An Actionability Gradient in LLM-Based Writing Feedback: Evidence from 196K Student Drafts in K–12 Argumentative Writing

Liujing Ren[1][0009−0003−1825−3612], Emily LeBlanc[1][0000−0003−1447−412X], Boxing Li[1][0009−0001−3635−0112], and Lorenzo Motta[1][0009−0001−1788−3052]

ThinkCERCA, Chicago IL 60611, USA

Abstract. Large language model (LLM)-based writing feedback is increasingly deployed in K–12 classrooms, but whether AI feedback is equally actionable across argumentation components remains untested. We analyze 196,169 consecutive draft pairs from 58,332 student assignments on a real-world GPT-4 writing platform, operationalizing actionability as the component-level rate at which the system's rating improves between consecutive drafts, with teachers endorsing these ratings at moderate-to-high rates (63–95%). We identify an *actionability gradient*: rating-improvement rates vary fivefold across six components (51.1% for evidence vs. 10.2% for audience appeal) and remain stable across grade levels. The gradient is consistent with differences in the cognitive demands of revision tasks across components. Because improvement is measured through system rating changes rather than independent assessments of writing quality, we treat it as a proxy for actionability rather than evidence of learning gains. These findings provide a data-driven basis for progress-aware feedback design and targeted teacher support by component, identifying where AI feedback is most readily actionable and where human intervention is likely required.

Keywords: Automated writing evaluation · LLM feedback · Argumentation · K–12 · Actionability · Learning analytics · Feedback design

Keywords: Automated writing evaluation · LLM feedback · Argumentation · K–12 · Actionability · Learning analytics · Feedback design

1 Introduction

LLM-based writing feedback systems are increasingly deployed in K–12 classrooms to deliver individualized feedback at scale and reduce teacher workload [14,21,22]. Yet feedback is rarely differentiated across the argumentation components, implicitly treating all components as equally actionable. We define *actionability* as the degree to which a feedback comment translates into a clear, executable revision path. For example, adding a supporting sentence in response to evidence feedback is a discrete, local operation while reconsidering

rhetorical stance for audience appeal lacks an obvious local revision target. If actionability varies across components, AI feedback may redistribute rather than reduce teacher workload—concentrating demand on components where students are least likely to improve autonomously.

We examine this question using 196,169 draft pairs from 58,332 student-lessons on a GPT-4-based argumentation platform (grades 3–12), operationalizing actionability as the component-level improvement rate between drafts. We identify what we term an *actionability gradient*: a fivefold variation across components (51.1% for evidence vs. 10.2% for audience appeal) that is consistently stable across all grade levels. Counterargument is the sole exception, rising from the bottom ranks at grades 3–5 to mid-range by grade 7, potentially reflecting the developmental emergence of argumentation skills.

These findings suggest component-level actionability as a principled basis for differentiating interface design and teacher attention—shifting the effective unit of feedback system design from the essay to the individual component. Where the gradient is high—evidence, reasoning—AI feedback appears most readily actionable for autonomous revision and trajectory monitoring is the priority. Where it is low—audience appeal and early-grade counterargument—interfaces need to scaffold revision more actively and teachers need to intervene directly. Prior design decisions have largely relied on intuition or small-scale studies and the actionability gradient grounds them in large-scale naturalistic data.

2 Related Work

AWE and LLM feedback Automated and LLM-generated feedback improves student writing across settings [14,10,7], but effects are concentrated on surface features with limited impact on higher-order argumentation [19,13]. AWE systems such as Writing Pal [17] and MI Write [15] have extended feedback beyond surface corrections, but their dimensions—such as organization, word choice, and style—apply across writing genres rather than targeting argumentation-specific components such as claim, evidence, and counterargument.

Argumentation component-level feedback K–12 AWE research has begun distinguishing argumentation components, primarily through the eRevise line, starting with evidence feedback [23,20,8] and expanding to reasoning [12], with annotation [1] and interface work [2] confirming that component-level distinctions improve revision outcomes. Key components of argumentation instruction such as claim and counterargument remain largely unexamined, and cross-component comparisons of improvement rates in naturalistic K–12 deployments are absent.

Theoretical accounts predict a gradient, untested at component level. Students consistently default to surface-level revision over global restructuring, as the latter involves more cognitive demands with less obvious local targets [18,9,11,5]. Bereiter and Scardamalia [4] attribute this to the distinction

between *knowledge-telling*—retrieving and inserting content—and *knowledge-transforming*—restructuring an argument for a reader. For argumentation feedback, this suggest that components feedback with clear local revision targets will be more actionable than those requiring global rhetorical restructuring. However, it remains unclear whether this results in a measurable gradient across different argument components in real K–12 settings.

