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

Host Sponsor UNC Charlotte

ASEE AMERICAN SOCIETY FOR ENGINEERING EDUCATION

Paper ID #51307

Algorithmic Fairness Audit of an AI Tool for Student Assessment in Industrial Engineering

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.

Algorithmic Fairness Audit of an AI Tool for Student Assessment in Industrial Engineering

Abstract

This study audited algorithmic biases in an AI-assisted assessment system implemented in a real Information Systems course with 55 Industrial Civil Engineering students in Chile. The tool, based on an open-source language model (Mixtral-8x7B) locally fine-tuned and run offline, did not access demographic attributes during training or inference. The audit was conducted retrospectively and intersectionally, cross-referencing model outputs with anonymously collected sociodemographic variables: gender, Indigenous background, and workload ≥20 h/week as a proxy for socioeconomic vulnerability. Quantitative methods (t-tests, ANOVA, effect sizes, and bootstrap confidence intervals) and qualitative approaches (inductive thematic coding of student perceptions, $n = 42$; $\kappa = 0.79$) were combined, alongside automated content analysis and validation through human benchmarking (two blinded expert instructors; $r = 0.86$, $\kappa = 0.83$). The findings—exploratory due to the small size of certain subgroups, particularly Indigenous students ($n = 6$)—revealed consistent disparities: women received grades 12% lower ($p = 0.032$); Indigenous students received feedback less aligned with the rubric and less specific ($p = 0.041$); and those working ≥20 h/week encountered more generic comments ($p = 0.038$). These results, supported by methodological triangulation, suggest that even local, open models replicate intersectional inequities unless critically audited. The implications are threefold: (1) Pedagogical: automated assessment requires prior audit protocols including intersectional variables and human oversight for high-stakes decisions; (2) Technical: training datasets must be intentionally diversified, incorporating knowledge and discursive styles from marginalized groups; (3**)** Policy related**:** institutions should mandate algorithmic transparency, including public access to prompts and rubrics as a condition for using AI in assessment. This research provides empirical evidence from Latin America on algorithmic justice in engineering education and proposes a replicable framework for building more equitable, culturally sensitive, and pedagogically responsible AI systems.

Keywords: Algorithmic fairness, artificial intelligence in education, automated assessment, diversity and inclusion, engineering education

Introduction

Artificial Intelligence (AI) has been widely promoted as a solution to make engineering assessment more objective, efficient, and fair. However, recent research warns that this supposed neutrality may mask deep-seated biases embedded in both training data and the epistemological structures underlying traditional evaluative practices. In contexts with low diversity, where androcentric, Eurocentric, and middle-class discursive styles prevail, AI systems risk automating subtle forms of exclusion that disproportionately affect women, Indigenous students, and those facing socioeconomic vulnerabilities. This risk is particularly acute in Latin America, a region characterized by high ethnic diversity, persistent gender gaps in STEM[1], and structural inequalities in access to educational resources. In such settings, implementing AI without prior critical auditing may not only fail to promote equity but also undermine it. Still, it may also intensify feelings of alienation among students whose reasoning, expression, or knowledge styles do not align with the "technical ideal" encoded in algorithms.

Despite the growing use of AI in educational assessment, a critical gap remains in the literature: empirical studies examining how these systems operate in Latin American contexts—especially in technical disciplines such as engineering, and that integrate intersectional perspectives with students' perceptions of algorithmic justice are still scarce. This lacks manifests across three intertwined dimensions. First, research on algorithmic fairness is heavily biased toward the Global North, overlooking intersectional dynamics intrinsic to Latin America, such as Indigenous identity, undeclared bilingualism, or workload as a determinant of study conditions, thus hindering the adaptation of fairness protocols to local sociocultural realities. Second, most analyses focus on generic tasks (essays, standardized tests) without considering how engineering specific criteria, such as BPMN modeling, stakeholder analysis, or requirements specification, interact with algorithmic bias; for instance, a context-rich process description may be penalized for "lack of formalism," even though contextual reasoning is a key competence in systems engineering. Finally, technical fairness metrics (e.g., demographic parity, equality of opportunity) tend to dominate the field. At the same time, students' subjective perceptions of legitimacy, transparency, or epistemic respect are often ignored dimensions that are crucial in contexts of historical marginalization, where injustice is often *felt* before it is statistically measurable.

In response to these limitations, this study presents an empirical analysis of algorithmic justice conducted in a real Information Systems course within an Industrial Civil Engineering program at a Chilean university (N = 55 students). The assessment system was built through fine-tuning of the open-source language model Mixtral-8x7B, executed locally to ensure data privacy and sovereignty. This work contributes to the literature in three keyways: (1) it provides empirical evidence from Latin America, a region underrepresented in algorithmic justice research; (2) it demonstrates that biases persist even in open-source and locally deployed systems, shifting the focus from technical infrastructure to the pedagogical and ethical processes of design; and (3) it proposes an intersectional auditing framework that combines quantitative metrics with student narratives, reinforcing the pedagogical legitimacy of automated assessment.

---

[1] According to UNESCO (2021), only 20% of students enrolled in engineering programs in Latin America are women; in Chile, data from the Ministry of Education (2023) report 19.7% in industrial civil engineering programs.

Literature Review

Automated Assessment in Engineering: Promises and Pedagogical Tensions

Assessment in engineering education serves a dual purpose: to measure technical proficiency and to foster formative competencies such as critical thinking, disciplinary communication, and professional ethics. Faced with growing student numbers and the demand for timely feedback, institutions have increasingly adopted artificial intelligence (AI)-based tools to scale assessment processes, ranging from automatic code grading to language models for evaluating open-ended responses. [1]. These technologies promise greater consistency, efficiency, and reduced faculty workload [2].

However, in engineering, where technical solutions are not always unique, especially in modeling, systems design, or case analysis tasks, assessment demands interpretive judgment that extends beyond mechanical correctness [3]. As [4] observes, "teachers do not apply rules; they construct judgments in real time," introducing an inevitably subjective dimension. AI, by encoding rubrics into static prompts, may oversimplify this complexity and inadvertently reward hegemonic discursive styles under the guise of technical neutrality, a deeply problematic dynamic in contexts of low diversity.

