

Composite Scores vs. Preference Rankings: Measuring Architecture Effects in LLM Feedback

Harvey Ngoe Kolle¹
kolle@ualberta.ca

Carrie Demmans Epp^{1,2}
cdemmansepp@ualberta.ca

Amna Liaquat³
aliaquat@gmu.edu

Maria Cutumisu^{1,4,5}
maria.cutumisu@mcgill.ca

¹Department of Computing Science, University of Alberta, Edmonton, AB, Canada

²Alberta Machine Intelligence Institute (Amii), Edmonton, AB, Canada

³George Mason University, Fairfax, VA, United States

⁴Department of Educational and Counselling Psychology, McGill University, Montreal, QC, Canada

⁵Mila – Quebec Artificial Intelligence Institute, Montreal, QC, Canada

Abstract

AI-generated feedback is often judged by quality ratings or preference rankings but rarely both. Comparing a multi-agent system, a single-agent system, and human feedback on student writing, we show that the two evaluation approaches can support different reported conclusions even when their underlying effects are nearly identical. These differences have consequences for how AI-generated feedback should be evaluated.

1 Background

Good feedback helps learners understand how their current performance compares with expected goals and what they need to do to improve (Hattie and Timperley, 2007; Shute, 2008). However, high quality feedback is hard to produce and is often sparse or delayed in practice (Gibbs and Simpson, 2005; Nicol and Macfarlane-Dick, 2006). Large language models (LLMs) offer one way to generate feedback at scale (Rüdian et al., 2025; Seßler et al., 2025; Lu et al., 2025) but the approaches can differ in quality. This creates a practical problem: how should competing methods for generating feedback be compared?

Most LLM systems generate feedback in a single step: one prompt has the model assess accuracy, pedagogical alignment, and tone all at once (Rüdian et al., 2025; Lu et al., 2025; Seßler et al., 2025), even though the feedback literature treats these as separate qualities (Hattie and Timperley, 2007; Shute, 2008). An alternative is to split generation into specialized agents, separating accuracy, pedagogical alignment, and tone (Kolle et al., 2026a; Cao et al., 2025). Whether this added complexity is worthwhile is an empirical question. It depends on

whether the multi-agent approach produces feedback that raters judge as better than a single-agent system built from the same model and prompts, and on whether the comparison can detect such a difference at all.

However, whether such a difference can be detected depends on how feedback quality is measured, and current practice has two weaknesses. First, studies usually report either an aggregate quality score or a preference ranking (Rüdian et al., 2025; Weidlich et al., 2025) but rarely both, so it is unclear whether the two would point to the same answer. Second, the rating instruments often have unreported reliability (Nunnally, 1975; Streiner, 2003), which weakens confidence in any detected differences. Thus, choosing the better approach is partly a question of what counts as adequate evidence; the same concern that underlies argument-based approaches to validation in educational measurement (Kane, 2013). The present study addresses both problems using a 14-item, theory-grounded instrument that captures ten dimensions of feedback quality (Kolle et al., 2026b).

2 Aims

This study investigates whether the two evaluation approaches, composite quality scores and pairwise preference rankings, consistently indicate which feedback raters judge to be of higher quality. We examine this with three feedback conditions for adult English-language-learner writing: a multi-agent system, a single-agent system using the same base model, and human instructor feedback. The question matters because the choice of measure could determine whether one approach is better than another.

3 Sample

Eighteen raters (15 women, two men, and one non-binary) participated in this study. These raters were instructors and pre-service teachers with relevant experience evaluating writing or providing feedback. Age was available for 11 raters and averaged 25.9 years ($SD = 6.3$, range 19–40). Most raters reported prior experience with a teaching placement (72%), giving writing feedback (94%), grading writing (89%), feedback training (72%), and pedagogy or assessment training (78%). Fifteen raters (83%) had used LLMs before.

Student writing was from a prior study with adult English language learners from the Language Instruction for Newcomers to Canada program (Li-aqat et al., 2019). Each rater evaluated three submissions and reviewed all three feedback versions for each submission, yielding 54 rated feedback instances. These 54 instances were drawn from 10 unique student submissions. Each submission was rated by between 2 and 8 raters across the three writing tasks.

4 Methods

We employed a within-subject design. The single-agent and multi-agent conditions used the same underlying language model (GPT-4.1) and task materials. Thus, the comparison isolated the effect of architecture rather than that of the model or task. The single-agent condition produced feedback in one step, using a single prompt that asked the model to address accuracy, pedagogical alignment, and tone. The multi-agent condition took that same initial output and passed it through a linear pipeline of specialized agents, each with its own prompt: successive agents checked grounding, corrected unsupported content, aligned feedback with task expectations, strengthened actionability, adjusted tone, and formatted the result. Both conditions shared the same base model and starting feedback, and differed only on whether that feedback was further processed by the pipeline. Full prompts and agent parameters are reported in the companion paper (Kolle et al., 2026a). The human condition used instructor-written comments from the original rubric-based feedback. For each submission, a rater reviewed three feedback versions. To reduce position effects, the version ordering was randomized and the source concealed.

