Beyond Surface Human-Likeness: AI–Mentor Feedback Alignment, Pedagogical Adaptation, and Student Engagement in Longitudinal Learning Data

Jing Fan[1], Jiseon Kim[2], Kimia Abedini[1], Daniel Koch-Truhponen[1], Charles Koutcheme[1] and Juho Leinonen[1]

[1]*Aalto University, Espoo, Finland*
[2]*KAIST Global Institute for Talented Education, Daejeon, Republic of Korea*

Abstract

Automated feedback is often evaluated by comparing it with human feedback, but surface resemblance may be an insufficient proxy for pedagogical quality. This extended abstract examines AI–mentor feedback alignment in the *KAIST Cyberbridge* program, a 10-week online science program for underserved gifted elementary students. The dataset includes 920 student-week observations from 92 students, with AI-generated feedback, AI next-session guidance, mentor-refined feedback, task scores, and participation records. We compared AI and mentor feedback using string similarity, token overlap, TF-IDF cosine similarity, sentence-embedding semantic similarity, and traditional lexicon-based sentiment analysis with TextBlob and VADER. Results show low surface similarity but moderate-to-high semantic similarity, suggesting that mentors often rewrote AI feedback while preserving its core meaning. TextBlob indicated a positive mentor-minus-AI shift, while VADER suggested that both AI and mentor feedback were already highly positive. Adaptation also differed by instructional group: Basic group students received feedback that was less similar to AI output and more positively reframed than Advanced group students. These findings suggest that automated feedback evaluation should move beyond surface human-likeness and consider semantic preservation, pedagogical adaptation, affective tone, learner level, and engagement-related indicators.

Keywords

Automated Feedback, Feedback Evaluation, Human-AI Collaboration, Pedagogical Adaptation, Student Engagement, Generative AI, Large Language Models

1. Introduction

Automated feedback is increasingly used in education, but its pedagogical quality cannot be evaluated only by fluency, correctness, or resemblance to human feedback. Formative feedback research emphasizes that useful feedback should be actionable, understandable, supportive, and usable by learners over time [1, 2, 3]. This is especially important for underserved learners, for whom feedback may also need to support confidence, belonging, and continued participation.

Recent work on LLM-based educational feedback has highlighted both the scalability of automated feedback and the need to evaluate it against pedagogical criteria rather than surface fluency alone [4, 5]. From a feedback design perspective, the value of feedback depends not only on what is delivered, but also on whether learners can interpret, use, and act on it over time [6, 3]. This motivates evaluation approaches that look beyond surface similarity to human feedback and examine how AI-generated feedback is adapted for pedagogical use.

This extended abstract examines a human–AI feedback workflow in the *KAIST Cyberbridge* program, an online science program for underserved gifted elementary students. In this program, AI-generated feedback was first produced from student submissions, after which human mentors could keep, modify, or replace it before delivery. This setting allows us to examine automated feedback not as a standalone

PEAF 2026: Workshop on Pedagogical Evaluation of Automated Feedback, June 28, 2026
jing.fan@aalto.fi (J. Fan); jskim315@kaist.ac.kr (J. Kim); kimia.abedini@aalto.fi (K. Abedini); daniel.koch-truhponen@aalto.fi (D. Koch-Truhponen); charles.koutcheme@aalto.fi (C. Koutcheme); juho.2.leinonen@aalto.fi (J. Leinonen)

© 2026 Copyright © 2026 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

AI output, but as a resource that human mentors adapt for pedagogical use. Our research questions for this work are:

- RQ1: How closely is mentor feedback aligned with AI-generated feedback at surface, lexical, and semantic levels?
- RQ2: How does mentor feedback differ from AI feedback in affective tone?
- RQ3: How do alignment and affective adjustment vary by instructional group and mentor, and how are they associated with engagement indicators?

2. Context and Data

We use anonymized data from the *KAIST Cyberbridge* program which integrates AI-assisted feedback with human mentoring to support underserved gifted students in science education. In this program, AI-generated feedback is first produced based on student reflection journals, after which human mentors review the AI output and may keep, modify, or replace the feedback before delivering it to students. When generating the feedback, the AI was given information about the concept students were learning that week and instructed to write a single paragraph of feedback related to the task (what the student did), process (feedback on their approach or strategy), and self-regulation based on the student's reflection journal. The system used OpenAI GPT-4 to generate the feedback. The exact prompts are not provided due to space limitations. The dataset contains 92 students across 10 instructional sessions, yielding 920 possible student–week rows. For each session, the data include student reflections, AI-generated feedback, AI next-session guidance, mentor-refined feedback, task scores, and submission records. The original wide-format dataset was reshaped into a student-week format, and AI feedback and AI next-session guidance were combined into one AI text field for comparison with mentor feedback. Across the 920 rows, 670 contained reflections, 656 contained AI feedback, 761 contained mentor feedback, and 652 contained both AI and mentor feedback. Because task scores were generally high and showed limited variability, we treated them as secondary descriptive indicators and focused on participation-based engagement indicators, including assignment submission, reflection participation, full submission, and next-week participation.

