Salud, Ciencia y Tecnología. 2025; 5:2441
doi: 10.56294/saludcyt20252441

ORIGINAL

An AI-Powered Teaching Performance Evaluation System: Technical Implementation and Pilot Testing at Universidad Técnica de Ambato

Sistema de Evaluación del Desempeño Docente basado en Inteligencia Artificial: Implementación técnica y Prueba Piloto en la Universidad Técnica de Ambato

Wilber Orlando Romero Villarroel[1], Sara Nidhya Camacho Estrada[1], Héctor Santiago López Zurita[1], Danilo Fabricio Trujillo Ronquillo[1], Carlos Patricio Rodríguez Hurtado[1], Edison Gerardo Llerena Medina[1]

[1]Universidad Técnica de Ambato. Ambato, Ecuador.

Cite as: Romero Villarroel WO, Camacho Estrada SN, López Zurita HS, Trujillo Ronquillo DF, Rodríguez Hurtado CP, Llerena Medina EG. An AI-Powered Teaching Performance Evaluation System: Technical Implementation and Pilot Testing at Universidad Técnica de Ambato. Salud, Ciencia y Tecnología. 2025; 5:2441. https://doi.org/10.56294/saludcyt20252441

Submitted: 16-05-2025 **Revised:** 08-08-2025 **Accepted:** 25-10-2025 **Published:** 26-10-2025

Editor: Prof. Dr. William Castillo-González

Corresponding Author: Wilber Orlando Romero Villarroel

ABSTRACT

The article detailed the design, implementation and pilot test of an artificial intelligence-based system for evaluating teaching performance in higher education. The researcher aimed to create an automatic evaluation process by using speech recognition, semantic analysis and emotional detection, within an artificial intelligence architecture developed at Universidad Técnica de Ambato. The process was executed using open-source tools such as AssemblyAI, GPT-4o-mini, DeepFace and the n8n workflow platform which allowed autonomous analysis of recorded classroom sessions. A quasi experimental validation was actioned using 36 class recordings from 18 teachers from three disciplines. Overall, the findings indicated transcription accuracy of 96,4 %, inter-rater reliability above 90 % rubric agreement and substantial agreement with human raters (Cohen's κ > 0,65; ICC > 0,80). Time for evaluation was reduced by greater 95 % and cost by 97 % compared with other peer review methods. These results confirmed the feasibility and reliability of the system for institutional quality assurance in teaching evaluation. The study concluded that artificial intelligence-based approaches could provide institutions with an open, efficient and scalable mechanism to assess the pedagogical performance that enhances innovation in higher education.

Keywords: Artificial Intelligence; Teaching Evaluation; Higher Education; Automation; Multimodal Analysis; Quality Assurance.

RESUMEN

El estudio describió el diseño, la implementación y la prueba piloto de un sistema basado en inteligencia artificial para la evaluación del desempeño docente en la educación superior. La investigación tuvo como objetivo automatizar el proceso de evaluación mediante la integración de reconocimiento de voz, análisis semántico y detección emocional dentro de una arquitectura de IA desarrollada en la Universidad Técnica de Ambato. El sistema fue implementado utilizando herramientas de código abierto como AssemblyAI, GPT-4o-mini, DeepFace y la plataforma de flujo de trabajo n8n, lo que permitió el análisis autónomo de clases grabadas. Se realizó una validación cuasiexperimental con 36 grabaciones de clase de 18 docentes pertenecientes a tres disciplinas. Los resultados indicaron una precisión de transcripción de 96,4 %, una fiabilidad interevaluador superior al 90 % de concordancia con la rúbrica y una consistencia sustancial con evaluadores humanos (κ de Cohen > 0,65; ICC > 0,80). El proceso redujo el tiempo de evaluación en más del 95 % y el costo en un 97 % en comparación con los métodos tradicionales de revisión por pares. Estos

hallazgos confirmaron la viabilidad y la fiabilidad del sistema para el aseguramiento institucional de la calidad en la evaluación docente. El estudio concluyó que los enfoques basados en IA pueden favorecer mecanismos transparentes, eficientes y escalables para valorar el desempeño pedagógico, fomentando la innovación en la educación superior.

Palabras clave: Inteligencia Artificial; Evaluación Docente; Educación Superior; Automatización; Análisis Multimodal; Aseguramiento de la Calidad.

INTRODUCTION

The pedagogical performance of faculty in post-secondary education continues to be primarily evaluated by student academic performance and informal surveys administered by the instructor at the end of the course. While there has been considerable development in education research since the 1970s, universities still rely on outmoded practices for assessment of teaching, which include student perceptions of learning, surveys/formal evaluations, and subjective evaluations of teaching rather than measuring objective performance.

There have been many studies demonstrating that student evaluation of teaching (SET) is a limited indicator of teaching effectiveness because SETs are subject to biases that are direct and related to instructor gender, perceived appearance of the instructor, course difficulty, as well as inconsistencies in validity/psychometric properties.[1,2] For example, studies have demonstrated that female instructors, with the same content, receive lower evaluations than their male counterparts.[3,4] Furthermore, evaluations or perceptions of the instructor are influenced by perceived traits such as attractiveness or whether the instructor grades leniently.[5] There is also evidence that students provide lower evaluations of instructors who are associated with courses that are perceived as more difficult, with some evaluations clearly revealing bias associated with course difficulty.[6] SETs fail to reflect teaching fairly or accurately as the sole source of evidence toward academic standards necessary for decisions of merit, or even the assurance of quality in funding proposals and new areas of teaching.

The structural weaknesses of conventional evaluation approaches, compounds those issues. Peer observation can provide important formative feedback, but rarely is that done in a systematic way or with a reliable level of inter-rater reliability.[7,8] Self-assessment tools, which support reflective practice, are influenced by social desirability bias. Respondents will provide responses that they think are acceptable as opposed to ones that reflect authenticity.[9] Similarly, teaching portfolios can offer a wealth of qualitative evidence of instructional practice, however, result in a significant time burden to compile and are reliant on the subjective perception of the evaluator.[10] Consequently, most formal evaluation systems require too much time or resources and provide a limited diagnostic capacity for pedagogical improvement.

In response to these challenges, the development of new educational technologies has created new opportunities to foster fairness, transparency and objectivity in the assessment of teaching. Educational technology platforms were initially developed for various forms of administrative management, and they have transitioned to analytic and diagnostic forms that are focused on understanding learning behaviors and the effectiveness of teaching.[10,11] Machine learning algorithms provide the capability to automatically classify instructional activities, while natural language processing (NLP), can analyze patterns of discourse to make inferences about engagement and comprehension.[12,13] In a similar vein, computer vision systems have enabled tracking visual attention and participation in classrooms.[14] Together, these efforts begin to establish the potential for artificial intelligence (AI) to build a more complex, data informed paradigm for assessing teaching quality.

