2026 ASEE ANNUAL CONFERENCE June 21–24, 2026 | Charlotte Convention Center, NC

Host Sponsor UNC Charlotte

Paper ID #51302

Artificial Intelligence-Assisted Academic Assessment: Reducing Subjectivity in Engineering Courses

Dr. Roberto Patricio Carú, Universidad Andres Bello

Professor with more than 20 years of experience at the University of Talca, FEN, Accounting and Auditor program, Computer Engineering. Director of the Computer Engineering program, UGM. Professor of Logistics, Costs, Management Control, Commercial Engineering, UGM; Professor of Technological Management in Business, Business Intelligence, University of Valparaíso. Professor of Management Systems, Industrial Civil Engineering, UNAB; Thesis Director, Computer Engineering, Business Administration Engineering, MBA, Utalca, UGM; Postgraduate Security, U.Chile.

Dr. Juan Felipe Calderón, Universidad Andres Bello, Viña del Mar, Chile

Juan Felipe Calderón received the bachelor's in computer science and MSc and PhD degrees in engineering sciences from the Pontificia Universidad Católica de Chile. He is an assistant professor in the Faculty of Engineering at the Universidad Andres Bello. His research interests are learning design supported by technology, innovation in engineering education, sustainability in cloud computing, technological infrastructure.

Artificial Intelligence–Assisted Academic Assessment: Reducing Subjectivity in Engineering Courses

Abstract

Manual grading of academic assignments, even when carried out by a single instructor, can be affected by subjective factors such as fatigue, fluctuating interpretations of criteria, and variations in mood, leading to inconsistent grades and limiting the quality of feedback for students. This study addressed these challenges through a semester-long training and continuous validation of multiple standard artificial intelligence (AI) models, including ChatGPT, Gemini, and Claude, aligned with disciplinary rubrics to automatically, objectively, and efficiently review and grade assessments.

A quasi-experimental study was conducted with 120 students from the information systems course in an industrial engineering program, distributed into two groups during the first semester of 2025. Both groups were taught by the same instructor, which enabled the isolation of intrarater subjectivity and the control of pedagogical variables. The control group ($n = 60$) was assessed using traditional methods, while the experimental group ($n = 60$) used an AI system for assessment and feedback. A mixed-methods approach was employed, combining comparative statistical analysis, measurement of grade variability, semi-structured interviews, and focus groups.

The results showed a significant reduction in grade variability in the AI-assessed course compared to the traditional method (standard deviation: 6.6 vs. 11.7 points, $F = 18.7$, $p < 0.001$). The correlation between initial and final grades was significantly higher in the AI group ($r = 0.89$ vs. $r = 0.62$, $p < 0.001$), indicating greater evaluative consistency. The average time to deliver feedback was reduced from 72 to 24 hours. In addition, AI achieved 99.5% evaluative accuracy, allowing the instructor to limit reviews to 1 out of every 100 assessments. Students assessed with AI reported higher satisfaction with feedback consistency (4.2 vs. 3.1 on a 5-point Likert scale, $p < 0.01$), though they valued personalized qualitative feedback less.

The implementation of artificial intelligence proved effective in reducing the subjectivity inherent in human assessment, even when performed by a single instructor, and in improving operational efficiency. However, opportunities were identified to personalize qualitative feedback and to integrate personalized qualitative feedback to optimize the educational experience. The findings suggest that a hybrid model that combines AI for objective assessment and human intervention for formative feedback could maximize the benefits of both approaches.

Keywords: Generative artificial intelligence, AI-assisted assessment, intra-rater subjectivity, systems modeling, formative feedback, teaching efficiency.

Introduction

Assessment constitutes a fundamental pillar of the educational process, not only as a mechanism for measuring learning, but also as a source of formative feedback that guides students' continuous improvement [1]. Within Information Systems courses, where process and requirements modeling demand interpretive judgment, and both technical expertise and the ability to solve complex problems are valued, the quality, consistency, and timeliness of assessment become critically important. However, manual assessment, even when carried out by a single experienced instructor, faces challenges inherent to human cognition: fatigue, workload, variable interpretation of rubrics, and mood fluctuations can generate unintended biases and inconsistencies [2].

These inconsistencies not only affect students perceived fairness but also weaken the pedagogical usefulness of feedback. When grades vary more due to contextual factors than to the actual merit of the work, trust in the assessment system and motivation for improvement are eroded [3]. In addition, the time required to mark and deliver detailed feedback in highenrollment courses often delays its delivery, reducing its formative impact [4].

In this context, artificial intelligence (AI) has emerged as a promising tool to address these challenges. Publicly available large language models (LLMs), such as ChatGPT, Gemini, or Claude, can be aligned with specific rubrics through prompt engineering and systematically applied to multiple submissions, ensuring a uniform application of assessment criteria [5] [6]. Unlike human evaluators, these systems do not experience fatigue or implicit biases, which allows for greater stability in grades over time.

This study is distinguished by having trained and compared multiple standard AI models over an entire academic semester, iteratively refining their prompts to maximize fidelity to the rubric. Furthermore, by having a single instructor for both groups, intra-rater subjectivity, a relatively underexplored dimension in the literature, is isolated, and it is shown that AI can mitigate even the fluctuations in the judgment of the same person.

Literature Review

Assessment in engineering education has been the subject of increasing scrutiny over the past two decades, especially in terms of its fairness, consistency, and formative value. Although it has traditionally been assumed that an experienced instructor applies criteria uniformly, recent research has shown that subjectivity persists even in single-marker assessment contexts [2] [7]. This section reviews three converging strands of literature: (1) the challenges of human assessment in engineering, (2) the nature and manifestations of intra-rater subjectivity, and (3) the emerging role of artificial intelligence (AI) as a tool to improve evaluative objectivity and efficiency.

