Democratizing Writing Support with AI: Insights from One Year of Real-World Interactions with an Open-Access Writing Feedback Tool

Babette Bühler[1,2], **Ivo Bueno**[1,2], **Enkelejda Kasneci**[1,2]

[1]Technical University of Munich, Germany
[2]Munich Center for Machine Learning (MCML), Munich, Germany
babette.buehler@tum.com, ivo.bueno@tum.com, enkelejda.kasneci@tum.com

Abstract

Writing is a foundational skill for educational, professional, and civic participation, yet access to frequent and timely writing feedback remains deeply unequal. Teachers face significant workload constraints, particularly in large classes, and many learners lack alternative sources of individualized feedback. While large language models (LLMs) offer the opportunity for scalable, adaptive support, little is known about how students engage with such feedback tools in real-world, self-directed settings. We present a large-scale, year-long analysis of 23,650 voluntary interactions with an open-access AI writing feedback system used by students across diverse educational contexts and age groups, conducted in accordance with strict data protection standards. Using a clustering approach, we identify 2,800 iterative revision chains and apply a validated LLM-based multidimensional scoring framework to assess text quality over time. Our findings reveal that students who revised their texts after receiving AI feedback demonstrated statistically significant, albeit modest, improvements across both content and language-related dimensions (overall writing quality: $\Delta = 0.067$, $p < .001$, $r = .17$), with the greatest gains observed among initially low-performing writers. Revision frequency was positively associated with improvement, particularly in higher-order writing skills. However, engagement was uneven, with higher usage among students in academically oriented schools. These results demonstrate both the technical feasibility and social potential of deploying generative AI for educational support at scale, while highlighting the need for inclusive infrastructure, accessible design, and targeted outreach to truly democratize educational benefits.

Code — https://gitlab.lrz.de/hctl/ai-writing-feedback/-/tree/9f32f35d999fd54a34ab5d16a60bf133c69ac3ab/

Introduction

Writing texts is a fundamental skill that enables individuals to participate fully in society, access educational and professional opportunities, and exercise their civic and personal rights (Freedman et al. 2016). One of the most powerful and widely recognized tools to improve student writing development is feedback (Graham, Hebert, and Harris 2015). Frequent, timely, and high-quality feedback helps learners identify gaps in understanding, refine their work, and engage in iterative cycles of drafting and revision, which are essential for mastering complex cognitive skills such as writing (Hattie and Timperley 2007; Graham, Hebert, and Harris 2015). Yet, around the world, access to such feedback remains deeply unequal. Students in under-resourced schools often lack sufficient support to develop writing proficiency, and even in well-funded educational systems, teachers face significant time restrictions that limit their ability to provide high-quality individualized feedback (Wiley 2006). These challenges are particularly acute for learners whose parents or caregivers may be unable to assist with writing tasks at home, further exacerbating educational inequalities.

There is a large body of research investigating feedback based on automated writing evaluation, leveraging natural language processing and machine learning (Fleckenstein, Liebenow, and Meyer 2023; Hahn et al. 2021); however, despite its demonstrated effectiveness (Fleckenstein, Liebenow, and Meyer 2023), automated writing evaluation has remained largely impractical for scalable use by teachers or learners due to its task-specific nature, the high cost and effort of generating annotated training data, and limited flexibility in evaluating diverse writing criteria (Ramesh and Sanampudi 2022; Jansen, Horbach, and Meyer 2025). The recent emergence of large language models marks a paradigm shift in how written language can be analyzed, generated, and supported by AI. Their capability in understanding and generating natural language enables them to provide detailed, adaptive feedback on student writing, which offers a promising avenue to address these challenges (Holmes, Miao et al. 2023; Kasneci et al. 2023). Initial experimental studies provide empirical evidence that GenAI-generated feedback can improve student motivation and facilitate writing improvement (Meyer et al. 2024; Lo, Wong, and Chan 2025; Seßler, Kepir, and Kasneci 2024). These tools might have the potential to democratize access to high-quality feedback, support self-regulated learning, and reduce teacher workload (Kasneci et al. 2023; Meyer et al. 2024).

