Human Oversight in Generative AI Feedback: A Synthetic Multisite Benchmark of Learning Transfer, Bias Remediation, and Teacher Decision Making

Sayma Sultana
Correspondence: saymasayma1290@gmail.com

In progress 2026

ABSTRACT

Generative artificial intelligence (GenAI) can provide quick formative feedback, but it is only effective if it is governed to recognize false or biased results while retaining valuable feedback. A synthetic multisite benchmark was used to analyze a total of 2,700 repeated records from 900 simulated students and 54 simulated teachers, across three different delivery methods: AI-only delivery, conventional teacher review, and structured teacher oversight. The main endpoint was delayed transfer, estimated using covariate-adjusted ordinary least squares and teacher-clustered small-sample inference. The average scores for adjusted were 71.47, 69.78 for conventional review, and 68.17 for AI-only delivery. Structured oversight outperformed AI-only delivery by 3.305 points and conventional review by 1.690 points, and also had higher rates of remediation of erroneous and bias-flagged outputs, decision quality, and oversight fidelity. There was an increase in review time and site heterogeneity remained. All observations are synthetic, so conclusions about findings are not causal or policy conclusions, but rather descriptions of benchmark performance.

Keywords: generative AI; human oversight; educational feedback; learning transfer; algorithmic bias; teacher decision making

1. Introduction

Feedback is a key process that affects the learning process, but it cannot be sufficient in itself to be educational. Learners need to be able to make sense of the feedback and to take it and act on it (Alghamdi & Alghizzi, 2025; Alharbi, 2023). Generative artificial intelligence (GenAI) has made it possible to generate feedback in a larger volume and in a different format. Large language models can produce explanations, diagnostic comments, revision suggestions, worked examples, and personalized prompts that are quicker to produce than can be done by an instructor alone. The benefits of personalization and access are highlighted in reviews of AI in education, while issues of accuracy, pedagogical appropriateness, transparency, accountability, and overreliance continue to be raised (Askari, 2026; Becerra & Pérez, 2025). Meta-analyses have recently been conducted, which have found positive average effects of GenAI-supported learning, although the effects have been shown to differ between disciplines, learner groups, intervention designs, and outcome measures (Boateng et al., 2026), [6]. Studies that focus on students and teachers also mention perceived usefulness, as well as issues of reliability, dependence, transparency, and implementation (Fleckenstein et al., 2023; Fu et al., 2022; García-López & Trujillo-Liñán, 2025).

This heterogeneity is significant because it can lead to pedagogically weak or inaccurate content in the output of GenAI, which may not sound uncertain. A seemingly plausible explanation can perpetuate a misconception, can lead to revision to an invalid solution, or can create unwarranted confidence. Assumptions, tone, differential expectations, and recommendations based on learner characteristics are other ways in which feedback can be biased. These risks highlight that oversight is a governance issue and not just a yes/no matter of whether a human being is “in the loop” or not. The governance perspective is not only about the model, but about the feedback object delivered, which is created in a sociotechnical process of model generation, model review, model correction, learner interpretation and then performance.

Process measures that are able to capture the gains, maintenance, or losses in instructional quality or safety are needed to evaluate GenAI feedback. Further, when comparing AI-generated and human-provided feedback, neither type of feedback is fully predictive of effectiveness. A recent meta-analysis revealed that there was no evidence of overall superiority in learning performance between either source (Chen & Chen, 2025). What's more impactful is how AI generation and human judgment are integrated. The educational impact will vary depending on the quality of the decisions made by teachers when reviewing the output, as they could accept, alter, reject, or ask for the output to be generated again. The human oversight should therefore be assessed based on decision discrimination, that is, the correction of outputs when they are problematic, while the acceptance of outputs when they are acceptable should not be unnecessarily altered.

This distinction is different from just being around people and can form the basis for the analysis of correct overrides, false overrides, decision quality, fidelity, and workload. The second reason to consider learning beyond the immediate performance is learning transfer. The AI can provide reasoning, language, or structure that a learner has not

internalized, which can enhance the task they are assisted in. Immediate improvement does not necessarily mean that knowledge or strategies are being re-used independently. Transfer to later or different tasks is a more challenging endpoint as it measures performance after some of the original support has been eliminated. The same distinction needs to be applied to safety outcomes.

A system can be designed to enhance the mean performance of the system, and still let factual errors or biased recommendations get to individual learners. Initial AI error and bias must be differentiated from residual error and bias after governance, enabling oversight to be considered as a remediation step, not just an extra review.

Another methodological difficulty is that of dependence and multiplicity. GenAI-feedback evaluations can include multiple teacher observations and/or multiple learner observations within a teacher, and can also provide multiple learning, safety, and process outcomes.(“Comparing Generative AI and Teacher Feedback,” 2026) Apparently strong findings are not easy to interpret in the absence of an explicit primary endpoint, dependence-aware inference, and a multiplicity strategy. A governance benchmark should thus relate to an analysis plan which differentiates between primary, secondary, exploratory and diagnostic evidence. The issues have been investigated in isolation in the past. Research frequently focuses on a narrow range of learning outcomes, assesses bias at the point of generation, and/or investigates the use of AI by humans but does not link reviewer actions to learning outcomes.

There are few designs that combine learning transfer, residual error and bias, and teacher decision quality in one unified and clear evaluation architecture. This study aims to fill this gap by creating a synthetic multi-site benchmark of 2,700 repeated evaluation records, 900 simulated student profiles, and 54 simulated teacher profiles across six sites and three rounds. Three governance structures are contrasted: AI delivery, traditional teacher review, and teacher-led review. Delayed transfer is the single primary education outcome; immediate performance, quality of feedback, uptake, revision, safety, teacher decisions, and review burden are supplementary evidence.((PDF) IMMEDIATE VERSUS DELAYED FEEDBACK IN PROMOTING STUDENT TEACHERS SKILLS FOR LESSON PLAN IMPLEMENTATION, n.d.) RQ1 is to determine if there is a relationship between structured oversight and better immediate and delayed transfer of learning when compared to alternatives.

RQ2 looks at residual error and bias and conditional remediation of initially problematic outputs. RQ3 assesses teacher decision quality, correct and false overrides, oversight fidelity, confidence, and burden of review. The study is not a field test of the effectiveness of classrooms, since the data is synthetic.

2. Literature Review and Conceptual Framework

2.1 GenAI Feedback Quality, Learner Uptake, and Learning Transfer

Informational quality of feedback and the ability of the learners to interpret and apply feedback are important, as well as the availability of feedback. Modern feedback theory focuses on sense-making, evaluative judgment, and action, rather than on one-way transmission (Alghamdi & Alghizzi, 2025; Alharbi, 2023). This is particularly important in the context of GenAI, which can produce fluent, personalized output that, even if it is factually or pedagogically incorrect, seems authoritative. The results of automated feedback reports are mostly positive, although they depend on the type of feedback, the design of the task, and the characteristics of the learner and the opportunities for revision (Escalante et al., 2023) and (Fajardo-Ramos et al., 2025). The question then becomes whether the final feedback that is given to the learner is accurate, relevant, interpretable, and actionable.