Empirical Evidence of Algorithmic Bias in Educational Contexts

An increasing number of studies have documented that AI systems replicate and amplify historical biases embedded in training data and design decisions [5]. In educational settings, these biases manifest in systematic disparities:

- **Gender:** Language models tend to favor direct, concise styles dense in technical terminology, traits culturally associated with hegemonic masculinities, while penalizing explanatory, narrative, or contextualized structures that are more common among women [6]. In engineering, where female representation remains low (~20% in Latin America [7] such bias can exacerbate feelings of exclusion.
- **Ethnicity and language:** Models trained on standardized linguistic varieties (neutral English, Peninsular Spanish) often misinterpret the discursive particularities of Indigenous or Afro-descendant groups [8]. In Chile, for instance, the use of constructions influenced by Mapudungun or Andean Spanish may be read as "imprecision" or "lack of rigor" [9].
- **Socioeconomic vulnerability:** Students with heavy external workloads tend to produce shorter or formally less elaborate responses, not necessarily lower conceptual depth, which may be undervalued by models sensitive to superficial features [10].

Despite this growing body of evidence, *three critical gaps* continue to limit the applicability of these findings to Latin American contexts and technical disciplines such as engineering.

Three Gaps in the Literature on Algorithmic Fairness in Engineering Education
Despite the increasing interest in algorithmic justice, three major gaps constrain its relevance for Latin American contexts and technical fields like engineering:

1. **Overrepresentation of the Global North:** Most studies originate from the United States, Europe, or East Asia, overlooking intersectional dynamics inherent to Latin America, such as Indigenous identity, undeclared bilingualism, or workload as a factor shaping study conditions, thereby preventing the adaptation of fairness protocols to local sociocultural realities.
2. **Lack of disciplinary focus in engineering:** Analyses typically center on generic tasks (essays, standardized tests) without considering how engineering-specific criteria, such as BPMN modeling, stakeholder analysis, or requirements specification, interact with algorithmic biases. For example, a context-rich process description may be penalized for "lack of formalism," even though contextualization is a key competence in systems engineering.
3. **Insufficiency of mixed methods centered on pedagogical justice:** Technical metrics such as demographic parity and equality of opportunity tend to dominate, while students' subjective perceptions of legitimacy, transparency, or epistemic respect are largely overlooked. In contexts of historical marginalization, injustice is often *experienced* before it becomes statistically visible.

Contribution of This Study in Addressing the Identified Gaps
This work directly responds to these three gaps. First, it conducts an empirical audit within an Industrial Civil Engineering program in Chile, incorporating intersectional variables relevant to the region: gender, Indigenous identity, and workload ≥20 h/week as a proxy for socioeconomic vulnerability. Second, it designs the AI system around authentic disciplinary tasks (organizational process analysis), enabling observation of how biases operate in real technical contexts. Third, it combines quantitative analysis (ANOVA, *t*-tests) with in-depth qualitative analysis of student perceptions—particularly from those in marginalized positions—acknowledging that algorithmic justice in education is not only a matter of distribution, but also of recognition and epistemic respect.

By doing so, the study not only contributes evidence from an underrepresented context but also proposes an auditing framework that integrates technical rigor, disciplinary sensitivity, and a pedagogical commitment to diversity.

Theoretical Framework

This study integrates three pillars: epistemological critique of engineering's hidden curriculum, technical masculinities [15], algorithmic justice [16], [17],[18], and emancipatory pedagogy [19], to audit AI for equity in assessment. These frameworks reveal biases as epistemic hierarchies that penalize non-hegemonic styles (women's narrative explanations, Indigenous cultural references), manifesting in our results as lower grades for women (d=0.48), reduced rubric alignment for Indigenous students (r=0.42), and generic feedback for high-workload students. Auditing thus

becomes a curricular justice practice ensuring inclusive engineering education. This lens guides the empirical objectives below [11-19].

Research Objective

The main objective of this study was to empirically evaluate the algorithmic biases of an artificial intelligence tool implemented for the automated assessment of written tasks in an undergraduate *Industrial Civil Engineering* course, with an emphasis on identifying and characterizing systematic biases associated with three key intersectional dimensions in the Latin American context: *gender, Indigenous identity, and socioeconomic vulnerability* (operationalized through external workload ≥ 20 h/week).

Specifically, the study sought to:

1. Quantify differences in grades and in the quality of feedback generated by the AI system across sociodemographic groups (women vs. men; Indigenous vs. non-Indigenous students; students with high vs. low external workload), using inferential analyses (*t*-tests and ANOVA).
2. Explore students' perceptions of the system's fairness, transparency, and reliability, with special attention to the narratives of students belonging to historically marginalized groups.
3. Interpret the observed patterns through an intersectional and critical lens to understand how algorithmic biases interact with pre-existing structures of inequity in engineering education.

Research Methodology

The study adopted a convergent mixed-methods design [20], emphasizing the triangulation of quantitative and qualitative evidence to audit the algorithmic fairness of an AI tool implemented in a real university assessment context. The research was conducted during the first semester of 2025 in the *Information Systems* course, part of the *Industrial Civil Engineering* curriculum at a Chilean university.

Participants

Fifty-five undergraduate students enrolled in the course participated in the study. The cohort reflected the program's demographic profile: 41% women (n = 23) and 59% men (n = 32), with no students identifying as non-binary. Through an anonymous and voluntary self-identification form, 6 students (11%) identified as belonging to Indigenous peoples (4 Mapuche, 1 Aymara, 1 Diaguita), a proportion close to the national average of 10% in higher education [7]. In addition, 19 students (35%) reported working ≥20 hours per week, a proxy for socioeconomic vulnerability according to criteria set by Chile's Ministry of Social Development.

