"Can You See Me Think?" Grounding LLM Feedback in Keystrokes and Revision Patterns

Samra Zafar
FAST NUCES
samrazafar003@gmail.com

Shifa Yousaf
FAST NUCES
l226542@lhr.nu.edu.pk

Abstract

As large language models (LLMs) increasingly assist in evaluating student writing, researchers have begun to explore whether these systems can attend not just to final drafts, but to the writing process itself. We examine how LLM feedback can be anchored in student writing processes, using keystroke logs and revision snapshots as cognitive proxies. We compare two conditions: C1 (final essay only) and C2 (final essay + process data), using an ablation study on 52 student essays. While rubric scores changed little, but process-aware feedback (C2) offered more explicit recognition of revisions and organization changes. These findings suggest that cognitively-grounded feedback from LLMs is more pedagogically aligned and reflective of actual student effort.

1 Introduction

LLMs are widely adopted in writing instruction, offering scalable feedback across learning environments. However, most implementations are limited to evaluating the final essay, ignoring the iterative and recursive nature of writing. Writing is not linear; it is inherently recursive, involving cycles of planning, pausing, rephrasing, and structural adjustments. (Flower and Hayes (1981)). Current LLM feedback may thus miss formative signals such as hesitation, rewriting, or organization changes. This work presents a pilot investigation into how revision behavior can inform process-aware language model feedback, serving as an initial feasibility study toward scalable, human-aligned writing support. We examine whether LLMs can integrate process signals to produce feedback that better reflects students' effort and revision strategies.

This study is guided by the following research questions:

RQ1: Does access to writing process data (snapshots and keylogs) significantly alter the LLM's rubric-based evaluation of student essays?

RQ2: What types of cognitive or revision-oriented feedback emerge when LLMs are exposed to the full writing trace?

RQ3: Does the process-aware condition (C2) produce richer and more targeted justification language compared to the final-only condition (C1)?

2 Related Work

Recent research shows that LLMs can support formative writing feedback, particularly when aligned with educational standards. Studies like Kinder et al. (2025) and Han et al. (2023) highlight the effectiveness of rubric-based LLM feedback in enhancing clarity and relevance, while a meta-analysis by Fleckenstein and Liebenow (2023) found moderate positive effects ($g \approx 0.55$) of LLM feedback on student writing. Other works Escalante et al. (2023); Dai et al. (2023) show that ChatGPT-style feedback can rival human tutor guidance, though Mah et al. (2025) notes that LLMs often focus on surface-level edits. To better ground feedback in student cognition, researchers have incorporated writing process data. Studies by Zhu et al. (2023) and Vandermeulen et al. (2023) demonstrate that

keystroke patterns correlate with writing ability and can support learning as effectively as classroom instruction. Process-tracing techniques like keylogging and document snapshots are increasingly used in feedback systems Swamy et al. (2024); Steinert et al. (2024), but no prior work has directly tested their effect on LLM feedback behavior—an empirical gap this paper addresses.

3 Methodology

3.1 Data Collection and Interface

We built a web-based essay platform (Python backend, JavaScript frontend) to capture essays and writing traces. Fifty-two non-native English undergraduates wrote 20-minute essays on randomly drawn prompts. JavaScript enabled event logging across browsers, with backspace releases longer than 3 s logged to mark revision episodes, and periodic snapshots preserving intermediate drafts. These records allowed reconstruction of pauses, deletions, and reordering.

3.2 Feedback Generation

Each essay was evaluated by the Gemini Flash LLM under two prompt conditions. In *C1* (final-only), the model scored four CEFR-aligned dimensions—Thesis, Organization, Language, and Engagement—with short justifications. In *C2* (process-aware), the model also received keylogs and snapshots, using the same rubric but prompted to consider revision behaviour. In the process-aware condition (C2), the model produced two outputs: Rubric Feedback (scores and justifications, directly comparable to C1), and Revision Feedback (commentary on revision behaviour). For fairness, only Rubric Feedback was used in statistical comparisons.

We selected Gemini Flash for rubric-based scoring due to its affordability and API accessibility, which made it feasible to evaluate all essays under both conditions. For thematic coding, we used GPT-4, following prior studies where it has been validated as a reliable coding assistant for qualitative analysis Braun and Clarke (2006); Han et al. (2023). This dual-model setup reflects pragmatic constraints but does not affect the core hypothesis, which is model-agnostic.

