data identified communication and transparency gaps in all four interviews, missing or inadequate feedback in all four interviews, question relevance issues in three interviews, pacing or timing issues in three interviews, and broader AI skepticism or trust concerns in three interviews. These findings established the major implementation problem for the pilot: the educational logic of Integrevise™ was generally

defensible, but its value was weakened when students experienced it as unexplained, disconnected from the assignment workflow, or ambiguously related to grading.

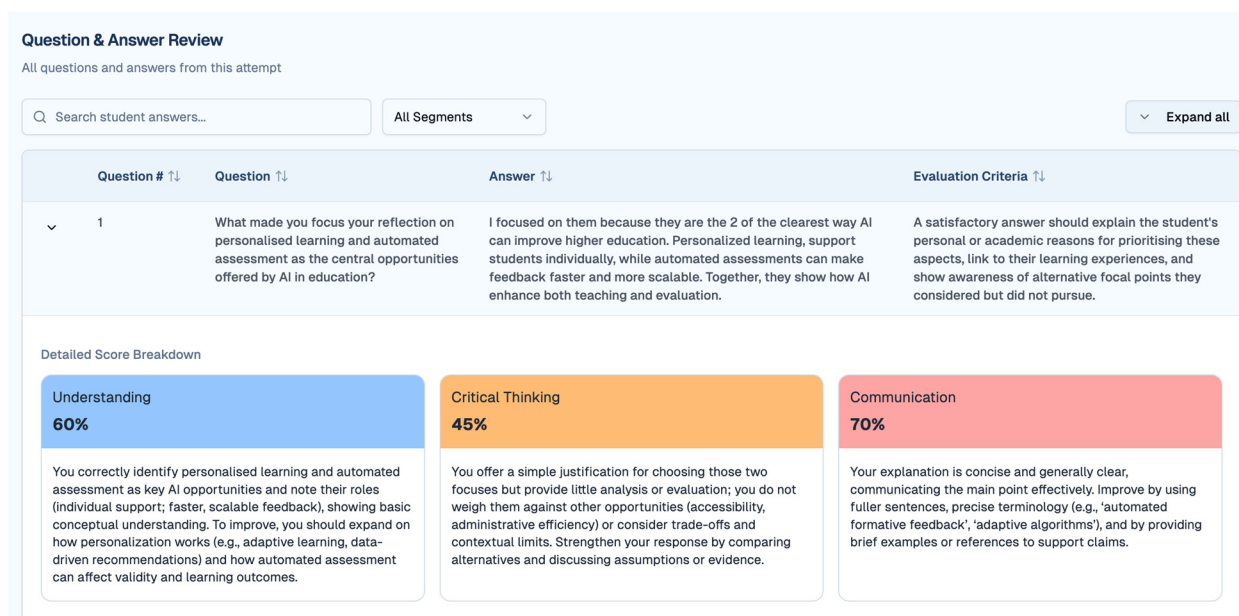


Figure 4. Response rating screen, included for illustrative purposes only and not treated as research evidence.

The Spring 2026 deployment provided the primary quantitative evidence of revised platform use, but this evidence should be interpreted descriptively. Platform records showed 52 completed vivas, with 22 positive feedback records, 10 negative feedback records, and 20 sessions with no feedback submitted. Expressed as proportions of total completed vivas, 42.3% generated positive feedback, 19.2% generated negative feedback, and 38.5% generated no feedback. The missing feedback cases materially affect interpretation. If students who did not submit feedback were less engaged, more dissatisfied, technically interrupted, or uncertain about the purpose of the viva, the observed positive and negative distribution could overstate acceptance; if nonrespondents omitted feedback for reasons unrelated to experience, the missingness may be less consequential. Because the pilot did not capture the reason for missing feedback, these 20 cases are treated as an important source of uncertainty rather than neutral absences. The qualitative feedback gives additional texture to the 32 sessions with student feedback. Positive comments emphasized reflection, deeper questioning, and the value of being prompted to think beyond the written submission. Representative themes included students noting that the viva helped them reflect on their paper, asked questions that went more deeply into the assignment, and required them to consider what their argument would mean under alternate conditions. Negative comments clustered around AI resistance, technical interruption, freezing, questions perceived as poorly matched to the assignment, voice or text lag, and discomfort with AI in academic contexts. The results, therefore, show neither simple endorsement nor simple rejection. Instead, they show conditional acceptance among respondents: students were most receptive when the viva felt relevant, technically stable, clearly explained, and visibly connected to learning rather than surveillance.