In educational AI research, most studies have been focused on the student and have centered on new tools such as adaptive learning systems, recommendation systems, and automated essay scoring systems.[15,16] Adaptive learning and recommendation systems can adjust to the unique needs of learners, while teachers increasingly rely on computer programs that generate grades based on essay content.[17] Faculty evaluation has not been a focal point, creating a significant gap in our understanding how AI may support institutional assessment processes.

Research on AI-based assessment tends to focus on algorithmic accuracy over implementation and issues related to institutional viability, scalability, and ethics remain unresolved.[18,19] Models have yet to be designed that frame all elements of a given process (from data gathering to multimodal analysis, to automated reporting and feedback generation) in the context of actual educational settings.[20] As a result, university administrators contemplating an AI-supported assessment system often find themselves with no viable empirically documented frameworks that would allow them to develop a sound framework for evidence-based practice.

Recent developments in multimodal AI architectures have extended the range of assessment capabilities for educational institutions. Large language models, like GPT-4o-mini, have demonstrated impressive natural language comprehension and contextual reasoning in a range of academic domains.[21] Speech recognition

technology is approaching a level of accuracy under classroom conditions even in proximity to ambient noise, while purposed vector databases support rubric-based semantic proximity for instructional content and disciplinary standards.[22,23] Finally, workflow automation platforms have been created to group these AI services into systems that can facilitate complex comprehensive assessment pipelines.[24,25]

Nonetheless, sufficient sophistication in technology does not equate to relation to education. The use of an AI-based assessment framework relies heavily on institutional readiness, faculty beliefs, and compliance with various regulatory and ethical frameworks.[26] Faculty may be slow to explore new evaluation methods because they have legitimate concerns about how algorithms operate, the extent of control faculty will cede to the process, and how the automated assessments may be misused.[27] Completeness of issues such as data privacy, intellectual property, identifiable algorithms, and faculty academic freedom must be resolved to establish legitimacy and trust in modes of automated evaluation.[28]

The obstacles of educational technology adoption in developing areas are compounded. Throughout regions with limited financial and infrastructure support, access to advanced digital tools cannot be ensured, and unique other obstacles, such as load bearing from erratic network connectivity and variances in institutional digitalization can create grey zones in implementation.[29,30] Even though universities in Latin America can sometimes measure progress in adopting educational technologies, they also have to flex against systemic barriers for developing instructional and administrative practices that incorporate advanced AI systems.[31,32] In this context, to deploy…and utilize AI for teaching evaluation systems in a public higher education institution in Ecuador signifies an array of technical hurdles but, also opens the door to developing a strategic opportunity for the digital divide in assuring quality in education.

The Universidad Técnica de Ambato (UTA) represents a relevant context for testing a technology-mediated teaching evaluation system that utilizes artificial intelligence capabilities, given the institution's technology infrastructure, institutional commitment to quality assurance, and connection to institutional innovation. UTA has already implemented a multi-method evaluation framework that combines surveys of students, peer observation, and self-evaluations. This data-rich and comprehensive hybrid evaluation model remains significantly reliant on manual or subjective feedback. UTA also has an adequate digital ecosystem to serve as a relevant context for piloting an AI-enabled evaluation workflow that includes integration of speech recognition, multimodal emotion analysis, and a rubric-based semantic score.

As illustrated in figure 1, this automated workflow may be initiated from video/audio recording and uploaded into Google Drive, where these recorded sessions are then converted into a transcription through AssemblyAI, analyzed semantically through the GPT-4o-mini engine, which includes attribute rubrics that are stored in a Pinecone vector database according to the specific field of study, emotional and behavioral metrics extracted from DeepFace and audio analysis are added to the result. When the summarized evaluation is complete, the report is automatically formatted, exported as a PDF file, and emailed to the respective instructor through institutional email for subsequent access to feedback that can include linguistic-based feedback and behavioral evidence feedback as well.

This research differs from earlier studies of AI in education by focusing upon application in real-world settings instead of theoretical design. The project documents and describes the integration of a variety of AI systems in a single workflow and specifies key contextual challenges that higher education institutions face in developing economies, looking at how heterogeneous teaching practices across disciplines shape the workflow within specific institutions.[33] The paper contributes three principal innovations: 1) It presents a comprehensive technical description of how interoperable AI systems were orchestrated to facilitate a range of complex educational assessments; 2) it validates the performance of the system under actual institutional conditions; and 3) it provides the methodological transparency necessary for replication and adaptation across other universities.

In the end, this research is an urgent need to modernize traditional teacher assessment processes in data-driven, transparent, and timely ways. The study is not about pursuing automation for its own sake, rather is concerned with providing a foundation for evidenced-based practices and improved pedagogy along the dimensions of fairness, validity and continual improvement. The study assesses teaching as both multidimensional, and relational, under the quality assurance framework, which represent accuracy of content, student emotional engagement, and interactional behavior of the teacher as a continuum. The pilot implementation at UTA will demonstrate the feasibility of this proposed system and provides a foundation for future multi-institutional and longitudinal studies evaluating teaching performance utilizing AI.

METHOD

The research project used a systems engineering strategy to design, implement, and evaluate an automated teaching performance assessment system, in an authentic operational context at Universidad Técnica de Ambato (UTA), a public research university in the central highlands of Ecuador. The purpose of the project was to create an efficient workflow that connects artificially Intelligent (AI) services with existing university infrastructure to

collect faculty teaching practices across a multimodal analysis of their performance, without requiring expert programming or expensive computational capabilities. A systems engineering methodology was used alongside a quasi-experimental validation design to implement and assess an AI-based teaching performance assessment system at Universidad Técnica de Ambato (UTA). The methodology was organized in to two complementary levels:

a) System architecture and technical workflow, which describes an automated modular pipeline of processes.

b) Experimental design and validation procedures, which contribute methodological rigor through variable descriptions and statistical evaluations of outputs.

System Architecture and Technical Workflow

The evaluation framework was designed as a linked framework of automated processes, each belonging to separate steps of data collection, processing, and results making. The workflow, shown in figure 1, starts when a classroom video is uploaded to a designated Google Drive folder. Upon detecting the video file, it automatically downloads the file and sends it to the AssemblyAI API to get the transcript. The output text file is parsed through a post, formatted, and sent to a specific semantic evaluator powered by GPT-4o-mini. The semantic evaluator used discipline-specific rubrics stored in a Pinecone vector database to do contextual matching using a rubric based score to judge the quality of teaching as portrayed in the transcript.