The rater assessed each version using a 14-item instrument spanning ten dimensions of feedback

quality (Table 1) and then ranked the three versions by overall preference. Composite scores, rescaled to 0–100, were analyzed with a linear mixed-effects model; condition was a fixed effect and rater and submission were random intercepts (Baayen et al., 2008). Preference rankings were analyzed with a Bradley-Terry model (Bradley and Terry, 1952). All pairwise comparisons were adjusted for multiple comparisons using Holm correction within each outcome.

5 Results

The 14-item instrument had high internal consistency, with $\omega = 0.917$, 95% CI [0.903, 0.931], bootstrapped with $B = 2000$ resamples (Kolle et al., 2026b; McDonald, 2013), supporting its use as a composite outcome measure. Composite scores were highest for multi-agent feedback, followed by single-agent feedback and human feedback (Table 4). Both AI conditions scored higher than human feedback. The composite-score difference between multi-agent and single-agent feedback did not survive Holm correction (Table 2).

The ranking analysis showed the same ordering (Table 3): multi-agent feedback was preferred to single-agent feedback, and both AI conditions were preferred over human feedback. At the dimension level, the only significant difference between the multi-agent and single-agent conditions was on supportive tone, with multi-agent feedback rated higher ($\Delta = 4.63$, $p_{adj} = .049$, $d = 0.31$).

For the multi-agent vs. single-agent contrast, the two measures produced the same ordering and only marginally different effect sizes (composite $d = 0.29$, ranking $r = .31$); they differed only in whether the contrast reached significance after correction. A direct comparison confirmed that the measures themselves did not differ: the two feedback versions were ranked in the same direction as their composite scores on 72% of instances (95% CI [0.62, 0.82]), and the standardized effects of the two measures did not differ significantly (bootstrap difference = 0.02, 95% CI [−0.32, 0.39]).

6 Discussion and Conclusions

Composite scores and preference rankings agreed that both AI conditions were rated and ranked above human feedback. For the multi-agent vs. single-agent contrast, they agreed on ordering and effect magnitude but differed on whether the contrast reached significance. This difference in signif-

Table 1: The 14-item instrument used to rate feedback quality, grouped into ten dimensions. Items marked [R] are reverse-scored before summing so that the composite consistently reflects higher perceived quality.

Dimension	Item
Specific & Clear	This feedback targets specific parts of the writing. This feedback is clearly written.
Goal-Referenced	This feedback explains how the writing meets the expected goals.
Actionable	This feedback offers clear guidance the student can follow. This feedback supports improvement in writing skills.
Supportive Tone	This feedback maintains a positive tone. This feedback is discouraging. ^[R]
Trustworthiness	This feedback is inaccurate. ^[R]
Objective	This feedback is based on evidence from the student’s writing. This feedback focuses on student traits. ^[R]
Non-Threatening	This feedback uses harsh language. ^[R]
Non-Revealing	This feedback suggests alternative phrasing.
Conciseness	This feedback avoids overly detailed elaboration.
Overall	This feedback would be appropriate to share directly with a student.

Table 2: Pairwise contrasts for composite feedback quality scores. 95% CIs are unadjusted; p_{adj} values are Holm-corrected.

Comparison	Diff	95% CI	p_{adj}	d
Multi-Agent vs. Single-Agent	2.46	[0.14, 4.78]	.114	0.29
Multi-Agent vs. Human	26.38	[22.48, 30.27]	< .001	1.85
Single-Agent vs. Human	23.92	[20.24, 27.59]	< .001	1.78

Table 3: Pairwise preference probabilities from the Bradley-Terry model.

Comparison	$P(\text{preferred})$	p_{adj}	r
Multi-Agent vs. Single-Agent	.657	.015	.31
Multi-Agent vs. Human	.957	< .001	.91
Single-Agent vs. Human	.921	< .001	.84

Table 4: Descriptive statistics for composite scores by condition (rescaled to 0–100).

Condition	M	SD	Range
Multi-Agent	87.0	11.2	54–100
Single-Agent	84.5	12.6	51–100
Human	60.6	13.5	40–91

icance is not itself evidence that the measures differ (Gelman and Stern, 2006), and a direct comparison found no significant difference between them.

The pattern of both AI conditions outperforming human feedback is consistent with prior reports that LLM-generated feedback can compare favorably with human feedback on some dimensions (Rüdian et al., 2025; Weidlich et al., 2025). However, the human feedback in this study was drawn from a prior instructional context, so the observed

difference should be interpreted cautiously.

The split in significance between the two AI conditions may reflect that a difference concentrated in one dimension (i.e., tone) is diluted once combined into a 14-item composite (Nunnally, 1975). However, with 18 raters, it may reflect that the ranking analysis was better powered to detect small effects. A larger-sample replication is planned to tease apart these explanations. If the tone difference persists, it would be consistent with the multi-agent pipeline’s design that uses a dedicated tone agent to revise feedback after its content and pedagogical structure have already been refined (Kolle et al., 2026a).