Student survey results reinforced the mixed nature of the experience and should be read in light of the very small formal survey sample. In the Spring 2026 survey sample ($n = 7$), five respondents (71.4%) reported that the use of vivas did not influence how they approached the assignment, while one respondent (14.3%) agreed that it did. Similarly,

only one respondent (14.3%) agreed that completing the oral assessment helped them think more deeply about the assignment, while four respondents (57.1%) disagreed or strongly disagreed and two respondents (28.6%) remained neutral. On the item asking whether the oral assessment helped students reflect on and deepen their understanding after submitting the assignment, one respondent (14.3%) agreed, three respondents (42.9%) were neutral, and three respondents (42.9%) disagreed or strongly disagreed. These data show that reflective benefit was not experienced uniformly across the Spring survey respondents (Figure 5). However, the feedback-review item was more encouraging: five respondents (71.4%) reported reviewing the AI-generated feedback always or most of the time, while two respondents (28.6%) reported never reviewing it. The usability and fairness items also remained mixed. For platform ease of use, two respondents (28.6%) agreed or strongly agreed, two respondents (28.6%) were neutral, and three respondents (42.9%) disagreed or strongly disagreed. For fairness as a way to verify understanding, one respondent (14.3%) strongly agreed, three respondents (42.9%) were neutral, and three respondents (42.9%) disagreed or strongly disagreed. These results indicate that the refined model improved access to feedback relative to the Fall interview findings, but that student trust, perceived relevance, and user experience remained unresolved variables requiring continued refinement.

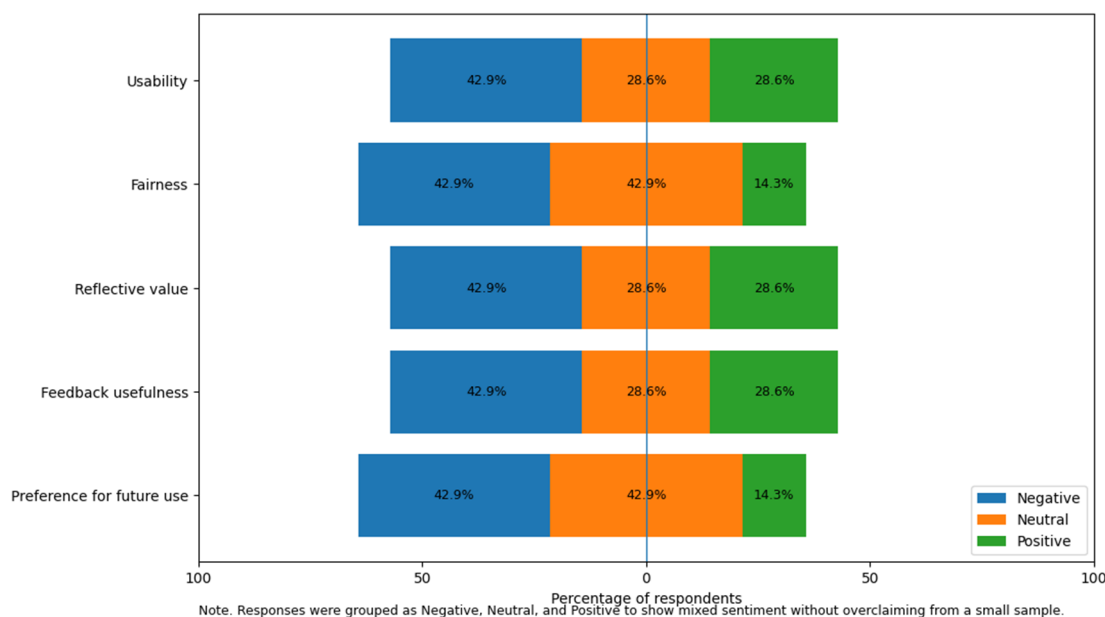


Figure 5. Spring 2026 Student Survey.

Complementary case-study impressions and tutor observations suggested a possible diagnostic effect that should be distinguished from the formal survey findings. These observations did not establish a prevalence rate and should not be interpreted as evidence that most students experienced deeper reflection. Rather, they indicated that the viva process sometimes surfaced comprehension gaps, incorrect applications, or disengagement risks that were not fully visible in a written submission. Tutor observations supported this narrower conclusion by identifying amber and red areas where selected students did not fully understand the material or its application. One illustrative case showed that a student had been working from the wrong supplemental material for weeks without realizing it; the issue surfaced during the assessment process itself rather than after final grading (Figure 6). This case is consequential as an implementation example, not as proof of generalizable efficacy. Written submissions can show the presence of an artifact, but the

viva may reveal whether selected students can explain, contextualize, and apply the work in ways that are instructionally actionable.