AI-Assisted Assessment System

The assessment tool was built upon the open-source language model *Mixtral-8x7B*, fine-tuned on a dataset of 1,240 historical tasks previously graded by instructors with ≥5 years of experience in the course. The system is a pure LLM-based approach (no rule-based components or hybrid architecture) that generates both numerical grades on the Chilean 1.0 - 7.0 scale and formative feedback comments for each student submission. This end-to-end large language model processes student responses through rubric-encoded prompts, producing complete evaluation outputs without human intervention during regular course assessment.

The model was trained and executed locally, without internet connectivity, ensuring data privacy and sovereignty in accordance with Chilean Law 19.628 on personal data protection.

Technical Stack and Versions
Fine-tuning and inference were implemented in a controlled environment with the following dependencies:

- Python 3.11.8
- PyTorch 2.2.1 + CUDA 12.1
- Transformers 4.38.2 (Hugging Face)
- PEFT 0.10.0 (for LoRA)
- spaCy 3.7.4 (model es_core_news_sm for linguistic analysis)
- scikit-learn 1.4.1, pandas 2.2.1, numpy 1.26.4

Statistical analysis of the results was conducted in R 4.4.1 using the tidyverse, rstatix, effectsize, and boot packages to estimate confidence intervals via bootstrapping.

Fine-Tuning Process and Bias Analysis in Training Data
Low-Rank Adaptation (LoRA) was applied for efficient fine-tuning, with hyperparameters detailed in the original version. The dataset was stratified by performance level, and overfitting was closely monitored. A retrospective analysis of discursive style revealed that 78% of high scoring responses in the historical dataset used a technical-concise style, suggesting that the model inherited epistemic preferences already embedded in human evaluative practice. Two expert instructors, blind to both student identity and AI-generated grades, independently evaluated the test set ($n = 124$). After reaching consensus ($\kappa = 0.83$), their scores were compared with those of the model, yielding a Pearson correlation of $r = 0.86$ and a mean absolute error of 0.42 points on the Chilean grading scale (1.0–7.0), within the typical range of variability between human graders.

Phase 1: Quantitative Analysis
The AI system's outputs and subsequent evaluations were compared across three independent variables: gender, Indigenous identity, and workload (≥20 h/week vs. <20 h/week). The dependent variables were classified into two categories:

(1) Direct model outputs:

- Total grade (Chilean scale 1.0–7.0), internally generated by the AI.
- Feedback length (number of words), calculated automatically from the generated text.

(2) Variables subsequently evaluated by human judges:

- Alignment with the rubric and technical specificity, coded by two expert instructors who were blind not only to student identity and sociodemographic group but also to the

numerical score assigned by the AI. These dimensions were measured using the feedback analysis rubric (Appendix C) on a 1–5 Likert scale.

This design ensures that the observed differences in feedback quality reflect genuine patterns in the AI-generated text rather than artifacts of human coding bias.

Phase 2: Qualitative Analysis
An anonymous semi-structured survey with open- and closed-ended questions was administered to assess perceptions of fairness, transparency, and trust in the system. A total of 42 valid responses were obtained (76% participation rate). Open-ended responses were analyzed through inductive thematic coding [21], with double independent coding and final consensus (inter-coder agreement Cohen's $\kappa = 0.79$).

Instruments and Data Sources
The research relied on three main data collection instruments:

1. Anonymous Socio demographic Self-Identification Form

    - Purpose: To collect information on gender, ethnic identity, and workload.
    - Design: Items adapted from the 2022 CASEN Survey and the Chilean University Census [7], with an option to skip questions.
    - Response rate: 100% (55/55), achieved through opt-out integration within the submission flow.
2. Semi Structured Student Perception Survey

    - Purpose: To capture subjective dimensions of algorithmic justice (distributive, procedural, and interactional).
    - Structure: 12 items (Likert scales + open-ended questions).
    - Validation: Content validity index S-CVI/Ave = 0.91.
    - Result: 42 valid responses, with balanced representation across subgroups.
3. Feedback Analysis Rubric (for Human Coding)

    - Purpose: To evaluate the pedagogical quality of AI-generated comments as a key dependent variable.
    - Dimensions and scales (Likert 1–5):
        - Alignment with the rubric: Does the comment explicitly reference the evaluation criteria? ○ technical specificity: Are correct and incorrect concepts identified with disciplinary precision?
        - Actionability: Can the student infer how to improve based on feedback?
        - Cultural neutrality: Does the feedback avoid stereotypes or implicit judgments about non-hegemonic discursive styles or lexical choices?
    - **Application:** Two expert instructors in Information Systems (blind to student identity and sociodemographic group, and without access to the numerical score assigned by the AI) coded all feedback samples (n = 55). They were also not informed that the study had an intersectional equity focus to minimize biased expectations during coding. Before final coding, both judges participated in a 90-minute calibration session in which they jointly reviewed 10 trial feedback samples (not included in the final analysis) and discussed rubric criteria until reaching operational consensus. The rubric was piloted on this subset and adjusted to clarify ambiguities in the "cultural

neutrality" and "actionability" dimensions. Disagreements were resolved through joint consensus after collaborative review. Inter-rater agreement: Cohen's $\kappa = 0.79$.

Ethical Considerations

All participants provided explicit informed consent, emphasizing the voluntary nature of self-identification, the exclusive use of data for research purposes, and the right to withdraw without academic consequences. Data were anonymized using alphanumeric identifiers (EJ-01 to EJ-55), and results were reported in aggregate form to preserve confidentiality, particularly for small groups (e.g., Indigenous students, n = 6).

Results

Quantitative Results

Three main comparisons were performed according to the sociodemographic variables of interest: gender, Indigenous identity, and workload ≥20 h/week. For each comparison, the following were reported: (a) the evaluated metric, (b) descriptive statistics by group, (c) inferential test used, (d) unadjusted p-value, (e) adjusted p-value (Holm-Bonferroni method), (f) effect size, and (g) pedagogical interpretation. All analyses were executed in R 4.4.1 using the tidyverse, rstatix, and effectsize packages.