These results suggest that the choice between aggregate scores and preference rankings is not just a reporting detail: it can determine whether a difference between AI systems is detected at all. More broadly, this implies that conclusions

about AI-generated feedback are partly an artifact of how feedback quality is measured, not just of the systems themselves. As these tools become more common, evaluations that only report one outcome measure run the risk of over- or under-stating real differences between systems. Therefore, comparisons of AI systems should report both where possible.

References

- R.H. Baayen, D.J. Davidson, and D.M. Bates. 2008. [Mixed-effects modeling with crossed random effects for subjects and items](#). *Journal of Memory and Language*, 59(4):390–412. Special Issue: Emerging Data Analysis.
- Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.
- Jie Cao, Chloe Qianhui Zhao, Xian Chen, Shuman Wang, Christian Schunn, Kenneth R Koedinger, and Jionghao Lin. 2025. [From first draft to final insight: A multi-agent approach for feedback generation](#). In *International Conference on Artificial Intelligence in Education*, pages 163–176. Springer-Verlag.
- Andrew Gelman and Hal Stern. 2006. [The difference between “significant” and “not significant” is not itself statistically significant](#). *The American Statistician*, 60(4):328–331.
- Graham Gibbs and Claire Simpson. 2005. Conditions under which assessment supports students’ learning. *Learning and Teaching in Higher Education*, (1):3–31.
- John Hattie and Helen Timperley. 2007. [The power of feedback](#). *Review of educational research*, 77(1):81–112.
- Michael T. Kane. 2013. [Validating the interpretations and uses of test scores](#). *Journal of Educational Measurement*, 50(1):1–73.
- Harvey Ngoe Kolle, Carrie Demmans Epp, Amna Liaqat, and Maria Cutumisu. 2026a. [Do Language Models Matter? Evaluating Model Choice in a Multi-agent Feedback Framework](#). In *Artificial Intelligence in Education. Late Breaking Results, WideAIED, Practitioners, Industry and Policies, Blue Sky, Doctoral Consortium, FoL Workshops and Tutorials, FoL Invited Papers*, volume 3031, pages 108–114, Cham. Springer Nature Switzerland. Series Title: Communications in Computer and Information Science.
- Harvey Ngoe Kolle, Carrie Demmans Epp, Amna Liaqat, and Maria Cutumisu. 2026b. Evaluating the pedagogical quality of LLM-generated feedback: A criterion-based and comparative study. In *Proceedings of the First International Workshop on Pedagogical Evaluation of Automated Feedback (PEAF 2026)*, CEUR Workshop Proceedings. CEUR-WS.org.
- Amna Liaqat, Gokce Akcayir, Carrie Demmans Epp, and Cosmin Munteanu. 2019. [Mature ELLs’ Perceptions Towards Automated and Peer Writing Feedback](#). In *Transforming Learning with Meaningful Technologies*, Lecture Notes in Computer Science, pages 266–279. Springer International Publishing.
- Xinyi Lu, Aditya Mahesh, Zejia Shen, Mitchell Dudley, Larissa Sano, and Xu Wang. 2025. [Exploring LLM-generated feedback for economics essays: How teaching assistants evaluate and envision its use](#). In *International Conference on Artificial Intelligence in Education*, pages 392–406, Berlin, Heidelberg. Springer-Verlag.
- Roderick P McDonald. 2013. *Test theory: A Unified Treatment*. Psychology Press.
- David J. Nicol and Debra Macfarlane-Dick. 2006. [Formative assessment and self-regulated learning: A model and seven principles of good feedback practice](#). *Studies in Higher Education*, 31(2):199–218.
- Jum C Nunnally. 1975. Psychometric theory—25 years ago and now. *Educational researcher*, 4(10):7–21.
- Sylvio Rüdian, Julia Podelo, Jakub Kužílek, and Niels Pinkwart. 2025. [Feedback on feedback: Students’ perceptions for feedback from teachers and few-shot LLMs](#). In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 82–92, New York, NY, USA. Association for Computing Machinery.
- Kathrin Seßler, Maurice Fürstenberg, Babette Bühler, and Enkelejda Kasneci. 2025. [Can AI grade your essays? a comparative analysis of large language models and teacher ratings in multidimensional essay scoring](#). In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 462–472, New York, NY, USA. Association for Computing Machinery.
- Valerie J. Shute. 2008. [Focus on formative feedback](#). *Review of Educational Research*, 78(1):153–189.
- David L. Streiner. 2003. [Starting at the beginning: An introduction to coefficient alpha and internal consistency](#). *Journal of Personality Assessment*, 80(1):99–103.
- Joshua Weidlich, Flurin Gotsch, Kai Schudel, Claudia Marusic-Würscher, Jennifer Mazzarella, Hannah Bolten, Dario Büttler, Simon Luger, Bettina Wohlfehnder, and Katharina Maag Merki. 2025. Teacher, peer, or AI? comparing effects of feedback sources in higher education. *Computers and Education Open*, page 100300.