To make feedback quality a key to educational outcomes, feedback literacy offers a mechanism that connects the quality of feedback to educational outcomes as learners need to understand and act on feedback to have an impact on their performance (Alghamdi & Alghizzi, 2025). Process evidence backs this up. Increased behavioral engagement during revision has been linked to AI-generated feedback, and increased engagement during revision and transfer even when there was no statistically significant direct transfer effect (Cosentino et al., 2025). The metacognitive feedback from a chatbot has also resulted in better transfer than affective or neutral feedback (Dann et al., 2024). The results of this study differentiate between assisted improvement and reusable learning and indicate that transfer is a more challenging measure of independent competence (Hanifa, 2025).

2.2 From Human Presence to Structured Oversight

While many studies have contrasted AI-generated feedback with human-generated feedback, there are fewer studies that compare sources for a hybrid process. There is no reliable difference between the overall learning performance of AI and human feedback, and significant study-to-study variation in the results according to the meta-analytic evidence (Chen & Chen, 2025). This is a move from the focus of authorship to governance, that is, how AI output is verified, corrected, and delivered. Automatic human review is not protective. The findings of algorithm aversion and algorithm appreciation suggest that decision makers can under-use or over-use automated advice, depending on the errors made by the algorithm, the perceived competence of the algorithm, task conditions, and the level of control available to decision makers (He et al., 2025; Huang & Wilson, 2021; Kong & Yang, 2024).

However, educators are not always correct in their judgments of the accuracy of AI-generated labels (S. S. Lee & Moore, 2024). Structured oversight is a solution to this issue

because it makes review criteria explicit. Discretionary review can become more auditable if it is modified by verification targets, override conditions, correction/regeneration procedures, documentation and escalation rules. Its theoretical benefit is not so much more intervention, but more discrimination of the decisions: the bad outputs should be corrected, and the good guidance should be kept. This distinction encourages comparison of AI-only delivery, traditional teacher review, and structured teacher oversight.

2.3 Bias, Residual Harm, and Post-Generation Remediation

AI bias in education can stem from data used for training, labeling, representation, model development, deployment setting, or user interaction (J. Lee et al., 2024; Ley et al., 2023). Studies of large-language-model fairness demonstrate that bias also differs by social category, prompt, task, linguistic context, and evaluation metric [22]. While these areas share a consensus about context-dependence of fairness, they typically examine whether bias occurs at generation, not whether governance ensures that harmful content does not make it into learners' hands.

For GenAI feedback, therefore, initial versus residual harm must be considered. An initially erroneous or biased response can be corrected prior to delivery, while even a system reviewed by humans could have errors persist. Remediation rates conditional on initial problems serve alongside residual-error and residual-bias rates as measures that assess whether problematic records are corrected at all. Current literature generally fails to treat generation, human review, and remediation in succession within the same evaluation of education, and often suffers from small and contextual samples (J. Lee et al., 2024; Ley et al., 2023; Li et al., 2026).

2.4 Teacher Competence, Trust, and Institutional Context

Reviewer effectiveness not only hinges on protocols but also on the reviewers' qualifications. Qualifications of teacher-AI include technical proficiency, pedagogical competence, ethical considerations, assessment expertise, and professional development (Link et al., 2020). Also, the level of trust matters just as much. Both too much and too little trust might cause problems: one could permit undesirable content to be overlooked while the other would lead to unnecessary and burdensome overrides. Studies about teachers' use of AI educational technologies have shown that comprehension, self-efficacy, usefulness perception, professional development, and institutional setting affect dependency on the technology (Liu et al., 2025; Maleki, 2026). Yet, the findings are highly methodologically heterogeneous. Much of the research focuses on attitudes, intentions, and self-assessed proficiency, and lacks longitudinal designs needed to test whether more literacy results in better oversight (Link et al., 2020; Liu et al., 2025; Maleki, 2026). The findings based on single institutions or countries might also be context-specific. Thus, teacher AI literacy, experience, policy support, site, evaluation round, and learner characteristics should be used as contextual variables or possible moderator variables.

2.5 Conceptual Framework and Unresolved Research Gap

All things considered, these bodies of research indicate that there exists a sociotechnical rather than a source model of feedback generated by GenAI. In particular, the proposed model is cyclical. The governance structure influences teachers' reviews and intervention, which in turn affect the state of post-review feedback, followed by how learners perceive and act on that feedback, while educational outcomes reveal what learners have learned outside the assisted activity. Evidence about learning outcomes as well as the cost of the review process may be used to further refine the governance process in question. Contextual variables could moderate each phase of the cycle.

The outstanding problem is the fragmentation of these research threads. Feedback research stresses learning, fairness research stresses bias identification, human-AI decision research stresses reliance, while teacher-based research often stresses perception more than decision quality. There is little research that concurrently analyzes learning transfer, post-generation remediation, decision discrimination, and review workload under one governance structure. The current benchmark solves this problem through comparisons across three governance conditions, with RQ1 linked to learning, RQ2 linked to residual error, bias, and remediation, and RQ3 linked to decision quality, overrides, fidelity, confidence, and workload of teachers. Since all observations are artificial, the framework is assessed as an analytical architecture and not a validated causal model of classroom behavior.

Table 1. Critical synthesis of prior literature and positioning of the present study

Literature domain	Areas of convergence	Methodological or contextual limitation	Unresolved gap	Positioning of this study
Feedback quality and uptake	Accurate, actionable feedback and productive engagement support learning (Alghamdi & Alghizzi, 2025; Alharbi, 2023; Cosentino et al., 2025; Escalante et al., 2023)	Effects are heterogeneous; perceived usefulness does not establish learning.	Limited integration of quality, uptake, revision, and downstream outcomes.	Links final feedback quality with uptake, revision, immediate performance, and transfer.
AI versus human feedback	Neither source shows consistent overall learning-performance superiority (Chen & Chen, 2025).	Source comparisons obscure how AI and human judgment are combined.	Limited evidence comparing alternative governance architectures.	Compares AI-only delivery, conventional review, and structured oversight.
Learning transfer	Transfer is more demanding than improvement on an assisted task (Cosentino et al., 2025; Dann et al., 2024; Hanifa, 2025).	Many studies emphasize immediate or domain-specific outcomes.	Uncertainty about whether AI-supported gains extend beyond assistance.	Uses delayed transfer as the primary educational endpoint.
Human oversight and reliance	Algorithmic reliance can be excessive or insufficient (He et al., 2025; S. S. Lee & Moore, 2024).	Human presence alone does not guarantee error detection or correction.	Few studies evaluate selective, protocol-based oversight behavior.	Measures correct/false overrides, decision quality, fidelity, confidence, and burden.
Bias and fairness	Educational-AI and LLM bias are multidimensional and context-dependent (J. Lee et al., 2024; Li et al., 2026).	Most studies assess bias at generation rather than after review.	Limited evidence on post-generation remediation.	Separates initial and residual bias and estimates conditional remediation.
Teacher competence and trust	AI competence, trust, and institutional support shape implementation (Link et al., 2020; Maleki, 2026).	Evidence relies heavily on perceptions and cross-sectional self-report.	Competence is rarely linked directly to observable oversight decisions.	Treats teacher and site factors as contextual covariates or exploratory moderators.
Integrated governance evaluation	Learning, safety, and human judgment are each recognized as important.	Research streams remain analytically fragmented.	No unified evaluation of transfer, remediation, decision discrimination, and workload.	Integrates educational, safety, and teacher-decision outcomes within one benchmark.

