

The Scoring Paradox: Multi-Agent Architectures for Unbiased Feedback Optimization

Okan Bulut¹, Bin Tan¹, Elisabetta Mazzullo¹, Cole Walsh²,

¹Measurement, Evaluation, and Data Science, University of Alberta

²Acuity Insights Inc.

Correspondence: bulut@ualberta.ca

Abstract

Does grade awareness bias automatically generated feedback? We present a multi-agent artificial intelligence (AI) framework that contrasts score-blind and score-informed evaluators and reconciles their critiques. Across 908 feedback blocks, score exposure systematically biased the model's critique: harsher at low grades, warmer at high (halo/horn effects). Isolating a score-blind judgment improved score-alignment without sacrificing faithfulness.

1 Introduction

Feedback is central to formative assessment because it helps students compare their current performance with expected standards and identify how to improve (Hattie and Timperley, 2007; Sadler, 1989). High-quality feedback should be timely, specific, actionable, supportive, and oriented toward learning (Bulut et al., 2026; Shute, 2008). Providing such feedback, however, is time-intensive for teachers, particularly when students submit open-ended responses that require individualized interpretation. Automated feedback generation (AFG) systems, including earlier rule-based and data-driven approaches, were initially developed to reduce this burden by generating feedback at scale. Although these systems improved efficiency, their feedback often remained rigid, generic, and insufficiently aligned with task content or students' individual support needs (Deeva et al., 2021). Recent AFG systems leveraging large language models (LLMs) have partly addressed these limitations, because LLMs can generate personalized, detailed, and context-relevant feedback on students' open-ended responses (Bulut et al., 2026; Dai et al., 2023; Jacobsen and Weber, 2023; Matelsky et al., 2023; Steiss et al., 2024).

Despite these advantages, LLM-based AFG systems are not always reliable in their consistency or accuracy (Seo et al., 2025), and their reliability

depends on factors such as model choice, prompt design, task domain, and evaluation criteria. One factor that has received comparatively little attention is *score awareness*. In many AFG applications, feedback is generated after a score has already been assigned to the student response. Score information may be useful, because it can help the model calibrate feedback to the student's apparent performance level. At the same time, it may alter how the model interprets the same written response because LLMs generate text by conditioning on the information included in the prompt. A numeric score can become a salient contextual cue that shapes the feedback narrative, biasing it toward coherence with the assigned score rather than fidelity to the student's actual work.

In this paper, we investigate this tension, which we term the *scoring paradox*, and ask whether awareness of a numeric score systematically biases the qualitative feedback an LLM produces. We propose and illustrate a modular, multi-agent pipeline that decouples a *score-blind* evaluation of a response from a *score-informed* one, reconciles the two, and regenerates optimized feedback. This parallel design serves two purposes: contrasting the blind and informed critiques exposes the magnitude and direction of any score-induced bias, including halo effects, in which high scores inflate the critique, and horn effects, in which low scores deflate it, while the reconciliation step provides a mechanism for mitigating that bias in the final feedback.

We further introduce an Evaluative Drift Index (EDI) to quantify the bias caused by score awareness, and we assess the pipeline through a multi-tier framework that combines an independent automated LLM judge, the EDI, and double-blind human expert validation. The specific research questions and objectives are detailed in Section 3.

2 Related Work

2.1 Feedback and AFG

Effective formative feedback helps learners close the gap between current and desired performance, and is most useful when it is specific, criterion-referenced, and actionable (Hattie and Timperley, 2007; Sadler, 1989; Shute, 2008). Delivering feedback of this quality by hand is labor-intensive, especially for open-ended responses, which motivated early AFG systems based on rule-based and data-driven methods. Such systems reduced teacher workload but tended to produce rigid and generic feedback that was poorly aligned with task content or individual learner needs (Deeva et al., 2021). LLMs have partly overcome these constraints, generating open-ended feedback that approaches human quality in some settings (Dai et al., 2023; Matelsky et al., 2023; Koutchene et al., 2024; Steiss et al., 2024), and this capability has been extended to criterion-referenced educational tasks such as item and feedback generation (Bulut et al., 2026; Mazzullo and Bulut, 2025).

2.2 Reliability of LLM-Based Evaluation

Despite these advances, LLM-based AFG systems are not uniformly reliable (Seo et al., 2025). Their reliability can be considered along three related dimensions. *Intra-model consistency* concerns whether the same LLM produces stable judgments, scores, or feedback for the same response across repeated runs. LLMs can be relatively stable under constrained conditions, but their outputs remain sensitive to prompt design, rubric specificity, and response characteristics (Chamieh et al., 2024; Lee et al., 2024).