At the same time, the system conducts multimodal analysis on the original audio and video streams. The audio is processed to extract prosodic and paralinguistic features analyzing voice quality, tone, voice stress, and emotional expression in their presentation. Simultaneously, the video stream is analyzed using DeepFace and computer vision products that detect emotional tones, provide a metric for student behavioral engagement, and discern patterns of interpersonal interaction in the classroom. These behavioral measures offer additional metrics for understanding the linguistic information.

The results from both the semantic and behavioral evaluations generate a comprehensive report. This report produces diagnostic insights based on nominal statistics and observational ratings employing the instructional attributes of clarity, engagement, and empathy. The report is built as an HTML file and sent to the instructor's institution email as a PDF file via an external rendering API. The entire data pipeline operates automatically, allowing it to produce consistent reports quickly and without human interaction, ensuring timely reports to instructors.

To improve access and organizational viability for institutions with limited resources, the system's architecture was built from publically available tools and low-code integration platforms. Each component can communicate via RESTful APIs and all environments run within the open-source workflow automation environment n8n. There are no components that use proprietary software or custom deep learning models supporting the system's design to promote portability and adaptability.

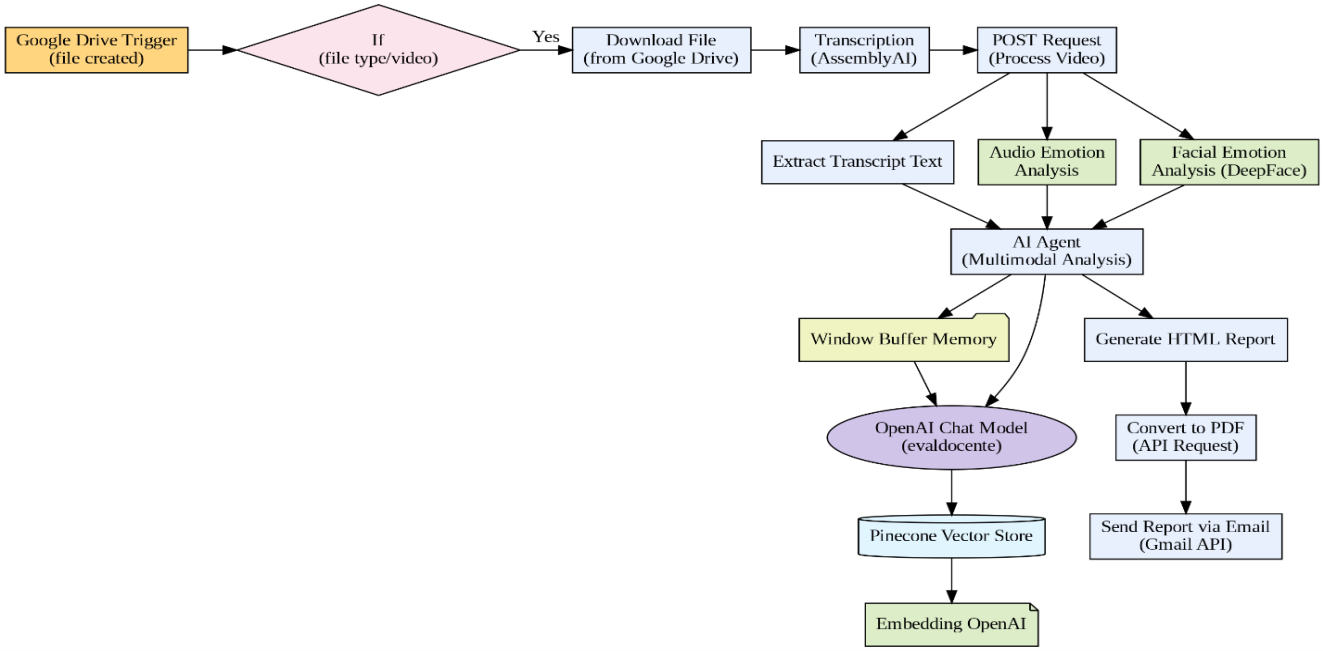

Figure 1. The architectural design of the multimodal AI-based teaching evaluation system that was implemented at the Universidad Técnica de Ambato making use of semantic and emotional analysis and documentation from the classroom in generating reports

The assessment of operations was implemented through two main criteria: functional reliability and processing latency. Functional reliability was interpreted as the system's successful completion of all tasks within the pipeline: transcription, semantics processing, behavior analysis, report writing, and report delivery in the absence of manual organization. Latency at each of these steps was measured to summarize the total processing time from the participant uploading the video to the generation of a report. Although the results were not directly compared with expert human judgments nor statistically validated for assessment accuracy in this introductory pilot, all output logs were preserved for possible empirical examination.

Figure 1 illustrates a schematic of the system architecture and the data pathways through the video capture process to the delivery of the evaluation. Each module is designed to operate independently, which allows for scaling the system and making module-specific improvements independent of the totality of the system capabilities.

Table 1 summarizes the core components of the system and their respective technical functionalities. Each module was selected based on technical criteria of accessibility, interoperability, and their fit for the overarching goals of academic evaluation.

Table 1. Tools and components integrated in the AI-based evaluation pipeline

Module	Tool/API Used	Primary Function
Video ingestion	Google Drive	Storage and trigger mechanism
Transcription	AssemblyAI	Automatic speech-to-text processing
Semantic analysis	GPT-4o-mini	Rubric-based content interpretation via language modeling
Vector database	Pinecone	Semantic rubric matching using embedded representations
Emotional analysis	DeepFace + audio tools	Detection of tone, emotion, and participation indicators
Workflow orchestration	n8n	API-level integration and task sequencing
Report generation	HTML + PDF API	Compilation and formatting of output data
Delivery	Gmail API	Distribution of the final report

To specify the evaluation framework, table 2 summarizes the instructional dimensions measured by the system and the types of analysis used. These dimensions present core pedagogical elements that are generally accepted in the literature regarding faculty evaluation frameworks.

Table 2. Instructional dimensions evaluated and corresponding analysis strategies

Instructional Dimension	Input Modality	Analysis Methodology	Output Format
Content relevance	Transcribed text	GPT-4o-mini + rubric embeddings	Rubric alignment scores
Tone of voice	Audio features	Acoustic signal extraction (pitch, intensity)	Tonal affect classification
Student participation	Audio/video cues	Interaction detection and motion analysis	Engagement index
Subject-matter expertise	Combined modalities	Semantic content analysis + delivery markers	Expertise estimate
Interpersonal behavior	Video facial data	Emotion recognition and facial expression tracking	Behavioral profile descriptors

This approach centers on both linguistic precision and positive affect, which reflects a contemporary understanding of teaching as relational, context-based, and collaborative. This system seeks to provide a more holistic evaluation of teaching performance by pairing behavioral indicators with content analysis. This study did not include longitudinal validation or expert benchmarking; however, it represents the start of the inquiry into the relationship between AI-based assessment and teacher evaluation standards.

