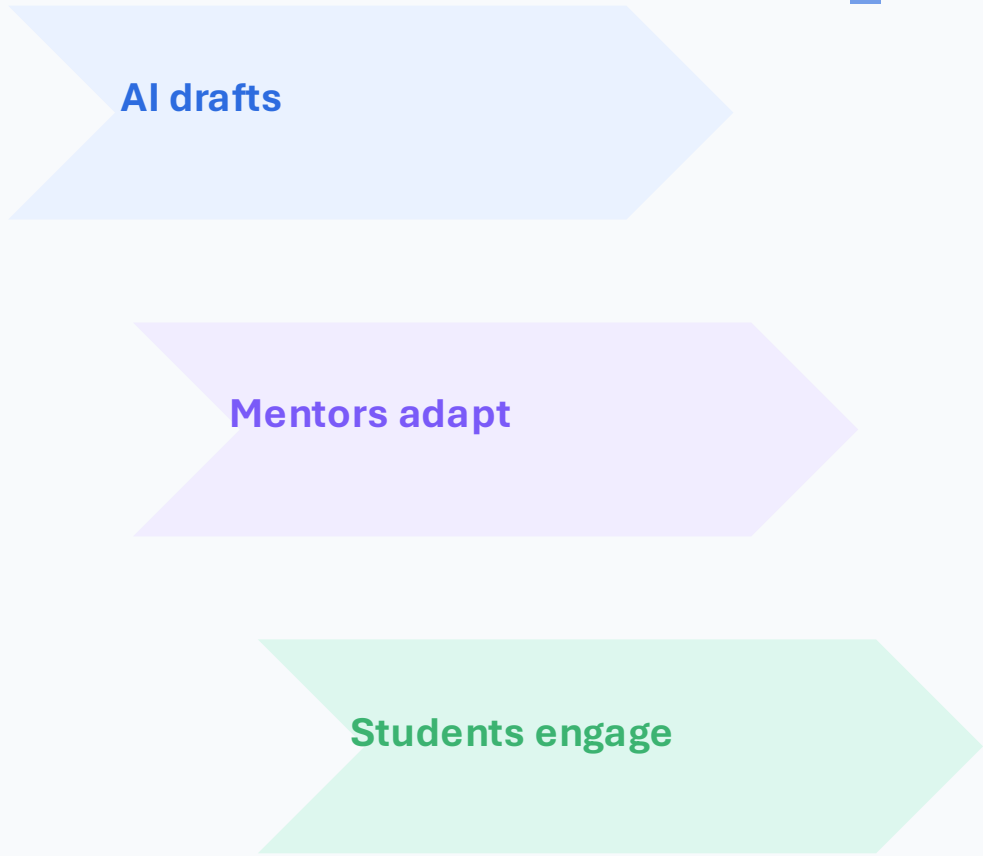
Beyond Surface Human-Likeness

AI–mentor feedback alignment, pedagogical adaptation, and student engagement in longitudinal learning data

Presentation deck

PEAF 2026



Core question: does human-likeness really mean better feedback?

The study moves from surface resemblance toward semantic preservation, pedagogical adaptation, and student engagement.

Automated feedback is often compared with human feedback, but “looking human-written” is not the same as being useful for learners.

Mentors may substantially rewrite wording, tone, and readability while preserving the instructional core of AI feedback.

Evaluation should account for learner context: Basic and Advanced groups may require different degrees of adaptation.

Research questions

RQ1: How closely is mentor feedback aligned with AI 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?

One-sentence takeaway

Do not only ask whether AI resembles mentors; ask whether AI can become a high-quality resource for mentor adaptation.

Study context and data: a 10-week online science program

The program served underserved gifted elementary students; AI drafted feedback, and mentors reviewed, modified, or replaced it.

92
students

10
sessions

920
student-week rows

652
rows with both AI + mentor feedback

Available field	Rows
Reflections	670
AI feedback	656
Mentor feedback	761
Both AI + mentor	652

Main unit of analysis: student-week. Since task scores were generally high, participation indicators were more informative than scores.

Method: unpacking “similarity” into four levels

Within each student-week row, the combined AI text was compared with mentor feedback.

Surface similarity

String similarity

Was the wording reused?

Token overlap

Jaccard overlap

Were similar word chunks used?

Lexical space

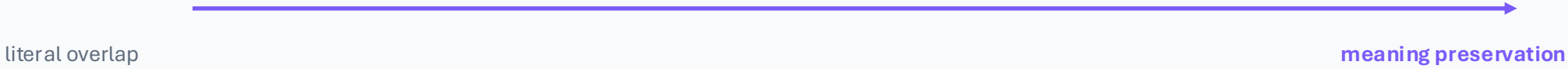
TF-IDF cosine

Were similar keyword structures preserved?

Semantic space

Sentence-embedding cosine

Was the meaning preserved?