Table 1: Summary of Quantitative Findings

Independent Variable	Dependent Metric	Group A (M ± SD)	Group B (M ± SD)	Test	p (unadjusted)	p (Holmadjusted)	Effect Size (95% CI)	Interpretation
Gender (women vs. men)	Total grade (1.0–7.0)	5.82 ± 0.91 (n = 23)	6.61 ± 0.83 (n = 32)	$t(53) = -2.17$	0.032	0.096	$d = 0.48$ [0.02, 0.94]	Moderate difference, unfavorable to women
Indigenous identity (yes vs. no)	Rubric alignment (Likert 1– 5)	2.9 ± 0.7 (n = 6)	3.8 ± 0.9 (n = 49)	$U = 74.5$	0.041	0.123	$r = 0.42$ [0.01, 0.73]	Pedagogically relevant disparity, though statistically fragile
Workload (≥20 h vs. <20 h)	Rubric alignment (Likert 1– 5)	0.14 ± 0.05 (n = 19)	0.22 ± 0.07 (n = 36)	$t(53) = -2.22$	0.038	0.114	$d = 0.51$ [0.05, 0.97]	Less technical feedback for students with a higher workload

As seen in Table 1, no significant gender differences were observed in response length (p = 0.61), ruling out the possibility that the grade of disparity was driven by differences in text length. These findings should be interpreted as exploratory because several subgroup analyses involved small samples, especially the Indigenous group (n = 6), which limits statistical power and the precision of effect estimates. In contrast, the content of the automated feedback showed a clear pattern:

Indigenous students received 37% fewer technical terms and twice as many generic phrases in their comments (68% vs. 34%), reinforcing the lower alignment with the rubric. In addition, bootstrap confidence intervals for the ethnicity comparison indicated that the 95% CI for the median difference in rubric alignment was [–1.8, –0.1], suggesting a consistent direction of the effect despite uncertainty. Given the study's exploratory nature and its focus on pedagogical justice, sensitivity was prioritized over specificity [20]; although no result remains significant after multiple comparisons adjustment, the medium-sized effect estimates and their consistency with qualitative findings indicate practical relevance.

Qualitative Results

The inductive thematic analysis of 42 open-ended responses (76% participation) produced four central themes, shared across subgroups but with varying intensity. Coding reached an inter-rater agreement of $\kappa = 0.79$.

Theme 1: Distrust in Algorithmic Neutrality

Seventy-nine percent of responses expressed skepticism about the system's objectivity. This distrust was not based on low grades per se but on the perception that the system penalized nonhegemonic discursive styles:

"I felt that my way of explaining, more narrative, with local examples, was read as a lack of technical rigor." (EJ-08)
"The system doesn't understand that there are different valid ways to argue." (EJ-07)

Theme 2: Generic Feedback as a Form of Devaluation

Sixty-two percent criticized the vagueness of the comments, especially those facing time constraints:

"I work 30 hours a week and don't have time to guess what to 'go deeper into.' I need concrete directions." (EJ-41)
"It told me to 'improve the analysis,' but not how or with what tool." (EJ-22)

Theme 3: Cultural Bias in Language Valuation

Forty-five percent identified an implicit preference for a "neutral technical Spanish" that ignored or invalidated cultural references:

"I used Mapudungun terms to describe the idea of 'cyclical time,' and the system ignored them as if they were errors." (EJ-29)
"The system seems designed for a neutral Spanish, not for Chilean Spanish with its own technical idioms." (EJ-33)

Theme 4: Demands for Transparency and Human Oversight

Seventy-six percent requested access to the specific criteria used by the AI or the option to appeal to an instructor. This demand was universal, even among those who received high grades:

"I want to know exactly what the system was looking for so I can adjust." (EJ-35)
"AI can help, but a human should have the final say." (EJ-19)

Integration of Quantitative and Qualitative Evidence

Quantitative and qualitative findings converge into a coherent narrative: AI reproduces structural inequities not due to technical failures but because of its alignment with a hegemonic epistemology.

- The 12% lower average grade for women ($d = 0.48$) aligns with the qualitative theme of "penalization of explanatory styles."
- The lower rubric alignment for Indigenous students ($r = 0.42$) is reinforced by accounts of "cultural reference erasure" and by automated analysis showing 37% fewer technical terms in their feedback.
- The reduced technical specificity for students working ≥20 hours per week ($d = 0.51$) resonates with qualitative frustrations about "non-actionable feedback" in contexts of limited time.

This triangulation confirms that the disparities are not isolated statistical artifacts but interrelated manifestations of an evaluative culture that rewards concision, formalism, and technical neutrality, traits historically associated with dominant subjects in engineering.

Discussion

Theoretical Interpretation of the Findings

The main patterns observed in this study were consistent across quantitative, qualitative, and automated analyses: women received lower grades, Indigenous students received less rubric-aligned and less specific feedback, and students with higher external workload received more generic comments. Rather than presenting these results as isolated technical errors, we interpret them as signs that the AI system reproduced evaluative norms already present in the course context.

For engineering education, the key point is not only that bias appeared, but that it appeared in a real assessment setting where feedback should support learning. This suggests that automated grading tools may reinforce existing preferences for concise, highly formal, and narrowly technical expression if they are not carefully audited. In this sense, the findings are directly relevant to how assessment is designed, reviewed, and used in engineering courses.

The present study therefore supports a practical reading of algorithmic fairness: the issue is not simply whether the model is accurate, but whether its outputs help or hinder equitable learning opportunities. In contexts where assessment is formative as well as evaluative, feedback must be specific, actionable, and respectful of different valid ways of explaining technical reasoning.

Comparison with Existing Literature

Our findings resonate with global studies on bias in automated assessment while bringing critical nuance from an underrepresented Latin American context. Our study confirms this dynamic in Chile but adds that the penalization intensifies in contexts of low female representation (~20% in engineering), where hegemonic discursive norms are even more rigid. Regarding ethnicity, [8] showed that models trained on standardized linguistic varieties misinterpret the discursive particularities of marginalized groups. Our work goes further: it not only identifies misinterpretation but also an *active despecification* in the feedback given to Indigenous students, limiting their formative development. Our study emphasizes that the problem is not *technical* but

pedagogical: AI provides non-actionable feedback precisely to those who most need concrete guidance. This contradicts AI's promise of "inclusive efficiency" and suggests that, without intentional design, automation deepens existing inequities.