Human assessment in engineering

In engineering education, assessment must balance technical rigor, where correct answers are often verifiable, with judgment about the quality of reasoning, creativity in problem solving, and communicative clarity [8]. This balance is complex: although many problems have unique

solutions, the path to reaching them allows for variability, which requires the instructor to exercise interpretive judgment rather than purely mechanical marking [9].

This judgment, however, is subject to cognitive constraints. Studies in high-workload contexts have shown that instructors tend to simplify their criteria or rely on heuristics when facing large volumes of student work [2]. In engineering, where courses often have high enrolments and multiple practical assignments, these pressures can compromise the quality and fairness of feedback.

Nature and manifestations of intra-rater subjectivity

Much of the early literature on evaluative inconsistency focused on inter-rater variability, that is, differences between instructors when applying the same rubric [10]. However, more recent research has shown that the same instructor can grade the same work inconsistently at different points in time. A seminal study on judicial decisions, showed that factors such as fatigue or hunger significantly affect human judgment, even among experts[11]. In education, [7]found that the order of marking influences assessor severity: papers reviewed at the end of a stack tend to receive lower or more generous grades, depending on the context.

This inconsistency is argued by [9], explaining that it is not an error but an inevitable consequence of the interpretive nature of assessment: “The instructor does not apply rules, but constructs judgments in real time, influenced by their cognitive and emotional state.” In engineering, where objectivity is highly valued, this subjectivity can generate perceptions of unfairness and diminish confidence in the assessment system [12].

The emerging role of artificial intelligence (AI) as a tool to improve evaluative objectivity and efficiency

In the face of these challenges, generative AI has emerged as a potential solution. Large language models (LLMs) such as ChatGPT, Gemini, or Claude can be trained through prompt engineering to apply rubrics in a systematic and reproducible way [13]. Unlike humans, these systems do not experience fatigue, are not affected by implicit biases, and provide feedback in minutes rather than days [14].

Recent studies have shown that AI can replicate human grades with high correlation in structured tasks [15], and that students particularly value the consistency and speed of automated feedback [6]. However, critical issues have also been identified: AI tends to generate generic comments, with limited ability to recognize innovative reasoning or subtle conceptual errors [16]. This has led to the proposal of hybrid models, in which AI handles objective grading and the instructor focuses on formative feedback with high pedagogical value [17].

In the context of engineering, where technical accuracy and critical thinking are central, this dual approach could offer the best of both worlds: algorithmic efficiency and human pedagogical sensitivity. However, there is still little empirical evidence in real engineering settings, especially in designs that control for the evaluator to isolate the effect of AI on intra-rater subjectivity, a gap that this study seeks to address.

Theoretical Framework

Formative assessment and equity in engineering education

In engineering education, assessment transcends its summative function to become a driver of continuous learning. According to [4], effective feedback must be timely, specific, aligned with clear criteria, and oriented toward closing the gap between current and desired performance. This formative approach is particularly relevant in technical disciplines, where conceptual or procedural errors must be identified and corrected early to prevent the consolidation of misunderstandings [8].

Equity in assessment, understood as the fair, consistent, and transparent application of criteria, is an ethical and pedagogical pillar in accredited engineering programs [18]. However, achieving this equity is a practical challenge, particularly in courses with high enrollment or multiple assignments, where the instructor's cognitive load can compromise the uniformity of evaluative judgment [9].

Subjectivity in Human Assessment

While much of the literature has emphasized inter-rater variability, differences between instructors when applying the same rubric, recent studies have highlighted that intra-rater inconsistency is equally problematic [2]. The same instructor can grade the same work differently at different times due to factors such as cognitive fatigue [11], accumulated workload, or even the order in which submissions are reviewed (the so-called "contrast effect [7]).

This subjectivity does not necessarily imply negligence, but rather inherent characteristics of human processing. As [9] notes, "evaluative judgment is an interpretive act, not merely a technical one," which introduces an inevitable dose of variability. In engineering contexts, where precision and objectivity are expected, this fluctuation can erode students' confidence in the fairness of the assessment process and diminish the usefulness of the feedback received ([3].

Artificial Intelligence as a Tool for Evaluative Consistency

Large language models (LLMs), such as ChatGPT, Gemini, or Claude, can be "aligned" with specific rubrics through prompt engineering, enabling systematic and repeatable application of assessment criteria [5] [6]. Unlike humans, these systems do not experience fatigue, are not affected by implicit biases, and apply the same standards to all submissions, regardless of timing or context.

Preliminary studies in higher education have shown that AI can replicate human grades with high reliability in structured tasks [15] and that students particularly value the consistency and speed of automated feedback [19]. However, a critical limitation has also been identified: AI tends to generate generic feedback that is insensitive to the subtleties of individual reasoning, reducing its formative impact in complex tasks [16].

This has led to the proposal of hybrid models, where AI handles objective grading and detection of common errors, while the instructor focuses on qualitative feedback, mentoring, and pedagogical decision-making [17] In engineering, where technical precision and critical thinking are central, this dual approach could balance efficiency, equity, and pedagogical depth.

Research Objective

The objective of this study was to analyze the impact of implementing artificial intelligence (AI) systems trained and validated over a complete academic semester on reducing evaluative subjectivity and improving the quality, consistency, efficiency, and effectiveness of feedback in industrial engineering courses, within a controlled context where a single instructor taught both courses. Specifically, it sought to:

1. Describe the manifestations and magnitudes of subjectivity in traditional assessments conducted manually by the instructor during the first semester of 2025, with emphasis on grade variability and inconsistency in criteria application.
2. Compare the objectivity, temporal consistency, and perceived quality of feedback between two assessment modalities: traditional (human) and AI-assisted, using quantitative metrics (variability, correlation, delivery time, effectiveness) and qualitative metrics (student satisfaction, perceived depth).
3. Explain the mechanisms through which AI contributed or failed to reduce inherent human judgment subjectivity, considering technical factors (alignment with rubrics, reproducibility) and pedagogical factors (ability to provide meaningful formative feedback).