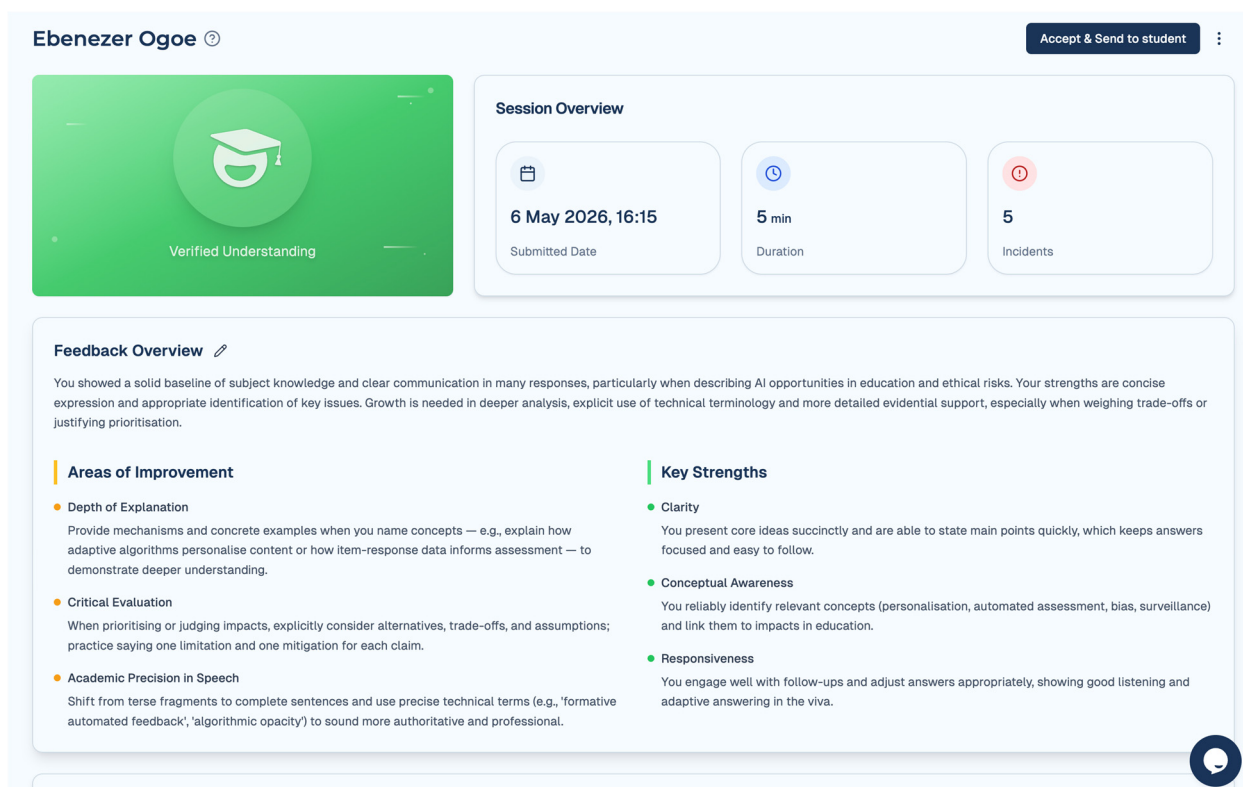


Figure 6. Feedback overview, included for illustrative purposes only and not treated as research evidence.

The cross-cycle comparison between Fall 2025 and Spring 2026 shows that implementation design materially shaped the results. In Fall 2025, the dominant barriers were late rollout, unclear purpose, lack of visible feedback, perceived AI surveillance, uncertainty about grading implications, and conversational pacing problems. In response, the Spring 2026 phase introduced clearer onboarding, more explicit student support materials, push-to-talk functionality, visible live transcription, refined question flow, improved feedback delivery, and a more deliberate formative-to-summative sequence. These revisions addressed several of the Fall barriers, particularly the feedback visibility problem, but did not eliminate broader resistance to AI-mediated assessment. Students who opposed AI in education remained skeptical even when the tool was positioned as supportive rather than punitive. Others accepted the concept but preferred human conversation, writing center support, or a blended model involving peer or instructor interaction. This pattern is important for interpreting the pilot's overall effectiveness. Integrevise™ did not function best as a replacement for human assessment; it functioned best as a possible verification layer that could identify where human follow-up, faculty attention, or student support might be needed.

