

How Well Does AI-Generated Feedback Work? Intrinsic and Extrinsic Evaluation across more than 20,000 EFL Essay Drafts

Steven Coyne^{1, 2} Diana Galvan-Sosa³ Ryan Spring¹ Machi Shimmei^{1, 2}
Michael Zock⁴ Keisuke Sakaguchi^{1, 2} Kentaro Inui^{5, 1, 2}

¹Tohoku University ²RIKEN ³ALTA Institute, Computer Laboratory, University of Cambridge
⁴CNRS, LIS, Aix-Marseille University ⁵MBZUAI



Overview

- We deploy written corrective feedback (WCF) systems in a large Academic English class
- We compare **expert rubric scoring (intrinsic eval)** and **student interaction data (extrinsic eval)**
- We find **no significant overlap** between the chosen metrics
- These measures represent **complementary perspectives of feedback quality**

WCF Approaches

One error highlight per WCF comment

Grammar Advice #4 / 13

If **all country** that have the sea can do that too then we can use tidal power everywhere and help save the earth.

What's wrong: The word 'country' is singular, but the context suggests you are talking about all countries, which requires the plural form.

What to do: Change 'country' to 'countries' to match the plural context.

Was this advice useful? ☐ Yes ☐ No

“What’s wrong” and “What to do”

Student Feedback



Shared Step:
Grammatical Error
Correction (GEC)

Source:

I likes the BTS.

Corrected:

I like $\emptyset$ BTS.

Template System

ERRANT Edits

likes $\rightarrow$ like the $\rightarrow$ $\emptyset$

R:VERB:SVA U:DET

Fill Templates

SVA WCF

Unnecessary
Article WCF

“Simple” LLM System

“Compare the source and corrected sentences and write WCF for each error.”



Write Feedback

SVA WCF

Unnecessary
Article WCF

Atomic Edit System

ERRANT Edits

R:VERB:SVA U:DET



Refine and Classify

SVA

Article: Proper Noun



Write Feedback

SVA WCF

Unnecessary
Article WCF

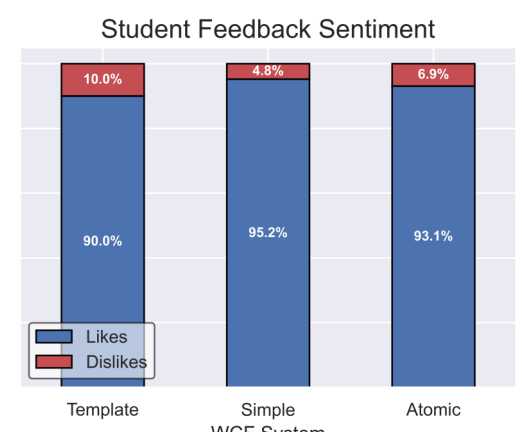
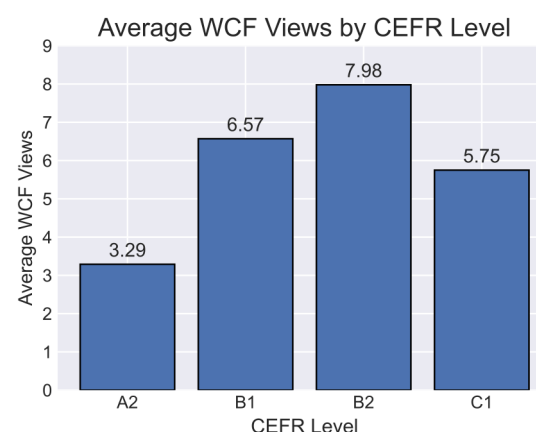
Intrinsic: Teacher Ratings

- Raters: EFL teachers with 9+ years of experience
- Applied the rubrics from Coyne et al. (2025)
- LLM Systems outperformed the template system

WCF Rating Criterion	IAA	Tmp.	Smpl.	Atm.
Correction is valid	0.70	0.92	0.98	0.94
Relevant to error	0.30	0.99	0.99	1.00
Factually correct	0.65	0.92	0.96	0.92
Has what's wrong & why	0.61	0.28	0.99	0.98
Has what to do	-0.004	0.99	1.00	1.00
Is Comprehensible	0.04	0.98	0.91	0.92
Has out-of-scope (↓)	0.22	0.01	0.006	0.005
Has direct correction	0.40	0.84	1.00	1.00
Highlight Quality	0.25	4.54	4.38	4.64
Overall Quality	0.45	3.14	4.51	4.36

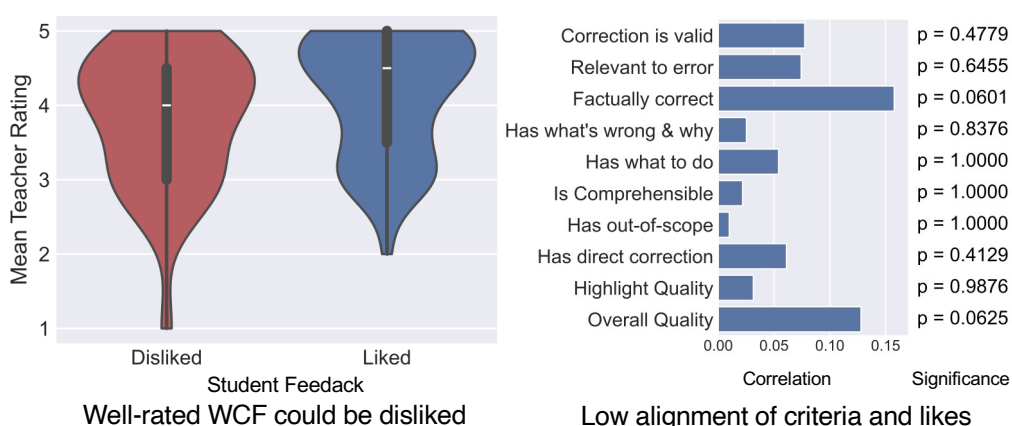
Extrinsic: Student Interactions

- Half of students (49.76%) never viewed WCF
- View rates vary by CEFR level, increasing until C1
- Student feedback is sparse (3% of views) and mostly likes



Findings: No Convergence

- We found **no significant relationships** between teacher-based and student-based metrics



Discussion

- These are complementary perspectives of WCF quality
- Students prioritize **immediate usability**
- Teachers prioritize **factual validity** and **careful phrasing**

Next Steps

- Investigate areas with low teacher agreement**
Annotation flaws vs. multiple pedagogical perspectives
- Inter-draft revision analysis**
What WCF factors lead to a successful revision?