Methodology

Study Design

A quasi-experimental design with two non-equivalent groups [20] was implemented, combined with a sequential explanatory mixed-methods approach [21]. This design was selected because, although random assignment of students was not possible, key variables were controlled, such as the instructor, course content, rubrics, and schedule, to isolate the effect of assessment modality. The control group (n = 60) received feedback and grading through traditional manual marking by the instructor. The experimental group (n = 60) was assessed by an artificial intelligence (AI) system aligned with the same rubric, with final instructor oversight to validate overall coherence (without modifying individual grades).

Both groups corresponded to sections of the same course taught by a single instructor during the first semester of 2025 in the Industrial Engineering program at a private university in Chile. This condition allowed the analysis to focus on intra-rater subjectivity, that is, fluctuations in one person's judgment when marking at different times, and to evaluate whether AI contributes to stabilizing criteria application. To strengthen the validity of the findings, internal triangulation was applied: quantitative results (variability, correlations, times) were systematically contrasted with qualitative data (interviews and focus groups), enabling a deeper and more coherent interpretation of AI's effects on academic assessment.

Selection of the Language Model

After comparing ChatGPT-4, Gemini 1.5 Pro, and Claude 3 Sonnet, Claude 3 Sonnet was selected as the base model due to its superior performance in structured reasoning tasks and low rate of technical hallucination (Anthropic, 2024). According to the MMMU benchmark (Yue et al., 2024), Sonnet achieves 64.2% in engineering, surpassing GPT-4 (61.8%) and Gemini (59.3%). Additionally, its 200K token context window allows processing complete submissions (including textually described diagrams) without truncation. Its training with Constitutional AI reduces biases in assessment [22] an approach that prioritizes truthfulness, coherence, and

absence of bias in responses, qualities critical in evaluative contexts where technical accuracy and fairness are essential.

Participants
The sample consisted of 120 students enrolled in two parallel sections of the Information Systems course in the Industrial Engineering program. There were no significant differences between the groups in terms of gender ($\chi^2 = 0.32$, $p = 0.57$), prior academic average ($t(118) = 0.89$, $p = 0.38$), or prior experience with digital tools ($t(118) = 1.02$, $p = 0.31$), which reinforces initial comparability. All participants signed an informed consent form.

Intervention: AI-assisted assessment system
During the first semester of 2025, three publicly available large language models (LLMs) were trained and compared: ChatGPT-4 (OpenAI), Gemini 1.5 Pro (Google), and Claude 3 Sonnet (Anthropic). Each model was configured with a structured prompt (see Appendix D) and subjected to an iterative validation process: each week, 10% of the evaluations generated by each AI were randomly reviewed, and examples and criteria in the prompt were adjusted to correct deviations.

The configuration process required several engineering steps: defining a detailed rubric, translating rubric criteria into structured prompt instructions, testing multiple language models, reviewing a representative sample of outputs each week, and refining examples and decision rules until the system stabilized. Although these steps were technically demanding at the beginning, they did not require advanced software development or model training from scratch. In practice, the process was more operational than computational, but it did require careful prompt design, repeated validation, and close coordination between the instructor and the AI system.

From week 12 onward, the consolidated system achieved 99.5% effectiveness, defined as the proportion of evaluations that required no correction after human review. This meant that, on average, the instructor only needed to intervene in 1 out of every 100 submissions.

Instruments and Data Sources
Quantitative and qualitative data were collected:

1. Initial and final grades: Two individual written assignments on Information Systems modeling, based on real business cases adapted to an academic context, were evaluated at the beginning and end of the semester. Each submission required:

    - Requirements analysis (actors, processes, functional/non-functional),
    - Textually described graphical modeling:
      ○ Use case diagram (UML), ○ BPMN 2.0 model of the main process (with start, end, decisions, and exceptions), ○ Class diagram (attributes, relationships, cardinality) ○ technical solution proposal with feasibility estimation (time, costs, technology).

2. Feedback time: Automatically recorded from submission until grade publication

3. Satisfaction survey: 5-point Likert scale (1 = very dissatisfied, 5 = very satisfied) on feedback consistency, clarity, and usefulness ($\alpha = 0.84$)

4. Semi-structured interviews: Conducted with 12 purposefully selected students (6 per group), focused on perceptions of fairness and formative utility
5. Focus groups: Two sessions (one per group, n = 8 per session) to explore collective dynamics regarding trust in the assessment system. Appendix E

Data Analysis

Quantitative data were analyzed using statistical software (R, version 4.3.2), employing the tidyverse, car, and psych packages. Descriptive statistics were calculated, along with Welch's independent t-tests to compare means between groups, Levene's test to assess homogeneity of variances, and Pearson correlations to examine evaluative consistency. Statistical significance was set at $p < 0.05$.

For qualitative data, an inductive thematic analysis was conducted [23] through an iterative reflexive review process: interview and focus group transcripts were read repeatedly to identify patterns, initial codes were generated and then grouped into central themes. To strengthen the credibility of the analysis, internal triangulation with quantitative data was used (e.g., contrasting student perceptions with their grades and satisfaction levels), which enabled validation of the coherence of the findings.

Results

The analysis of data collected during the first semester of 2025 revealed key findings regarding evaluative consistency, operational efficiency, and student perception. The results are presented below, organized according to the study's central dimensions: variability in grades, temporal consistency, feedback time, student satisfaction, and operational effectiveness.