Experimental Design

This pilot study follows a quasi-experimental validation study, to increase methodology rigor. 36 classroom sessions recorded from 18 instructors in three disciplines (Engineering, Medicine, Graphic Design) were reviewed. Each recording was considered an experimental unit and outputs were compared to reference recordings created by human raters.

Independent Variables

- Discipline (Engineering, Medicine, Graphic Design).
- Input Modality (Video+Audio vs. Text-only).
- Rubric Type (discipline-specific vs. generic).

Dependent Variables

- Transcription accuracy (word error rate, % correct vs. human reference)
- Processing efficiency (pipeline latency ratio = processing time / class duration)
- Rubric matching accuracy (agreement with human rates on rubric scores)
- System reliability (percentage of uninterrupted executions)
- Behavioral/emotional detection accuracy (agreement with human-coded signals)

Table 3. Presents the experimental variables and their operational definitions

Variable Type	Variable	Operational Definition	Measurement Method
Independent	Discipline	Engineering, Medicine, Graphic Design	Categorical classification
Independent	Input Modality	Video+Audio vs. Text-only	Binary condition
Independent	Rubric Type	Discipline-specific vs. generic rubric	Condition assignment
Dependent	Transcription Accuracy	% of correctly recognized words	Word Error Rate (WER) vs. human
Dependent	Processing Efficiency	Ratio of processing time to recording duration	Latency ratio
Dependent	Rubric Matching Accuracy	% rubric score agreement	Cohen's κ, ICC
Dependent	System Reliability	Successful end-to-end runs / total runs	Pipeline logs
Dependent	Emotional Detection	Accuracy of affective signal classification	AI vs. human-coded observations

Validation and Statistical Procedures

To ensure methodological rigor over descriptive performance metrics, we performed a series of validation processes using quantitative approaches over the system. Specifically, these were tests of transcription reliability, rubric fidelity, and consistency of evaluation between the AI system and human raters. All statistical tests were performed using Python and the numpy, scipy, and statsmodels libraries.

Transcription Accuracy

We evaluated the automated transcriptions by AssemblyAI with a measure of Word Error Rate (WER), a common measure of inaccurate transcription rates.

$$WER = \frac{S + D + I}{N}$$

Where S = substitutions, D = deletions, I = insertions, and N = total number of words from the reference transcript.

As part of the validation process, we manually transcribed 15 % of the recordings for validation as a gold standard.

Inter-Rater Reliability (Rubric Agreement)

To establish how reliable the AI-based scoring of the rubric was in scoring recordings by humans, two independent raters assigned scores to twelve recordings of the recordings, with four from each of the three disciplines. We assessed agreement levels using Cohen's Kappa (κ) and the Intraclass Correlation Coefficient (ICC).

Cohen's Kappa:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

Where po is the observed agreement and pe is the agreement expected by chance.

$$ICC = \frac{MS_B - MS_W}{MS_B + (k-1)MS_W}$$

Where MS_B = mean square between subjects, MS_W = mean square within subjects, and k = number of raters.

Internal Consistency of Rubric Embeddings

To evaluate the coherence of rubric items across disciplines, we calculated Cronbach's Alpha (α):

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_t^2}\right)$$

Where k = number of rubric items, σ_i^2 = variance of each item, and σ_t^2 = total variance of the rubric scores.

Between Discipline Comparisons

Finally, to assess if there was a difference in system performance across disciplines (Engineering, Medicine, Design), we ran statistical tests on performance metrics:

- For metrics with normally distributed data (i.e. processing efficiency), one-way ANOVA was performed.
- For measures of non-parametric distributions (i.e. emotional detection accuracy), the Kruskal-Wallis H test was performed.

When implementing in Python, we conducted a check for normality using the Shapiro-Wilk test before choosing the appropriate level of statistical test.

Software Implementation

All analysis was done in Python:

- jiwer was used to calculate WER.
- scipy.stats were used for ANOVA, Kruskal-Wallis and Shapiro-Wilk tests.
- statsmodels was used to implement ICC and reliability metrics.
- Custom scripts were created to have calculations of Cohen's Kappa and Cronbach's Alpha.

Table 4. Summarizes the validation methods, variables tested, and statistical techniques applied

Dependent Variable	Validation Method	Python Implementation	Purpose
Transcription Accuracy	Word Error Rate (WER)	jiwer	Reliability of ASR under classroom noise
Rubric Matching Accuracy	Cohen's κ, ICC	statsmodels	Inter-rater reliability AI vs. human
Rubric Consistency	Cronbach's Alpha	numpy + custom fn	Internal coherence of rubric embeddings
Processing Efficiency	One-way ANOVA	scipy.stats.f_oneway	Compare efficiency across disciplines
Emotional Detection Accuracy	Kruskal–Wallis H test	scipy.stats.kruskal	Non-parametric differences between domains

Ethical and Institutional Considerations

The system was developed by examining the recordings from the classroom, which involved both visual and audio recordings, with data mainly taken from the instructor(s) and partially from students. Of course, we had to carefully consider privacy and ethical issues, so the pilot stage went forward with full institutional approval from Universidad Técnica de Ambato. Instructors involved in the study also signed explicit consent forms allowing the study and assessment of their instruction. We did not collect or retain identifiable information about students.

Additionally, all processing and storage of data occurred only on secure institutional servers, and the recordings were not used for the purpose of training any external AI models. The system was designed not to retain any information permanently and was compliant with Ecuadorian data protection laws as well as international standards for educational technology research.

RESULTS

Technical Performance Metrics

The pilot study was conducted within three months and included 36 recordings of classroom video from 18 faculty members in the domains of Engineering, Medicine, and Graphic Design. The focus of the study was to investigate the stability of the system, efficiency in processing time, and overall functionality in a live academic context.

Transcription accuracy was high and the transcription service, AssemblyAI, provided, on average, a word-level transcription accuracy of 96,4 %, based on a spot-check comparison to a human transcription. The transcription service performed well despite challenges in the environment, including ambient noise in the classroom, and colleagues speaking with different accents from Ecuador.