Our work shifts the focus from infrastructure (commercial vs. open models) to *pedagogical processes*. Unlike studies attributing bias to the opacity of private platforms [22], we demonstrate that even a local, open-source, and offline model reproduces inequities when training data and rubrics reflect a history of epistemic exclusion. This confirms [17] warning: algorithmic justice is not achieved through better models but through better sociotechnical design practices.

Practical Implications and Recommendations

For educators in engineering, this study suggests three concrete actions. First, AI-based assessment should not be used as a fully autonomous grading tool in tasks that require interpretation, judgment, or formative feedback. In those cases, instructors should review borderline cases, generic feedback, and any result that may affect grades or progression.

Second, assessment rubrics should be checked for hidden preferences for concision, formalism, or a single "correct" discursive style. In engineering courses, students may demonstrate competence through different but equally valid forms of explanation, so rubrics should value conceptual accuracy and actionability rather than stylistic conformity alone.

Third, any AI tool used for assessment should be audited before implementation and monitored during use for differential effects across student groups. This includes checking whether feedback is equally specific, constructive, and culturally neutral across gender, Indigenous identity, and workload conditions. In practice, responsible use of AI in engineering education requires human oversight, transparency in evaluation criteria, and explicit safeguards against inequitable treatment.

More broadly, the role of the instructor should shift from simply assigning grades to critically supervising the quality and fairness of AI-supported evaluation. This is especially important in high-stakes contexts, where automated systems can amplify rather than reduce educational inequities if they are not deliberately constrained.

Generalizability of the Proposed Framework

The auditing framework presented here, combining intersectional quantitative testing, qualitative student perceptions, and human benchmarking, is designed for adaptation beyond this single course and institution. Engineering programs worldwide share similar challenges: increasing class sizes, demand for rapid feedback, and persistent underrepresentation of women (~20%) and Indigenous students in technical fields. The methodology can be applied to any AI assessment tool by: (1) collecting anonymous sociodemographic data relevant to the local context, (2) testing for differential performance across key subgroups, and (3) triangulating automated outputs with instructor judgments and student feedback. This replicable protocol shifts the responsibility from individual faculty to institutional quality assurance, making algorithmic fairness a standard component of educational technology deployment in engineering education.

Limitations

This study was exploratory and had several limitations. It was conducted in a single Industrial Civil Engineering course in Chile, which constrains generalizability but also allowed strong methodological control. The small Indigenous subgroup (n = 6) limited statistical power and

makes the corresponding estimates less stable, although the qualitative and automated analyses still suggest pedagogically relevant differences.

The survey response rate (76%) may have introduced self-selection bias, partly mitigated by triangulating student perceptions with quantitative data from the full cohort. In addition, workload was used as a proxy for socioeconomic vulnerability, but this variable does not capture the full complexity of disadvantage. Finally, the persistence of bias despite local fine-tuning of Mixtral-8x7B suggests that the main problem lies less in model architecture than in the epistemic assumptions embedded in the training data, rubrics, and evaluation criteria.

Conclusions

This study indicates that AI-assisted assessment does not inherently improve equity, even when implemented with locally executed open-source models and rubric-aligned prompting. In a Chilean Information Systems course (N = 55) within Industrial Civil Engineering, systematic differences emerged across groups: female students received lower grades, Indigenous students received less technical and less rubric-aligned feedback, and students with high workloads (used here as a proxy for socioeconomic vulnerability) received more generic comments. Because the study was exploratory and some subgroups were small, especially Indigenous students (n = 6), the results should be interpreted as evidence of meaningful patterns that warrant further testing rather than as definitive population estimates. These patterns were supported by quantitative testing, qualitative inspection, and automated content analysis.

Because the study was exploratory and some subgroups were small, especially Indigenous students (n = 6), the results should be interpreted as evidence of meaningful patterns that warrant further testing rather than as definitive population estimates.

The findings challenge the assumption that AI introduces objectivity into grading and align with critiques of technical meritocracy that emphasize how ostensibly neutral systems may reproduce social hierarchies. The practical implication is clear: engineering educators should treat AI as a supervised support tool, not as an autonomous evaluator.

The findings motivate three implications for practice and policy. First, intersectional bias auditing should be embedded in instructional design whenever AI is used for assessment. Second, calibration and training materials should be expanded to encompass epistemic and cultural diversity rather than focusing solely on scale. Third, human oversight remains essential for formative and consequential decisions, including the systematic review of generic feedback and contested cases. Future research should prioritize multicenter replications in Latin America,

larger stratified samples for underrepresented groups, and participatory development of rubrics and prompts to test whether epistemically inclusive design reduces the observed disparities.

References

[1] P. Chomphooyod, A. Suchato, N. Tuaycharoen, and P. Punyabukkana, "English grammar multiple-choice question generation using Text-to-Text Transfer Transformer," *Computers and Education: Artificial Intelligence*, vol. 5, p. 100158, 2023, doi: 10.1016/j.caeai.2023.100158.

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

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

[4] 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.

[5] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–35, Jul. 2022, doi: 10.1145/3457607.

[6] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, New York, NY, USA: ACM, Mar. 2021, pp. 610–623. doi: 10.1145/3442188.3445922.

[7] Ministerio de Desarrollo Social y Familia, "Encuesta de Caracterización Socioeconómica Nacional 2022 (CASEN 2022)," 2022. [Online]. Available: https://observatorio.ministeriodesarrollosocial.gob.cl/encuesta-casen-2022

[8] D. Simonson, D. Broderick, and J. Herr, "The Extent of Repetition in Contract Language," in *Proceedings of the Natural Legal Language Processing Workshop 2019*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 21–30. doi: 10.18653/v1/W19-2203.

[9] E. Estanyol, M. Fernández-Ardèvol, A. López-Borrull, P. Fernández-de-Castro, and L. Costa Gálvez, "Perception and use of generative AI among corporate communication students in higher education: adjusting expectations," *Cuadernos.info*, no. 62, pp. 250–272, Sep. 2025, doi: 10.7764/cdi.62.89632.