Variability in final grades

No significant differences were observed in average performance between the groups ($M = 68.5$ for both), indicating that AI did not artificially alter grades toward higher or lower values. However, variability was notably lower in the AI-assessed group. The standard deviation in the control group was 11.7 points, while in the experimental group it was 6.6 points. Levene's test confirmed that this difference in variances was statistically significant ($W = 18.7$, $p < 0.001$), suggesting greater stability in the application of assessment criteria by the AI system.

In addition to reducing variance, the AI system showed high grading precision when compared against instructor validation. After the iterative refinement phase, only 0.5% of the evaluations required human correction, corresponding to 2 out of 480 total assessments. This indicates that the system was able to apply the rubric consistently and with near-human reliability in structured engineering tasks, although exceptional cases involving creative reasoning or hidden contextual constraints still required instructor intervention.

These results are summarized in Table 1.

Table 1: Descriptive statistics of final grades by group (0–100 scale)

Group	N	Mean	Standard Deviation	Mínimum	Máximum
Control	60	68.5	11.7	49.8	96.2

Experimental	60	68.5	6.6	52.0	94.0

Note: Grades correspond to the second course deliverable

Evaluative consistency over time

The Pearson correlation between initial and final grades, a measure of coherence in evaluative judgment, was significantly higher in the experimental group ($r = 0.90$) than in the control group ($r = 0.58$). Fisher's r-to-z test confirmed that this difference was statistically significant ($z = 4.12$, $p < 0.001$). This indicates that while the AI system maintained a highly consistent application of the rubric across both assessments, the instructor showed greater fluctuation in judgments, even when grading the same students in similar contexts (Table 2).

Table 2: Correlations between initial and final grades by group

Group	**r**	**p**
Control	0.58	<0.001
Experimental	0.90	<0.001

Efficiency and Student Satisfaction

The average time between work submission and feedback publication was drastically reduced in the experimental group. While the control group waited an average of 72.1 hours (SD = 2.6), the AI group received feedback in 24.3 hours (SD = 1.4). The difference was highly significant ($t(118) = 128.4$, $p < 0.001$, Welch's t-test).

Students in the experimental group reported greater satisfaction with the consistency of the feedback received ($M = 4.2$, $SD = 0.4$) than did students in the control group ($M = 2.9$, $SD = 0.7$), $t(118) = –12.3$, $p < 0.001$. However, in semi-structured interviews and focus groups, several AI group students noted that while the feedback was clear and timely, it lacked personalized qualitative nuances. One student commented: "The AI told me it was wrong but didn't understand why I did it that way or help me think differently." In contrast, control group students valued the instructor's ability to provide contextualized comments, although they acknowledged these arrived late and varied in depth. These findings are summarized in Table 3.

Table 3: Efficiency and student satisfaction

Variable	**Control Group**	**Experimental Group**	**t(118)**	**p**
Feedback time (hours)	M = 72.1, SD = 2.6	M = 24.3, SD = 1.4	128.4	<0.001
Satisfaction with consistency (1-5)	M = 2.9, SD = 0.7	M = 4.2, SD = 0.4	-12.3	<0.001

Note: Satisfaction was measured using a 5-point Likert scale (1 = very dissatisfied, 5 = very satisfied). Appendix A.

Operational effectiveness of the AI system
The AI system achieved 99.5% effectiveness in autonomous grading, measured as the proportion of submissions that required no correction after the instructor's random review. This reduced the instructor's workload from 60 weekly corrections to less than 1, freeing up time for higher-value pedagogical formative activities (Table 4).

Table 4: Cases requiring human intervention (0.5% of evaluations)

Case	Solution Type	AI Error/Limitation	Instructor Judgment	Complexity Category
Case 1	Use of the *Event Sourcing* pattern for clinical history (instead of traditional CRUD)	AI was graded as "incomplete" for not following the standard BPMN flow	Recognized as a **creative and technically solid solution**, aligned with modern architectures	Non-linear reasoning / conceptual innovation
Case 2	24/7 high availability proposal without a budget for replication	AI failed to detect the **contradiction between functional requirements and organizational constraints**	Penalized for lack of technical realism, but the analysis of needs was valued	Implicit requirements conflict / organizational context

0.5% of cases that required human intervention (2 out of 480 total evaluations) corresponded to:

- (case 1) Non-standard solution that the AI graded as "incomplete" for deviating from expected patterns, but which the instructor recognized as creative (e.g., use of Event Sourcing pattern instead of CRUD for clinical history management).
- (case 2) Context error: the AI failed to detect a contradiction between functional requirements and stated organizational constraints (e.g., "24/7 high availability" vs. "no budget for replication").

In all cases, the AI generated technically correct feedback but underestimated the pedagogical value of alternative reasoning or failed to identify implicit assumptions in students' responses, a limitation consistent with discussions about its lack of deep contextual understanding.

Discussion

The findings of this study provide strong empirical evidence on the potential of artificial intelligence (AI) to mitigate one of the most persistent tensions in engineering assessment: the subjectivity inherent in human judgment, even when a single instructor applies the criteria. Unlike previous studies that have attributed variability to differences between evaluators [10] [2]

this work demonstrates that inconsistency also arises within the same evaluator, likely due to factors such as cognitive fatigue, workload, or fluctuating interpretation of the rubric [7],[9]. In contrast, AI applied the same standards reproducibly, reducing the grade standard deviation by nearly 44% (from 11.7 to 6.6 points) and raising the assessment correlation to near-perfect consistency ($r = 0.90$).

This result is particularly relevant for engineering education, where assessment systems are expected to reflect the discipline values of precision, transparency, and equity [18]. The high intra-rater correlation in the AI group suggests that students received predictable feedback aligned with their prior performance, reinforcing perceptions of procedural fairness, a key factor in student motivation and engagement [12].