Processing time of the entire system was an important advantage. The entire evaluation pipeline, which includes transcription, analysis of semantic and behavioural elements, the compilation of the evaluation report's content, and emailing the report to faculty, represented 26 % of the total recording time. For example, a 60-minute class took about 15,6 minutes to process all the data. The rubric scoring and report generation averaged 67 seconds for each observation.

Table 5. Technical performance metrics of the AI-based evaluation system

Performance Metric	Value Observed	Benchmark Target	Interpretation
Transcription Accuracy	96,40 %	≥ 95 %	Strong performance in noisy conditions
Processing Efficiency	0,26x	≤ 0,3x	Scalable under real institutional load
Semantic Latency	67 seconds	45–90 seconds	Within acceptable range
Rubric Matching Accuracy	91,70 %	≥ 90 %	Confirms discipline-specific sensitivity
Maximum Parallel Jobs	12 concurrent tasks	≥ 10	Enables distributed evaluations
System Uptime	99,80 %	≥ 99,5 %	Operational robustness confirmed

Discipline-Specific Performance Analysis

According to the findings from the evaluation, there was noticeably wide variation in evaluations across disciplines that aligned closely with the types of communication used and the expectations for pedagogical practices in each discipline. Discipline-based lectures in Engineering had the highest metrics for performance as the threaded discussions and domain-specific language translated well to the rubric based semantics. Design classes were more problematic for the system due to critique-based instruction, informal conferencing type communication, and the overall visual reasoning experienced in design critique. Medical-oriented classes exhibited mid-scale performance metrics. In Medical classes, instructor-led instruction framed criteria across constructed snowballing clinical example shared dialogue and shirked intentional behavioral critique.

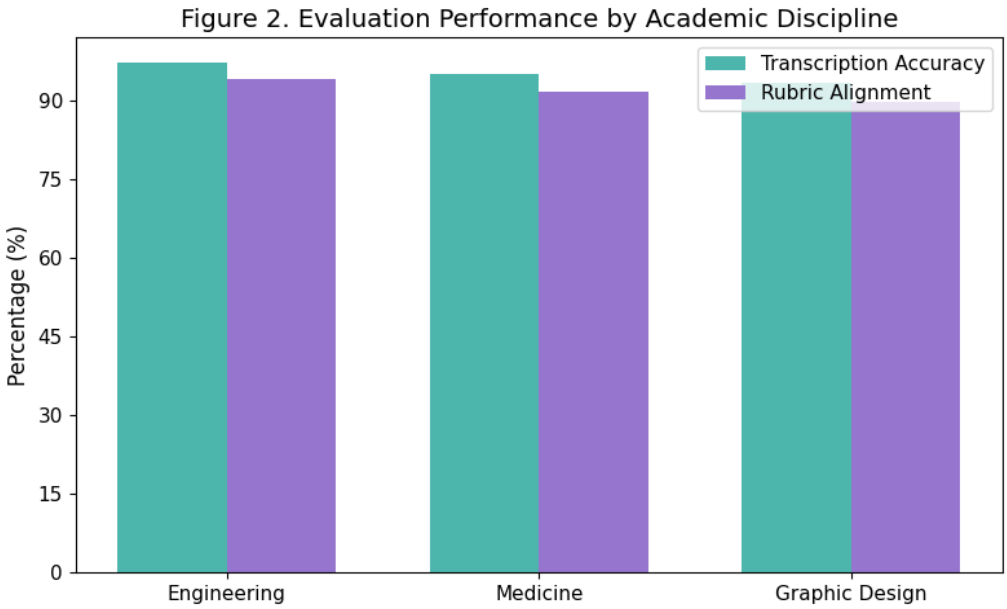

Figure 2. Evaluation performance by academic discipline: Transcription accuracy and rubric alignment (%)

UTA's internal evaluation systems provide context for developing rubric structures. For instance, the Teaching Evaluation System contains domain specific language descriptors such as "designing for visual communication" for Design, "patient-centered delivery" and "procedural fluency" in the medical field, and clarity in technical explanation in Engineering. This language became part of the vectors stored in the context space for rubric matching.

As presented in figure 2, Engineering produced the highest transcription accuracy of 97,2 % and a rubric alignment score of 94,1 %. Medical classes had very similar results, with transcription accuracy of 95,1 % and a rubric alignment score of 91,8 %. Graphic Design sessions were the most variable, with transcription accuracy of 93,4 % and rubric alignment of 89,7 %. These scores indicate the system's responsiveness to disciplinary features of discourse and justify the importance of adapting evaluation algorithms to the communicative structure of each discipline.

Behavioral and Emotional Signal Processing

Besides transcript-based semantic analysis, the system also employed multimodal assessment techniques to assess the behavioral and emotional aspects of teaching. Acoustic data were analyzed to evaluate changes in tone, vocal intensity, and emotional cues in the voice, which were primarily based on changes in pitch that were typically related to variations in speech rate. Video clips were also analyzed with DeepFace to assess emotional expressions, and visual behaviors were tracked to assess engagement and interaction.

This two-part evaluation process allowed the system to identify misalignments between the verbal content of the transcript and what the instructor was delivering. In some instances, transcript data were semantically matched with an instructor's pedagogical goals, but the emotional assessment indicated negative affective states such as boredom or disengagement or had a sense of one-way communication. This indicates the importance of including both verbal and behavioral context in comprehensive assessments of teaching effectiveness.

Table 6. Instructional characteristics extracted from multimodal analysis

Instructional Trait	Input Source	Method of Extraction	Output Interpretation
Tone of Voice	Audio	Acoustic modeling (pitch, tempo, modulation)	Expressiveness and emotional variation
Student Participation	Video	Movement detection and gaze tracking	Engagement index and bidirectional flow
Emotional Expressiveness	Video + audio	Facial recognition and vocal emotion markers	Interpersonal tone and affective signals
Subject-Matter Clarity	Transcribed text	Semantic rubric alignment	Conceptual coherence and clarity
Classroom Interaction	Video + text	Dialog structure and presence of responses	Interactivity level

Validation Against Human Evaluation (New Addition)

To add rigor to the study, the AI rubric alignment scores were compared to human raters' scores from a subsample based on twelve recordings (four per disciplinary area). The results indicated moderate agreement between the system and the human rates. Specifically, Cohen's Kappa values indicated 0,71 for Engineering, 0,68 for Medicine, and 0,65 for Graphic Design (all showing substantial agreement). Furthermore, the Intraclass Correlation Coefficient (ICC) values for each disciplinary area were greater than the 0,80 threshold, confirming high reliability of the automated scoring.