Inter-model consistency concerns whether different LLMs produce comparable evaluations of the same response under standardized instructions. This form is often weaker, because differences in architecture and training data lead models to emphasize different textual features. For instance, Xue et al. (2026) found stronger intra-model than inter-model consistency in automatic scoring.

Agreement with human evaluators concerns whether LLM-generated scores or feedback-quality judgments align with expert ratings. Zhang et al. (2024) showed that detailed rubrics can improve LLMs' grading consistency, yet absolute agreement with human scores remained limited. Similarly, Koutchene et al. (2024) found that GPT-4, used to judge feedback quality, achieved only mod-

erate agreement with human raters and tended to rate feedback positively.

2.3 Bias in Human and Machine Evaluation

The concern that score awareness may bias LLM-generated feedback is rooted in two converging lines of literature. First, human evaluative judgment is vulnerable to prior information: halo and horn effects, in which a favorable or unfavorable prior impression of a student inflates or deflates judgments of their subsequent work, are well documented in grading and motivate practices such as anonymous marking (Malouff et al., 2013; Malouff and Thorsteinsson, 2016). Relatedly, anchoring and confirmation biases describe how an initial numeric value or prior belief shapes the interpretation of later evidence, so that judgments drift toward the anchor rather than the evidence itself (Nickerson, 1998; Tversky and Kahneman, 1974).

Second, LLM-based evaluators reproduce analogues of these biases: LLM-as-a-judge systems exhibit position bias (Shi et al., 2025) as well as prompt sensitivity, skewed rating patterns, and other judgmental inconsistencies (Stureborg et al., 2024). Furthermore, LLMs may produce plausible rationales that do not faithfully reflect the actual basis of their judgments, particularly when their outputs are shaped by biasing cues (Madsen et al., 2024; Turpin et al., 2023).

In the context of AFG, these strands imply that a score-informed model may generate feedback that strives to remain coherent with the assigned score; the evaluative frame supplied by the score may shape how the model interprets the response, leaving the feedback less independently grounded in the student's textual evidence and limiting its value for personalized, learning-oriented feedback. Despite this convergence, score-induced bias in LLM-generated feedback has not been directly isolated or quantified, which is the main gap the present study addresses.

3 Current Study

Although the two literatures reviewed above converge on the expectation that a numeric score can bias an LLM's evaluative judgment, score-induced bias in LLM-generated feedback has not been directly isolated, quantified, or mitigated so far. The present study addresses this gap by leveraging a modular, multi-agent AI architecture that decouples a score-blind evaluation of a student response

from a score-informed one, contrasts the two to expose the influence of the score, and reconciles them into optimized feedback. The two objectives underlying this study are:

1. **To quantify confirmation bias in score-aware feedback.** We measure the semantic and evaluative drift between score-blind and score-informed agents that produce personalized feedback. By analyzing how the presence of scores alters the AI’s critique of the same student response, we seek to isolate the grading effect inherent in single-prompt LLM evaluations, including its direction across the performance range (i.e., halo effects at higher scores and horn effects at lower scores).
2. **To evaluate the efficacy of the reconciler agent.** We evaluate a Reconciler Agent tasked with synthesizing the conflicting outputs of the blind and informed agents. Specifically, we assess the Reconciler’s capacity to generate optimized feedback that balances pedagogical objectivity with accurate, contextual constraints (i.e., feedback that remains grounded in the student’s textual evidence while staying realistic to the student’s actual performance level).

To address these objectives, we evaluate the pipeline through a multi-tier framework that pairs an independent automated LLM judge and the proposed EDI metric with double-blind human expert validation, allowing us both to measure the bias and to test whether the reconciliation process mitigates it without degrading feedback quality.

4 Methods

4.1 Sample

The dataset for this study was obtained from the Professional Skills Development¹ assessment, a formative evaluation platform developed by Acuity Insights. This assessment presents learners with 30 situational, open-ended workplace or social dilemmas, prompting them to construct behavioral responses grounded in their individual lived experiences. Each item on the assessment targets one of the four foundational skills: Critical Thinking, Social and Ethical Responsibility, Intrapersonal Skills, and Interpersonal Skills.

¹<https://acuityinsights.com/products/professional-skills-development/>

The sample used in this study consists of 841 post-secondary students located across the United States and Canada. The dataset encompasses item-level text responses, item-level scores (ranging from 0 to 5), skill-level composite scores (i.e., average item-level scores at the skill level), and historical automated feedback generated via GPT-5 Mini based on criterion-level reasoning (i.e., without considering any scores).