Taken together, the results suggest four preliminary findings. First, Integrevise™ showed potential diagnostic value by making aspects of student understanding more visible than artifact-only assessment in selected cases, especially through oral explanation and adaptive questioning. Second, student acceptance depended heavily on communication, transparency, and trust, with confusion about grading and AI evaluation producing resistance. Third, platform usability remained a significant factor, particularly around timing, interruptions, question relevance, and the perceived need for more visible or ac-

tionable feedback. Fourth, the pilot identifies a candidate model for scalable authentic assessment, but the present evidence does not establish scalability, efficacy, or broad student endorsement. The evidence therefore supports continued development of Integrevise™ as a process-rich assessment intervention, while indicating that future studies should use larger samples, clearer comparison groups, rubric-aligned outcome measures, systematic treatment of missing data, and more robust analysis of learning gains across modalities.

5. Discussion

This pilot contributes to the emerging literature on AI-era assessment by reframing the central problem from detection to verification, while remaining cautious about the strength of its empirical claims. The literature review established that generative AI has weakened the evidentiary relationship between submitted artifacts and student learning, especially where conventional essays, written responses, reports, coding tasks, or reflections are treated as sufficient proof of understanding. Scarfe et al. [3] showed that AI-generated submissions could pass through real university assessment systems with little detection, while Cotton, Cotton, and Shipway [2] argued that ChatGPT challenges academic integrity because AI-generated work does not conform to traditional plagiarism models. The present study extends that discussion by examining how a viva-based assessment layer can generate additional evidence beyond the artifact itself. Integrevise™ did not attempt to determine whether students used AI; instead, it asked whether students could explain, apply, and defend the work they submitted. This distinction is significant because it shifts assessment away from forensic suspicion and toward construct validity: the question becomes whether the assessment actually reveals the learning it claims to measure. However, because the formal Spring 2026 survey included only seven respondents and platform feedback was missing for 38.5% of completed vivas, the findings should be interpreted as preliminary evidence from a bounded pilot rather than as a generalizable validation of the platform. In this sense, the pilot aligns with broader scholarly arguments against reliance on post hoc detection, especially given the unreliability and equity concerns associated with AI-detection tools [5,6], but it does not by itself prove that viva-based verification solves the assessment validity problem.

A key insight from the study is that viva-based verification may make otherwise hidden gaps in understanding visible while there is still time for instructional response, although this claim rests primarily on qualitative and observational evidence. This finding aligns with formative assessment theory, which emphasizes that assessment becomes pedagogically powerful when it provides actionable evidence that can be used to modify teaching and learning [40,41]. The pilot's most compelling value was not that students uniformly liked the platform, since the formal survey results were mixed to negative on several items, but that the viva process sometimes surfaced comprehension, uncertainty, and disengagement signals that were not apparent in written submissions. Tutor observations and one illustrative case suggested that oral explanation may function as both an assessment event and a metacognitive prompt in selected circumstances. This is consistent with Hattie and Timperley's [42] account of feedback as most useful when it helps learners understand where they are, where they need to go, and how they might close the gap. Integrevise™ is therefore best understood not as a replacement for instructor judgment nor as an already validated scalable solution, but as a mechanism that may generate timely diagnostic evidence for instructors and support systems to evaluate further. The case in which a student had been working from the wrong supplemental material for weeks illustrates the potential value of early visibility, while also underscoring the need for larger studies that can determine how often such diagnostic events occur and under what instructional conditions.

The pilot also suggests that scalable authentic assessment requires more than simply adding oral assessment to existing assignments. The literature has increasingly called for authentic and AI-inclusive assessment, but the Integrevise™ results show that implementation conditions determine whether students experience such assessment as meaningful, fair, or punitive. Perkins et al. [7] argue that AI-inclusive assessment requires explicit communication about permitted tool use, while Furze et al. [8] show that structured frameworks can reduce ambiguity in implementation. The Fall 2025 findings strongly support this point: students frequently misunderstood the purpose of the platform, worried that AI was grading them, or perceived the activity as disconnected from the learning process. After Spring 2026 refinements, including clearer onboarding, push-to-talk functionality, live transcription, improved feedback delivery, and a formative-to-summative sequence, the model became more coherent, although student sentiment remained mixed. This matters because trust is a validity condition in AI-mediated assessment. If students believe that an AI system is judging them opaquely, recording them unfairly, or functioning as a hidden detector, then their responses may be shaped by anxiety rather than by authentic demonstration of understanding. The pilot therefore suggests that the design of the assessment ecology, including explanation, timing, integration, feedback, and instructor framing, is as important as the technical function of the tool itself.