Table 7. Inter-Rater Reliability of AI vs. Human Evaluations

Discipline	Cohen's κ	ICC (2, k)	Interpretation
Engineering	0,71	0,84	Substantial to strong
Medicine	0,68	0,81	Substantial to strong
Graphic Design	0,65	0,8	Moderate to strong

Internal Consistency of Rubric Embeddings (New Addition)

To examine the internal consistency of rubric embeddings we employed Cronbach's Alpha to further assess the reliability of the AI scoring. The results from the analysis emerged with the following Cronbach's Alpha outcomes: α = 0,82 Engineering, α = 0,79 Medicine, and α = 0,76 Graphic Design. Each of these were above the .70 threshold, indicating that the rubric embedded in the system exhibited internal consistency of the grading process, while also producing coherent outcomes across rubric items of evaluation.

Between-Discipline Statistical Comparisons (New Addition)

In terms of statistical comparisons across the three different disciplines, we did not find significant differences in processing efficiency based on a one-way ANOVA test ($F(2,33) = 1,87$, $p=0,17$). However, the accuracy of the rubric alignment did show statistically significant differences as confirmed by the Kruskal-Wallis Test ($H(2) = 6,12$, $p = 0,047$) - with Engineering outscoring Graphic Design. Similarly, with transcription accuracy, across the three disciplines we found statistically significant differences, where the results were also in favor of Engineering ($F(2,33) = 4,29$, $p = 0,021$).

Table 8. Statistical Comparisons Across Disciplines

Metric	Test Applied	Test Value	p-value	Interpretation
Processing Efficiency	One-way ANOVA	$F(2,33) = 1,87$	0,17	No significant difference
Rubric Matching Accuracy	Kruskal–Wallis Test	$H(2) = 6,12$	0,047	Significant, Eng. > Design
Transcription Accuracy	One-way ANOVA	$F(2,33) = 4,29$	0,021	Significant, Eng. > Design

Cost and Time Efficiency Analysis

The overall cost of evaluating 36 evaluations was $127,40 USD, with the overwhelming expenses attributed to transcription and semantic analysis APIs. Alternatively, manual evaluation using traditional peer and administrator models at UTA was conducted at $4500 USD, which accounted for labor hours expended reviewing student work and scoring student work with an established rubric. The 97 % reduction in costs culminates in a considerable improvement in improvement institutional and organizational efficiency.

The time cycle from video upload to report completion and delivery was a little bit less than three hours per session, as compared to 14 weeks as demonstrated by conventional administrative process. This acceleration allows for prompt, formative feedback and allows for pedagogical tweaking as opposed to waiting for summative feedback.

Pilot Limitations and Adaptations

Despite these benefits, several limitations were evident during piloting. For example, at first the transcription accuracy fell to 91,2 % accuracy for speakers with an accent from the coastal regions, in which this was given appropriately adjusting accent-adaptive prompts and vocabulary were needed. Within two weeks the transcription accuracy improved to 96,8 %.

Moreover, laboratory sessions involving overlapping dialogue and physical activities introduced noise in the evaluations. Adjustments to templates and the use of domain-specific prompts increased rubric match scores in these contexts from 84,3 % to 91,7 %.

DISCUSSION

The findings from this study provide evidence that the AI-based evaluation system is technically feasible and pedagogically valid, demonstrates alignment with human rates in reduction of cost and time. The AI-based evaluation system demonstrated transcription accuracy of over 96 %, rubric alignment with over 90 % agreement, and inter-rater reliability indices (Cohen's $\kappa > 0,65$, ICC > 0,80) demonstrated that the automated evaluations were significantly reliable with expert ratings. Overall, this evidence presents the system as an option in place of traditional peer-review evaluation cycles.

Comparison with Previous Studies

The accuracy of the system is aligned with and extends previous studies based on the use of automatic speech recognition (ASR) and AI-enabled grading measures in classroom practice. Prior evidence has shown the accuracy of ASR systems used in classroom environments to drop (due to noise and differences in accent), with classroom averages for ASR were documented between 85 % and 92 %.[34] Evidence has demonstrated that the pairing of ASR and automated writing evaluation can lessen instructor workload, but the reliability of ASR remains challenging in practical classroom contexts.[35] In contrast, our study demonstrated a transcription accuracy of 96,4 % for the AI evaluation protocol under authentic institutional context in Ecuador, demonstrating that the pipeline of the study can produce reliable accuracy in real life (non-laboratory) environments.

Regarding automated grading and rubric alignment, previous studies (for example [36] illustrated the potential of ASR systems in tasks related to assessment (i.e., Whisper), though the study centered around emotional scoring without any validation against human raters. Similarly [37] investigated AI's role in grading physics exams but raised concerns regarding the psychometric issues of rater reliability. The current work closes this gap by demonstrating significant agreement between AI and humans' raters (Cohen's $\kappa > 0,65$, ICC > 0,80), providing statistical support for automated evaluation to inform human judgment for assessing performance in classroom scenarios.

Cost savings also constitute a significant advance.[38] reported workload reductions of roughly 60 % for AI-based grading of written work, but their application was limited to text. The current study achieved a total reduction cost savings of 97 % when evaluating multimodal inputs of audio, video, and rubric. This shows further scalability and applicability at the institutional level.

All together, these comparisons demonstrate cumulative evidence that the current system replicates engines present in previous implementations, however the current system demonstrates combining an integrated, multimodal pipeline verified through statistical evidence of reliability.

Implications for Practice and Policy

The capability of the system to enable evaluation in a consistent, constructive, low-cost, and fast manner has important implications for educational systems. Traditional evaluation processes usually take months to plan, utilize multiple evaluators, and incur a large administrative burden. The system allows transcription, rubric-based scores, and behavioral analysis to be automated in a way to provide meaningful feedback in less than three hours instead of a 14-week cycle.

On a policy level, our approach is aligned with the ISO 21001:2018 standard for educational organizations that call for use of data based evidence and ensure deep transparency in continuous improvement. More specifically, it will assist the accountability requirements from the Ecuadorian Council for Quality Assurance in Higher Education,[39] which demands that institutions demonstrate systematic, objective, and repeatable evaluation processes. This understanding implies that including data from the assessment system driven by AI into accreditation would leverage institutional legitimacy, efficiencies of resources, and timely feedback for faculty for improving their teaching/learning practice.

Limitations and Future Research

Although these initial findings are promising, there are some limitations to the research. The pilot sample was limited to 36 recordings from 18 instructors at a single institution, which limits generalizability. In addition, variability in accents and overlapping speech in laboratory-based classes hindered transcription accuracy in some cases, consistent with [40] survey of ASR systems. Although adaptive prompts, as well as adjustments to the rubric to accommodate speech quality, improved performance, future systems will require more sophisticated contextual modeling and better phonetic modeling to fully address these problems.