4.2 Multi-Agent Framework

To investigate the scoring paradox and eliminate evaluative bias, this study implemented a modular, multi-agent AI pipeline using Qwen3-235B via the Vulcan API at the University of Alberta². The pipeline audits, refines, and optimizes historical student feedback (i.e., existing human or AI-generated feedback). The multi-agent workflow is structured into four distinct procedural phases (see Figure 1).

4.2.1 Phase 1: Score-Blind Evaluation

Student responses, historical feedback, and the assessment rubric are routed to Agent A, with all item- and skill-level scores withheld. Acting as an unbiased observer, Agent A critiques the alignment between the written evidence and historical feedback against rubric criteria alone.

4.2.2 Phase 2: Score-Informed Evaluation

Concurrently, Agent B receives the same input alongside the numeric scores, verifying whether the historical feedback accurately justifies the assigned grades and flagging divergences between tone and recorded performance.

4.2.3 Phase 3: Algorithmic Reconciliation

Agent C cross-examines both critiques to isolate confirmation bias. Where evaluations conflict, it defaults to Agent A’s textual objectivity while respecting Agent B’s score boundaries, ensuring the reconciled narrative reflects the student’s actual performance tier.

4.2.4 Phase 4: Tone Polishing and Final Optimization

The de-biased content is then passed to Agent D, which applies growth-mindset scaffolding and an empathetic, professional tone to produce the final four-block feedback output (i.e., personalized feedback per skill).

²<https://inference.vulcan.alliancecan.ca/>

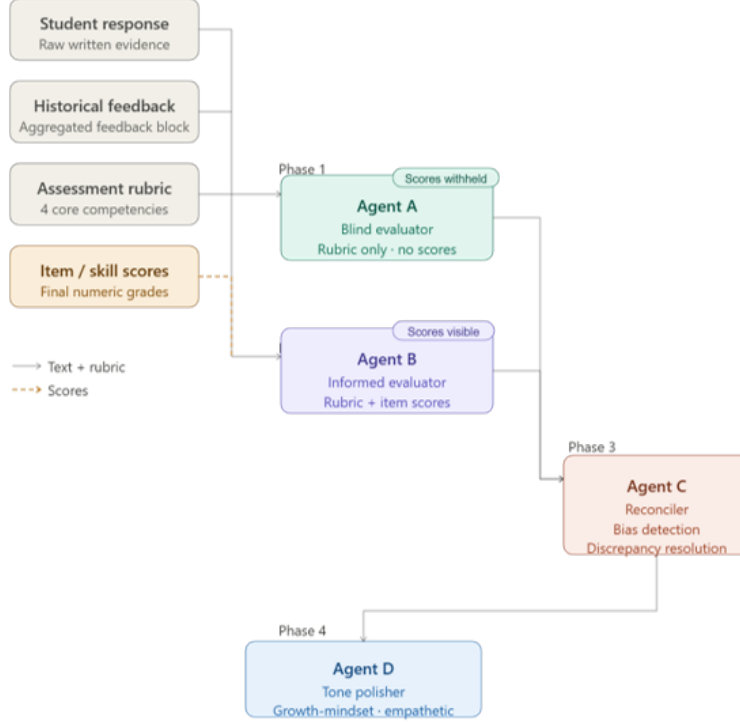


Figure 1: Multi-agent pipeline for auditing and optimizing historical student feedback.

4.3 Evaluation Metrics

We evaluated the multi-agent pipeline’s performance through a four-tier framework (see Figure 2) that combines automated quality scoring, an internal drift metric, human expert validation, and statistical comparison.

First, an independent frontier model (Claude Haiku 4.5), decoupled from the generation pipeline, scored both the historical (baseline) and optimized feedback blocks on a 5-point scale across four dimensions: (1) faithfulness and grounding, (2) score-alignment consistency, (3) pedagogical actionability, and (4) affective tone. For each case, the two blocks were presented to the judge in randomized order and without provenance labels (blind paired scoring), preventing any systematic preference for the optimized output.

Second, we used EDI to quantify score-induced confirmation bias by measuring the divergence between the intermediate critiques of the blind evaluator (Agent A) and the informed evaluator (Agent B). For a given feedback case, let $\mathbf{e}_A$ and $\mathbf{e}_B$ denote the sentence-embedding representations (SBERT, all-mpnet-base-v2) of the blind and informed critiques, and let $s_A, s_B \in [-1, 1]$ denote their sentiment polarity (i.e., VADER compound scores). The EDI combines a semantic and a sentiment com-

ponent as follows:

$$\text{EDI} = w_{\text{sem}} \underbrace{\left(1 - \frac{\mathbf{e}_A \cdot \mathbf{e}_B}{\|\mathbf{e}_A\| \|\mathbf{e}_B\|} \right)}_{\text{semantic drift}} + w_{\text{sent}} \underbrace{\frac{|s_B - s_A|}{2}}_{\text{sentiment drift}},$$

where the semantic component is the cosine distance between the two critique embeddings, the sentiment component is the absolute difference in polarity normalized to $[0, 1]$, and $w_{\text{sem}} = w_{\text{sent}} = 0.5$. The EDI is bounded on $[0, 1]$; higher values indicate greater score-induced divergence between the blind and score-informed critiques. To recover the direction of the bias, we additionally examine the signed sentiment drift, $s_B - s_A$, where negative values indicate that score exposure made the critique harsher (horn effect) and positive values indicate that it made the critique more lenient (halo effect).

Third, a stratified sample of feedback blocks ($n = 89$) drawn proportionally within each of the four skill domains underwent double-blind review by two assessment experts, both to validate the automated findings and to capture qualities that automated metrics cannot reliably assess. For each sampled case, experts were shown the original and optimized feedback blocks in randomized order and without provenance (labeled only “Feedback A” and “Feedback B”) and rated each on a 5-point

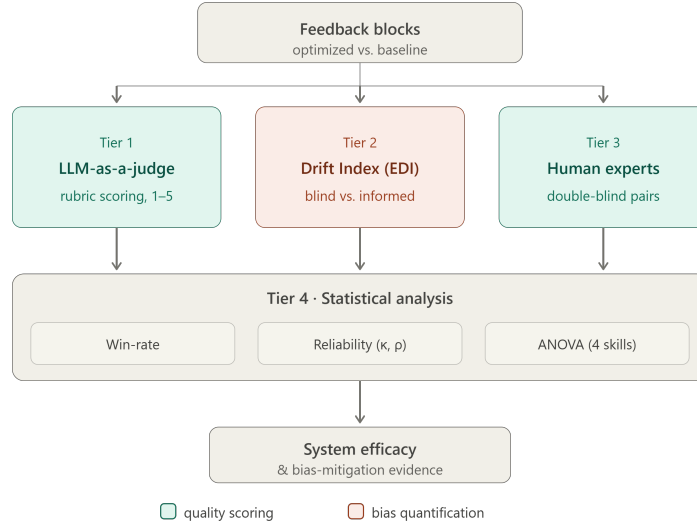


Figure 2: Multi-tiered evaluation framework

scale across four dimensions: clarity and readability, motivational quality, pedagogical usefulness, and affective tone; they additionally recorded a holistic forced-choice preference indicating which block they would give the student. To keep the human ratings free of the score-induced bias under investigation, reviewers were not shown the numeric scores. Two of these dimensions (i.e., pedagogical usefulness and affective tone) correspond to dimensions scored by the automated LLM judge, enabling a direct human–model agreement analysis, whereas the remaining two (i.e., clarity and motivational quality) capture subjective, learner-facing qualities that require professional judgment and are not objectively verifiable by a model.

Finally, we synthesized the preceding tiers through three statistical analyses. (i) A win-rate analysis computed the proportion of blocks on which the optimized feedback outscored the baseline, per dimension and for the composite, under both the automated LLM judge and the human raters. (ii) Inter-rater reliability quantified the agreement between the automated LLM judge and the human experts on the two shared dimensions using quadratic-weighted Cohen’s κ (absolute agreement) and Spearman’s ρ (rank correlation). (iii) A one-way analysis of variance (ANOVA) tested whether the magnitude of the optimization effect and of the evaluative drift differed significantly across the four skill domains (critical thinking, social and ethical responsibility, intrapersonal, and interpersonal skills).

5 Results

Tier 1 findings showed that the multi-agent pipeline significantly improved score-alignment consistency ($+0.28$, $p < .001$) and preserved faithfulness, confirming the reconciliation mechanism works, but reduced pedagogical actionability (-1.07 , $p < .001$), making the baseline feedback more preferable overall (see Figure 3).

Figure 4 shows the distribution of the EDI across all feedback blocks, where the horizontal axis captures how far the score-informed critique (Agent B) drifted from the score-blind critique (Agent A) for that block, while the vertical axis is the number of feedback blocks. The mass of the distribution sits at low EDI values, suggesting that the two agents largely agreed (i.e., score awareness did not dramatically change the critique). However, a long tail stretches to the right toward higher drift, suggesting that a meaningful subset of feedback blocks shows large drift—cases where exposure to the score substantially reshaped the informed agent’s critique of the same response. Overall, Figure 4 indicates that the bias is real and non-trivially distributed rather than a rounding artifact.