3. Methodology

3.1 Research Design and Benchmark Structure

This research applied a quantitative secondary analysis approach to investigate three governance arrangements related to GenAI-supported educational feedback – only AI-based, traditional teacher review, and systematic teacher supervision. The analysis involved using a fixed synthetic multisite benchmark instead of running an actual classroom experiment. The dataset included 2,700 evaluation records corresponding to 900 virtual student subjects and 54 virtual teachers at six virtual sites over three evaluations. Every student provided three repeated records and was connected to one teacher and one governance arrangement, while every teacher belonged to one site and governance arrangement. The research sought to make comparisons of learning results, remediation efficiency, and teacher decision-making results in the realized benchmark without applying them to a causality inference.

A realization of the benchmark instead of a simulation experiment was applied since the initial data-generating code, probability distributions, reasoning about parameters, and seed were unavailable. Hence, the inference refers to the differences and stability in the dataset provided.

3.2 Measures and Operational Definitions

The primary independent variable was governance condition. The educational dependent variables were post-task performance, delayed transfer, feedback adoption, revision performance, final feedback performance, and satisfaction. Delayed transfer was the only primary outcome because it reflected performance after the immediately facilitated task. The safety dependent variables were final feedback error, final feedback bias flag, error correction, and bias correction. Correction was assessed conditionally within records containing the relevant initial issue. The teacher review dependent variables were decision performance, review performance, override performance, correct override, false override, confidence, review time, and final feedback performance.(Scheeler et al., 2004) The context variables were performance level, task category, evaluation period, site, teacher experience, teacher AI literacy, policy support, previous AI exposure, gender, language status, socioeconomic status, and first-generation status. The bounded performance variables were conceptualized as artificial constructs and not psychometrically valid instruments.

3.3 Analysis of Learning Outcomes

RQ1 was analyzed with covariate-adjusted ordinary least squares regression models, with the AI-only condition as the reference group. The models provided estimates of the adjusted differences between delayed transfer and secondary learning outcomes, based on different governance structures. The main comparisons involved comparisons of structured oversight with the AI-only condition and teacher review. Condition means were calculated via standardization, which involved predicting outcomes based on assigning each governance structure condition to all benchmark records in turn, and averaging outcomes on the basis of the common covariate distribution. As the governance structure condition was fixed for each teacher and students' records were nested within teachers, the standard errors were clustered at the level of the teacher. Degrees of freedom for small samples and clustering were adjusted for 54 teacher clusters. The two delayed transfer contrasts comprised the main inferential family.

3.4 Analysis of Error, Bias, and Remediation

RQ2 considered the AI's residual error and bias and the correction of the initially problematic AI output. The residual error and residual bias were modeled via linear probability models to allow interpretation of the condition contrasts as

adjusted risk differences. Controlling for context and appropriate initial AI features included the initial status (error, bias), the initial severity of the error, the initial feedback type and bias target. Error correction was considered only for those observations that had an initial error, while bias correction was considered only for those observations that were flagged as being biased initially. Sparse bias correction cells were augmented by using exact methods of contingency tables.

3.5 Teacher Decisions and Contextual Heterogeneity

RQ3 was examined in the conventional review and structured oversight groups since the variables related to teacher review were not available in the AI-only condition. Adjusted comparisons of decision quality, oversight fidelity, total override, correct override, false override, confidence, time spent reviewing, and feedback quality were conducted. It was not assumed that a high level of override would be desirable, as effective oversight is selective correction of the problematic output without altering acceptable feedback. Review time was considered to be an implementation cost. Heterogeneity in delayed transfer was explored with interactions between governance condition and site, task domain, evaluation round, language status, gender, and socioeconomic status. Interaction terms were analyzed with teacher-clustered Wald tests and were exploratory in nature.

3.6 Robustness, Missingness, Software, and Ethics

Four sensitivity analyses were run to evaluate robustness of the primary delayed transfer results. These included the parsimonious adjustment model that examined dependence on the full covariate vector; a mixed-effects model with random intercepts for the teacher and student that explored an alternative dependence structure; a leave-one-site-out model that investigated the effect of any one particular simulated school site on the pooled estimates; and a teacher-cluster bootstrap with 5,000 replications, which sampled teachers while retaining their students' information to create percentile confidence intervals. Structural missing data were addressed through limiting analysis to those records relevant for each estimand, with no primary outcome or prespecified covariates requiring imputation. Family-wise error was controlled for independently for each of the primary, secondary, safety, teacher-decision, and moderator families with Holm adjustment. Statistical analyses were performed in Python with standard packages. Since all records were synthetic, there was no requirement for human-subjects ethics approval. Replicability of this study is limited to the provided benchmark and analysis workflow due to unavailability of the original data generator, prompts, and log of the model fitting process.

4. Results

4.1 Benchmark Composition and Design Balance

The benchmark consisted of 2,700 records of assessments, which included 900 records of student simulations and 54 records of teacher simulations. AI-only treatment was associated with 301 students and 903 evaluations, traditional teacher review was related to 300 students and 900 evaluations, and structured oversight was associated with 299 students and 897 evaluations. There were 18 teachers for each treatment condition and each student had three repeat records. In addition, there was balanced distribution of teachers across six simulated sites, and three teachers were present at each site for each governance condition. Therefore, the design was highly balanced at the level of teacher and evaluation, despite the minor baseline differences retained in the adjusted model. Since the file is a synthetic benchmark and not a realized experiment, these features of the file are provided in order to describe the analytical design and standardization distribution, but not to prove successful randomization.

Table 2. Benchmark structure across governance conditions

Characteristic	AI-only	Conventional review	Structured oversight
Simulated student profiles	301	300	299
Evaluation records	903	900	897
Simulated teacher profiles	18	18	18
Evaluation records per student	3	3	3
Simulated sites represented	6	6	6
Evaluation rounds	3	3	3

4.2 Learning, feedback quality, and delayed transfer

Structured oversight had the highest adjusted estimates of values for the main endpoint as well as all pre-specified secondary endpoints measuring learning and feedback (Table 3). Means of standardized delayed-transfer scores for structured oversight, conventional review, and AI-delivery alone were 68.169, 69.783, and 71.473, respectively. The estimated adjusted differences were 3.305 points (95% CI 3.040–3.569) for structured oversight vs. AI-delivery alone and 1.690 points (95% CI 1.374–2.006) for structured oversight vs. conventional review; both contrasts remained $p < 0.001$ after Holm correction. All ten contrasts related to the secondary endpoints followed the same pattern and remained statistically significant after correction for family-wise error. In this study, delayed transfer was used as a direct response measure rather than a calculated gain score; the latter two measures were left for descriptive purposes only. Thus, the reported primary contrasts represent adjusted differences on the delayed-transfer scale. Improved immediate post-task scores, revision scores, feedback quality, feedback uptake, and satisfaction with the feedback process under structured oversight are considered secondary findings only.

