Evaluating the Pedagogical Quality of LLM-Generated Feedback: A Criterion-Based and Comparative Study

Harvey Ngoe Kolle[1], Carrie Demmans Epp[1,2], Amna Liaqat[3] and Maria Cutumisu[1,4,5]

[1]Department of Computing Science, University of Alberta, Edmonton, AB, Canada
[2]Alberta Machine Intelligence Institute (Amii), Edmonton, AB, Canada
[3]George Mason University, Fairfax, VA, United States
[4]Department of Educational and Counselling Psychology, McGill University, Montreal, QC, Canada
[5]Mila – Quebec Artificial Intelligence Institute, Montreal, QC, Canada

Abstract

Evaluating automated feedback on pedagogical grounds requires more than a single holistic judgment. We developed a multi-dimension 14-item rating instrument grounded in formative feedback theory and used it to compare feedback from a multi-agent system, a single-agent system, and an instructor. Eighteen evaluators—instructors and pre-service teachers—rated and ranked 54 feedback instances that were tied to adult English language learner writing. Both automated conditions received scores that were more than 20 points higher than human feedback, though this significant difference should be interpreted with caution, as the human feedback originated from a real teaching context and was not produced for this study. The only significant difference between the two automated conditions was on the supportive tone dimension, where the multi-agent condition received higher ratings than the single-agent condition. These results show that a multidimensional approach identifies differences in feedback quality that a single overall rating could miss.

Keywords

Automated Feedback, Multidimensional Evaluation, Formative Feedback, Large Language Models, Feedback Quality

1. Introduction

Automated feedback is now widely used in education, but most evaluations of this feedback rely on a single holistic rating or brief student surveys [1, 2]. This raises a practical question: are single holistic ratings sufficient for evaluating the quality of automated feedback?

Research on formative feedback points to several qualities that matter for learning: feedback should be trustworthy, tied to learning goals, specific, actionable, and supportive in tone [3, 4, 5]. These ideas are well established theoretically but rarely used to evaluate automated feedback [6, 7].

Therefore, the present study argues for evaluating automated feedback across multiple theory-grounded dimensions rather than relying on one overall judgment. We operationalised dimensions from the formative feedback literature into 14 rating items and used them in a study comparing automated and human-written feedback. We show that this approach surfaces differences in quality that a single overall rating would miss.

2. Rating Dimensions and Items

To evaluate feedback, we operationalised the many dimensions of formative feedback [3, 4, 5] into 14 rating items, each scored on a five-point Likert scale. The items are grouped into 10 dimensions (Table 1). To ensure consistency in scoring direction, negatively-worded items (discouraging, inaccurate,

PEAF 2026: First International Workshop on Pedagogical Evaluation of Automated Feedback
kolle@ualberta.ca (H. N. Kolle); cdemmansepp@ualberta.ca (C. Demmans Epp); aliaqat@gmu.edu (A. Liaqat); maria.cutumisu@mcgill.ca (M. Cutumisu)
0009-0002-0280-4996 (H. N. Kolle); 0000-0001-9079-4921 (C. Demmans Epp); 0000-0002-5170-1945 (A. Liaqat); 0000-0003-2475-9647 (M. Cutumisu)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Table 1
Rating dimensions and items for evaluating student writing feedback, derived from formative feedback theory.

Dimension	Item statement
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.

[R] Reverse-scored item.

trait-focused, and harsh) were reverse-scored before summing so that the composite score (range 14–70) consistently reflects higher perceived quality.

3. Study Design

3.1. Context and Feedback Conditions

We used the proposed instrument (Table 1) to compare three sources of feedback on adult English language learner writing. The student texts originated from a prior study in the Language Instruction for Newcomers to Canada programme [8], where learners completed three informal writing assignments and received rubric-based instructor feedback. That instructor feedback served as the **human** condition.