The findings further complicate the assumption that student satisfaction is the primary indicator of assessment value, while also cautioning against treating diagnostic anecdotes as substitutes for formal outcome evidence. The Spring 2026 survey results showed mixed or negative perceptions on several items, including usability, fairness, reflective value, feedback usefulness, and preference for future use. Those results must be interpreted alongside platform records, targeted interviews, tutor observations, and case-study evidence, but they should not be overridden by those sources. The mixed response does not negate the possible diagnostic value of the platform; rather, it reveals the friction that accompanies a new assessment practice in a cultural moment when students are already anxious about AI, surveillance, academic integrity, and grading. This tension mirrors the broader literature, where AI-inclusive assessment is often presented as necessary but institutionally difficult. Sullivan, Kelly, and McLaughlan [17] warned that academic integrity responses must preserve student learning rather than merely increase compliance pressure, and the Integrevise™ findings support that caution. Students who opposed AI in education often resisted the platform regardless of its stated purpose, while others accepted the idea of oral verification but preferred human conversation or clearer feedback. Thus, the study indicates that viva-based assessment should not be framed as automation replacing faculty but as a possible triage and verification layer that helps instructors identify where human attention may be needed.

The results also point toward a more nuanced understanding of authenticity in AI-era assessment, but they do so as design implications rather than established effectiveness findings. Before generative AI, authentic assessment often meant asking students to complete tasks resembling disciplinary or professional practice. In the current environment, however, professional practice increasingly includes AI-supported drafting, analysis, and revision, which means authenticity cannot simply mean banning AI or recreating pre-digital assessment conditions. Integrevise™ offers one possible response: keep the authentic artifact, but add a verification moment in which students must demonstrate understanding in real time. This model preserves the value of extended written, research, or project-based work while reducing the risk that polished artifacts mask weak comprehension. It also avoids the limitations of a wholesale return to closed-book examinations, which may improve security but can narrow the kinds of learning being assessed. The pilot therefore points toward a hybrid assessment architecture in which students produce substantive

work, instructors clarify allowable AI use, the platform elicits oral explanation, diagnostic signals identify possible misunderstanding or disengagement risk, and faculty or support staff intervene where needed. Such a model aligns with the literature's movement away from artifact-only assessment and toward process-rich evidence systems, but it requires more rigorous empirical evaluation before claims of scalability or efficacy can be made.

The study has several limitations that should temper interpretation of the findings. First, this was a bounded pilot within one institutional context, so the results should be understood as evidence from a situated implementation rather than as generalizable proof of efficacy across higher education. Second, the research design was repeated cross-sectional rather than longitudinal; Fall 2025 and Spring 2026 involved different student groups, so the study cannot support claims about individual change over time. Third, the formal Spring 2026 student survey sample was very small ($n = 7$), which means percentage findings should be read as descriptive summaries of a small pilot sample rather than inferential evidence. Fourth, 20 of 52 completed Spring 2026 vivas produced no student feedback record, creating 38.5% missing feedback data; because the reasons for missingness were not captured, the validity and interpretation of feedback proportions are uncertain. Fifth, faculty survey items were piloted for implementation reflection, but faculty survey results were not analyzed as outcome data, limiting claims about instructor experience. Sixth, the survey instrument was developed specifically for this pilot and was not psychometrically validated; reliability statistics were not calculated because the sample was too small and the items served heterogeneous implementation purposes. Seventh, qualitative responses and interviews were coded thematically, but the pilot did not include independent double-coding or interrater reliability testing. Eighth, the study combined surveys, interviews, platform records, tutor observations, postmortem analysis, and case-study evidence, but it did not include a randomized comparison group or a controlled measure of learning gains. Finally, students' perceptions were shaped by broader attitudes toward AI in education, technical usability, course-specific communication, and uncertainty about grading, all of which may have influenced responses independently of the assessment model itself. Future research should therefore test Integrevise™ across larger samples, additional disciplines, multiple assignment types, and program-level deployments, while examining rubric alignment, interrater interpretation of diagnostic signals, student learning outcomes, faculty workload, equity effects, systematic missing-data procedures, and retention-related indicators over time.

6. Conclusions

The Integrevise™ pilot suggests that one urgent assessment challenge in the age of generative AI is not simply whether students are using these tools, but whether institutions can still generate credible evidence that students understand, can explain, and can apply the work they submit. As the literature review established, artifact-only assessment has become increasingly unstable because written products, code, summaries, research syntheses, and reflective responses can now be produced or substantially revised through generative systems with limited visibility into student cognition. Detection-based responses offer only a partial and often unreliable solution, while a wholesale return to closed examinations risks narrowing assessment to decontextualized performance under artificial constraints. The findings from this pilot point toward a possible path: assessment systems may preserve authentic written, research, and project-based work while adding a verification layer that makes understanding more visible. Integrevise™ operationalizes this approach by pairing submitted work with a short adaptive viva, thereby shifting emphasis from the artifact alone to the student's capacity to explain reasoning, clarify concepts, and demonstrate comprehension in their own words.