3 Data and Platform

The study uses an on-demand GPT-4-powered writing platform deployed in K–12 classrooms to evaluate student argumentative writing in six components: claim, evidence, reasoning, counterargument, coherence, and audience appeal. The platform prompts students to reflect on their writing at the component level rather than directing them to repair specific sentences.The platform assigns each component an internal rating of Proficient, Developing, or Not Yet, presented to students as an encouraging title paired with a lesson-specific feedback comment rather than a numeric score—For example, a Developing rating on Reasoning might be accompanied by: "Your reasoning is present but could be clearer—try explaining how each piece of evidence supports your claim. For example, how does the Winnie-the-Pooh socks incident show unfairness?". Students may scan a new draft after making at least one character change. We define a *draft pair* as two consecutive drafts by the same student within a lesson.

Data were drawn from platform logs collected from March 31, 2025 through April 14, 2026. We kept student assignments with at least two consecutive drafts, resulting in 58,332 assignments and 196,169 consecutive draft pairs (See Table 1). Responses from narrative writing assignments using a different rubric structure were excluded (734 draft pairs). Grade-level data, obtained via platform assignment metadata, were available for 98.4% of pairs from grades 3–12 with a primary concentration of grades 6–8.

Table 1. Dataset statistics.

Metric	Value
Student-assignment ($\geq$2 drafts)	58,332
Consecutive draft pairs	196,169
Average drafts per student-lesson	4.34
Pairs with grade data	193,073 (98.4%)
Grade range	3–12

Teacher Validation of AI Ratings. To assess the perceived correctness of LLM-generated ratings, five teachers reviewed 170 randomly sampled responses across ten lessons from grades 3 to 12. For each response, teachers were shown the system's rating and feedback and asked whether they endorsed the rating, whether the identified span was accurate, and whether the feedback was helpful. The endorsement of the system's ratings ranged from 63% to 95% across

components, and was consistently higher for Not Yet ratings (90–95%) than for Proficient ratings (61–85%). Over 90% of feedback comments were rated as helpful or somewhat helpful across all components. Because teachers reviewed the system's rating before responding, we report these as endorsement rates rather than independent agreement (see Section 7).

4 Method

For each consecutive draft pair, we measure whether the LLM's rating for each component improved. Components already rated Proficient in the prior draft are excluded as Already-Proficient cases, since no upward rating change is possible and inclusion would introduce ceiling bias. We analyze improvement rates per component and whether each student improved at least once within their assignment. We additionally compute regression rates for Already-Proficient and non-Already-Proficient cases, and co-regression rates defined as the proportion of draft pairs in which two components simultaneously decline from Proficient. Together, these measures capture both upward and downward dynamics of component-level learning.

Primary Measure. We use component-level rating change as the primary improvement measure. Because the platform prompts reflection at the component level, we treat rating change as the primary measure of improvement rather than local text proximity. To validate this, we conducted a cross-check using fuzzy string matching on feedback-annotated sentences: off-target improvement (27.5%) substantially exceeded on-target improvement (6.2%), suggesting that rating change is a more appropriate proxy for component-level improvement than local text proximity in this system.

5 Results

Improvement Analysis Improvement rates vary fivefold across six components (Table 2). Evidence has the highest improvement rate (51.1%) and the highest Already-Proficient rate (84.5%), while audience appeal has the lowest improvement rate (10.2%) and the lowest Already-Proficient rate (19.2%). Across 58,332 student assignments, 61.7% included at least one improvement event, while 38.3% showed no improvement across any draft pair.

Of the 58,332 student-assignment instances, 61.7% included at least one improvement event while 38.3% showed no improvement across any draft pair.

Grade-Level Stability The component ordering holds across grades 3–12 (Fig. 1), with evidence consistently ranking first and audience appeal consistently ranking last. Overall improvement rates increase monotonically with grade from 32.2% at grade 3 to 46.3% at grade 12, and the evidence-to-audience-appeal ratio narrows from 6.6× to 3.9× . Counterargument is the sole component whose

Table 2. Component improvement rates (Already-Proficient excluded per component).

Component	Already-Proficient	Improvement%	n
Evidence	84.5%	**51.1%**	30,499
Reasoning	35.2%	**22.6%**	127,019
Claim	53.0%	**19.0%**	92,197
Counterargument	29.6%	**18.0%**	71,626
Coherence	31.4%	**15.9%**	119,879
Audience Appeal	19.2%	**10.2%**	141,044

rank changes substantially across grades, rising from ranks 5–6 at grades 3–5 to ranks 3–4 by grades 7–8, a pattern we return to in the Discussion.