However, critical questions remain about how learners interact with AI feedback systems in practice. We identify three key research gaps in the study of AI-generated writing feedback: (1) Prior work predominantly relies on small-scale, controlled experiments focused on specific school types or age groups, limiting ecological validity. Large-scale analyses of unsupervised, real-world use remain rare, par-

ticularly across diverse learner contexts spanning grades 1–13 (ages 6–18). (2) Most existing studies focus either on classroom-based deployments or the usage of AI tools embedded in assignments and supervised by teachers (Meyer et al. 2024; Jansen, Horbach, and Meyer 2025; Meyer, Jansen, and Fleckenstein 2025) and higher education settings characterized by higher levels of student autonomy (Kinder et al. 2025; Lo, Wong, and Chan 2025; Escalante, Pack, and Barrett 2023). While informative, these contexts do not capture how learners engage with feedback tools in voluntary, self-directed settings, where motivation, agency, and support structures vary widely. Moreover, scalable feedback alone does not ensure effective use: learners frequently ignore or superficially apply feedback, limiting its impact on learning (Attali 2004). (3) Writing is inherently iterative, yet prior research often conceptualizes engagement with AI tools as a one-time interaction (Meyer et al. 2024). It remains unclear whether learners revise and improve their texts over multiple iterations, an essential condition for realizing the educational potential of AI feedback.

Contributions. To address these gaps, we analyzed a full year of real-world interactions (n = 23,650) with an open-access Gen-AI feedback tool used voluntarily in various educational settings. Our contributions are:

- Methodological insights for analyzing large-scale, non-experimental data from authentic, self-directed AI tool use, highlighting challenges and opportunities for evaluating social impact.
- A detailed analysis of learner engagement patterns across text types, school types, grade levels, and temporal usage trends.
- A privacy-preserving approach to modeling iterative writing behavior at scale, using anonymized data and similarity-based clustering.
- An evaluation of writing improvement across revisions using a validated LLM-based multidimensional scoring framework aligned with feedback criteria (e.g., narrative coherence, stylistic quality, mechanics) (Seßler et al. 2025).
- An investigation into how revision frequency and initial writing quality shape improvement trajectories across up to five rounds of revision.

By investigating these questions, we provide the first large-scale, ecologically valid analysis of how students engage with generative AI feedback tools in voluntary, unsupervised settings, illustrating what can be learned from privacy-compliant, real-world educational data. Our findings shed light on engagement trajectories, revision patterns, and associated changes in writing quality, informing the design of scalable, equitable feedback systems.

Related Work

Research on LLM–generated writing feedback has evolved along three interrelated dimensions: the quality of LLM-generated feedback, its impact on writing outcomes, and the extent which students engage meaningfully with such feedback. Taken together, these strands underscore the promise and limits of LLMs in supporting writing development.

Feedback Quality. Studies benchmarking LLM-generated feedback often compare it to instructor, expert, or peer alternatives. Dai et al. (2023) found GPT's comments on student reports were similar in polarity to instructor sentiment, but deeper comparisons reveal quality gaps. Steiss et al. (2024) demonstrated that while ChatGPT slightly outperformed teachers in terms of criteria alignment, teachers excelled in clarity, accuracy, tone, and prioritizing revisions. Banihashem et al. (2024) found that ChatGPT provided more descriptive feedback than its peers, who were better at identifying issues; neither approach significantly improved writing. In creative writing, Rashkin et al. (2025) found LLMs (GPT-4, Claude, Gemini) produced fluent but vague comments, missing narrative-level issues, and trailing experts in helpfulness ($d = 0.86$) and specificity ($d = 0.72$). These findings suggest that while LLM feedback may be less pedagogically precise than expert input, it remains potentially useful, especially in contexts where human feedback is unavailable.

Feedback Effectiveness. A second line of work has tested whether LLM-generated feedback translates into measurable gains in student writing. In a randomized controlled trial with 459 Grade 10 students in German academic-track schools, Meyer et al. (2024) found that GPT-3.5-turbo–generated feedback, designed based on writing research and tailored to each student's draft, improved revision quality ($d = 0.19$), task motivation ($d = 0.36$), and positive affect ($d = 0.34$) relative to a no-feedback control. In this case, all students were explicitly prompted to revise. Other studies show more mixed effects. Escalante, Pack, and Barrett (2023) conducted two six-week longitudinal studies with university-level English learners. In Study 1, students received feedback from either GPT-4 or a trained human tutor. While both groups revised after receiving feedback, no significant difference in learning gains was found. Study 2 revealed that students viewed both sources as helpful but complementary, with preferences split. In a study of 269 pre-service teachers, Kinder et al. (2025) compared adaptive GPT-4 feedback to static expert-written feedback during a diagnostic reasoning task. LLM feedback led to higher justification quality ($d = 0.31$), longer responses (d = 0.39), and greater perceived usefulness (d = 0.51), although there was no significant effect on decision accuracy. Finally, Lo, Wong, and Chan (2025) conducted a large-scale RCT with 1,102 university students in Hong Kong. Compared to a no-feedback control, students receiving ChatGPT feedback showed significantly larger improvements in essay quality ($\Delta = 3.34$, p = .0031, $r = .21$), higher motivation ($r = .31$), and increased engagement ($r = .16$). Interviews highlighted the feedback's usability and specificity, though emotional responses were mixed and concerns about over-reliance emerged. These studies suggest that while LLM-generated feedback may not outperform expert human instruction, it can still support learning gains, particularly when feedback is adaptive and embedded in revision cycles.