3.3 Ablation Study Design

All essays were evaluated in both conditions, producing paired observations. For RQ1 we compared rubric scores ($\Delta = C1 - C2$) via two-tailed paired t-tests. For RQ2 we counted behavioural mentions and revision verbs in Rubric Feedback justifications. For RQ3 we checked those mentions against logs to confirm accuracy. This design separated score changes from shifts in explanatory focus and also assessed whether behaviour mentions were factually supported by observed logs (i.e., avoiding unsupported inferences).
To analyse C2 Revision Feedback, we used a reflexive thematic approach with GPT-4 following prior literature (e.g., Braun and Clarke (2006) inspired reflexive thematic pipelines) where GPT-4 has been validated as a coding assistant. assisting in clustering behaviour codes.Human researchers reviewed and merged codes, requiring each to be grounded in a direct quote and verified. Reliability was assessed on 20 excerpts by two coders applying six behavioural tags (lexical edits, pauses, uncertainty, expansion, structure, fluency). Agreement was 75% with Cohen's $\kappa = 0.72$, indicating substantial consistency in the GPT–human pipeline.

4 Results

4.1 Rubric score analysis

We first ask whether access to process traces alters rubric scores (RQ1). In our within–subject design, the *same* essay is scored under two conditions, so differences are attributable to process data. Table 1 shows two-tailed paired t–tests across four CEFR–aligned dimensions. Scores for *Thesis*, *Language*, and *Engagement* remain comparable across conditions ($p \geq 0.57$). By contrast, *Organization* improves significantly ($+0.50$, $p < 0.01$), suggesting that keylogs and snapshots help the model recognise planning and reordering. These findings align with cognitive theories where coherence emerges from revision, while other dimensions depend more on the submitted product.

Table 1: Paired comparison of rubric scores across $N=52$ essays. Δ is C2–C1.

Dimension	Mean Δ	Improved	Unchanged	Declined	p	Sig.
Thesis	+0.18	2	35	15	1.0000	No
Organization	+0.50	10	27	15	0.0046	**Yes**
Language Use	-0.05	5	42	5	0.5758	No
Engagement	+0.05	2	37	13	0.7147	No

To illustrate how access to process traces affects the justification language, we include paired excerpts for the same essays under C1 (final-only) and C2 (process-aware). These examples highlight a clear shift: C1 comments focus on the visible structure of the final draft, while C2 comments incorporate trace-grounded signals such as pauses, expansions, and reorganizing episodes. This side-by-side comparison clarifies how process data leads to more targeted Organization feedback.

Table 2: Paired examples showing how process-aware (C2) feedback adds trace-grounded comments beyond final-only (C1) feedback for the Organization dimension.

Essay ID	C1: Final-only feedback (Organization)	C2: Process-aware feedback (Organization)
o8ap	"Follows a basic structure with an introduction, body paragraphs, and a conclusion, but transitions between paragraphs are weak and abrupt, hindering logical flow. The essay lacks a clear thematic progression, making the connection between the impacts of social media and student attention spans uneven."	"Organization still feels somewhat disorganized: positive and negative aspects are presented in separate blocks and the conclusion restates the introduction rather than synthesizing. Snapshots show a gradual expansion of the essay, initially focusing on negative impacts before incorporating the positive after a significant pause around the 15-minute mark, possibly indicating a strategic shift in approach."
60im	"Lacks a clear organizational structure beyond a chronological recounting of events. The introduction is rambling and does not effectively set the stage for the main point, and the concluding paragraph abruptly shifts from a discussion of friendship to a focus on career goals."	"Structure remains somewhat disjointed: the introduction fails to concisely establish the essay's focus and the conclusion does not effectively summarize the main points or restate the thesis. Keylogs reveal numerous backspace-revision events and lengthy pauses in the introduction and conclusion, and the evolution from the 3-minute snapshot to the final draft shows a growing clarity of the central narrative, but also highlights the need for pre-writing strategies to strengthen overall organization."