3. Method

We compared the combined AI text with mentor feedback at the student-week level. Surface alignment was measured using sequence-based string similarity and token-level Jaccard overlap. Lexical alignment was measured using TF-IDF cosine similarity [7], and semantic alignment was measured using cosine similarity between multilingual sentence embeddings generated with `paraphrase-multilingual-MiniLM-L12-v2`, following Sentence-BERT and multilingual sentence-embedding methods [8, 9]. Affective tone was analyzed using TextBlob for polarity and subjectivity [10] and VADER for compound sentiment [11]. For each sentiment measure, we computed mentor-minus-AI differences.

Because task scores showed limited variability, we summarized engagement indicators at the student, mentor, and instructional-group levels. We also compared Basic and Advanced groups to examine whether feedback adaptation varied by learner level. To examine mentor-level differences, we used Kruskal–Wallis tests [12] for similarity, sentiment-adjustment, and participation indicators because these variables were not assumed to be normally distributed across mentors. To explore associations between feedback indicators and student engagement, we conducted two sets of exploratory correlation analyses. At the student level, we computed Pearson and Spearman correlations between mean feedback indicators and engagement outcomes, including submission count, reflection count, submission consistency, and reflection consistency. At the row level, we examined whether greater feedback modification, operationalized as lower similarity scores or as $1-$ similarity, was associated with next-week reflection participation. These analyses were exploratory, and statistical significance was evaluated using an alpha threshold of .05.

4. Findings

Table 1
Main quantitative patterns by instructional group. Sentiment values are mentor-minus-AI differences; similarity and sentiment values are computed only for student-week rows where both the combined AI text and mentor feedback were available.

Measure	Overall	Basic	Advanced
String similarity	.149	.078	.217
TF-IDF cosine	.394	.227	.553
Semantic similarity	.690	.561	.813
TextBlob diff.	.089	.161	.020
VADER diff.	.003	.013	-.006
Submission rate	.903	.865	.947
Reflection rate	.728	.684	.779

Table 2
Exploratory student-level association between affective adjustment and reflection engagement.

Association	Coef.	p
Pearson: VADER diff. vs. reflection count	.243	.021
Spearman: VADER diff. vs. reflection count	.248	.018

Low surface similarity but moderate semantic similarity. Across the 652 student-week rows with both combined AI text and mentor feedback, mentor feedback showed low surface similarity to AI feedback: mean string similarity was 0.149, token Jaccard similarity was 0.435, and TF-IDF cosine similarity was 0.394. However, mean sentence-embedding semantic similarity was 0.690. This contrast suggests that mentors often changed the wording substantially while preserving the core meaning of AI-generated feedback.

Affective tone showed mixed but informative patterns. TextBlob indicated that mentor feedback was more positive than AI feedback: mean AI polarity was 0.185, mean mentor polarity was 0.277, and the mentor-minus-AI difference was 0.089; this paired difference was statistically significant ($p < .001$). VADER produced a more conservative pattern: both AI and mentor feedback had high compound scores, and the paired mentor-minus-AI difference was only 0.003. This suggests a sentiment ceiling, where AI feedback was already highly positive and mentor adaptation may have involved subtle reframing rather than a simple shift from negative to positive.

Adaptation varied by learner level and mentor. As shown in Table 1, Basic group feedback had lower semantic similarity to AI output than Advanced group feedback (0.561 vs. 0.813) and stronger positive TextBlob adjustment (0.161 vs. 0.020). This suggests that mentors adapted AI feedback more strongly for Basic group students, whereas Advanced group feedback remained closer to the AI-generated version. Kruskal–Wallis tests indicated mentor-level differences in string similarity, TF-IDF similarity, semantic similarity, TextBlob sentiment adjustment, VADER sentiment adjustment, submission participation, and reflection participation (all $p < .001$, $\alpha = .05$). This suggests that human intervention varied by mentoring strategy.

Engagement associations were exploratory. Task scores were high on average (mean = 95.16), suggesting a ceiling effect. Participation indicators were more informative: the overall submission rate was 90.3%, the reflection rate was 72.8%, and the full-submission rate was 68.5%. Student-level association analyses showed limited evidence that feedback indicators were related to engagement. The

clearest pattern was a weak positive association between mean VADER mentor-minus-AI sentiment difference and reflection count (Pearson $r = .243$, $p = .021$; Spearman $\rho = .248$, $p = .018$; Table 2).