Furthermore, reducing feedback time from 72 to 24 hours not only improves operational efficiency but also enhances the formative value of assessment. As [4] note, "feedback loses effectiveness if it is not timely." In courses with frequent submissions, such as engineering, this acceleration allows students to correct errors before they become entrenched, fostering iterative, practice-based learning.

The 99.5% effectiveness achieved after 12 weeks of iterative training demonstrates that standard AI models, while not perfect, can reach near-human reliability levels in structured engineering tasks. This finding is particularly relevant for institutions with limited resources, where instructor overload compromises the quality of feedback. The fact that only 1 review was required per 100 assessments suggests that AI not only improves equity but also transforms the sustainability of the assessment process.

Qualitative data, however, reveal a key opportunity: while AI delivers clear, fair, and consistent feedback, its potential is enhanced when complemented by the instructor's pedagogical depth, especially for guiding alternative reasoning, unstated intentions, or subtle conceptual errors. Many students noted that AI-generated comments failed to capture their intentions, alternative reasoning, or subtle conceptual errors, aspects that the instructor did recognize. This finding supports [16] and [6]"Language *models can reproduce surface patterns with high reliability but still cannot emulate the contextual understanding and pedagogical sensitivity that characterize quality human feedback."*

This should not be interpreted as an insurmountable limitation of AI, but rather as an opportunity to redesign the instructor's role. Instead of spending hours on mechanical grading, the instructor could focus on what AI cannot do: guiding critical thinking, fostering technical creativity, and providing formative personalized feedback. This hybrid approach, AI for objective grading, human for professional formation, aligns with emerging recommendations in the literature [17], [19] and with student-centered teaching principles.

As illustrated in Table 4, the cases that escaped automated evaluation involved high-cognitive complexity solutions, where creativity or context exceeded the structured patterns that AI can recognize.

Finally, it is important to highlight that the AI system's effectiveness critically depended on rigorous alignment with the course of rubric through prompt engineering. This result shows that high reliability is not automatic: it depends on careful engineering of the assessment workflow, including rubric alignment, prompt iteration, human review thresholds, and explicit handling of

ambiguous cases. Without this calibration, the same AI system could produce less stable results, especially in tasks that involve open-ended reasoning, creativity, or implicit assumptions. Therefore, the proposed approach should be understood as a carefully engineered assessment pipeline rather than a fully autonomous grading tool.

This underscores that AI is not a "magic black box," but a tool whose pedagogical value is mediated by instructional design. The instructor's responsibility does not disappear; it transforms.

Overall, this study contributes to the literature by demonstrating that AI can address intra-rater subjectivity, an underexplored dimension, and by proposing a viable hybrid model for engineering education. The results support the hypothesis that automation does not replace the instructor but frees them to perform higher-value pedagogical functions.

Conclusions

This study demonstrates that implementing artificial intelligence (AI) systems in engineering course assessment can significantly reduce the subjectivity inherent in human judgment, even when a single instructor is responsible for both assessment modalities. The results confirm that AI does not alter students' average performance but improves the consistency, equity, and efficiency of the assessment process. The reduction in standard deviation (from 11.7 to 6.6 points), the increase in intra-assessment correlation ($r = 0.90$), and the acceleration of feedback time (from 72 to 24 hours) constitute robust quantitative evidence of AI's value as a teaching support tool.

Additionally, students explicitly recognized the greater consistency of AI-generated feedback, reinforcing their perception of fairness in the assessment process. However, they also identified a key limitation: the lack of personalized qualitative depth, a dimension where human intervention remains irreplaceable.

Overall, these findings support a hybrid assessment model in which AI handles objective, standardized grading, while the instructor focuses on high-value, pedagogical formative feedback. This approach not only optimizes instructor time but also aligns assessment with the principles of equity, transparency, and continuous improvement that govern engineering accreditation [18]. In this context, AI does not replace the engineer-instructor but empowers them.

Limitations

This study presents several critical considerations for interpreting its findings. First, the quasiexperimental design, although it controlled key variables (same instructor, same content, same rubric), did not allow random assignment of students, which could introduce unobserved biases. Second, the intervention was limited to a single intermediate-level course in the Industrial Engineering program at a higher education institution in Chile, restricting the generalizability of results to other contexts (e.g., introductory courses, non-technical disciplines, or institutions with different assessment cultures).

Additionally, although the prompt was refined weekly throughout the semester, it remains a static prompt engineering approach, without integration of automated feedback loops or model finetuning. Future implementations could incorporate systems that learn from instructor corrections in real time. Finally, student satisfaction was measured using a self-reported scale,

which may be subject to social desirability bias or immediate perceptions, and may not capture the long-term impact on learning.

Although the present study was conducted in an Information Systems course within an Industrial Engineering program, the proposed AI-assisted assessment workflow may be transferable to other engineering disciplines and class sizes only when the rubric is sufficiently explicit, the assessment criteria are structured, and instructor oversight is maintained. In larger cohorts or less structured disciplines, the model would likely require additional validation, calibration, and workload controls before broad adoption. From an ethical and governance perspective, institutions implementing AI-assisted grading at scale should define clear transparency rules regarding how grades are produced, provide students with access to appeal decisions, and maintain documented human review procedures for exceptional or contested cases.

The effectiveness achieved reflects not only technical precision but also AI's limits in contexts of high ambiguity or creativity. The cases requiring human intervention illustrate that, while AI excels at applying structured rubrics to conventional solutions, it requires instructor support when encountering:

- Non-linear reasoning (e.g., disruptive solutions based on advanced design patterns),
- Implicit requirements conflicts (e.g., balancing security and usability),
- Organizational context interpretation errors (e.g., technical feasibility vs. budget constraints).