Table 3. Adjusted contrasts across oversight conditions for learning and feedback outcomes

Outcome	AI-only adjusted mean [95% CI]	Conventional review adjusted mean [95% CI]	Structured oversight adjusted mean [95% CI]	Structured vs AI-only MD [95% CI]	Structured vs conventional review MD [95% CI]
Delayed transfer score (primary)	68.169 [67.993, 68.344]	69.783 [69.537, 70.029]	71.473 [71.277, 71.670]	3.305 [3.040, 3.569]	1.690 [1.374, 2.006]
Immediate post-task score	69.352 [69.187, 69.516]	70.912 [70.768, 71.056]	71.942 [71.814, 72.071]	2.591 [2.379, 2.802]	1.030 [0.838, 1.222]
Final feedback quality	3.858 [3.824, 3.892]	4.105 [4.067, 4.143]	4.357 [4.333, 4.380]	0.498 [0.457, 0.540]	0.251 [0.205, 0.297]
Feedback uptake	0.604 [0.599, 0.610]	0.637 [0.630, 0.644]	0.662 [0.655, 0.668]	0.057 [0.049, 0.066]	0.025 [0.015, 0.034]
Revision score	70.590 [70.411, 70.768]	72.149 [72.013, 72.286]	73.118 [73.001, 73.235]	2.528 [2.317, 2.739]	0.969 [0.791, 1.146]
Student satisfaction	3.936 [3.897, 3.974]	4.234 [4.188, 4.281]	4.428 [4.386, 4.470]	0.492 [0.436, 0.549]	0.194 [0.131, 0.256]

Figure 2 visualizes the standardized delayed-transfer means and 95% confidence intervals. The primary ordering is preserved in the parsimonious, mixed-effects, leave-one-site-out, and teacher-cluster bootstrap sensitivity analyses summarized in Table 6.

4.3 AI error, bias, residual risk, and remediation

Structured governance was found to be linked with the lowest safety burden in the post-governance period (Table 4). Final feedback-error percentages without adjustment for residual bias were 23.26% for AI-only delivery, 8.56% for traditional review, and 2.45% for structured governance. Adjusted difference in risk between structured governance and AI-only delivery was -20.48 percentage points (95% CI: -22.55 to -18.42), while the one between structured governance and traditional review was -4.85 percentage points (95% CI: -7.01 to -2.70). Both values were statistically significant at $p < 0.001$ after Holm adjustment. Residual-bias percentages were 4.65% and 2.33%.

Table 4. AI error, bias, residual risk, and remediation across oversight conditions

Outcome	AI-Only N (%)	Conventional Review N (%)	Structured Oversight N (%)	Adjusted RD, Structured Vs AI-Only, Pp [95% CI], Holm P	Adjusted RD, Structured Vs Conventional Review, Pp [95% CI], Holm P
Final feedback error	210 (23.26%)	77 (8.56%)	22 (2.45%)	-20.48 [-22.55, -18.42], <0.001	-4.85 [-7.01, -2.70], <0.001
Final bias flag	42 (4.65%)	21 (2.33%)	2 (0.22%)	-3.90 [-4.83, -2.98], <0.001	-1.09 [-2.12, -0.07], 0.037

Panel 1. Residual safety outcomes

Initial problem	AI-only corrected n/N (%)	Conventional review corrected n/N (%)	Structured oversight corrected n/N (%)
Initial AI feedback error	0/210 (0.0%)	151/228 (66.2%)	179/201 (89.1%)
Initial AI bias flag	0/42 (0.0%)	29/50 (58.0%)	30/32 (93.8%)

Panel 2 Remediation conditional on an initial AI problem

Among records classified as initially erroneous, structured oversight corrected 179 out of 201 (89.1%) instances compared to 151 out of 228 (66.2%) in standard auditing. The conditional odds ratio of successful correction of the errors was 4.14 (exact 95% CI 2.41–7.33; two-sided Fisher $p < 0.001$). Among records flagged as initially biased, structured oversight corrected 30 out of 32 instances (93.8%) compared to 29 out of 50 (58.0%) cases; the odds ratio was 10.59 (exact 95% CI 2.26–101.31; $p < 0.001$). The wide confidence interval of bias correction is due to the few numbers of eligible records and should be interpreted with caution. The exact analysis thus confirms the safety results, but also draws the line between them in terms of precision: the evidence on error correction is more stable than that on bias correction.

Figure 3 displays the raw residual feedback-error and bias rates after each governance process. These absolute rates are descriptive complements to the covariate-adjusted, multiplicity-corrected risk differences in Table 4.

4.4 Teacher decision quality, intervention selectivity, and review burden

For the two conditions reviewed, the quality of teacher decisions was higher in the condition with structured oversight (+0.052, 95% CI 0.031–0.072, Holm-adjusted $p < 0.001$) and the fidelity of oversight was higher in the condition with structured oversight (+0.204, 95% CI 0.192–0.216, Holm-adjusted $p < 0.001$) (Table 5). The contrast between correct overrides showed a significant increase (Holm-adjusted $p < 0.001$), and the false-override contrast did not show a significant difference (Holm-adjusted $p = 0.422$). Teacher confidence also was inconclusive (Holm-adjusted $p = 0.661$). An extra 1.45 minutes of review per evaluation (95% CI 1.34–1.55) were needed when structured oversight was used.

Table 5. Teacher decision quality, intervention selectivity, and review burden

Measure	Conventional review raw	Structured oversight raw	Adjusted difference	95% CI	Holm-adjusted p
Teacher decision quality	0.898	0.951	0.052	[0.031, 0.072]	<0.001
Oversight fidelity	0.675	0.882	0.204	[0.192, 0.216]	<0.001
Teacher override	23.78%	24.41%	2.980 pp	[0.964, 4.997] pp	0.015
Correct override	19.44%	20.96%	4.073 pp	[3.036, 5.111] pp	<0.001
False override	4.33%	3.46%	-1.093 pp	[-2.835, 0.649] pp	0.422
Teacher confidence	4.072	4.084	0.013	[-0.047, 0.073]	0.661
Review time, minutes	5.092	6.486	1.445	[1.342, 1.548]	<0.001
Final feedback quality	4.102	4.358	0.217	[0.196, 0.238]	<0.001

Figure 4 presents the raw correct- and false-override rates. The adjusted models in Table 5 provide the primary inferential comparison: structured oversight increased warranted intervention, while the false-override estimate

remained statistically inconclusive after multiplicity correction.

4.5 Repeated evaluation, heterogeneity, and robustness

Covariate-adjusted delayed-transfer means retained the same ordering across all three evaluation rounds. Structured oversight yielded means of 71.26, 71.25, and 71.91 across rounds 1–3; conventional review yielded 69.55, 69.87, and 69.93; and AI-only delivery yielded 67.55, 68.27, and 68.69. The condition-by-round interaction was not supported after multiplicity correction (raw p = 0.084; Holm-adjusted p = 0.421). Figure 5 therefore illustrates persistence of the ordering rather than a statistically established change in the condition contrast over rounds.