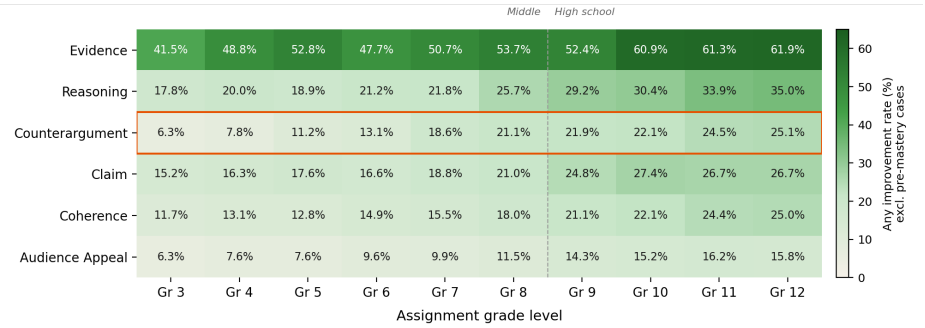

Fig. 1. Component improvement rates by grade (Already-Proficient excluded). Counterargument is the only component that changes rank substantially across grades.

Regression and Co-regression 20.9% of draft pairs show that at least one component rating decreased from Proficient. Among all component ratings, 9.7% Proficient-before ratings regressed, which is more than three times the rate of Developing-before ratings (2.8%). Co-regression is most pronounced between coherence and audience appeal (35%), coherence and evidence (28%), and coherence and reasoning (27%).

6 Discussion

6.1 The Actionability Gradient

The fivefold improvement rate gradient is consistent with Bereiter and Scardamalia's [4] distinction between *knowledge-telling* and *knowledge-transforming*. Feedback that requires a small, specific change, like adding a supporting sentence, maps onto knowledge-telling and imposes lower cognitive demands. In

contrast, feedback asking students to rethink their overall argument or logical flow maps onto knowledge-transforming, leaving no obvious local revision target even when the feedback is understood.

The data provide evidence against two alternative explanations. If the gradient is driven primarily by feedback quality, the audience appeal which received the highest teacher rated helpfulness (92%) should show the highest improvement rate, but it shows the lowest. This suggests that perceived feedback quality does not directly translate into actionable revision outcomes. If the gradient is driven primarily by student maturity, younger students should struggle more with abstract components and older students should close the gap. However, the actionability gradient holds from grade 3 through grade 12, with the evidence-to-audience-appeal ratio narrowing from 6.6× to 3.9×. This ordering is consistent with a characteristic of the revision task.

Counterargument is the sole component whose rank changes substantially across grades. Developmentally, spontaneous counterargument use is rare before early-to-mid adolescence [16]. Instructionally, counterargument becomes an explicit writing requirement only in middle school [6]. Students below this level have little reason to act on feedback for a component that is not yet part of their writing goals. Together, these factors suggest that feedback actionability depends not only on cognitive difficulty but also on curricular exposure and instructional timing.

6.2 Implications for Interface and Teacher Support

The stability of the actionability gradient across grades suggests that a uniform feedback interface is insufficient, as component-level differences persist across developmental stages. The implications derive from the differences in actionability across argumentation components and unfold across three key areas: monitoring students where AI feedback works, protecting gains that global revision may threaten, and increasing teacher support when AI feedback alone is insufficient.

Trajectory-based monitoring. Where AI feedback is most readily actionable –evidence and reasoning–interfaces should surface revision trajectories rather than snapshot ratings. Of the 58,332 assignments, 38.3% showed no improvement event. A per-student, per-component trend view in the teacher dashboard would allow teachers to distinguish students who need intervention from those who have disengaged.

Progress-aware feedback. Global revision introduces an additional risk that it does not currently address. In 20.9% of draft pairs, at least one Already-Proficient component declines. Coherence is particularly vulnerable, co-declining with evidence at 28% and reasoning at 27%. Because each draft is evaluated independently with no reference to past progress, students receive no signal when a previously proficient component regresses. When this occurs, feedback could explain the decline instead of just reporting the new rating. When a component

is rated Proficient, the interface could highlight what the student did well and encourage them to maintain that progress. This approach shifts feedback from an isolated assessment to continuity-aware scaffolding, which recent work on revision scaffolding interfaces has started to explore [24].

Curriculum-aligned coverage. When AI feedback cannot manage everything, the required intervention varies by component. For audience appeal, little improvement despite high teacher-validated helpfulness ratings (92%) suggests that the task itself lacks a clear revision path and teachers need to model perspective-taking. For early-grade counterargument, the conceptual foundation must be established before AI feedback can be useful. Both cases reflect the principle that actionability is constrained not only by cognitive demands but also by instructional availability of revision strategies. Interface that integrate grade-level standards to adjust active components, or give teachers direct control over the rubric—a direction recent work has begun to explore [3], positioning teachers as co-designers rather than passive recipients.