This empirically corroborates the proposal of a hybrid model in which AI serves as the first filter (99.5% of cases), and the instructor handles the 0.5% of high-cognitive-complexity cases, thereby optimizing assessment process sustainability without sacrificing pedagogical depth.

Future Work

The findings of this study open multiple avenues for future research. First, it is recommended to replicate the design across diverse engineering contexts and curricular levels to assess the robustness of the hybrid model. Second, more sophisticated AI architectures are suggested, such as adaptive feedback systems that combine LLMs with domain-specific rules or symbolic reasoning models, or ensemble strategies that experiment with different algorithms to optimize results according to varying data models and task types (e.g., structured vs. creative submissions) to improve the ability to detect deep conceptual errors.

Additionally, it would be valuable to investigate the long-term impact of hybrid feedback on learning, self-regulation, and persistence in engineering careers. From a pedagogical perspective, future studies could design and implement AI-instructor co-assessment protocols, where both agents generate complementary feedback, and students learn to contrast and synthesize multiple evaluative sources, a key competency in the digital era.

Finally, developing ethical frameworks and governance for AI implementation in academic assessment is recommended, including algorithmic transparency, appeal rights, and instructor training in algorithmic literacy. The future of engineering assessment is not human versus machine, but human with machine.

References

[1] P. Black and D. Wiliam, “Assessment and Classroom Learning,” *Assess. Educ.*, vol. 5, no. 1, pp. 7–74, Mar. 1998, doi: 10.1080/0969595980050102.

[2] S. Bloxham, B. den-Outer, J. Hudson, and M. Price, “Let’s stop the pretence of consistent marking: exploring the multiple limitations of assessment criteria,” *Assess. Eval. High. Educ.*, vol. 41, no. 3, pp. 466–481, Apr. 2016, doi: 10.1080/02602938.2015.1024607.

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

[4] J. Hattie and H. Timperley, “The Power of Feedback,” *Rev. Educ. Res.*, vol. 77, no. 1, pp. 81–112, Mar. 2007, doi: 10.3102/003465430298487.

[5] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, “Systematic review of research on artificial intelligence applications in higher education—Where are the educators?,” *International Journal of Educational Technology in Higher Education*, vol. 16, p. 39, 2019.

[6] R. Luckin, K. George, and M. Cukurova, *AI for School Teachers*. Boca Raton: CRC Press, 2022. doi: 10.1201/9781003193173.

[7] E. Jones, M. Priestley, L. Brewster, S. J. Wilbraham, G. Hughes, and L. Spanner, “Student wellbeing and assessment in higher education: the balancing act,” *Assess. Eval. High. Educ.*, vol. 46, no. 3, pp. 438–450, Apr. 2021, doi: 10.1080/02602938.2020.1782344.

[8] R. M. . Felder and Rebecca. Brent, *Teaching and learning STEM : a practical guide*. Jossey-Bass, 2016.

[9] D. R. Sadler, “Beyond feedback: developing student capability in complex appraisal,” *Assess. Eval. High. Educ.*, vol. 35, no. 5, pp. 535–550, Aug. 2010, doi: 10.1080/02602930903541015.

[10] M. Price *, “Assessment standards: the role of communities of practice and the scholarship of assessment,” *Assess. Eval. High. Educ.*, vol. 30, no. 3, pp. 215–230, Jun. 2005, doi: 10.1080/02602930500063793.

[11] S. Danziger, J. Levav, and L. Avnaim-Pesso, “Extraneous factors in judicial decisions,” *Proceedings of the National Academy of Sciences*, vol. 108, no. 17, pp. 6889–6892, Apr. 2011, doi: 10.1073/pnas.1018033108.

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

[13] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, “Systematic review of research on artificial intelligence applications in higher education – where are the educators?,” *International Journal of Educational Technology in Higher Education*, vol. 16, no. 1, p. 39, Dec. 2019, doi: 10.1186/s41239-019-0171-0.

[14] L. Kohnke, B. L. Moorhouse, and D. Zou, “Exploring generative artificial intelligence preparedness among university language instructors: A case study,” *Computers and Education: Artificial Intelligence*, vol. 5, p. 100156, 2023, doi: 10.1016/j.caeai.2023.100156.

[15] L. Zhao, “Blockchain adoption and contract coordination driven by supplier encroachment and retail service: An analysis from consumers’ information traceability awareness,” *Technol. Soc.*, vol. 73, p. 102237, May 2023, doi: 10.1016/j.techsoc.2023.102237.

[16] E. Kasneci *et al.*, “ChatGPT for good? On opportunities and challenges of large language models for education,” *Learn. Individ. Differ.*, vol. 103, p. 102274, Apr. 2023, doi: 10.1016/j.lindif.2023.102274.

[17] I. Roll and R. Wylie, “Evolution and Revolution in Artificial Intelligence in Education,” *Int. J. Artif. Intell. Educ.*, vol. 26, no. 2, pp. 582–599, Jun. 2016, doi: 10.1007/s40593016-0110-3.

[18] ABET, “Criteria for Accrediting Engineering Programs,” 2023, *Baltimore, MD*. [Online]. Available: https://www.abet.org/wp-content/uploads/2023/05/20242025_EAC_Criteria.pdf

[19] L. Kohnke, B. L. Moorhouse, and D. Zou, “Exploring generative artificial intelligence preparedness among university language instructors: A case study,” *Computers and Education: Artificial Intelligence*, vol. 5, p. 100156, 2023, doi: 10.1016/j.caeai.2023.100156.

[20] W. R. . Shadish, T. D. . Cook, and D. T. . Campbell, *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth/Cengage Learning, 2002.

[21] J. Creswell and V. Plano Clark, *Designing and conducting mixed methods research*. Sage publications, 2011.

[22] S. Huang *et al.*, “Collective Constitutional AI: Aligning a Language Model with Public Input,” in *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, New York, NY, USA: ACM, Jun. 2024, pp. 1395–1417. doi: 10.1145/3630106.3658979.