Decomposing the EDI revealed that the divergence was predominantly semantic ($M = 0.193$) rather than tonal ($M = 0.099$), indicating that score awareness altered *what* the informed evaluator attended to in the response, not merely the warmth with which it was expressed. The drift also varied systematically with performance: it was significantly larger for low-scoring responses

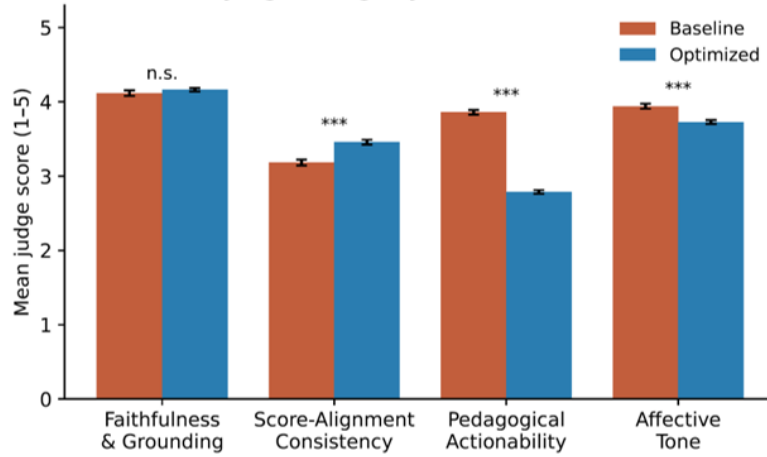


Figure 3: Mean judge ratings by evaluation dimensions in Tier 1. (Note: n.s. refers to non-significant results, *** $p < .001$)

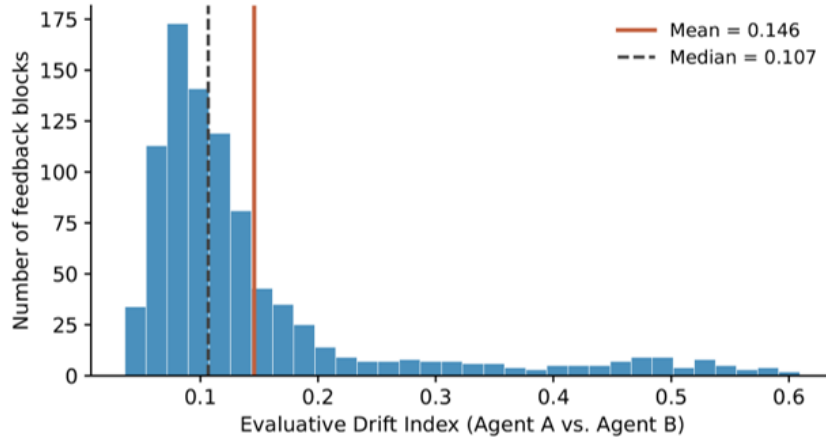


Figure 4: Distribution of blind-informed evaluative drift.

and declined monotonically as scores increased ($\rho = -.235$, $p < .001$), and it differed across the four skill domains (one-way ANOVA, $F = 10.8$, $p < .001$), being greatest for critical thinking.

Critically, the bias was not only larger at low scores but *directional*. The signed sentiment drift (informed – blind) shifted from negative at the low tier to positive at the high tier, and correlated positively with the assigned score ($\rho = +.151$, $p < .001$; see Figure 5). In other words, relative to the score-blind baseline, the informed evaluator’s critique became harsher than the textual evidence warranted for low-scoring responses and more lenient for high-scoring ones. This pattern is a direct empirical signature of the horn effect (i.e., Agent B’s critique turning harsher for low scores) and the halo effect (Agent B’s critique turning warmer for high scores), establishing that the assigned score—not the student’s writing alone—systematically shaped the LLM’s qualitative evalu-

ation of feedback.

Synthesizing across tiers confirmed and sharpened the automated findings. The win-rate analysis showed that the optimized feedback was outscored by the baseline on every dimension under both evaluators. Under the automated LLM judge, the baseline was preferred on 62.3% of blocks overall, driven primarily by pedagogical actionability (79.0% baseline wins); under the human experts, the baseline was preferred even more strongly, on 86.5% of blocks overall, with pedagogical usefulness showing the largest decline ($M_{\text{original}} = 4.65$, $M_{\text{optimized}} = 2.90$, $\Delta = -1.75$, $p < .001$; 92.1% baseline wins; see Figures 6 and 7). This pattern was corroborated by a forced-choice preference judgment, in which experts selected the original feedback for 79.8% of blocks and the optimized feedback for only 5.6% while having no preference for the remaining blocks (14.6%).