4.2 Themes and frequencies in process-aware feedback

Inductive coding of C2 Part 2 feedback yielded 37 codes grouped into six themes: *Cognitive Effort* – mentions of hesitation, uncertainty or difficulty progressing, often inferred from long pauses or frequent deletions (e.g., "frequent pauses before drafting the introduction"). *Revision Type* – explicit notes of rewriting, content expansion or low-level adjustments. *Revision Timing* – references to when revisions occurred (early, mid-task or late). *Structural Focus* – comments on organizational shifts, repositioned thesis statements or changes in argument order. *Outcome-Oriented* – evaluation of increased coherence, clarity or argument strength attributable to revision behaviour. *Process Markers* – explicit references to backspacing bursts or long pause sequences. A compact, side-by-side visual pairs the most frequent codes with a distribution of behaviour tags (Figure 2). Together, these views show that process access shifts the feedback from purely product-oriented textual critique to a process-aware account that captures when and how writers revised, and what cognitive pressure points (pauses, uncertainty) accompanied that activity. The dominance of lexical/structural codes (rewriting, expansion) alongside high-frequency pause/uncertainty tags confirms that adding process traces broadens feedback beyond surface text, supporting our claim that LLMs can reliably infer planning and cognitive effort when time-stamped signals are available.

Figure 1: Most frequent revision codes in C2 Revision Feedback feedback.

Code	Count
Sentence rewriting	11
Content expansion	9
Backspacing	7
Early revisions	6
Mid–task revisions	6
Expression struggles	6

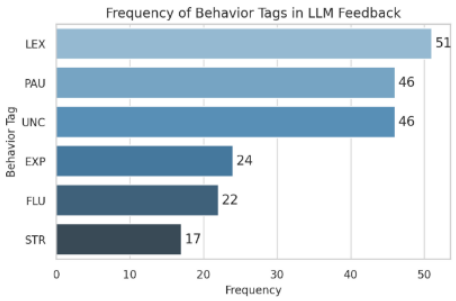

Figure 2: Distribution of behaviour tags in C2 feedback.

4.3 Behavioural mentions and alignment in comparable justifications

Even without being prompted to discuss behaviour, the model sometimes incorporates process evidence into Part 1 justifications when traces are compelling. Across the 52 pairs, 12 C2 justifications (23%) reference revision behaviour (e.g., hesitation, backspacing bursts, structural reordering), whereas C1 never does. The mean number of revision-related verbs is 0.6 in C2 (max 6) versus 0 in C1. Spot checks against logs and snapshots verified each sampled mention, and a reliability study on 20 C2 Part 2 excerpts yielded 75% agreement with Cohen's $\kappa = 0.72$, indicating substantial consistency of the behaviour tags. Taken together, the results suggest that access to traces induces selective but consistent incorporation of behavioural evidence: the model does not over-interpret noise, but when signals (pauses, deletions, re-sequencing) are strong, it integrates them into both its process commentary and, at times, into product-oriented judgments. The presence of verified behavioural mentions only in C2 justifications, plus substantial IRR, substantiates our claim that access to process traces yields grounded, non-hallucinatory behavioural reasoning that the model invokes conservatively when evidence is strong. The exemplar timeline provides concrete, per-essay

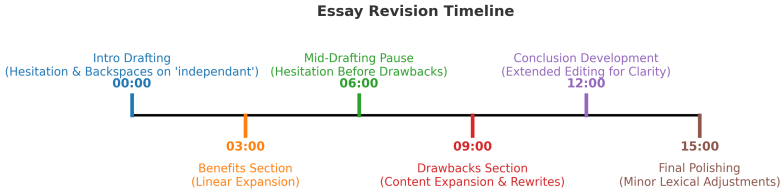

Figure 3: Revision timeline of one essay example with key stages and timestamps.

alignment between observed pauses/deletions and the model's behavioural claims, reinforcing that our behaviour-aware interpretations are trace-grounded rather than speculative.

5 Discussion

Our results show that incorporating process data (keystrokes and revision snapshots) enables LLMs to provide more structurally sensitive feedback. The significant improvement in Organization scores under the process-aware (C2) condition suggests that revision behaviors such as restructuring and expansion are recognized by the model, while other dimensions (Thesis, Language, Engagement) remain largely determined by the final text.