Two automated conditions were generated using GPT-4.1 with few-shot prompting [9]:

- **Single-agent**: feedback produced in a single prompt step.
- **Multi-agent**: the same initial output was then revised by a sequence of six specialised agents (verification, correction, subject-expert, pedagogy, tone, and formatting), each focused on a different aspect of quality.

The multi-agent pipeline follows a fixed sequential order in which each agent receives the output of the previous stage and refines it along a specific quality dimension [10]. Factual verification precedes domain review, which precedes pedagogical evaluation, tone adjustment, and formatting. This ordering is deliberate: downstream agents refine only verified content, and tone is adjusted after pedagogical structure is in place to avoid concealing substantive weaknesses behind encouraging language. Both automated conditions used the same foundation model and materials; the single-agent condition used a single prompt, while the multi-agent condition extended this with six additional agent-specific prompts, one per pipeline stage. Therefore, any differences in ratings reflect differences in how the feedback was generated rather than differences in the underlying foundation model.

3.2. Raters and Procedure

Eighteen raters participated in the study (15 women, 2 men, 1 non-binary; mean age 25.9 years, *SD* = 6.3). All had relevant background: 94% had given feedback on writing, 89% had graded writing, and 78%

had pedagogy or assessment training. Each rater evaluated three student submissions, one per writing task. For each submission, raters used the 14-item instrument to rate all three feedback versions and then ranked them from most to least preferred. The feedback source was not disclosed to raters. The study received research ethics board approval and all raters provided informed consent. Each study session took approximately one hour, and raters were compensated with a $30 gift card or PayPal transfer of equivalent value.

3.3. Analysis

Internal consistency of the 14 items was assessed using McDonald's omega, which is more appropriate than Cronbach's alpha for ordinal rating data as it does not assume equal item contributions to the composite score [11]. Composite scores were analysed with a linear mixed-effects model, with feedback condition as a fixed effect and rater and submission as random intercepts [12]. Rankings were analysed with a Bradley-Terry model [13]. All pairwise contrasts used Holm correction for multiple comparisons.

4. Results

4.1. Composite Scores

The 14 items demonstrated high internal consistency, with $\omega = 0.917$ (95% CI $[0.903, 0.931]$, bootstrapped with $B = 2000$ resamples), indicating that the items reliably measure a common underlying construct [14, 15].

Composite scores (scaled to 0–100) were highest for multi-agent feedback ($M = 87.0$, $SD = 11.2$, range: 54–100), followed by single-agent feedback ($M = 84.5$, $SD = 12.6$, range: 51–100), and human feedback ($M = 60.6$, $SD = 13.5$, range: 40–91).

Table 2 shows the pairwise contrasts for composite feedback quality scores. Both automated feedback conditions were rated substantially higher than the human feedback, with large effect sizes (Cohen's $d > 1.7$) [16]. No difference was found between the two automated feedback conditions.

Table 2
Pairwise contrasts for composite feedback quality scores (14-item instrument, scores rescaled to 0–100).

Comparison	Diff	95% CI	p_{adj}	Cohen's 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

4.2. Preference Rankings

The ranking analysis confirmed the same ordering (Table 3). Raters preferred the multi-agent feedback over the single-agent feedback, and both automated conditions were strongly preferred over human feedback.

Table 3
Pairwise preference probabilities from the Bradley-Terry model showing which feedback condition was preferred in each head-to-head comparison.

Comparison	P(Top-ranked)	p_{adj}	r
Multi-Agent vs Single-Agent	.657	**.015**	0.31
Multi-Agent vs Human	.957	$< .001$	0.91
Single-Agent vs Human	.921	$< .001$	0.84

Table 4
Descriptive statistics by dimension and condition (scores rescaled to 0–100).