The study's results suggest that viva-based verification can surface learning gaps, uncertainty, and disengagement risk in selected cases, but the evidence remains preliminary. The formal Spring 2026 survey included only seven respondents and showed mixed perceptions, including only one respondent (14.3%) agreeing that the oral assessment helped them think more deeply about the assignment and four respondents (57.1%) disagreeing or strongly disagreeing. Platform feedback data were also incomplete, with 38.5% of completed vivas producing no feedback record. These limitations mean that the pilot cannot establish broad efficacy, student endorsement, or scalability. Its value lies instead in identifying a plausible diagnostic mechanism and the implementation conditions that must be addressed in subsequent research. The platform appears most useful when instructors explain its purpose clearly, connect it to learning rather than surveillance, provide visible feedback, and treat outputs as prompts for human judgment rather than as autonomous determinations of student understanding.

The broader contribution of this study is to position Integrevise™ as a preliminary model for process-rich assessment in a post-generative AI environment. The platform responds to the limitations of both artifact-only assessment and detection-led integrity systems by creating a structured opportunity for students to demonstrate understanding after submission. It does not eliminate the need for faculty expertise, clear AI-use policies, high-quality assignment design, validated measures, or equitable support structures. Rather, it provides a candidate layer through which instructors may gather additional evidence, identify misunderstandings earlier, and intervene more meaningfully. Future research should extend this pilot across larger samples, additional disciplines, multiple assignment types, and program-level assessment cycles, with particular attention to rubric alignment, reliability of diagnostic signals, student learning outcomes, faculty workload, equity implications, retention indicators, and missing-data procedures. The central implication is therefore cautious but important: if generative AI has made polished academic production too easy to treat as sufficient evidence of learning, assessment must move toward models that require explanation, judgment, and visible reasoning while validating those models through larger and more systematic empirical study.

Author Contributions: Conceptualization, K.P.; Methodology, J.H.; Software, K.P., E.O. and K.A.A.; Validation, J.H. and K.A.A.; Formal analysis, K.P., E.O. and K.A.A.; Investigation, J.H. and K.P.; Resources, K.P.; Data curation, J.H. and K.A.A.; Writing—original draft, J.H.; Writing—review and editing, J.H.; Visualization, K.P. and E.O.; Supervision, J.H.; Project administration, J.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki and approved by the Lindenwood University IRB on 12 August 2025 (IRB-26-7).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: Authors Kyle Poyer, Ebenezer Ogoe, and Kelvin Atologun are affiliated with Integrevise Ltd., the developer of the Integrevise™ platform evaluated in this study. Their affiliations are disclosed for transparency. James Hutson has no financial interest in Integrevise Ltd. The study was conducted as an independent academic evaluation of the platform. The authors affirm that the design of the study, data collection, analysis, interpretation of the findings, and preparation of the manuscript were undertaken in accordance with accepted scholarly standards, and the company did not exercise editorial control over the reporting of the results.