Inter-rater reliability on the two dimensions

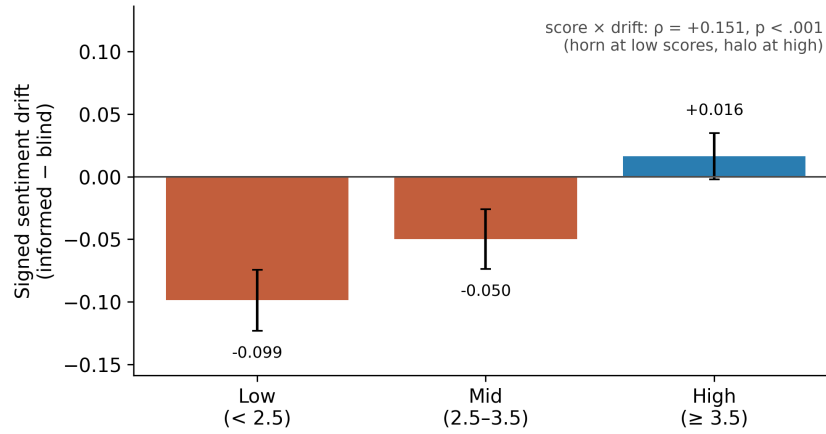


Figure 5: Signed sentiment drift (Informed - Blind) by performance categories.

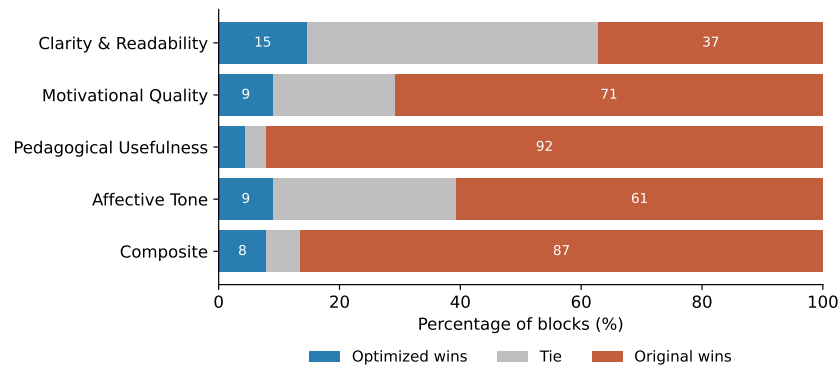


Figure 6: Win, tie, and loss rates by dimension based on human expert ratings.

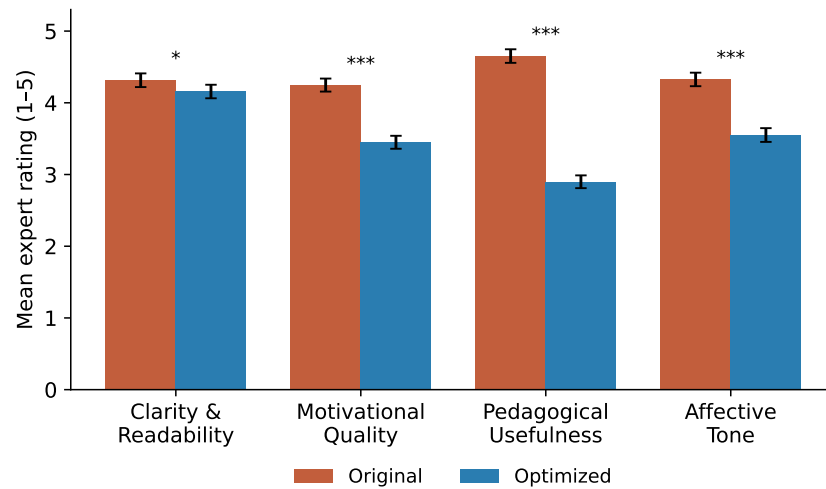


Figure 7: Mean human expert ratings for original vs. optimized feedback by dimension (*Note*: Error bars = ± 1 SE; significance from Wilcoxon signed-rank tests; * $p < .05$, ** $p < .01$, *** $p < .001$).

shared between the automated LLM judge and the human experts showed substantial agreement: pedagogical usefulness/actionability yielded a quadratic-weighted $\kappa = .60$ and Spearman's $\rho = .65$, and affective tone yielded $\kappa = .44$ and $\rho = .29$ (both $p < .001$). Human-human agree-

ment on the 21 doubly-rated pairs was similarly strong for usefulness ($\kappa = .85$, $\rho = .89$), validating the automated LLM judge as a reliable proxy for expert judgment on the dimensions most central to the scoring paradox.