This shift highlights an equity implication: process traces make visible the revision-heavy effort of students who may otherwise be undervalued when only polished drafts are judged. Such recognition is particularly relevant for non-native speakers and learners from underrepresented groups, including many Muslim students who engage with English as an additional language. By accounting for cognitive effort, process-aware systems help reduce structural biases in automated evaluation.

Finally, the substantial agreement ($\kappa = 0.72$) between human coders and LLM-inferred behaviors shows that models, when given authentic traces, can provide grounded rather than speculative feedback. This positions process-aware LLMs as viable tools for inclusive, low-stakes educational applications.

6 Limitations and Future Work

The present study is a pilot investigation with several inherent limitations. Importantly, our aim was not to compare different language models (e.g., GPT-4 vs. Gemini Flash), but to examine how revision behavior itself—as a cognitive and behavioral signal—can inform feedback generation through large language models. The dataset of 52 essays, drawn from non-native English undergraduates, constrains generalizability across populations and genres. Additionally, while GPT-4 and Gemini Flash were used to operationalize the framework, their role was purely instrumental rather than comparative; the use of these commercial APIs also limits reproducibility and transparency, as model updates and proprietary behavior cannot be fully controlled. Moreover, while keystroke logs offer a robust proxy for writing process data, they capture only part of the revision effort. Finally, the focus on short, single-assignment essays prevents observation of longitudinal effects or feedback adaptation over time. Collectively, these factors position the present study as an initial feasibility probe establishing methodological groundwork rather than definitive empirical claims.

Future research will extend this pilot in three directions. First, by comparing process-aware LLM feedback with human instructor and peer feedback, we can assess alignment and complementary strengths. Second, expanding and diversifying the dataset—and incorporating open-source, reproducible models—will enable more robust and transparent analysis. Third, long-term classroom studies can evaluate the sustained pedagogical impact of process-grounded feedback and examine whether iterative exposure to revision data fosters adaptive *LLM learning*. Together, these directions aim toward scalable, transparent, and human-aligned writing feedback systems.

7 Ethical Considerations

All participants voluntarily consented to the collection of keystrokes and document snapshots within a low-stakes educational setting. No grading, assessment, or other consequential outcomes were attached to participation. All data were anonymized, with 4-digit IDs generated at random for organization. Participants were explicitly informed that their writing would be used solely for research purposes, with the option to withdraw at any time. These design choices—low-stakes context, anonymization, and opt-in consent—were implemented to minimize the risks of surveillance or coercion. While process data can raise privacy concerns, our approach demonstrates that pedagogically valuable insights can be obtained without exposing sensitive information. A public ethics statement and supporting documentation are available at: `https://doi.org/10.6084/m9.figshare.29927414`.

References

Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006. doi: 10.1191/1478088706qp063oa.

Jinlong Dai, Zhuoran Li, Xinyi Chen, and Junxian Gao. Using gpt-based feedback to improve undergraduate essay drafts: A controlled lab study. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, 2023.

Jorge Escalante, Yi-Ting Huang, and Meera Patel. How good is chatgpt's feedback? evaluating ai-powered revision advice in undergraduate writing. *Frontiers in Artificial Intelligence*, 6:112, 2023.

Johanna Fleckenstein and Lucas Liebenow. Meta-analysis of ai-supported writing feedback in k–12 education. *Educational Technology Journal*, 40(4):67–89, 2023.

Linda Flower and John R. Hayes. A cognitive process theory of writing. *College Composition and Communication*, 32(4):365–387, 1981.

Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, and Alice Oh. Llm-as-a-tutor in efl writing education: Focusing on evaluation of student–llm interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada, 2023. Association for Computational Linguistics.

Annette Kinder, Fiona J. Briese, Marius Jacobs, Niclas Dern, Niels Glodny, Simon Jacobs, and Samuel Leßmann. Effects of adaptive feedback generated by a large language model: A case study in teacher education. *Computers and Education: Artificial Intelligence*, 8:100349, 2025.

Alisha Mah, Samiya Qureshi, and Hoang Tran. Surface-level or dialogic? a comparative study of human vs llm feedback in college essays. *EdWorkingPapers*, (25-109), 2025.

Gabriel Steinert, Sergio Avila, Sascha Küchemann, et al. Harnessing large language models to develop research-based learning assistants for formative feedback. *Smart Learning Environments*, 11(62), 2024.

