Evaluating the Pedagogical Quality of LLM-Generated Feedback

A criterion-based and comparative study

Harvey Ngoe Kolle presenting
with Carrie Demmans Epp · Amna Liaqat · Maria Cutumisu

University of Alberta · Alberta Machine Intelligence Institute (Amii) · George Mason University · McGill University · Mila, Quebec Artificial Intelligence Institute

Most evaluations of automated feedback rest on a single holistic rating.

1

SCORE = QUALITY?

One number. One verdict. A brief survey at best, for something as layered as teaching a student how to write.

02 · THE QUESTION

Is one overall rating enough to judge feedback quality?

Formative-feedback literature says quality is several things at once:

trustworthy · goal-tied · specific · actionable · supportive

A 14-item rating instrument, grounded in theory.

14
rating items,
five-point Likert

10
theory-grounded
dimensions

14-70
composite score
ω = 0.92

01 Specific & Clear
02 Goal-Referenced
03 Actionable
04 Supportive Tone
05 Trustworthiness
06 Objective
07 Non-Threatening
08 Non-Revealing
09 Conciseness
10 Overall

Three masked sources of feedback.

HUMAN

Real instructor feedback

Rubric-based comments written in an actual LINC writing classroom.

SINGLE-AGENT

GPT-4.1, one prompt

Feedback produced in a single generation step with few-shot prompting.

MULTI-AGENT

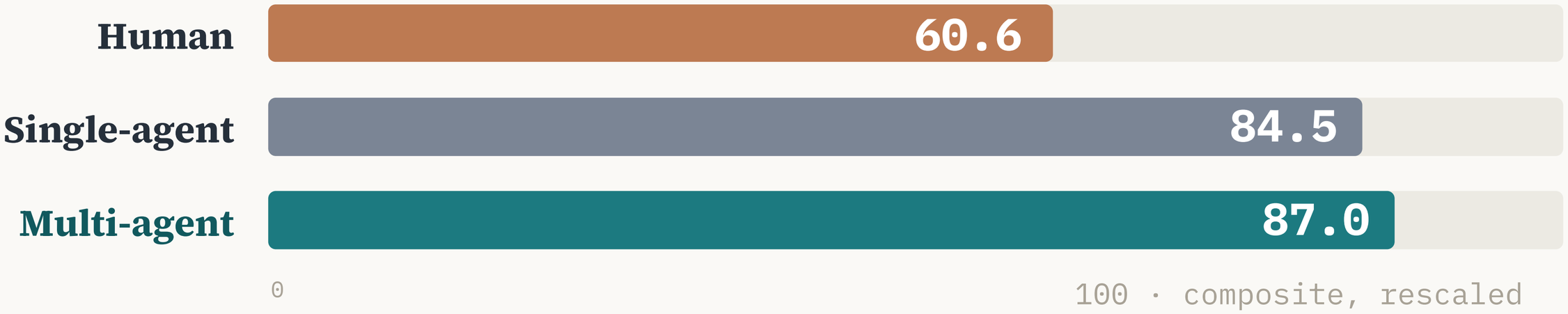
Six specialised agents

verify → correct → subject → pedagogy → tone → format, each refining one quality.

18 raters (instructors and pre-service teachers), masked to the source, rated and ranked 54 feedback instances.

05 · THE SCORES

Both automated conditions beat human feedback by 20+ points.



Rankings caught what the scores missed.

BY COMPOSITE SCORE

Multi ≈ Single

A 2.5-point gap, not statistically significant. One number says they are approximately the same.

BY RANKING & DIMENSION

65.7% of head-to-head matchups favored multi-agent (p = .015)

Supportive Tone

the **one** dimension where multi beat single

p = .049

07 · THE TAKEAWAY

A single overall rating would have hidden this difference.

Multiple theory-grounded dimensions, paired with preference rankings, surface real distinctions in feedback quality that one number cannot.

08 · A CAVEAT

Read the human comparison with care.

The instructor feedback came from a real classroom; it was never written for this study. The 20-point gap reflects different contexts and purposes , not just human versus machine.

Ground evaluation in theory, and the picture gets clearer.

Other domains

Do these dimensions hold for coding and mathematics feedback?

Computable metrics

Turn individual items into automatic, scalable measures.

Student outcomes

Do rated-quality judgments predict actual learning?

Thank you.

Multidimensional evaluation surfaces what a single rating hides.

CONTACT

Harvey Ngoe Kolle

kolle@ualberta.ca

SUPPLEMENTARY

Full study design

bit.ly/4bPmBcF

SUPPLEMENTARY · THE FULL INSTRUMENT

All 14 items across 10 dimensions.

[R] reverse-scored

01 · Specific & Clear

This feedback targets specific parts of the writing.
This feedback is clearly written.

02 · Goal-Referenced

This feedback explains how the writing meets the expected goals.

03 · Actionable

This feedback offers clear guidance the student can follow.
This feedback supports improvement in writing skills.

04 · Supportive Tone

This feedback maintains a positive tone.
This feedback is discouraging. [R]

05 · Trustworthiness

This feedback is inaccurate. [R]

06 · Objective

This feedback is based on evidence from the student's writing.
This feedback focuses on student traits. [R]

07 · Non-Threatening

This feedback uses harsh language. [R]

08 · Non-Revealing

This feedback suggests alternative phrasing.

09 · Conciseness

This feedback avoids overly detailed elaboration.

10 · Overall

This feedback would be appropriate to share directly with a student.