The one-way ANOVA indicated that the magni-

tude of the optimization effect on human composite ratings varied significantly across the four skill domains ($F(3, 85) = 4.28, p = .007$), ranging from a near-negligible decline for critical thinking ($\Delta = -0.22, 22.7\%$ optimized win rate) to the most severe decline for interpersonal skills ($\Delta = -1.35, 0\%$ optimized win rate). Notably, this ordering did not track the magnitude of evaluative drift reported in Tier 2, for which critical thinking exhibited the *largest* EDI ($M = 0.181, F(3, 904) = 10.82, p < .001$) rather than the smallest. The skill domain most susceptible to score-induced bias in the blind-versus-informed critique was not the domain where reconciliation degraded feedback quality most. This dissociation suggests that the magnitude of confirmation bias detected by the EDI and the pipeline’s capacity to preserve feedback quality during reconciliation are governed by at least partially distinct mechanisms.

5.1 Discussion

This study investigated the *scoring paradox*: whether awareness of a numeric score systematically biases the qualitative feedback an LLM generates about the same student response. Our results answer this question in the affirmative and extend it in an unexpected direction. Consistent with the human grading literature on halo and horn effects (Malouff et al., 2013; Malouff and Thorsteinsson, 2016) and on anchoring and confirmation bias (Nickerson, 1998; Tversky and Kahneman, 1974), we found that a score-informed LLM evaluator systematically drifted from its score-blind counterpart, becoming harsher for low scoring responses and more lenient for high-scoring ones. This directional pattern is a machine analogue of biases long documented in human raters, and it converges with reports that LLM-as-a-judge systems are themselves susceptible to systematic rating biases (Shi et al., 2025; Stureborg et al., 2024). The divergence was predominantly semantic rather than tonal further, suggesting that the score does not merely color an evaluator’s language, it reshapes what the evaluator attends to in the response, a possibility consistent with concerns that LLM outputs can be shaped by biasing cues the model does not disclose in its stated rationale (Madsen et al., 2024; Turpin et al., 2023).

At the same time, our reconciliation mechanism only partly delivered on its aim. Isolating a score-blind judgment significantly improved score-alignment consistency without sacrificing faithful-

ness, showing that an unbiased textual anchor can correct a specific, measurable distortion. However, the optimized feedback was judged less pedagogically useful than the original by both an independent LLM judge and, more strongly, by two human experts, indicating that bias correction and feedback quality are not automatically co-optimized in a multi-agent pipeline. Notably, the judge and human experts converged on this conclusion (usefulness/actionability $\kappa = .60$; affective tone $\kappa = .44$), which contrasts with prior evidence that LLM judges tend to rate AI-generated feedback more positively than human raters do (Koutchme et al., 2024). This finding suggests that decoupling the judge from the generating pipeline appears to have preserved its critical validity rather than inflating it.

5.2 Limitations and Future Research

This study has several limitations. First, the multi-agent pipeline was evaluated on a single assessment platform, a single student population, and a single generator–judge model pairing (Qwen3-235B; Claude Haiku 4.5), so the magnitude of the observed bias and the actionability cost of reconciliation may not generalize to other LLMs, task domains, or learner populations. Future work should replicate the multi-agent framework across additional AFG contexts and model backbones to test whether the scoring paradox, and its correction cost, are architecture-general or specific to this pipeline.

Second, our design isolates that the tone-polishing stage degrades pedagogical usefulness but not *why*: we did not ablate Agent D’s prompting or intermediate outputs, so the mechanism driving the loss of actionable detail remains unknown. Future studies should conduct component-level ablations of the Reconciler and Polisher agents to identify which transformations trade away usefulness, and should explore alternative polishing strategies (e.g., explicit constraints to preserve concrete next steps) that might retain the bias-correction gains without this cost.

Lastly, the human validation, though convergent with the automated judge, drew on a modest sample (89 blocks), limiting statistical power for the skill-domain analyses and yielding low reliability for one dimension (clarity) due to a ceiling effect in expert ratings. Larger and more fully-crossed human panels would strengthen both the precision of these estimates and the generalizability of the human–model agreement findings.