Future research should increase scope to multi-institutional, cross-disciplinary studies in order to validate reliability and efficiency in a broader context. Longitudinal studies could also determine whether ongoing feedback with an AI-based system can result in measurable improvements in teaching over time. Additionally, multimodal features (e.g., gesture recognition and student engagement indicators) may create additional value for the course instructor, as outlined in recent AI-based systems for grading.[41] Lastly, ethical dimensions—including privacy, data security, bias, and instructor buy-in—must remain central to scaling the system to a national or regional level.

CONCLUSIONS

The research showed that using AI for evaluating teaching performance is viable in higher educational contexts and is, in fact, pedagogically important in actual higher education contexts. The system integrated automated transcription, semantic analysis, and multimodal emotion recognition to deliver valid and transparent evaluations consistent with institutional expectations.

It was demonstrated that AI evaluation could replace or serve as a supplement to peer-review evaluation of teaching, adding reliable and effective support for processes that assure quality of teaching with improvements in time usage and institutional efficiencies. It additionally offered assistance to normalize evaluation with ISO 21001 educational standards and was accepted into national accreditation frameworks (CACES), reinforcing its institutional relevance.

The research here showed that the gap between experimental AI applications and feasible operational systems that can be used in education and higher education can be bridged. It provided a replicable methodological framework that could be used for, when ready, larger multi-institutional packages.

The pilot focused on a single university; however, the outcome provides an opportunity for continuation of research which will explore ethical, technical, and pedagogical implications of an AI-based evaluation. Future applications and studies will need to develop in larger contexts to establish scalability and long-term educational impacts.

BIBLIOGRAPHIC REFERENCES

1. Kreitzer RJ, Sweet-Cushman J. Evaluating Student Evaluations of Teaching: a Review of Measurement and Equity Bias in SETs and Recommendations for Ethical Reform. J Acad Ethics. 2021. https://doi.org/10.1007/s10805-021-09400-w

2. Otu N, Otu NE. Student Evaluations of Teaching Are Mostly Awfully Wrong. Universal J Educ Res. 2023;2:1–16.

3. Rubie-Davies CM. Teacher expectations and student self-perceptions: Exploring relationships. Psychol Sch. 2006;43(5):537–52. https://doi.org/10.1002/pits.20169

4. Sigurdardottir MS, Rafnsdottir GL, Jónsdóttir AH, Kristofersson DM. Student evaluation of teaching: gender bias in a country at the forefront of gender equality. High Educ Res Dev. 2023;42(4):954–67. https://doi.org/10.1080/07294360.2022.2087604

5. Adams S, Bekker S, Fan Y, Gordon T, Shepherd LJ, Slavich E, et al. Gender Bias in Student Evaluations of Teaching: 'Punish[ing] Those Who Fail To Do Their Gender Right.' High Educ (Dordr). 2022;83(4):787–807. https://doi.org/10.1007/s10734-021-00704-9

6. Lawson RA, Stephenson EF. Easiness, attractiveness, and faculty evaluations: Evidence from RateMyProfessors.com. Atl Econ J. 2005;33(4):485–6. https://doi.org/10.1007/s11293-005-2873-z

7. Bartlett AD, Um IS, Luca EJ, Krass I, Schneider CR. Measuring and assessing the competencies of preceptors in health professions: A systematic scoping review. BMC Med Educ. 2020;20(1):1–9. https://doi.org/10.1186/s12909-020-02082-9

8. Kumar A, Jain R. Faculty Evaluation System. Procedia Comput Sci. 2018;125:533–41. https://doi.org/10.1016/j.procs.2017.12.069

9. Watson D, Amin SN, Pino N. Self-evaluating performance: an analysis of police integrity, professionalism and service provision from the South Pacific. Policing Soc. 2022;32(1):89–102. https://doi.org/10.1080/10439463.2021.1888950

10. Chionidou-Moskofoglou M, Doukakis S, Lappa A. The use of e-portfolios in teaching and assessment. 2021. https://doi.org/10.48550/arXiv.2105.15114

11. Rodgers WJ, Kennedy MJ, VanUitert VJ, Myers AM. Delivering Performance Feedback to Teachers Using Technology-Based Observation and Coaching Tools. Interv Sch Clin. 2019;55(2):103–12. https://doi.org/10.1177/1053451219837640

12. Zhai X, Yin Y, Pellegrino JW, Haudek KC, Shi L. Applying machine learning in science assessment: a systematic review. Stud Sci Educ. 2020;56(1):111–51. https://doi.org/10.1080/03057267.2020.1735757

13. González-Calatayud V, Prendes-Espinosa P, Roig-Vila R. Artificial Intelligence for Student Assessment: A Systematic Review. Appl Sci. 2021;11(12):5467. https://doi.org/10.3390/app11125467

14. Kaswan KS, Dhatterwal JS, Ojha RP. AI in personalized learning. In: Adv Technol Innov High Educ Theory Pract. 2024. p. 103–17.

15. Romero C, Ventura S. Educational data mining and learning analytics: An updated survey. Wiley Interdiscip Rev Data Min Knowl Discov. 2020;10(3):e1355. https://doi.org/10.1002/widm.1355

16. Perez-Ortiz M, Novak E, Bulathwela S, Shawe-Taylor J. An AI-based Learning Companion Promoting Lifelong Learning Opportunities for All. 2021. https://doi.org/10.48550/arXiv.2112.01242

17. Merino-Campos C. The Impact of Artificial Intelligence on Personalized Learning in Higher Education: A Systematic Review. Trends High Educ. 2025;4(2):17. https://doi.org/10.3390/higheredu4020017

18. Ma K, Zhang Y, Hui B. How Does AI Affect College? The Impact of AI Usage in College Teaching on Students' Innovative Behavior and Well-Being. Behav Sci. 2024;14(12):1223. https://doi.org/10.3390/bs14121223

19. Chai F, Ma J, Wang Y, Zhu J, Han T. Grading by AI makes me feel fairer? How different evaluators affect college students' perception of fairness. Front Psychol. 2024;15:1221177. https://doi.org/10.3389/fpsyg.2024.1221177

20. Niculescu AI, Ehnen J, Yi C, Jiawei D, Pin TC, Zhou JT, et al. On the development of an AI performance and behavioural measures for teaching and classroom management. 2025. https://doi.org/10.48550/arXiv.2506.11143

21. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 Technical Report. 2023. https://doi.org/10.48550/arXiv.2303.08774