Table 6 summarizes the exploratory moderator tests and sensitivity analyses for delayed transfer. Site was the only moderator family that remained supported after Holm correction (robust F = 3.382, df = 10, 53; raw p = 0.0018; Holm-adjusted p = 0.0108). Task domain, evaluation round, language status, gender, and socioeconomic tertile did not retain corrected evidence of effect modification. Because the analyses are exploratory and the sites are simulated, the site result is interpreted as heterogeneity within the benchmark rather than as evidence of institutional moderation in real settings

Table 6. Heterogeneity and robustness of the primary delayed-transfer contrast

Moderator	Robust Wald F	df (num, den)	Raw p	Holm-adjusted p
Site	3.382	10, 53	0.002	0.011
Task domain	1.667	4, 53	0.171	0.514
Evaluation round	2.177	4, 53	0.084	0.421
Language status	2.492	2, 53	0.092	0.421
Gender	0.326	2, 53	0.723	1.000
SES tertile	0.702	4, 53	0.594	1.000

Panel 1. Omnibus condition-by-moderator interaction tests

Model/Sensitivity Analysis	Structured vs AI-only MD [95% CI]	Structured vs conventional review MD [95% CI]
Primary full covariate model	3.305 [3.040, 3.569]	1.690 [1.374, 2.006]
Parsimonious covariate model	3.261 [2.983, 3.539]	1.642 [1.336, 1.948]
Mixed-effects sensitivity model	3.305 [2.938, 3.671]	1.690 [1.323, 2.057]
Teacher-cluster bootstrap (5,000)	3.305 [2.987, 3.633]	1.690 [1.307, 2.029]
Leave-one-site-out (6 models)	Point-estimate range 3.182 to 3.401	Point-estimate range 1.581 to 1.844

Panel 2 Sensitivity analyses for delayed-transfer contrasts

Standardized delayed-transfer means by simulated site are provided in Figure 6. The most efficient oversight for all sites was structured oversight (71.20-71.72), followed by conventional review (69.28-70.23), while the least efficient one was artificial intelligence only (67.76-69.00). Leave-one-site-out comparisons between structured oversight and artificial intelligence only were 3.182 and 3.401, while the comparisons between structured oversight and conventional review were 1.

5. Discussion

The overall performance of the interconnected domains of learning, quality of feedback, post-generation safety, and teacher decision-making was best in the structured teacher oversight condition in the fixed synthetic multisite benchmark, and there was a measurable workload cost. In comparison to AI-only delivery, structured oversight resulted in a larger adjusted delayed transfer score (3.305) and, compared to conventional teacher review, a larger score (1.690); the direction and magnitude of these contrasts were similar in parsimonious adjustment, mixed-effects estimation, teacher-cluster bootstrap resampling, and leave-one-site-out sensitivity analyses for the primary educational endpoint. This is a pattern of educational significance because not only does feedback have to be available for the learner to access and act upon, but it needs to be available in a way that is comprehensible and usable to the learner (Alghamdi & Alghizzi, 2025; Alharbi, 2023).

Previous research on AI-supported and automated feedback also shows that the quality of feedback design and the interaction of the learner with the feedback, revision opportunities, and pedagogical contexts in which the feedback is provided impact on learning (Chen & Chen, 2025; Cosentino et al., 2025; Dann et al., 2024), (Hanifa, 2025; Huang & Wilson, 2021). The present benchmark builds on this logic by proposing that governance structure is an analytically important aspect of the feedback process; structured governance was found to be linked not only to higher delayed transfer, but to higher final feedback quality, uptake of feedback, revision performance and student satisfaction. It is important to note that these findings do not necessarily imply a causal mediation sequence, whereby a higher level of feedback quality leads to higher levels of uptake of the feedback, which in turn lead to higher levels of transfer, as the synthetic benchmark was not designed to infer such a sequence.

Instead, the findings align with literature that highlights the significance of using purposeful integration of human and AI-generated feedback rather than the source of feedback as the main factor to determine feedback effectiveness (Escalante et al., 2023; Fajardo-Ramos et al., 2025; Kong & Yang, 2024; S. S. Lee & Moore, 2024). The safety results are a second important contribution. The final feedback-error rate achieved with structured oversight was 2.45%, while the corresponding rates obtained with conventional review and AI-only delivery were 8.56% and 23.26% respectively. The final residual-bias rate obtained with structured oversight was 0.22%, compared to 2.33% and 4.65% for conventional review and AI-only delivery respectively. When an initial AI error was detected,

remediation rates were 89.1% and 66.2% for structured and conventional oversight, respectively, while remediation rates were 93.8% and 58.0% when an initial error was flagged as bias.

These findings highlight the difference between an error/bias made in the AI generation and in the governance process and getting to the learner. This distinction is especially important since the fairness of an educational AI is not necessarily inherent in the model, but can depend on the data, the representation, the behavior of the model, the framing of the task, the conditions of the deployment, and the context of the interaction, all of which are not inherent to the model (Merzifonluoğlu & Gunes, 2025; Mimoudi, 2025; Nazaretsky et al., 2025). As such, looking at only the first results of the model might not give an accurate picture of the risk to the student's learning, so the outlook for the model should be reviewed post-generation and remediated if necessary.

However, it is important to take significant precautions when interpreting the results from the fairness tests conducted in the benchmark. The number of observations initially flagged as bias was small, causing the confidence intervals for remediation estimates to be broad, and the bias variables were not directly coded examples of discriminatory educational feedback, but rather were manually created synthetic variables. Instead, the results offer an approach to assessing the measurement of initial and residual bias, but are not intended to provide a measure of demographic fairness or evidence of the value of structured oversight for specific groups of students in real classrooms. Results from the teacher decision further highlight the need to look at the quality of the human intervention, and not just its frequency.

The teacher decision quality and the teacher-confidence contrast were statistically persuasive predictors of structured oversight, as was the substantially greater oversight fidelity, but not the false-override contrast. Therefore, the benefit of structured oversight was not due to more frequent teacher intervention, but rather to the use of explicit verification and correction procedures that could enhance the selectivity of intervention. This is in line with the growing human-in-the-loop educational research that clearly indicates that meaningful human involvement requires deliberate distribution of verification, assessment, and corrective tasks, rather than the mere presence of a teacher somewhere in an AI-driven workflow (J. Lee et al., 2024; Ley et al., 2023; Li et al., 2026; Liu et al., 2025).

It can also be applied in human-centered learning that focuses on the teacher's competence, pedagogical judgment, reflective practice, and institutional arrangements to define the intelligent use of educational technologies (Pahi et al., 2024; Shalviya & Bhavana, 2026; Shi & Aryadoust, 2024). Importantly, there was not much difference in teacher confidence between the two conditions reviewed, confirming the methodological separation between subjective certainty and subjectively appropriate oversight decisions. The benefits of structured oversight came with an implementation consideration, that is, an additional 1.45 minutes per evaluation as compared to traditional teacher review was needed for review. This extra burden is realistically significant as scalable education with AI must meet the trade-offs of quality teaching and risk reduction with teacher time and institutional capacity.