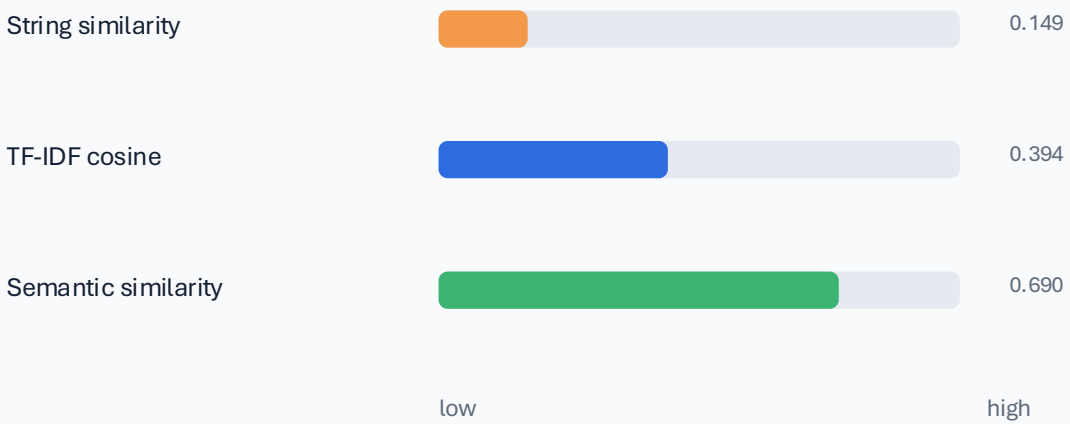


Additional analyses: TextBlob and VADER mentor-minus-AI sentiment differences; Basic vs. Advanced and mentor-level comparisons; exploratory correlations with engagement.

Finding 1: low surface resemblance, moderate semantic alignment

Mentors often rewrote the expression rather than simply copying AI output.

Overall similarity means



Key interpretation: low surface similarity does not mean AI feedback was irrelevant. Mentors may preserve instructional meaning while improving tone, emphasis, and readability.

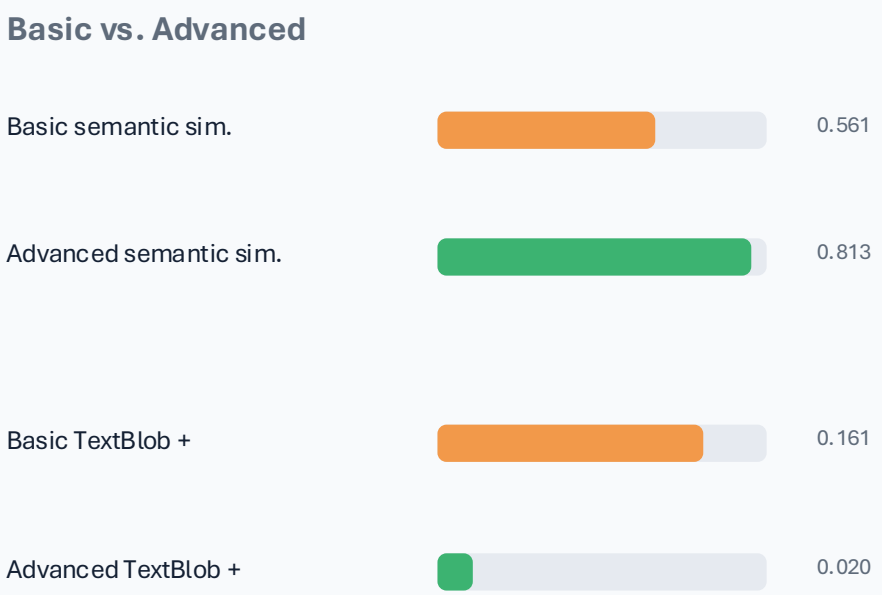
Evaluation implication

Do not collapse human-likeness into one score. Separate wording changes, meaning preservation, and pedagogical adaptation.

Finding 2: mentors adapted more strongly for the Basic group

Basic group feedback moved farther from the AI wording and was more positively reframed; Advanced feedback stayed closer to AI output.

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
Submission rate	.903	.865	.947
Reflection rate	.728	.684	.779



Conclusion: adaptation was not random rewriting; it appeared linked to learner level and mentoring strategy.

Finding 3: engagement signals exist, but causality is not supported

Task scores showed a ceiling effect; participation behavior was more informative.

95.16
mean task score; limited variation

90.3%
submission rate

72.8%
reflection rate

68.5%
full-submission rate

Clearest exploratory association

VADER mentor-minus-AI sentiment difference was weakly positively associated with reflection count.

Pearson r = .243, p = .021
Spearman ρ = .248, p = .018

Simple conclusion not supported

Neither "more AI-like is better" nor "more modification is better" received robust support.

Interpretation: feedback quality may emerge through affective support, readability, actionability, and sustained participation over time.

From human-likeness to an adaptable-feedback evaluation framework

Treat AI feedback as a resource that mentors can edit, transform, and use to support students.

Semantic preservation

Does the mentor still use the instructional core of the AI feedback?

Pedagogical adaptation

Can mentors adjust tone, emphasis, and readability for learner level?

Affective tone

Does feedback support confidence, belonging, and continued participation?

Engagement indicators

Do reflection, submission, and next-week participation provide observable signals?

Evaluation question: not “Can AI imitate human feedback?” but “Can AI support meaningful human adaptation and learner-centered use?”

Conclusion and next steps

The study supports a more fine-grained view of automated feedback evaluation.

Three conclusions

Mentor feedback was lexically different from AI feedback but moderately aligned at the semantic level.
Basic group students received stronger rewriting and more positive reframing, suggesting learner-level adaptation.
Engagement results were exploratory: affective adjustment was weakly associated with reflection count, but causal claims are not warranted.

Future work

Analyze representative feedback examples in depth.
Develop richer outcomes: reflection quality, feedback uptake, help-seeking, and sustained participation.
Move beyond TextBlob and VADER to capture encouragement, rapport, cultural context, and pedagogical warmth.

Final message: evaluate not only whether AI “sounds human,” but whether it helps mentors make better pedagogical adaptations.