	Human		Single-Agent		Multi-Agent	
Dimension	M	SD	M	SD	M	SD
Specific & Clear	71.5	20.7	88.2	20.8	90.7	17.6
Goal-Referenced	58.2	26.7	87.0	22.4	89.3	18.1
Actionable	44.4	21.9	85.7	18.5	87.8	17.8
Supportive Tone	58.0	20.5	87.4	15.3	92.0	12.0
Trustworthiness	64.8	29.8	87.4	19.9	90.0	18.5
Objective	72.2	12.5	75.0	13.8	74.1	13.5
Non-Threatening	74.4	24.4	88.2	19.2	92.2	15.3
Non-Revealing	35.2	23.0	81.1	23.8	84.4	22.2
Conciseness	78.9	25.3	79.6	20.0	83.7	22.0
Overall	44.4	27.1	87.0	16.6	88.5	14.9

4.3. Dimension-Level Results

The two automated feedback conditions scored similarly across all dimensions except **supportive tone**, where the multi-agent condition received higher ratings ($\Delta = 4.63$, $p_{adj} = .049$, $d = 0.31$). Both the single-agent and multi-agent feedback conditions scored higher than the human feedback on most dimensions (see Table 4 for descriptive statistics), with the largest gaps on actionable, non-revealing, and overall (all $p < .001$). No measurable differences were observed on the conciseness and objective dimensions across all three feedback conditions. Full dimension-level results are presented in the supplementary materials.[1]

5. Discussion

5.1. The Value of Multidimensional Evaluation

When looking at composite scores alone, the two automated feedback conditions appear similar. The dimension-level breakdown reveals more: the two conditions diverged only on supportive tone. This is a good example of what multi-dimensional instruments can do. Here, they surface distinctions that a single overall rating would hide [7].

The findings revealed a large gap between automated and human feedback across most dimensions, with the notable exception of conciseness and objective, where scores were similar across all three feedback conditions. These similarities may reflect a shared tendency among human instructors and automated systems to ground feedback in the student's work and focus on relevant aspects of the writing [17].

The human feedback originated from a real teaching context and was not written for this comparison, which limits how far we can push that interpretation. Still, the pattern fits with other work showing that automatically generated feedback can hold up well against human feedback on certain qualities [1, 18].

5.2. Composite Scores Versus Preference Rankings

Composite scores did not help to distinguish the two automated conditions, but the ranking task did. Raters noticed a difference that the composite score did not capture. This points to the value of collecting both ratings and rankings when evaluating feedback quality, and it suggests that the qualities of the feedback matter to evaluators even when the difference is subtle.

At the dimension level, the only difference between the two automated conditions was on supportive tone, which is consistent with the multi-agent design, where a dedicated tone agent refines this quality

[1] https://bit.ly/4bPmBcF

after pedagogical structure is in place.

5.3. Implications

Grounding evaluation items in feedback theory provides a clearer picture of where automated feedback succeeds or falls short. Future work could examine whether these dimensions hold across different subject areas and feedback contexts such as coding or mathematics, and whether individual items could be turned into computable metrics. Future work should also examine how students perceive and respond to automated feedback and whether rater-based quality judgments predict actual learning outcomes or improvements in submitted student work. Meanwhile, this study constitutes a starting point for building richer evaluation approaches for automatically-generated feedback.

6. Conclusion

We argued for evaluating automated feedback across multiple theory-grounded dimensions rather than relying on a single overall rating. Using 14 items derived from formative feedback theory [3, 4, 5], we compared automated and human-written feedback on adult writing. There were many differences in the qualities of the human and the generated feedback, but only tone showed a difference between the two automated approaches. It would not have been possible to identify these differences using a single, global rating. Both automated conditions scored substantially higher than the human feedback, though this comparison should be interpreted cautiously given the different contexts in which each was produced. Moreover, the dimension breakdown identified supportive tone as the one area where the two automated conditions differed. Importantly, while composite scores did not significantly differentiate the two automated conditions, the ranking task reliably distinguished them, underscoring the value of collecting preference judgments alongside rating scores.