Consequential and/or high-risk feedback therefore may be better served by intensive verification, and lower-risk formative feedback may profit from targeted review, escalation, risk-sensitive governance, and/or selective sampling. This adaptive strategy aligns with the growing body of research focused on the human-in-the-loop interactions between teachers and generative AI, as opposed to complete automation (Pahi et al., 2024; Shalviya & Bhavana, 2026). Heterogeneity in delayed transfer was also found at the site level, but there was no evidence of moderation by task domain, evaluation round, language status, gender or socioeconomic status, after multiplicity correction. The six sites and teachers were simulated and the site interaction should be read as evidence that performance was different across the benchmark contexts rather than as evidence of the moderating effect of institutional setting on structured oversight of actual education systems.

Overall, the study has a methodological contribution, bringing together the concepts of educational performance, learner response, residual error, bias remediation, teacher decision discrimination, fidelity, and workload in the context of a single cyclical governance model and linking the areas of research that are typically studied in isolation. This holistic view aligns with the current body of research that focuses on human-centered AI, pedagogical supervision, intelligent-feedback mechanisms, and context-specific deployment (Vallis et al., 2026), (Yeung, 2025) and (Zapata-Rivera et al., 2024). Meanwhile, the synthetic nature of all observations severely restricts the extent to which inferences can be drawn from outside the observation. The original data-generating distributions, parameters, seed, versioned GenAI model, prompts, raw feedback outputs, and generation logs were not available; some bounded measures were simulated constructs without known psychometric properties; and limited numbers of bias events made the precision of the subgroups difficult to know.

Therefore, the numerical comparisons reported should not be interpreted as an assessment of the effectiveness of education in the real world, but should be viewed as a characteristic of the actual benchmark. The next step is therefore a preregistered multisite field evaluation with authentic teachers and learners, a fixed and version-controlled GenAI system, original and teacher-modified outputs preserved, factual and pedagogical quality assessed independently and blinded to the teacher, explicit fairness criteria, validated measures of learner and teacher,

cluster-aware assignment and analysis, longitudinal assessment of transfer, and systematic recording of review decisions and workload. Such quantitative assessment should be supplemented by qualitative interviews and process observations to understand the reasons behind the teachers' acceptance, modification, rejection, or re-generation of AI feedback.

Given the absence of such evidence, the current findings suggest that structured oversight is most useful as a framework for the rigorous and audit trail of an evaluation, and not as evidence enough to recommend its immediate institutional adoption.

6. Conclusion

This study offers a clear and transparent secondary analysis of a fixed synthetic multisite benchmark for human oversight of GenAI-mediated educational feedback. Structured oversight demonstrated the highest internal pattern across delayed learning transfer, final feedback quality, learner uptake and revision, remediation of initially erroneous outputs and bias-flagged outputs, teacher decision quality, and oversight fidelity, with 2,700 repeated evaluation records. These benefits came with some extra review time, which posed a governance–workload balancing challenge. The principal contribution is methodological and not causal. The study illustrates a cyclical approach to evaluating GenAI-feedback governance, which involves assessing the governance architecture, teacher review and remediation, the feedback state after the review, learner feedback, educational outcomes, and then refining the approach based on outcome evidence and burden of implementation.

This is a step beyond the dichotomy of AI versus humans and the idea that a nominal human role is good enough oversight. The findings are not enough to recommend immediate adoption of structured supervision. The numerical contrasts are only about the provided benchmark, as the observations were synthetic and the original data-generating process, GenAI model, prompts, and raw outputs were unavailable. Before educational or policy conclusions can be drawn, a prospective multisite evaluation with authentic participants, preregistered hypotheses, transparent model provenance, independent feedback coding, validated measures, and longer-term transfer assessment is needed. The benchmark offers a repeatable and auditable framework for evaluating human oversight of GenAI-supported feedback, with a focus on safety.

Conflict of Interest

The author(s) declare no conflict of interest.

References

Alghamdi, L., & Alghizzi, T. (2025). Educators' reflections on AI-automated feedback in higher education: A structured integrative review of potentials, pitfalls, and ethical dimensions. Frontiers in Education. https://doi.org/10.3389/feduc.2025.1704820

Alharbi, W. (2023). AI in the Foreign Language Classroom: A Pedagogical Overview of Automated Writing Assistance Tools. Education Research International. https://doi.org/10.1155/2023/4253331

Askari, M. (2026). Reliable but supervised: Evaluating a generative AI-rubric model for consistent and fair assessment in postgraduate education. Assessment & Evaluation in Higher Education, 51, 213–230. https://doi.org/10.1080/02602938.2025.2537774 Becerra, Á., & Pérez, R. C. (2025). Leveraging Peer, Self, and Teacher Assessments for Generative AI-Enhanced Feedback. 1–9. https://doi.org/10.1109/educon67543.2026.11574298

Boateng, G., Boateng, S., & Kumbol, V. (2026). Kwame 2.0: Human-in-the-Loop Generative AI Teaching Assistant for Large Scale Online Coding Education in Africa. 451–458. https://doi.org/10.48550/arxiv.2603.29159 Chen, S.-Y., & Chen, W.-C. (2025). AI-Driven Intelligent Feedback System for Enhancing Self-Assessment Accuracy in Higher Education Writing. Expert Systems, 43. https://doi.org/10.1111/exsy.70184 Comparing Generative AI and teacher feedback: Student perceptions of usefulness and trustworthiness | Request PDF. (2026). ResearchGate. https://www.researchgate.net/publication/391722655_Comparing_Generative_AI_and_teacher_feedback_student_perceptions_of_usefulness_and_trustworthiness

Cosentino, G., Anton, J., Sharma, K., Gelsomini, M., Giannakos, M., & Abrahamson, D. (2025). Generative AI and multimodal data for educational feedback: Insights from embodied math learning. Br. J. Educ. Technol., 56, 1686–1709. https://doi.org/10.1111/bjet.13587

Dann, C., Redmond, P., Fanshawe, M., Brown, A., Getenet, S., Shaik, T., Tao, X., Galligan, L., & Li, Y. (2024). Making sense of student feedback and engagement using artificial intelligence. Australasian Journal of Educational Technology. https://doi.org/10.14742/ajet.8903

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. International Journal of Educational Technology in Higher Education, 20, 1–20. https://doi.org/10.1186/s41239-023-00425-2

Fajardo-Ramos, D. C., Chiappe, A., & Mella-Norambuena, J. (2025). Human-in-the-loop assessment with AI: implications for teacher education in Ibero-American universities. Frontiers in Education. https://doi.org/10.3389/feduc.2025.1710992

Fleckenstein, J., Liebenow, L., & Meyer, J. (2023). Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. Frontiers in Artificial Intelligence, 6. https://doi.org/10.3389/frai.2023.1162454 Fu, Q.-K., Zou, D., Xie, H., & Cheng, G. (2022). A review of AWE feedback: Types, learning outcomes, and implications. Computer Assisted Language Learning, 37, 179–221. https://doi.org/10.1080/09588221.2022.2033787

García-López, I., & Trujillo-Liñán, L. (2025). Ethical and regulatory challenges of Generative AI in education: A systematic review. Frontiers in Education. https://doi.org/10.3389/feduc.2025.1565938