7 Limitations

Several limitations bear on interpreting the actionability gradient. Improvement is operationalized as a change in the system's own rating rather than an independent measure of writing quality. We therefore treat the gradient as capturing feedback actionability rather than learning gains, and note that the same LLM both generates feedback and judges improvement. The gradient may also reflect system-specific design choices such as the coarse three-level scale and LLM feedback framing, so replication across platforms is needed. Causal interpretations regarding cognitive demands or instructional context should be treated as tentative. Finally, teacher validation involved five raters who endorsed system ratings rather than rating independently, supporting perceived correctness of static ratings but not the validity of rating changes. Future work pairing rating-change measures with independent teacher evaluation and learning outcome data would establish whether the gradient reflects genuine writing improvement.

8 Conclusion

Across 196,169 consecutive draft pairs from 58,332 K–12 student assignments, we observe a stable fivefold variation in component-level improvement rates that holds across grades 3 through 12. The actionability gradient reflects the interaction between cognitive demands, instructional context, and revision task structure in LLM-mediated writing environments. This implies that system design should differentiate interfaces: trajectory monitoring where feedback is readily actionable, continuity-aware scaffolding where regression risk is real, and explicit teacher intervention where revision lacks a clear operational path. As LLM-based writing feedback scales across K–12 classrooms, component-level actionability offers a structural basis for moving beyond one-size-fits-all feedback design.

References

1. Afrin, T., et al.: Annotation and classification of evidence and reasoning revisions. In: BEA Workshop (2020)
2. Afrin, T., et al.: Effective interfaces for student-driven revision sessions for argumentative writing. In: CHI 2021 (2021)
3. Bai, J., et al.: iRULER: Intelligible rubric-based LLM evaluation for revision. In: CHI 2026 (2026)
4. Bereiter, C., Scardamalia, M.: The Psychology of Written Composition. Erlbaum, Hillsdale, NJ (1987)
5. Butler, J.A., Britt, M.A.: Investigating instruction for improving revision of argumentative essays. Written Communication **28**(1), 70–96 (2011)
6. National Governors Association: Common Core State Standards for English Language Arts. Washington, DC (2010). Available at http://www.corestandards.org
7. Chen, A., et al.: A systematic review of AI-enabled assessment in language learning. J. Computer Assisted Learning **41**(1) (2025)
8. Correnti, R., et al.: Building a validity argument for eRevise as a formative assessment. Computers and Education Open **3**, 100084 (2022)
9. Faigley, L., Witte, S.: Analyzing revision. College Composition and Communication **32**(4), 400–414 (1981)
10. Fleckenstein, J., et al.: Automated feedback and writing: A multi-level meta-analysis. Frontiers in Artificial Intelligence **6**, 1162454 (2023)
11. Hayes, J.R., et al.: Cognitive processes in revision. In: Advances in Applied Psycholinguistics, vol. 2, pp. 176–240. Cambridge Univ. Press (1987)
12. Liu, Z., et al.: eRevise+RF: A writing evaluation system for assessing student essay revisions. In: NAACL-HLT 2025, pp. 173–190 (2025)
13. McCarthy, K.S., et al.: Automated writing evaluation: Does spelling and grammar feedback support revision? Assessing Writing **52**, 100608 (2022)
14. Meyer, J., et al.: Using LLMs to bring evidence-based feedback into the classroom. Computers and Education: AI **6**, 100199 (2024)
15. Palermo, C., Wilson, J.: Implementing automated writing evaluation in different instructional contexts. J. Writing Research **12**(1), 63–108 (2020)
16. Ricco, R.: Argumentative reasoning: Development, training, and relevance to academic outcomes. Behavioral Sciences **15**(12), 1700 (2025)
17. Roscoe, R.D., McNamara, D.S.: Writing Pal: Feasibility of an intelligent writing strategy tutor. J. Educational Psychology **105**(4), 1010–1025 (2013)
18. Sommers, N.: Revision strategies of student writers and experienced adult writers. College Composition and Communication **31**(4), 378–388 (1980)
19. Tseng, W.T., Chen, S., Dong, T.H.: A meta-synthesis of automatic writing evaluation research. Educational Technology R&D (2026)
20. Wang, E.L., et al.: eRevis(ing): Students' revision of text evidence use in an AWE system. Assessing Writing **44**, 100449 (2020)
21. Wilson, J., Czik, A.: Automated essay evaluation software in English language arts classrooms. Computers & Education **100**, 94–109 (2016)
22. Wilson, J., Roscoe, R.D.: Automated writing evaluation and feedback: Multiple metrics of efficacy. J. Educational Computing Research **58**(1), 87–125 (2020)
23. Zhang, H., et al.: eRevise: Using NLP to provide formative feedback on text evidence. In: AAAI, vol. 33, pp. 9619–9625 (2019)
24. Zhang, C., et al.: Friction: Deciphering writing feedback into revisions through LLM-assisted reflection. In: CHI 2025, pp. 1–27 (2025)