Row-level analyses of next-week reflection participation showed only weak and inconsistent associations with the extent of AI-feedback modification. When modification was operationalized as 1−similarity, next-week reflection was weakly associated with TF-IDF-based modification extent (Pearson $r = .082$, $p = .046$), but this association was not robust in the Spearman correlation ($\rho = .061$, $p = .135$). Modification extent based on string similarity (Pearson $r = .056$, $p = .170$) and semantic similarity (Pearson $r = -.049$, $p = .236$) was not significantly associated with next-week reflection. Overall, the results do not support a simple claim that higher AI–mentor similarity or greater modification predicts better engagement outcomes.

5. Discussion

These findings suggest that human-likeness is a multi-level construct. Low surface similarity should not be interpreted as evidence that AI feedback was irrelevant or unusable: mentors may preserve the instructional meaning of AI feedback while changing wording, tone, emphasis, or readability. Automated feedback evaluation should therefore distinguish surface resemblance from semantic preservation and pedagogical adaptability.

The results also suggest that feedback evaluation should account for learner context. Basic group students received feedback that was less similar to AI output and more positively adjusted, whereas Advanced group feedback remained closer to the AI-generated version. Engagement analyses were exploratory: reflection participation showed the clearest signal, while next-week participation showed only weak and inconsistent associations with feedback modification. Overall, these findings suggest that evaluation frameworks should consider at least three dimensions: semantic preservation, pedagogical adaptation, and learner engagement.

6. Limitations and Future Work

This study is observational and should not be interpreted causally. Task scores showed limited variability, which restricted their usefulness as learning outcome measures. TextBlob and VADER provide coarse sentiment indicators and may not capture culturally specific encouragement, rapport, or pedagogical warmth. Future work should examine representative feedback examples in depth and develop richer outcome measures such as reflection quality, feedback uptake, help-seeking, and sustained participation.

7. Conclusion

This extended abstract shows that AI–mentor feedback alignment is not well captured by surface human-likeness alone. Mentor feedback was lexically different from AI feedback but moderately aligned at the semantic level, suggesting that mentors used AI feedback as a resource for pedagogical adaptation rather than as a final message. For automated feedback systems, the key evaluation question is not only whether AI feedback resembles human feedback, but whether it can support meaningful human adaptation and learner-centered use.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT for language polishing. The authors reviewed and edited the content as needed and take full responsibility for the publication's content. The study also analyzes AI-generated feedback produced by GPT-4 as part of the educational feedback workflow described in the paper.

References

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

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

[3] D. Carless, D. Boud, The development of student feedback literacy: Enabling uptake of feedback, Assessment & Evaluation in Higher Education 43 (2018) 1315–1325. doi:10.1080/02602938.2018.1463354.

[4] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, G. Kasneci, Chatgpt for good? on opportunities and challenges of large language models for education, Learning and Individual Differences 103 (2023) 102274. doi:10.1016/j.lindif.2023.102274.

[5] J. Steiss, T. Tate, S. Graham, J. Cruz, M. Hebert, J. Wang, Y. Moon, W. Tseng, M. Warschauer, C. B. Olson, Comparing the quality of human and chatgpt feedback on students' writing, Learning and Instruction 91 (2024) 101894. doi:10.1016/j.learninstruc.2024.101894.

[6] D. Boud, E. Molloy, Rethinking models of feedback for learning: The challenge of design, Assessment & Evaluation in Higher Education 38 (2013) 698–712. doi:10.1080/02602938.2012.691462.

[7] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information Processing & Management 24 (1988) 513–523. doi:10.1016/0306-4573(88)90021-0.

[8] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410/. doi:10.18653/v1/D19-1410.

[9] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 4512–4525. URL: https://aclanthology.org/2020.emnlp-main.365/. doi:10.18653/v1/2020.emnlp-main.365.

[10] S. Loria, Textblob: Simplified text processing, https://textblob.readthedocs.io/, 2025. Python library documentation, accessed 2026-05-15.

[11] C. J. Hutto, E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, Proceedings of the International AAAI Conference on Web and Social Media 8 (2014) 216–225. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14550. doi:10.1609/icwsm.v8i1.14550.

[12] W. H. Kruskal, W. A. Wallis, Use of ranks in one-criterion variance analysis, Journal of the American Statistical Association 47 (1952) 583–621. doi:10.1080/01621459.1952.10483441.