Hanifa, F. N. (2025). Assessing Algorithmic Fairness: Bias Analysis and Equity Principles in AI-Based Assessment and Personalized Learning Systems. Zawayatul Fikr: Journal of Islamic Education. https://doi.org/10.67198/zafie.v1i2.20

He, Z., Huang, S., & Shen, Y. (2025, December 12). TeachBench-AI: Design and Educational Applications of an Artificial Intelligence-Driven Classroom Teaching Analysis Platform. Proceedings of the 2025 3rd International Conference on Information Education and Artificial Intelligence. https://doi.org/10.1145/3799457.3799477

Huang, Y., & Wilson, J. (2021). Using automated feedback to develop writing proficiency. Computers and Composition. https://doi.org/10.1016/j.compcom.2021.102675 Kong, S.-C., & Yang, Y. (2024). A Human-Centered Learning and Teaching Framework Using Generative Artificial Intelligence for Self-Regulated Learning Development Through Domain Knowledge Learning in K–12 Settings. IEEE Transactions on Learning Technologies, 17, 1588–1599. https://doi.org/10.1109/tlt.2024.3392830

Lee, J., Hicke, Y., Yu, R., Brooks, C., & Kizilcec, R. (2024). The life cycle of large language models in education: A framework for understanding sources of bias. Br. J. Educ. Technol., 55, 1982–2002. https://doi.org/10.1111/bjet.13505

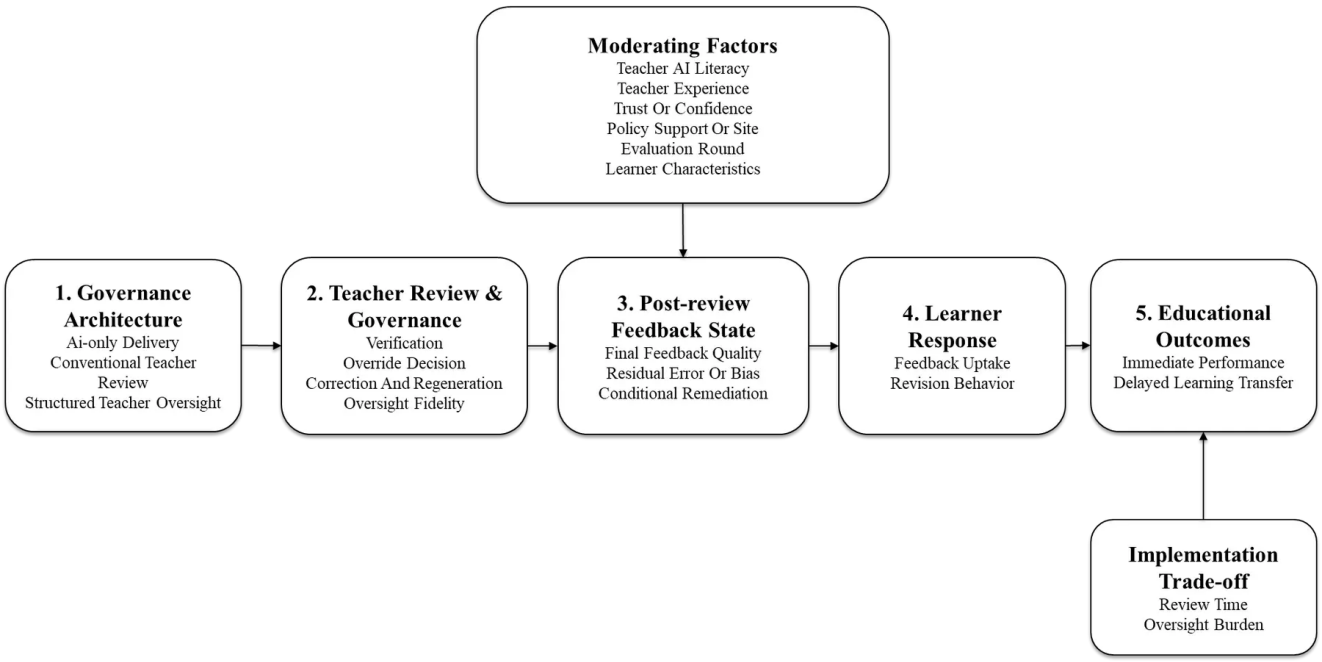


Fig. 1. Cyclical conceptual framework for human oversight in GenAI-mediated feedback.

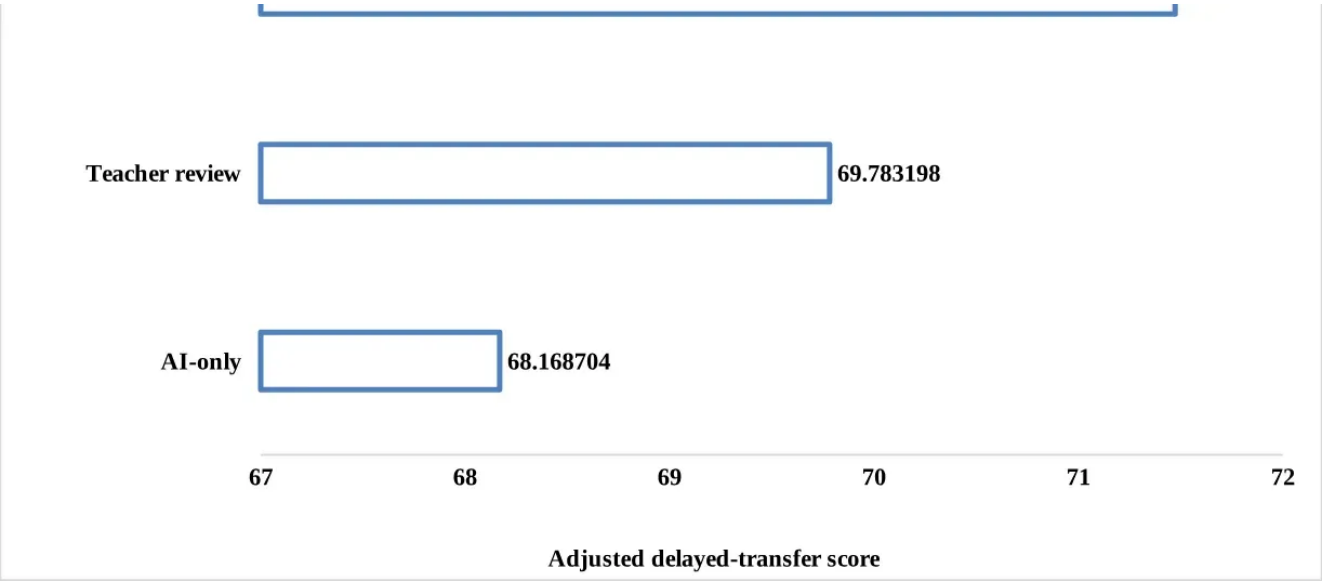


Fig. 2. Covariate-adjusted delayed-transfer score by oversight condition with 95% confidence intervals.