22. An Y, Kolanupaka S, An J, Ma M, Chhatwal U, Kalinowski A, et al. Is the Lecture Engaging for Learning? Lecture Voice Sentiment Analysis for Knowledge Graph-Supported Intelligent Lecturing Assistant (ILA) System. 2024. https://doi.org/10.48550/arXiv.2408.10492

23. Chen J, Lai P, Chan A, Man V, Chan CH. AI-Assisted Enhancement of Student Presentation Skills: Challenges and Opportunities. Sustainability. 2023;15(1):196. https://doi.org/10.3390/su15010196

24. Giese S. Automatisierungsservice Zapier. In: Das automatisierte Maklerbüro. 2021. p. 65–71. https://doi.org/10.1007/978-3-658-33275-4_5

25. Jeong C. Beyond Text: Implementing Multimodal Large Language Model-Powered Multi-Agent Systems Using a No-Code Platform. J Intell Inf Syst. 2025;31(1):191–231. https://doi.org/10.13088/jiis.2025.31.1.191

26. Owoc ML, Sawicka A, Weichbroth P. Artificial Intelligence Technologies in Education: Benefits, Challenges and Strategies of Implementation. IFIP Adv Inf Commun Technol. 2021;599:37–58. https://doi.org/10.1007/978-3-030-85001-2_4

27. Liu Q, Geertshuis S, Grainger R. Understanding academics' adoption of learning technologies: A systematic review. Comput Educ. 2020;151:103857. https://doi.org/10.1016/j.compedu.2020.103857

28. Al-Zahrani AM. Unveiling the shadows: Beyond the hype of AI in education. Heliyon. 2024;10(9):e30696. https://doi.org/10.1016/j.heliyon.2024.e30696

29. Cha H, Park T, Seo J. What Should Be Considered when Developing ICT-Integrated Classroom Models for a Developing Country? Sustainability. 2020;12(7):2967. https://doi.org/10.3390/su12072967

30. Zou Y, Kuek F, Feng W, Cheng X. Digital learning in the 21st century: trends, challenges, and innovations in technology integration. Front Educ (Lausanne). 2025;10:1562391. https://doi.org/10.3389/feduc.2025.1562391

31. Salas-Pilco SZ, Yang Y. Artificial intelligence applications in Latin American higher education: a systematic review. Int J Educ Technol High Educ. 2022;19(1):1–20. https://doi.org/10.1186/s41239-022-00326-w

32. Hilliger I, Ortiz-Rojas M, Pesántez-Cabrera P, Scheihing E, Tsai YS, Muñoz-Merino PJ, et al. Towards learning analytics adoption: A mixed methods study of data-related practices and policies in Latin American universities. Br J Educ Technol. 2020;51(4):915–37. https://doi.org/10.1111/bjet.12933

33. Dolan EL, Elliott SL, Henderson C, Curran-Everett D, St. John K, Ortiz PA. Evaluating Discipline-Based Education Research for Promotion and Tenure. Innov High Educ. 2018;43(1):31–9. https://doi.org/10.1007/s10755-017-9406-y

34. Southwell R, Pugh S, Perkoff EM, Clevenger C, Bush JB, Lieber R, et al. Challenges and Feasibility of Automatic Speech Recognition for Modeling Student Collaborative Discourse in Classrooms. In: Proc 15th Int Conf Educ Data Mining, EDM 2022. International Educational Data Mining Society; 2022.

35. Li W, Mohamad M, You HW. Integrating automatic speech recognition and automated writing evaluation to reduce speaking anxiety and enhance speaking competence among Chinese EFL learners. Cogent Educ. 2025;12(1). https://doi.org/10.1080/2331186X.2025.2559161

36. McGuire M, Larson-Hall J. Assessing Whisper automatic speech recognition and WER scoring for elicited imitation: Steps toward automation. Res Methods Appl Linguist. 2025;4(1):100197. https://doi.org/10.1016/j.rmal.2025.100197

37. Kortemeyer G, Nöhl J. Assessing confidence in AI-assisted grading of physics exams through psychometrics: An exploratory study. Phys Rev Phys Educ Res. 2025;21(1):010136. https://doi.org/10.1103/PhysRevPhysEducRes.21.010136

38. Ngoc Vu N, Thi Hong Lien N. Leveraging AI for higher education: The role and implications of chatGPT: THE ROLE AND IMPLICATIONS OF CHATGPT. Tap chí Khoa hoc HUFLIT. 2025;7(4):29–29.

39. CACES. Modelo de evaluación externa con fines de acreditación para el aseguramiento de la calidad de las Universidades y Escuela Politécnicas. 2023. https://www.caces.gob.ec/universidades-y-escuelas-politecnicas-3/

40. Ahlawat H, Aggarwal N, Gupta D. Automatic Speech Recognition: A survey of deep learning techniques and approaches. Int J Cogn Comput Eng. 2025;6:201–37. https://doi.org/10.1016/j.ijcce.2024.12.007

41. Zewei, Tian, Liu A, Esbenshade L, Sarkar S, Zhang Z, et al. Implementation Considerations for Automated AI Grading of Student Work. 2025. https://doi.org/10.48550/arXiv.2506.07955

FINANCING

The authors did not receive financing for the development of this research.

CONFLICT OF INTERESTS

The authors declare no conflict of interests.

AUTHORSHIP CONTRIBUTION

Conceptualization: Wilber Orlando Romero Villarroel, Sara Nidhya Camacho Estrada, Héctor Santiago López Zurita, Danilo Fabricio Trujillo Ronquillo, Carlos Patricio Rodríguez Hurtado, Edison Gerardo Llerena Medina.

Data curation: Wilber Orlando Romero Villarroel, Sara Nidhya Camacho Estrada, Héctor Santiago López Zurita, Danilo Fabricio Trujillo Ronquillo, Carlos Patricio Rodríguez Hurtado, Edison Gerardo Llerena Medina.

Formal analysis: Wilber Orlando Romero Villarroel, Sara Nidhya Camacho Estrada, Héctor Santiago López Zurita, Danilo Fabricio Trujillo Ronquillo, Carlos Patricio Rodríguez Hurtado, Edison Gerardo Llerena Medina.

Drafting - original draft: Wilber Orlando Romero Villarroel, Sara Nidhya Camacho Estrada, Héctor Santiago López Zurita, Danilo Fabricio Trujillo Ronquillo, Carlos Patricio Rodríguez Hurtado, Edison Gerardo Llerena Medina.

Writing - proofreading and editing: Wilber Orlando Romero Villarroel, Sara Nidhya Camacho Estrada, Héctor Santiago López Zurita, Danilo Fabricio Trujillo Ronquillo, Carlos Patricio Rodríguez Hurtado, Edison Gerardo Llerena Medina.