[23] V. Braun and V. Clarke, “Using thematic analysis in psychology,” *Qual. Res. Psychol.*, vol. 3, no. 2, pp. 77–101, Jan. 2006, doi: 10.1191/1478088706qp063oa.

APPENDICES

Appendix A: Student Satisfaction Survey on Feedback
Instructions: Please respond to the following questions about the feedback received in this course. Use a scale of 1 to 5, where:
1 = Very dissatisfied / Strongly disagree
2 = Dissatisfied / Disagree
3 = Neutral
4 = Satisfied / Agree
5 = Very satisfied / Strongly agree

Item	Statement	Note
1	The feedback I received was clear and easy to understand.	
2	The feedback was consistent with my previous submissions.	
3	The feedback helped me identify my technical errors.	
4	The feedback included helpful suggestions for improvement.	
5	The feedback arrived in time to be useful for my learning.	
6	Overall, I am satisfied with the quality of the feedback received.	

Demographic Data:

- Gender: ☐ Male ☐ Female ☐ Other ☐
- Prior academic average: _____
- Prior experience with AI tools: ☐ None ☐ Basic ☐ Intermediate ☐ Advanced

Appendix B: Assessment Rubric (Used by both the instructor and the AI system)

Criterion	Excellent (90-100)	Good (75-89)	Acceptable (60-74)	Insufficient (<60)
Case and requirements analysis	Identifies all actors, processes, functional/nonfunctional requirements, and constraints clearly and completely.	Omits 1–2 minor elements (e.g., one nonfunctional requirement).	Partial analysis; missing key actors or processes.	Superficial or incorrect analysis; problem not understood.

Systems modeling quality	Uses BPMN/UML notation correctly, without errors; complete, coherent, and wellstructured models.	Minor notation errors or redundancies, but functionally valid.	Significant errors affect understanding of flow or structure.	Incorrect, incoherent, or missing models.
Value proposition and feasibility	Clearly explains how the solution improves processes, with realistic estimates of benefits and technical/organizational feasibility.	Reasonable proposal but lacks depth in impact or feasibility.	Generic or unrealistic proposal.	No value proposition or out of context.

Final Grade: _____ / 100

Comments (required for instructor; AI-generated in experimental group):

Appendix C: Semi-Structured Interview

Objective: Explore perceptions of fairness, utility, and quality of the feedback received. **Key Questions:**

1. How would you describe the feedback you received on this course?
2. Did you feel the assessment was fair and consistent throughout the semester? Why?
3. To what extent did the feedback help you improve in subsequent submissions?
4. Did you notice differences in how you were graded compared to peers (if applicable)?
5. If the feedback was AI-generated: Did it feel impersonal? What was missing?
6. If the feedback was provided by the instructor: Was it timely? Was it in-depth?
7. What type of feedback do you prefer for future courses: fast and consistent, or slower but personalized?
8. Do you have suggestions for improving the assessment process in engineering courses?

Notes for the Interviewer:

- Maintain neutrality; do not reveal study hypotheses.
- Probe deeper with "why?", "can you give me an example?".
- Record with consent (audio or notes).

Appendix D: Structured Prompt Used for AI-Assisted Assessment. Information Systems Course

Instructions for the AI Model (ChatGPT, Gemini, Claude, or other LLM):
Act as an "expert Information Systems evaluator." You will grade an undergraduate student's academic submission for the "Information Systems" course. Follow the instructions below strictly:

1. Read the business case and the student's proposed solution (including textual descriptions of diagrams, models, or tables).
2. Evaluate the submission using only the attached detailed rubric.
3. Assign a numerical grade on a scale of 0 to 100 points. For each evaluation item, assign a score proportional to the student's achievement.
4. Generate written feedback in three clear sections:
    - **Strengths:** What aspects of the solution are correct, complete, or well-founded?
    - **Areas for improvement:** Where are these errors, omissions, inconsistencies, or poor modeling practices?
    - **Specific suggestions:** What should the student do to improve in future submissions?

Evaluation Rubric (total: 100 points)

Criterion	**Maximum Score**	**Description**
Case and requirements analysis	35	Clear identification of actors, business processes, functional/non-functional needs, and system constraints. - Identification of relevant actors (users, roles, external systems): 10 points - Clear description of affected business processes: 10 points - Specification of functional and non-functional requirements: 10 points - Recognition of technical or organizational constraints: 5 points

Systems modeling quality	40	Correct use of standard notation (BPMN, UML, flowcharts, etc.), logical coherence, completeness, and alignment with the business case. **Use case diagram:** Well-defined actors and cases, no redundancies: 8 points **BPMN model (business process):** Correct use of events, tasks, gateways, and flows: 12 points Completeness (start, end, decisions, exceptions): 8 points **Class diagram or data flow (depending on approach):** Correct UML/DFD notation, coherent attributes and relationships: 12 points
Value proposition and feasibility	25	Clarity on how the solution improves processes, reduces costs, or adds value, technical and organizational realism. Clear explanation of how the solution improves current processes: 10 points Realistic estimation of benefits (time, costs, errors): 8 points Consideration of technical and organizational feasibility: 7 points

Methodological Notes (not included in the prompt, but relevant to the study):

- This prompt was refined weekly during the 16 weeks of the 2025-1 semester, based on random review of 10% of AI-generated evaluations.
- Three standard AI models were tested: ChatGPT-4, Gemini 1.5 Pro, and Claude 3 Sonnet.
- The consolidated system (based on Claude 3 Sonnet) achieved 99.5% effectiveness, reducing instructor intervention to 1 review per 100 submissions.
- The same rubric was applied to both the control group (manual assessment) and the experimental group (AI assessment), ensuring comparability.