References

- Okan Bulut, Bin Tan, and Elisabetta Mazzullo. 2026. [Applications of large language models in automatic item and feedback generation for early elementary education](#). In Stamatis Papadakis, editor, *AI Applications in Preschool and Primary Education*, pages 209–225. Springer.
- Imran Chamieh, Torsten Zesch, and Klaus Giebertmann. 2024. [LLMs in short answer scoring: Limitations and promise of zero-shot and few-shot approaches](#). In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 309–315. Association for Computational Linguistics.
- Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. 2023. [Can large language models provide feedback to students? a case study on ChatGPT](#). In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*.
- G. Deeva, D. Bogdanova, E. Serral, M. Snoeck, and J. De Weerd. 2021. [A review of automated feedback systems for learners: Classification framework, challenges and opportunities](#). *Computers & Education*, 162:104094.
- J. Hattie and H. Timperley. 2007. [The power of feedback](#). *Review of Educational Research*, 77(1):81–112.
- L. J. Jacobsen and K. E. Weber. 2023. [The promises and pitfalls of ChatGPT as a feedback provider in higher education: An exploratory study of prompt engineering and the quality of AI-driven feedback](#). Preprint.
- C. Koutchme, N. Dainese, S. Sarsa, A. Hellas, J. Leinonen, and P. Denny. 2024. [Open source language models can provide feedback: Evaluating LLMs’ ability to help students using GPT-4-as-a-judge](#). Preprint, arXiv:2405.05253.
- Gyeong-Geon Lee, Ehsan Latif, Xuansheng Wu, Ninghao Liu, and Xiaoming Zhai. 2024. [Applying large language models and chain-of-thought for automatic scoring](#). *Computers and Education: Artificial Intelligence*, 6:100213.
- A. Madsen, S. Chandar, and S. Reddy. 2024. [Are self-explanations from large language models faithful?](#) In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 295–337. Association for Computational Linguistics.
- J. M. Malouff and E. B. Thorsteinsson. 2016. [Bias in grading: A meta-analysis of experimental research findings](#). *Australian Journal of Education*, 60(3):245–256.
- John M. Malouff, Sarah J. Stein, Lodewyk N. Bothma, Kimberley Coulter, and Ashley J. Emmerton. 2013. [The risk of a halo bias as a reason to keep students anonymous during grading](#). *Teaching of Psychology*, 40(3):233–237.
- J. K. Matelsky, F. Parodi, T. Liu, R. D. Lange, and K. P. Kording. 2023. [A large language model-assisted education tool to provide feedback on open-ended responses](#). Preprint, arXiv:2308.02439.
- Elisabetta Mazzullo and Okan Bulut. 2025. Automated feedback generation for open-ended questions: Insights from fine-tuned LLMs. In *Proceedings of Large Foundation Models for Educational Assessment*, volume 264 of *Proceedings of Machine Learning Research*, pages 103–120.
- R. S. Nickerson. 1998. [Confirmation bias: A ubiquitous phenomenon in many guises](#). *Review of General Psychology*, 2(2):175–220.
- D. R. Sadler. 1989. [Formative assessment and the design of instructional systems](#). *Instructional Science*, 18:119–144.
- Hyein Seo, Taewook Hwang, Jeesu Jung, Hyeonseok Kang, Hyuk Namgoong, Yohan Lee, and Sangkeun Jung. 2025. [Large language models as evaluators in education: Verification of feedback consistency and accuracy](#). *Applied Sciences*, 15(2):671.
- Lin Shi, Chiyu Ma, Weicheng Liang, Xinyu Diao, Wenhao Ma, and Soroush Vosoughi. 2025. [Judging the judges: A systematic study of position bias in LLM-as-a-judge](#). In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 292–314. Asian Federation of Natural Language Processing and Association for Computational Linguistics.
- V. J. Shute. 2008. [Focus on formative feedback](#). *Review of Educational Research*, 78(1):153–189.
- J. Steiss, T. Tate, S. Graham, J. Cruz, M. Hebert, J. Wang, Y. Moon, W. Tseng, M. Warschauer, and C. B. Olson. 2024. [Comparing the quality of human and ChatGPT feedback of students’ writing](#). *Learning and Instruction*, 91:101894.
- Rickard Stureborg, Dimitris Alikaniotis, and Yoshiko Suhara. 2024. [Large language models are inconsistent and biased evaluators](#). Preprint, arXiv:2405.01724.
- M. Turpin, J. Michael, E. Perez, and S. R. Bowman. 2023. [Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting](#). Preprint, arXiv:2305.04388.
- A. Tversky and D. Kahneman. 1974. [Judgment under uncertainty: Heuristics and biases](#). *Science*, 185(4157):1124–1131.
- M. Xue, X. Xiao, Y. Liu, and M. Wilson. 2026. [On the consistency of automatic scoring with large language models](#). *Educational and Psychological Measurement*. Advance online publication.

D. W. Zhang, M. Boey, Y. Y. Tan, and A. H. S. Jia. 2024.
[Evaluating large language models for criterion-based
grading from agreement to consistency.](#) *npj Science
of Learning*, 9(1):79.