AI-Assisted Evaluation of Revision Patterns in Young Students' Argument Writing

Tianwen Li[1], **Lindsay Clare Matsumura**[1], **Elaine L. Wang**[2], **Diane Litman**[1,3], **Zhexiong Liu**[3], **Richard Correnti**[1], **Michelle Hong**[1]

[1]Learning Research and Development Center, University of Pittsburgh [2]RAND Corporation [3]Department of Computer Science, University of Pittsburgh

Background

- Young students struggle with writing and revision.
- Teachers feel underprepared to teach writing and revising.
- They feel unsure how to provide effective feedback on student writing.

How formative assessment of revision can support teachers' writing instruction:

- Builds teachers' content knowledge by making the criteria for effective revision explicit.
- Provides information about students' revision quality to guide instruction and provide targeted feedback.

LLM & Writing Assessment

- Large Language Models (LLMs) show promise in writing evaluation
- Zero-shot and few-shot prompting
- **Chain-of-Thought (CoT)** prompting → Mixed results

Research Questions

RQ1 How **internally consistent** is GPT-4.1 in assessing revision patterns in students' writing across different prompting strategies?

- Each prompt was run three times
- Exact agreement and Quadratic Weighted Kappa (QWK)

RQ2 How **accurate** is GPT-4.1 in assessing revision patterns in students' writing across different prompting strategies?

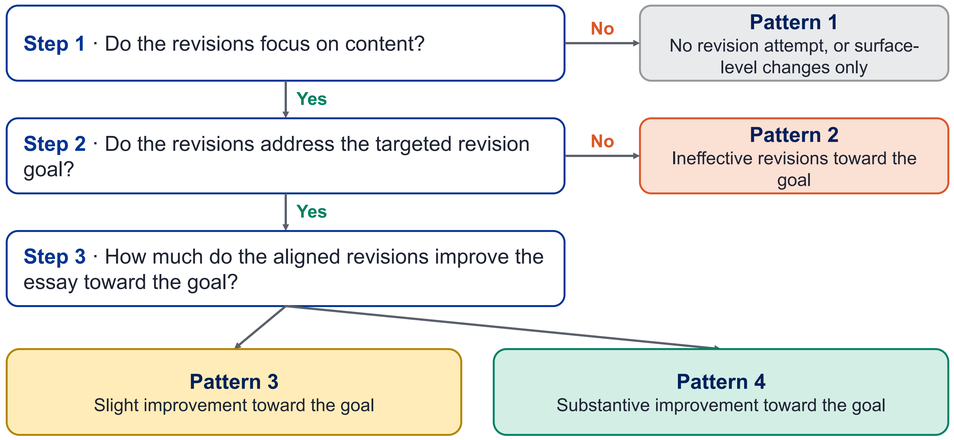
- Majority-vote approach
- Exact agreement and Quadratic Weighted Kappa (QWK)

Study Context & Data

- **eRevise+RF**: text-based argument writing
- **330 essay pairs** (first and revised drafts) from Grades 4–8 in six teachers' classrooms.
- After draft 1, each student got feedback on **one revision goal**:

134	**Goal 1 · Evidence** Add more evidence from the article
59	**Goal 2 · Details** Add text-based detail to existing evidence
137	**Goal 3 · Explanations** Explain how evidence supports the claim

Taxonomy of Revision Patterns



Patterns form an ordinal scale and manifest differently by goal (adapted from Wang et al., 2020).

Human Coding Benchmark

- All essay pairs were double-coded by two human coders.

Revision goal	Exact agreement	QWK
Goal 1 · Evidence	87.50%	0.87
Goal 2 · Details	84.91%	0.91
Goal 3 · Explanations	**79.67%**	**0.77**

Prompt Design · GPT-4.1

Both prompts begin with the task description (Using Goal 1 example):
You are a writing teacher who works with students from 4th grade to 8th grade. Students are working toward the revision goal of adding additional evidence from different parts of the text.

Few-shot Prompt (Baseline)

Below are four revision patterns; select the one that best represents the student's revisions:

1) Make no revision attempt or make surface-level revisions **(Definition + Example)**
2) Make ineffective revisions toward the targeted revision goal **(Definition + Example)**
3) Make revisions with slight improvement toward the targeted revision goal **(Definition + Example)**
4) Make revisions with substantive improvement toward the targeted revision goal **(Definition + Example)**

Few-shot CoT Prompt

To determine the appropriate revision pattern, follow these steps:

Step 1: Does the revision focus on content? If no, output Pattern 1; if yes, proceed to Step 2. **(Example of Pattern 1)**

Step 2: Does the revision effectively address the targeted revision goal? If no, output Pattern 2; if yes, proceed to Step 3. **(Example of Pattern 2)**

Step 3: To what extent does the revision improve essay quality with respect to the targeted revision goal? Pattern 3 or Pattern 4. **(Examples of Patterns 3 & 4)**

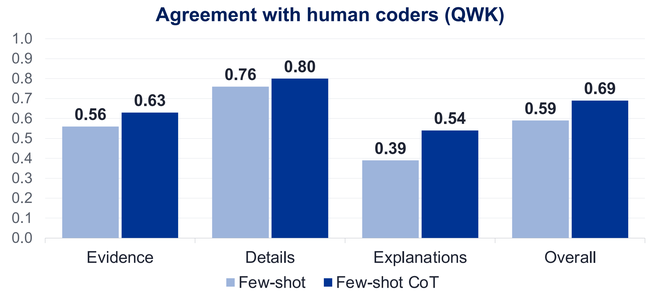
Separate, goal-specific prompts were written for each revision goal (GPT-4.1, temperature = 0).

RQ1 · Internal Consistency

ICC ≥ .94 Excellent intra-rater reliability for every goal under both prompts

Goal	Baseline		CoT	
	Exact	ICC	Exact	ICC
Evidence	90.30%	0.94	90.30%	0.94
Details	89.83%	0.96	88.14%	0.94
Explanations	89.78%	0.94	92.70%	0.96

RQ2 · Accuracy



- CoT lifted overall QWK from **0.59 to 0.69** (moderate → substantial).
- Biggest gain on **explanations**: exact agreement 40.9% → 52.6%, weighted F1 0.40 → 0.51.
- Driven by spotting **ineffective revisions** (Pattern 2 F1: 0.34 → 0.63).
- On **details**, exact agreement fell (67.8% → 61.0%): extra steps may overcomplicate simpler judgments.

Conclusions & Next Steps

- GPT-4.1 shows strong potential for **formative assessment of revision**: highly consistent and moderately to substantially aligned with human coders.
- Making evaluation steps explicit helps most when judgments are **subjective and complex**.
- Stepwise rubrics can make scoring more transparent for both LLMs and human coders.
- **Next:** test validity, i.e., how teachers use revision patterns to guide instruction.

Li et al. (2026). Frontiers in Education, 11:1736489. doi:10.3389/feduc.2026.1736489
Supported by NSF Award #2202347. Contact: Tianwen.li@pitt.edu