References

1. Belkin, D. Princeton Changes Its 133-Year-Old Honor Code over AI Cheating Fears. *The Wall Street Journal*. 2026. Available online: <https://www.wsj.com/us-news/education/princeton-cheating-ai-proctors-2a1cf62e> (accessed on 2 July 2026).
2. Cotton, D.R.E.; Cotton, P.A.; Shipway, J.R. Chatting and Cheating: Ensuring Academic Integrity in the Era of ChatGPT. *Innov. Educ. Teach. Int.* **2024**, *61*, 228–239. [\[CrossRef\]](#)
3. Scarfe, P.; Watcham, K.; Clarke, A.; Roesch, E. A Real-World Test of Artificial Intelligence Infiltration of a University Examinations System: A ‘Turing Test’ Case Study. *PLoS ONE* **2024**, *19*, e0305354. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Stephenson, R.; Armstrong, C. *Student Generative Artificial Intelligence Survey 2026*; Higher Education Policy Institute: Oxford, UK, 2026. Available online: <https://www.hepi.ac.uk/reports/student-generative-ai-survey-2026/> (accessed on 2 July 2026).
5. Liang, W.; Yuksekogonul, M.; Mao, Y.; Wu, E.; Zou, J. GPT Detectors Are Biased against Non-Native English Writers. *Patterns* **2023**, *4*, 100779. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Weber-Wulff, D.; Anohina-Naumeca, A.; Bjelobaba, S.; Foltýnek, T.; Guerrero-Dib, J.; Popoola, O.; Šigut, P.; Waddington, L. Testing of Detection Tools for AI-Generated Text. *Int. J. Educ. Integr.* **2023**, *19*, 26. [\[CrossRef\]](#)
7. Perkins, M.; Furze, L.; Roe, J.; MacVaugh, J. The AI Assessment Scale: A Framework for Ethical Integration of Generative AI in Educational Assessment. *J. Univ. Teach. Learn. Pract.* **2024**, *21*, 6. [\[CrossRef\]](#)
8. Furze, L.; Perkins, M.; Roe, J.; MacVaugh, J. The AI Assessment Scale in Action: A Pilot Implementation of GenAI-Supported Assessment. *Australas. J. Educ. Technol.* **2024**, *40*, 38–55. [\[CrossRef\]](#)
9. Ipeirotis, P.; Rizakos, K. Scalable and Personalized Oral Assessments Using Voice AI. *arXiv* **2026**. [\[CrossRef\]](#)
10. Barba, L.A.; Stegner, L. The Conversational Exam: A Scalable Assessment Design for the AI Era. *arXiv* **2026**. [\[CrossRef\]](#)
11. Swiecki, Z.; Khosravi, H.; Chen, G.; Martinez-Maldonado, R.; Lodge, J.M.; Milligan, S.; Selwyn, N.; Gašević, D. Assessment in the Age of Artificial Intelligence. *Comput. Educ. Artif. Intell.* **2022**, *3*, 100075. [\[CrossRef\]](#)
12. Chan, C.K.Y. A Comprehensive AI Policy Education Framework for University Teaching and Learning. *Int. J. Educ. Technol. High. Educ.* **2023**, *20*, 38. [\[CrossRef\]](#)
13. Freeman, J. *Student Generative AI Survey 2025*; Higher Education Policy Institute: Oxford, UK, 2025. Available online: <https://www.hepi.ac.uk/reports/student-generative-ai-survey-2025/> (accessed on 2 July 2026).
14. Fan, Y.; Tang, L.; Le, H.; Shen, K.; Tan, S.; Zhao, Y.; Shen, Y.; Li, X.; Gašević, D. Beware of Metacognitive Laziness: Effects of Generative Artificial Intelligence on Learning Motivation, Processes, and Performance. *arXiv* **2024**. [\[CrossRef\]](#)
15. Kasneci, E.; Sessler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günemann, S.; Hüllermeier, E.; et al. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learn. Individ. Differ.* **2023**, *103*, 102274. [\[CrossRef\]](#)
16. Rudolph, J.; Tan, S.; Tan, S. ChatGPT: Bullshit Spewer or the End of Traditional Assessments in Higher Education? *J. Appl. Learn. Teach.* **2023**, *6*, 342–363. [\[CrossRef\]](#)
17. Sullivan, M.; Kelly, A.; McLaughlan, P. ChatGPT in Higher Education: Considerations for Academic Integrity and Student Learning. *J. Appl. Learn. Teach.* **2023**, *6*, 31–40. [\[CrossRef\]](#)
18. Yeadon, W.; Inyang, O.-O.; Mizouri, A.; Peach, A.; Testrow, C. The Death of the Short-Form Physics Essay in the Coming AI Revolution. *Phys. Educ.* **2023**, *58*, 035027. [\[CrossRef\]](#)
19. Nikolic, S.; Daniel, S.; Haque, R.; Belkina, M.; Hassan, G.M.; Grundy, S.; Lyden, S.; Neal, P.; Sandison, C. ChatGPT versus Engineering Education Assessment: A Multidisciplinary and Multi-Institutional Benchmarking and Analysis of This Generative Artificial Intelligence Tool to Investigate Assessment Integrity. *Eur. J. Eng. Educ.* **2023**, *48*, 559–614. [\[CrossRef\]](#)
20. Ardito, C.G. Contra Generative AI Detection in Higher Education Assessments. *arXiv* **2023**. [\[CrossRef\]](#)
21. Kirchner, J.H.; Ahmad, L.; Aaronson, S.; Leike, J. New AI Classifier for Indicating AI-Written Text. OpenAI. 2023. Available online: <https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text/> (accessed on 2 July 2026).
22. Elkhatat, A.M.; Elsaid, K.; Almeer, S. Evaluating the Efficacy of AI Content Detection Tools in Differentiating between Human and AI-Generated Text. *Int. J. Educ. Integr.* **2023**, *19*, 17. [\[CrossRef\]](#)
23. Perkins, M.; Roe, J.; Furze, L.; Vu, B.H.; Postma, D.; Hickerson, D.; McGaughan, J.; Khuat, H.Q. GenAI Detection Tools, Adversarial Techniques and Implications for Inclusivity in Higher Education. *arXiv* **2024**. [\[CrossRef\]](#)
24. Sadasivan, V.S.; Kumar, A.; Balasubramanian, S.; Wang, W.; Feizi, S. Can AI-Generated Text Be Reliably Detected? *arXiv* **2023**. [\[CrossRef\]](#)
25. Goodier, M. Revealed: Thousands of UK University Students Caught Cheating Using AI. *The Guardian*. 2025. Available online: <https://www.theguardian.com/education/2025/jun/15/thousands-of-uk-university-students-caught-cheating-using-ai-artificial-intelligence-survey> (accessed on 2 July 2026).
26. Perkins, M.; Roe, J.; Furze, L. The AI Assessment Scale Revisited: A Framework for Educational Assessment. *arXiv* **2024**. [\[CrossRef\]](#)
27. Wang, H.; Dang, A.; Wu, Z.; Mac, S. Generative AI in Higher Education: Seeing ChatGPT through Universities’ Policies, Resources, and Guidelines. *arXiv* **2023**. [\[CrossRef\]](#)