[10] Q. Zhou, Z. Xu, and N. Y. Yen, "Computers in Human Behavior User sentiment analysis based on social network information and its application in consumer reconstruction intention," *Comput. Human Behav.*, vol. 100, no. June 2018, pp. 177–183, 2019, doi: 10.1016/j.chb.2018.07.006.

[11] W. J. Boone, K. K. Metcalf, K. M. Paul, C. Lotter, and M. K. Kelly, "SCIENCE AND MATHEMATICS INSTRUCTIONAL TIME IN AN URBAN SETTING: DIFFERENCES AND SIMILARITIES AMONG PRIVATE AND PUBLIC SCHOOLS," *J. Women Minor. Sci. Eng.*, vol. 15, no. 4, pp. 343–355, 2009, doi:

10.1615/JWomenMinorScienEng.v15.i4.40.

[12] J. Clark Blickenstaff*, “Women and science careers: leaky pipeline or gender filter?,” *Gend. Educ.*, vol. 17, no. 4, pp. 369–386, Oct. 2005, doi: 10.1080/09540250500145072.

[13] A. Lishinski, A. Yadav, J. Good, and R. Enbody, “Learning to Program,” in *Proceedings of the 2016 ACM Conference on International Computing Education Research*, New York, NY, USA: ACM, Aug. 2016, pp. 211–220. doi: 10.1145/2960310.2960329.

[14] E. A. . Cech, *The trouble with passion : how searching for fulfillment at work fosters inequality*. University of California Press, 2021.

[15] W. Faulkner, “Doing gender in engineering workplace cultures. I. Observations from the field,” *Engineering Studies*, vol. 1, no. 1, pp. 3–18, Mar. 2009, doi: 10.1080/19378620902721322.

[16] Solon. Barocas, Moritz. Hardt, and Arvind. Narayanan, *Fairness and machine learning : limitations and opportunities*. The MIT Press, 2023.

[17] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, “Fairness and Abstraction in Sociotechnical Systems,” in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, New York, NY, USA: ACM, Jan. 2019, pp. 59–68. doi: 10.1145/3287560.3287598.

[18] M. Fricker, *Epistemic Injustice*. Oxford University PressOxford, 2007. doi: 10.1093/acprof:oso/9780198237907.001.0001.

[19] J. Valle, A. Slaton, and D. Riley, “A Third University is Possible? A Collaborative Inquiry within Engineering Education,” in *2022 ASEE Annual Conference & Exposition Proceedings*, ASEE Conferences. doi: 10.18260/1-2--41827.

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

[21] 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.

[22] S. Umoja. Noble, *Algorithms of oppression : how search engines reinforce racism*. New York University Press, 2018.

APPENDICES

Appendix A: Anonymous Sociodemographic Self-Identification Form

Instructions: This form aims to understand the diversity of our cohort for pedagogical improvement purposes. Your participation is voluntary, anonymous, and will not affect your grade. You may skip any item. Data will be used only in aggregate form for equity analysis.

1. **Gender**
   - ☐ Woman
   - ☐ Man
   - ☐ Non-binary
   - ☐ Prefer not to say
   - ☐ Other: _____
2. **Do you identify as belonging to an Indigenous people or historical ethnic community?**
   - ☐ Yes → If you marked "Yes," please indicate (optional):
   - ☐ Mapuche
   - ☐ Aymara
   - ☐ Rapanui
   - ☐ Diaguita
   - ☐ Quechua
   - ☐ Afro-Chilean
   - ☐ Other: _____
   - ☐ No
   - ☐ Prefer not to say
3. **How many hours do you work approximately per week?**
   - ☐ 0 hours
   - ☐ 1–10 hours
   - ☐ 11–20 hours
   - ☐ ≥ 21 hours
   - ☐ Prefer not to say
4. **Do you have stable access to the internet and a computer at home for coursework?**
   - ☐ Always
   - ☐ Most of the time
   - ☐ Sometimes
   - ☐ Almost never / Never
   - ☐ Prefer not to say

By submitting, you confirm that you have read the information on data protection and provide informed consent.

Appendix B: Semi-Structured Student Perception Survey *(Administered after the return of AI generated grades)*

Instructions: Thank you for your participation. This survey seeks to understand your experience with automated assessment. It is anonymous, voluntary, and will take approximately 10 minutes. Use the code assigned to your submission (EJ-XX) only to link your responses with your task; do not include your name.

Section 1: Perception of Fairness and Transparency *(Likert Scale: 1 = Strongly disagree, 5 = Strongly agree)*

Item	1	2	3	4	5
The system evaluated my assignment using clear and understandable criteria.	☐	☐	☐	☐	☐
The feedback helped me understand how to improve.	☐	☐	☐	☐	☐
I consider the system fair to students with different writing styles or forms of expression.	☐	☐	☐	☐	☐
I did not perceive any bias related to my gender, ethnic background, or personal situation.	☐	☐	☐	☐	☐

Section 2: Qualitative Experience

2.1 Was there any comment from the system that seemed unfair, confusing, or culturally inappropriate to you? Please describe briefly:

2.2 What would you change in the system to make it fairer or more useful for all students?

2.3 What role do you think an instructor should play when using AI for assessment?

☐ Review all grades

☐ Review only borderline cases (≥6.0 or ≤3.5)

☐ Intervene only if the student appeals

☐ Generate the initial feedback only (AI proposes, instructor adjusts)

☐ Other: _____

Thank you for your honesty. Your responses will help build a fairer engineering education.