Acknowledgments

This work was supported in part by funding from the Social Sciences and Humanities Research Council of Canada and the Natural Sciences and Engineering Research Council of Canada (NSERC) RGPIN-2019-07014 and RGPIN-2026-07040.

Declaration on Generative AI

The authors have not employed any Generative AI tools in the reported analyses or writing of this paper.

References

[1] S. Rüdian, J. Podelo, J. Kužílek, N. Pinkwart, Feedback on feedback: Students' perceptions for feedback from teachers and few-shot llms, in: Proceedings of the 15th International Learning Analytics and Knowledge Conference, Association for Computing Machinery, New York, NY, USA, 2025, pp. 82–92. doi:10.1145/3706468.3706479.

[2] K. Seßler, M. Fürstenberg, B. Bühler, E. Kasneci, 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, Association for Computing Machinery, New York, NY, USA, 2025, pp. 462–472. doi:10.1145/3706468.3706527.

[3] J. Hattie, H. Timperley, The power of feedback, Review of educational research 77 (2007) 81–112. doi:10.3102/003465430298487.

[4] V. J. Shute, Focus on formative feedback, Review of Educational Research 78 (2008) 153–189. doi:10.3102/0034654307313795.

[5] D. J. Nicol, D. Macfarlane-Dick, Formative assessment and self-regulated learning: A model and seven principles of good feedback practice, Studies in Higher Education 31 (2006) 199–218. doi:10.1080/03075070600572090.
[6] H. Nguyen, W. Xiong, D. Litman, Iterative design and classroom evaluation of automated formative feedback for improving peer feedback localization, International Journal of Artificial Intelligence in Education 27 (2017) 582–622. doi:10.1007/s40593-016-0136-6.
[7] H. Shi, V. Aryadoust, A systematic review of ai-based automated written feedback research, ReCALL 36 (2024) 187–209. doi:10.1017/S0958344023000265.
[8] A. Liaqat, G. Akcayir, C. Demmans Epp, C. Munteanu, Mature ells' perceptions towards automated and peer writing feedback, in: European Conference on Technology Enhanced Learning, Springer, 2019, pp. 266–279.
[9] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 1877–1901.
[10] H. N. Kolle, C. Demmans Epp, A. Liaqat, M. Cutumisu, Do language models matter? Evaluating model choice in a multi-agent feedback framework, in: International Conference on Artificial Intelligence in Education, 2026. Late-Breaking Results.
[11] R. P. McDonald, Test theory: A Unified Treatment, Psychology Press, 2013.
[12] R. Baayen, D. Davidson, D. Bates, Mixed-effects modeling with crossed random effects for subjects and items, Journal of Memory and Language 59 (2008) 390–412. doi:10.1016/j.jml.2007.12.005, special Issue: Emerging Data Analysis.
[13] R. A. Bradley, M. E. Terry, Rank analysis of incomplete block designs: I. the method of paired comparisons, Biometrika 39 (1952) 324–345.
[14] J. C. Nunnally, Psychometric theory—25 years ago and now, Educational researcher 4 (1975) 7–21.
[15] D. L. Streiner, Starting at the beginning: An introduction to coefficient alpha and internal consistency, Journal of Personality Assessment 80 (2003) 99–103. doi:10.1207/S15327752JPA8001_18.
[16] J.-C. Goulet-Pelletier, D. Cousineau, A review of effect sizes and their confidence intervals, part i: The cohen'sd family, The Quantitative Methods for Psychology 14 (2018) 242–265.
[17] W. S. Pearson, A typology of the characteristics of teachers' written feedback comments on second language writing, Cogent Education 9 (2022) 2024937. doi:10.1080/2331186X.2021.2024937.
[18] J. Weidlich, F. Gotsch, K. Schudel, C. Marusic-Würscher, J. Mazzarella, H. Bolten, D. Bütler, S. Luger, B. Wohlfehnder, K. M. Merki, Teacher, peer, or ai? comparing effects of feedback sources in higher education, Computers and Education Open (2025) 100300.