28. Ghimire, A.; Edwards, J. From Guidelines to Governance: A Study of AI Policies in Education. *arXiv* **2024**. [CrossRef]
29. UNESCO. *Guidance for Generative AI in Education and Research*; UNESCO: Paris, France, 2023. Available online: <https://unesdoc.unesco.org/ark:/48223/pf0000386693> (accessed on 2 July 2026).
30. Georgiou, G.P. ChatGPT Produces More ‘Lazy’ Thinkers: Evidence of Cognitive Engagement Decline. *arXiv* **2025**. [CrossRef]
31. Lo, C.K. What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. *Educ. Sci.* **2023**, *13*, 410. [CrossRef]
32. Li, M.; Enkhtur, A.; Cheng, F.; Yamamoto, B.A. Ethical Implications of ChatGPT in Higher Education: A Scoping Review. *arXiv* **2023**. [CrossRef]
33. Gao, R.; Merzdorf, H.E.; Anwar, S.; Hipwell, M.C.; Srinivasa, A. Automatic Assessment of Text-Based Responses in Post-Secondary Education: A Systematic Review. *arXiv* **2023**. [CrossRef]
34. Yan, L.; Sha, L.; Zhao, L.; Li, Y.; Martinez-Maldonado, R.; Chen, G.; Li, X.; Jin, Y.; Gašević, D. Practical and Ethical Challenges of Large Language Models in Education: A Systematic Scoping Review. *Br. J. Educ. Technol.* **2024**, *55*, 90–112. [CrossRef]
35. Dempere, J.; Modugu, K.; Hesham, A.; Ramasamy, L.K. The Impact of ChatGPT on Higher Education. *Front. Educ.* **2023**, *8*, 1206936. [CrossRef]
36. Gill, S.S.; Xu, M.; Patros, P.; Wu, H.; Kaur, R.; Kaur, K.; Fuller, S.; Singh, M.; Arora, P.; Parlikad, A.K.; et al. Transformative Effects of ChatGPT on Modern Education: Emerging Era of AI Chatbots. *arXiv* **2023**. [CrossRef]
37. Mahmoud, Q.H. Reimagining Assessment in the Age of Generative AI: Lessons from Open-Book Exams with ChatGPT. *arXiv* **2026**. [CrossRef]
38. Weale, S. UK Universities Warned to ‘Stress-Test’ Assessments as 92% of Students Use AI. *The Guardian*. 2025. Available online: <https://www.theguardian.com/education/2025/feb/26/uk-universities-warned-to-stress-test-assessments-as-92-of-students-use-ai> (accessed on 2 July 2026).
39. Bridgeman, A.; Liu, D.; Weeks, R. Program Level Assessment Design and the Two-Lane Approach. *Teaching@Sydney*. 2024. Available online: <https://educational-innovation.sydney.edu.au/teaching@sydney/program-level-assessment-two-lane/> (accessed on 2 July 2026).
40. Black, P.; Wiliam, D. Assessment and Classroom Learning. *Assess. Educ. Princ. Policy Pract.* **1998**, *5*, 7–74. [CrossRef]
41. Sadler, D.R. Formative Assessment and the Design of Instructional Systems. *Instr. Sci.* **1989**, *18*, 119–144. [CrossRef]
42. Hattie, J.; Timperley, H. The Power of Feedback. *Rev. Educ. Res.* **2007**, *77*, 81–112. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.