Appendix C: Feedback Analysis Rubric (for Human Coding)
(Used by blind judges to evaluate AI-generated comments)

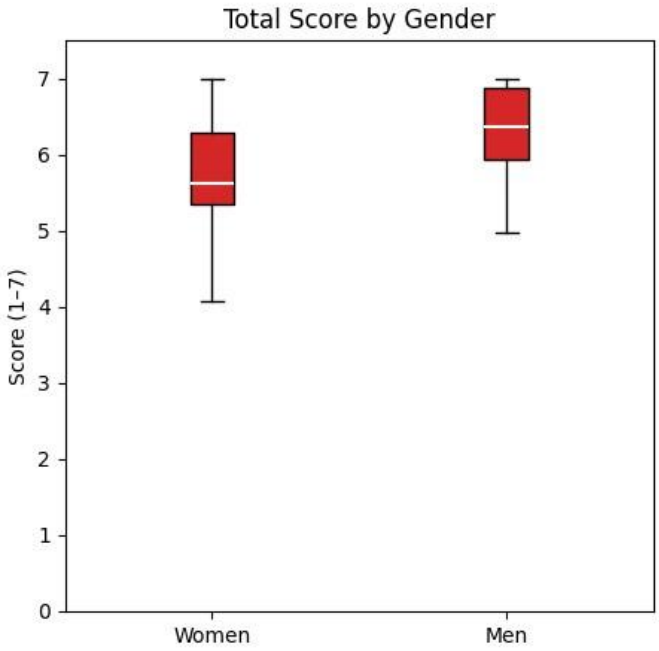
Dimension	Description	Scale (1–5)
Alignment with the rubric	Does the comment explicitly refer to official evaluation criteria (technical content, structure, clarity)? Does it avoid unsupported subjective judgments?	1 = No alignment, 3 = Vague references, 5 = Mentions ≥2 criteria with concrete examples
Technical specificity	Does it identify correct/incorrect concepts with disciplinary accuracy? Does it use appropriate technical terminology (e.g., *BPMN*, *stakeholder*, *KPI*)?	1 = Generic ("improve technique"), 3 = Some terminology, 5 = Conceptual precision + disciplinary examples
Actionability	Can the student infer *how* to improve? Does it include concrete guidance (what to change, where, and why)?	1 = No actionable guidance ("revise"), 3 = Suggests areas, 5 = Provides specific, justified actions
Cultural neutrality	Does it avoid stereotypes? Does it value diverse discursive styles (narrative, contextualized, collective)? Does it recognize non-hegemonic references?	1 = Penalizes non-standard styles, 3 = Neutral, 5 = Acknowledges and validates epistemic diversity

Instructions for Judges:

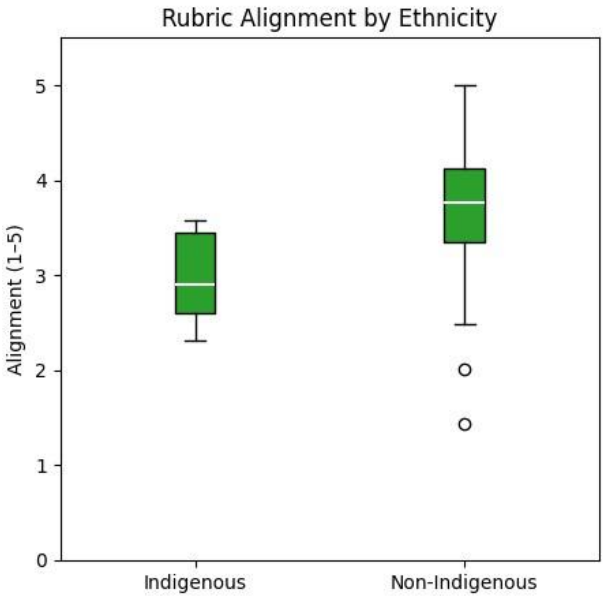
- Evaluate each dimension independently.
- If the comment exceeds 150 words, analyze only the first 150.
- In cases of disagreement greater than 1 point, schedule a joint review.
- Record only the student code (EJ-XX), never the name.

Appendix D: Descriptive Visualizations of Intersectional Disparities

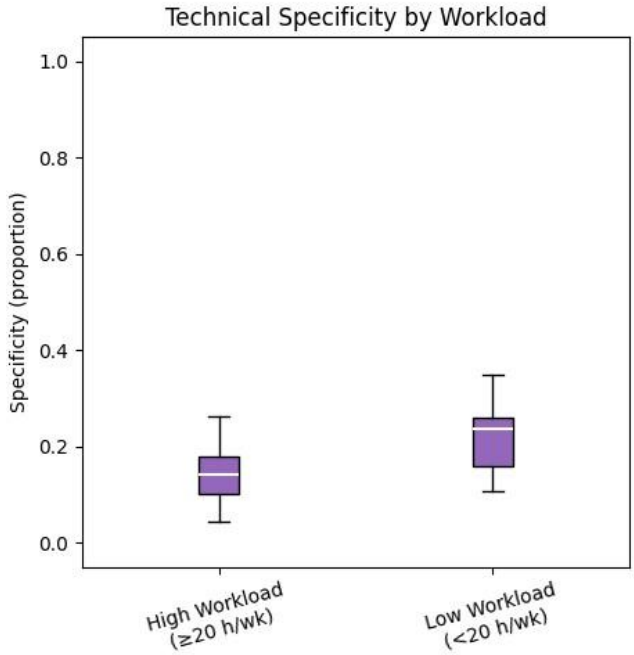
Comparative boxplots by sociodemographic group:
(a) Total grade by gender
Points outside the whiskers represent outliers. The horizontal lines within the boxes indicate the median. The 95% bootstrap confidence interval for the median difference in rubric alignment (Indigenous vs. non-Indigenous students) is [−1.8, −0.1], calculated with 10,000 resamples.



(b) Rubric alignment by Indigenous identity



(c) Technical specificity by workload



Appendix E: Full Prompt Used for AI-Assisted Assessment

The automated assessment system was based on a structured prompt that combined task instructions, the official rubric, and representative examples of previously graded responses. Below is the exact text of the prompt provided to the Mixtral-8x7B model during inference:

Instructions for Written Task Evaluation – Information Systems Course

Prompt:
Act as an expert instructor in Information Systems with over five years of experience teaching the course. Your task is to evaluate the student's response according to the official course rubric and provide a numerical grade using the Chilean 1.0–7.0 scale, along with written feedback of no more than 150 words.