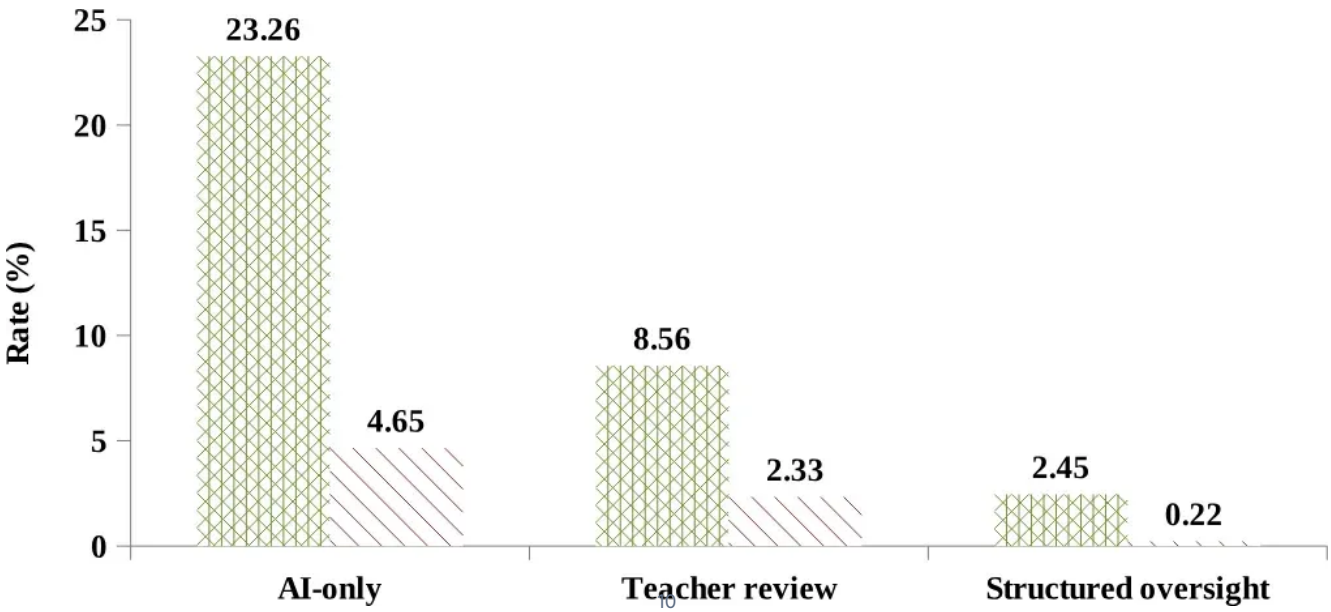


Fig. 3. Raw residual feedback-error and bias rates after each governance process.

Lee, S. S., & Moore, R. (2024). Harnessing Generative AI (GenAI) for Automated Feedback in Higher Education: A Systematic Review. Online Learning. https://doi.org/10.24059/olj.v28i3.4593

Ley, T., Tammets, K., Pishtari, G., Chejara, P., Kasepalu, R., Khalil, M., Saar, M., Tuvi, I., Väljataga, T., & Wasson, B. (2023). Towards a partnership of teachers and intelligent learning technology: A systematic literature review of model-based learning analytics. J. Comput. Assist. Learn., 39, 1397–1417. https://doi.org/10.1111/jcal.12844

Li, X., Chen, D., Tuffley, D., Scott, G., Tuxworth, G., Yao, T., & Torrisi-Steele, G. (2026). Evaluating AI-Generated Feedback for Formative Assessment in Higher Education. International Journal of AI in Pedagogy, Innovation, and Learning Futures. https://doi.org/10.46787/ijaipil.v2026i1.6962

Link, S., Mehrzad, M., & Rahimi, M. (2020). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. Computer Assisted Language Learning, 35, 605–634. https://doi.org/10.1080/09588221.2020.1743323

Liu, Z., Xing, W., Jiao, X., & Li, C. (2025). Exploring Fairness and Explainability in LLM-Generated Support for Online Learning Discussion Forums. J. Learn. Anal., 12, 8–33. https://doi.org/10.18608/jla.2025.8885

Maleki, A. (2026). Effects of AI generated and teacher feedback on EFL learners writing performance and emotional experience. Discov. Artif. Intell., 6, 199. https://doi.org/10.1007/s44163-026-00935-8

Merzifonluoğlu, A., & Gunes, H. (2025). Shifting Dynamics: Who Holds the Reins in Decision-Making With Artificial Intelligence Tools? Perspectives of Gen Z Pre-Service Teachers. European Journal of Education. https://doi.org/10.1111/ejed.70053

Mimoudi, A. (2025). Generative AI to bridge the educational divide: Personalized learning and challenges. Social Sciences & Humanities Open. https://doi.org/10.1016/j.ssaho.2025.102140

Nazaretsky, T., Mejia-Domenzain, P., Swamy, V., Frej, J., & Käser, T. (2025). Who Gives Feedback Matters: Student Biases Towards Human and AI -Generated Formative Feedback. J. Comput. Assist. Learn., 42. https://doi.org/10.1111/jcal.70153

Pahi, K., Hawlader, S., Hicks, E., Zaman, A., & Phan, V. (2024). Enhancing Active Learning through Collaboration Between Human Teachers and Generative AI. Computers and Education Open. https://doi.org/10.1016/j.caeo.2024.100183 (PDF) IMMEDIATE VERSUS DELAYED FEEDBACK IN PROMOTING STUDENT TEACHERS SKILLS FOR LESSON PLAN IMPLEMENTATION. (n.d.). ResearchGate. Retrieved August 31, 2026, from https://www.researchgate.net/publication/317098703_IMMEDIATE_VERSUS_DELAYED_FEEDBACK_IN_PROMOTING_STUDENT_TEACHERS_SKILLS_FOR_LESSON_PLAN_IMPLEMENTATION

Scheeler, M. C., Ruhl, K. L., & McAfee, J. K. (2004). Providing Performance Feedback to Teachers: A Review. Teacher Education and Special Education, 27(4), 396–407. https://doi.org/10.1177/088840640402700407

Shalviya, J., & Bhavana, A. R. (2026). Role of AI-Driven Teaching in Enhancing Classroom Decision-Making Skills of Prospective Teachers. Shanlax International Journal of Arts, Science and Humanities. https://doi.org/10.34293/sijash.v13is3-mar.10487

Shi, H., & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. ReCALL. https://doi.org/10.1017/s0958344023000265

Vallis, C., Huber, E., Sun, J., & Bryant, P. (2026). Hand in hand: Designing and integrating human and AI-generated feedback in business education. Journal of Work-Applied Management. https://doi.org/10.1108/jwam-03-2026-0062

Yeung, S. (2025). University students' engagement with generative AI-supported automated writing evaluation (AWE) feedback. Journal of Second Language Writing. https://doi.org/10.1016/j.jslw.2025.101203 Zapata-Rivera, J.-D., Torre, I., Lee, C.-S., Sarasa-Cabezuelo, A., Ghergulescu, I., & Libbrecht, P. (2024). Editorial: Generative AI in education. Frontiers in Artificial Intelligence, 7. https://doi.org/10.3389/frai.2024.1532896