Vinitra Swamy, Davide Romano, Bhargav Srinivasa Desikan, Oana-Maria Camburu, and Tanja Käser. Illuminate: An llm-xai framework leveraging social science explanation theories towards actionable student performance feedback. *arXiv preprint arXiv:2409.08027*, 2024.

Anna Vandermeulen, Jamal Nasir, and Helen Zhao. Real-time writing analytics in classrooms: A year-long intervention using keystroke comparisons. In *Proceedings of the 2023 International Conference on Learning Analytics & Knowledge (LAK)*, 2023.

Wenjia Zhu, Huan Lim, and Wei-Chun Tan. Keystroke dynamics as predictors of writing ability: Evidence from middle school classrooms. *Journal of Writing Research*, 15(2):45–68, 2023.

A Prompts Used for Thematic Coding

A.1 LLM Coding Instructions

The following prompts were used to guide GPT-4 in performing reflexive thematic analysis of revision-aware feedback, adapted from Braun & Clarke's SAGE framework.

SAGE Steps I–II: Familiarization and Initial Code Generation

You are performing a thematic analysis of AI-generated feedback that comments on a student's revision behavior during a timed writing task. Your goal is to extract initial codes that capture how the LLM interpreted the student's writing process, revision patterns, and signs of cognitive effort. For each revision behavior comment, do the following:

- Identify meaningful quote(s) that express a distinct idea
- Explain what that quote refers to or suggests
- Assign a grounded, descriptive code to the idea

Use this format:

```
'{quoted text from feedback}' refers to/mentions '{definition of the idea}'.
Therefore, we get a code: '{CODE NAME}'
```

SAGE Steps III–VI: Theme Construction, Review, Naming, and Mapping

You are now in the next phase of reflexive thematic analysis (SAGE/Braun & Clarke).Based on the following list of qualitative codes (from C2 LLM revision feedback), your task is to:

- Group similar codes into higher level candidate themes
- Name each theme concisely (max 6 words)

For each theme, provide:

- A definition of what it captures
- The list of codes grouped under it
- A summary of what the theme suggests about how the LLM interprets student revision behavior
- 1–2 representative quotes

B Behaviour Tag Scheme

Table 3: Behaviour tags used for alignment and reliability.

Tag	Interpretation
LEX	Lexical edits / phrasing / word choice
PAU	Pauses or long hesitations
UNC	Uncertainty / cognitive struggle
EXP	Expansion or elaboration of ideas
STR	Structural rearrangement
FLU	Fluent/linear writing with minimal revision

C Behaviour–Feedback Alignment Examples

Table 4: Examples where LLM's behavioural mentions align with observed process traces (snapshots/keylogs).

Essay ID	Observed Trace Pattern	LLM Feedback (excerpt)	Aligned?
18	No major restructuring after 12 min; later edits are additive	"Small changes after 12 minutes suggest the writer reached a stable structure."	Yes
12	3–6 min: content expanded with precise examples	"Pauses and revisions between 3–6 minutes indicate careful elaboration."	Yes
41	Third paragraph substantially elaborated by 12 min	"Later snapshots show clear addition of details and varied benefits."	Yes

D Thematic Codebook (Summary)

Table 5: Concise definitions of six high-level themes.

Theme	Definition
Cognitive Effort	Pauses, hesitation, uncertainty, difficulty progressing.
Revision Type	Rewriting, expansion, surface-level adjustments.
Revision Timing	Early, mid-task, or late-phase concentration of edits.
Structural Focus	Reordered arguments, refining thesis, transitions.
Outcome-Oriented	Gains in coherence, clarity, or argument strength.
Process Markers	Backspacing bursts, long pauses, trace-level signals.

Table 6: Representative codes and frequencies illustrating each theme.

Code	Short Definition	Freq.
Cognitive Hesitation	Frequent deletions/pauses indicating indecision	18
Struggles with Expression	Difficulty articulating or phrasing ideas	21
Sentence Rewriting	Rewording/restructuring existing content	24
Content Expansion	Adding elaboration/examples	19
Early Revisions	Edits made in early snapshots	14
Mid-Task Revisions	Reorganizations mid-way through essay	9
Organization Improvements	Better transitions/paragraph structure	13
Increased Clarity	Writing improves across revisions	10
Backspacing Behavior	Frequent use of delete/backspace	15