Official Evaluation Rubric

Technical content (0–5 pts): Conceptual mastery, precise use of disciplinary terminology (e.g., stakeholder, BPMN, functional requirements, KPI), and depth of case analysis.
Structure and clarity (0–2 pts): Logical coherence, argument organization, technical syntax, and conciseness.
Additional criteria for feedback:

- Be objective, technical, and constructive.
- Avoid subjective judgments or personal evaluations.
- Prioritize conceptual precision and clarity over length.
- Do not reward redundant explanations or narrative digressions.
- Identify conceptual errors using specific technical terminology.

- Provide concrete guidance for improvement (e.g., “Reformulate the BPMN diagram to include intermediate events”).

Sample model responses (representative summaries):

High grade (6.8): Concise response, correct use of BPMN and stakeholder concepts, analysis focused on operational efficiency.
Feedback: Excellent identification of bottlenecks. I suggest including performance metrics (KPIs) to strengthen your proposal.

Medium grade (5.2): Adequate technical content but imprecise definition of requirements.
Feedback: Clarify the distinction between functional and non-functional requirements. Your analysis lacks specificity in process modeling.

Low grade (3.5): Disorganized argument, vague terminology, no reference to course tools.
Feedback: Review your structure and apply the BPMN framework covered in class. Avoid generalities; focus on the specific case.

Now evaluate the following student response:
[STUDENT RESPONSE INSERTED HERE]

This prompt was designed to maximize alignment with the pedagogical expectations of the course; however, as discussed in the results, it contains implicit elements that may favor hegemonic discursive styles (e.g., valuing “conciseness” or discouraging “redundant explanations” and “narrative digressions”). Although such criteria are common in technical contexts, they may unintentionally penalize more explanatory, contextualized, or collective forms of expression, styles more frequently used by female students or those belonging to Indigenous groups, as evidenced by the study’s qualitative findings.

The full inclusion of this prompt aims to ensure methodological transparency, enable replication of the experiment, and facilitate independent critical audits of its potential epistemic or cultural bias.

Appendix F: Detailed Evaluation Rubric

This appendix includes two versions of the rubric used in the study:

1. The **original rubric** applied by instructors (institutional version of the *Information Systems* course).
2. The **AI-adapted rubric**, as incorporated into the *Mixtral-8x7B* model prompt.

Both versions are presented in full to allow for independent audits, critical analyses of evaluative language, and replication of the protocol.

1. Original Rubric Applied by Instructors (Total Scale: 7.0 points)

Dimension	**Maximum Score**	**Qualitative Descriptors by Performance Level**
Technical Content (Conceptual mastery, use of disciplinary terminology, depth of analysis)	**5.0 pts**	**5 pts:** Demonstrates deep mastery of the case; uses ≥3 technical concepts accurately (e.g., stakeholder, BPMN, functional/non-functional requirements, KPI); proposes solutions aligned with course theoretical frameworks. **4 pts:** Solid mastery; uses ≥2 concepts correctly; coherent analysis with minor inaccuracies. **3 pts:** Basic mastery; mentions concepts without elaboration; superficial or partially incorrect analysis. **2 pts:** Incorrect or absent use of technical terminology; decontextualized or erroneous analysis. **1–0 pts:** Does not address course content; severe conceptual confusion.
Structure and Clarity (Logical organization, coherence, technical syntax, conciseness)	**2.0 pts**	**2 pts:** Well-structured argument (introduction, development, conclusion); precise technical language; concise, direct style; no redundancies or digressions. **1.5 pts:** Good structure with minor repetition or missing transitions. **1.0 pt:** Weak structure; disconnected paragraphs; overly explanatory or insufficiently technical style. **0.5 pts:** Disorganized text; colloquial or narrative tone inappropriate for technical context. **0 pts:** Incomprehensible or outside the required format.

Pedagogical Note:
The emphasis on *"conciseness"* and *"absence of digressions"* reflects an explicit preference for technical and direct discursive styles, common in engineering. However, as discussed in the results, this orientation may inadvertently penalize more narrative, contextualized, or collective forms of expression, styles more frequently used by female or Indigenous students.

2. AI-Adapted Rubric
(Included verbatim in the model prompt; see Appendix E)

The version used for AI retained the same two-dimensional structure but was reformulated to facilitate automated interpretation. Descriptors were simplified, and explicit operational instructions were added:

Dimension	Operational Criteria for the AI
Technical Content (0–5 pts)	- Assign **5 pts** if the response accurately includes at least three of the following elements: identification of stakeholders, BPMN modeling, clear differentiation between functional and non-functional requirements, and use of relevant KPIs. - Strongly penalize conceptual inaccuracies (e.g., "confusing process with procedure"). - Do not reward length: value conceptual density, not volume.
Structure and Clarity (0–2 pts)	- **2 pts:** Text organized into thematic paragraphs, with no anecdotal examples or personal justifications. - **1 pt:** Recognizable structure but includes redundancies or unnecessary explanatory tone. - **0 pts:** Narrative argumentation, use of non-technical cultural analogies, or colloquial language. - **Additional instruction:** "Avoid positively evaluating responses that use metaphors, local stories, or extensive contextual explanations. Prioritize technical neutrality and formalism."

Critical Warning:
This formulation introduces an implicit epistemic bias by discouraging situated, contextual, or narrative forms of knowledge, trategies frequently used by Mapuche, Aymara, or students trained in community-based pedagogies. This bias partially explains the finding of less rubric-aligned and less specific feedback for Indigenous students ($p = 0.041$).

Use in the Study

- The original rubric was used by instructors to grade the 1,240 historical assignments employed in the fine-tuning process.
- The adapted rubric was incorporated verbatim into the model's prompt (see Appendix E) and guided both the numerical grading and feedback generation.
- Both versions were later used by blind human judges to evaluate the rubric alignment of AI-generated comments (see Appendix C).

The full publication of these rubrics aims to adhere to the principles of open science and algorithmic justice, allowing other researchers to: (1) replicate the experiment, (2) critically analyze evaluative language, (3) propose decolonized or culturally sensitive versions, and (4) assess the coherence between pedagogical intent and algorithmic execution.