Appendix E: Focus Groups Guide

Objective: Explore collective perceptions about the fairness, utility, and quality of the feedback received in the course.
Duration: 45–60 minutes
Participants: 6–8 students per group (one per section: control and experimental)
Recording: With informed consent

Guide Questions:

- How would you generally describe the feedback you received on this course?
- Did you feel the grades fairly reflected your performance? Why?
- Did you notice differences in feedback between submissions (first vs. second)?
- What aspects of the feedback were most helpful for improvement?
- If your group used AI: Did you trust its grades? What do you think of its style?
- If your group was assessed by the instructor: Was it timely? Was it sufficiently detailed?
- Do you prefer fast and consistent feedback, or slower but personalized?
- Do you have suggestions for improving the assessment system for future courses?

Notes for the Moderator:

- Encourage interaction between participants ("Does anyone else think the same?").
- Avoid leading toward expected responses.
- Record non-verbal observations (hesitations, tacit agreements, etc.).

Appendix F: Example of AI Feedback for the Student

Grade: 82 / 100

Strengths

- You correctly identified the main actors in the case (customer, seller, inventory system), demonstrating solid understanding of the business context.
- Your use case diagram is clear and covers essential functionalities (search product, place order, check stock).
- The improvement proposal includes a reasonable estimate of reduced service times (30%), showing value-oriented thinking.

Areas for Improvement

- Incomplete BPMN model: Missing explicit start event and exception handling (e.g., "product out of stock"). This reduces process robustness.
- Non-functional requirements omitted: No mention of security, response time, or usability, key elements in real systems.
- Class diagram with notation errors: Attributes lack data types (e.g., "name" without specifying string), and cardinality is missing in relationships.
- Technical feasibility mentioned generically ("can be implemented with current technology"), without referencing specific languages, frameworks, or integrations.

Specific Suggestions for Improvement

- Review the BPMN 2.0 notation guide and ensure start/end events, decision gateways, and exception flows are included in your processes.
- Complete your analysis with at least three non-functional requirements (e.g., "System must respond in under 2 seconds," "Must authenticate users with secure credentials").
- In the class diagram, specify data types for each attribute (string, int, date, etc.) and define multiplicity in each relationship (1..1, 0..*, etc.).
- For the value proposition, research and mention concrete technologies (e.g., "Implemented with React for frontend and Node.js + MongoDB backend") to strengthen feasibility.

With these adjustments, your model will not only be technically stronger but also closer to industry expectations.

Appendix G: Detailed Business Case, Medical Appointment Management System at a University Clinic

1. Organizational Context

The "Salud Integral" University Clinic serves a population of 18,000 students and 2,500 staff, with limited staffing of 4 general practitioners, 2 nurses, and 3 administrative personnel, distributed across 2 daily shifts (morning and afternoon, 4 hours each). Currently, appointment management is done manually using paper spreadsheets and phone calls, which generates:

- Long in-person queues (average: 45 minutes wait),
- Work overload for administrative staff (up to 80 calls/day),
- High no-show rates (32% on average),
- Difficulty rescheduling or managing emergencies.

2. Objective of the Proposed System

Design a medical appointment management system that:

- Reduces average waiting time to ≤ 15 minutes,
- Decreases phone calls by ≥ 70%,
- Minimizes no-show losses through automated reminders,
- It is operable with current infrastructure (1 local server, 3 desktop computers, 10 Mbps symmetric internet connection).

3.Functional Requirements (minimum expected) The system must allow:

ID	Requirement
RF-01	User registration (students/staff) with ID number, program/position, and contact information (email + phone).
RF-02	Display of doctors' availability by specialty and schedule (in 15-minute blocks).
RF-03	Appointment booking, modification, and cancellation by the user.
RF-04	Automatic notification via SMS and email 24 hours before the appointment.
RF-05	Daily report generation: occupancy, no-show list, pending requests.
RF-06	Automatic waitlist management: if a patient cancels, the first person on the list is notified.

4.Non-functional requirements (critical for assessment)

Category	Requirement
Availability	The system must operate 24/7, with fault tolerance of ≥ 99.5% (maximum 3.6 hours/month downtime).
Category	**Requirement**
Performance	Response time ≤ 2 seconds for critical operations (search, booking) with 200 concurrent users.
Security	Authentication using institutional credentials (LDAP), TLS encryption, and basic HIPAA compliance (protection of health data).
Usability	Accessible interface (WCAG 2.1 AA), with support for users with low technological familiarity (e.g., older staff).
Maintainability	Modular design allows the addition of future modules (e.g., laboratory, medical history) without rewriting the core.

5.Technical and organizational constraints

Category	Requirement
Budget	≤ USD 5,000 for software/tools (only open-source or free educational licenses are allowed).
Infrastructure	There is no possibility of using cloud services (due to institutional policy); all components must run on-premises.
Integration	Must interoperate with the university's identity system (LDAP) and the existing patient database (MySQL 5.7).
Personnel	Only one system administrator is available, with a shared workload (20 hours/week for maintenance).

Expected Deliverable (for evaluation)
Students must submit a document of 800–1,200 words that includes:

1. **Case analysis:** Identification of stakeholders, critical processes, and constraints.
2. **Graphical modeling (described textually):**

- **Use case diagram (UML):** Including actors and main/alternative flows.
- **BPMN 2.0 model** of the process “Medical Appointment Management,” including start/end events, decisions, and exception handling (no-shows, cancellations, emergencies).
- **Class diagram:** With attributes, data types, relationships, and multiplicity (1..1, 0..*, etc.).

3. **Technical proposal:**
   - Feasibility estimation (development time, required resources). o Proposed technologies (frontend, backend, database).
   - Plan to mitigate constraints (e.g., replication of high availability with limited hardware).