Engagement with Feedback. Recent work has emphasized that simply providing AI-generated feedback does not guarantee that students will meaningfully engage with it, especially in self-directed or minimally scaffolded con-

texts, where engagement becomes a key mediator of impact (Fleckenstein et al. 2024). In a study of 14,236 secondary students across 655 classrooms using GPT-4–generated feedback, Jansen, Horbach, and Meyer (2025) found that 48% of students made no revisions, despite receiving feedback and being prompted to revise. Engagement rates were unaffected by grade level, task type, or the linguistic features of feedback or writing, suggesting that deeper motivational or contextual barriers may be at play. Similarly, Meyer, Jansen, and Fleckenstein (2025) studied 937 German students (Grades 7–9) using automated feedback in a controlled revision task. Although all students were asked to revise, 20% did not revise at all, and 47% revised without improving their performance. Non-engagement was more likely among male students ($d = 0.24$) and those with lower cognitive ability ($d = 0.48$), while unsuccessful engagement was associated with weaker grades and lower task value. By contrast, intrinsic task value predicted successful engagement ($d = 0.29$). These results parallel earlier findings on rule-based feedback systems: for example, Attali (2004) found that in the Criterion system (used in U.S. middle and high schools), 71% of students submitted only once, despite access to feedback. Still, those who did revise tended to improve, reinforcing the potential of automated feedback systems when they succeed in eliciting engagement.

Taken together, this body of work suggests that while LLM-generated feedback is not yet equivalent to expert human instruction in terms of diagnostic quality or impact, it represents a scalable and adaptive alternative, especially in under-resourced contexts where human feedback is unavailable. Evidence suggests that LLM feedback can support writing improvement, particularly when integrated into structured revision workflows. However, consistent with broader learning science research, the mere provision of feedback does not ensure its uptake: motivation, task framing, and learner supports remain critical levers. Yet, most existing studies have been conducted in structured, often short-term classroom deployments (Meyer et al. 2024; Jansen, Horbach, and Meyer 2025), primarily with older students in higher education (Kinder et al. 2025; Lo, Wong, and Chan 2025; Escalante, Pack, and Barrett 2023) or late secondary settings (Meyer et al. 2024). Thus, findings may not generalize to younger learners or to voluntary, unsupervised use. Moreover, while studies often capture single-shot interactions, writing is inherently iterative; little is known about how learners revise across multiple feedback rounds, or whether revision quality improves with repeated usage.

Methods

AI Feedback System Description

Our study focuses on a publicly accessible AI writing feedback platform PEER (Seßler et al. 2023)[1], designed to provide personalized, constructive feedback to students across a range of school levels and writing tasks. The system accepts student texts via direct input or OCR-based image upload and returns detailed feedback tailored to text type, grade level, and other metadata. Unlike traditional automated text scoring systems that prioritize rubric-based grading, the tool aims to replicate the function of a human teacher: offering encouragement, highlighting strengths, and suggesting areas for improvement in a supportive tone. Students can revise and resubmit their texts for new feedback, enabling iterative improvement aligned with process-oriented writing pedagogy. The tool utilizes GPT-3 in a zero-shot framework, generating feedback through optimized instructional prompts selected via a weighted lottery based on Elo scores (Seßler et al. 2023). Each feedback type is produced with a distinct German-language prompt, tailored to emulate a teacher's tone and include the text type, grade level, task description, and student text. This prompt design ensures context-aware, curriculum-aligned feedback across three dimensions: (1) a general qualitative evaluation, (2) a criteria-based assessment, and (3) actionable suggestions for improvement.

Dataset

Since its public launch in early 2023, the platform has seen widespread usage. Users consented to the use of their anonymized data for research by confirming the platform's data protection notice, in compliance with German data privacy laws. Collected data included text submissions and non-identifying metadata. No persistent user identifiers were stored. This paper focuses on interactions between December 2023 and November 2024. The dataset comprises 23,650 writing tasks submitted by learners across various educational contexts in Germany, including both the original student texts and the AI-generated feedback provided by the system. Each record contains metadata about the submission, including the text category (e.g., report, discussion), study year (grade level), school type (e.g., primary, academic secondary), a state within Germany, and a timestamp for each interaction.

Data Pre-Processing

Before analysis, the dataset was cleaned to include only authentic learner submissions and meaningful feedback interactions. From the original 30,755 entries, we removed 2,494 texts with fewer than 35 words, approximately five sentences of seven words each (Brügelmann and Richter 1994), to exclude placeholders and incomplete drafts. We also removed 4,063 exact duplicates, retaining only the first occurrence to prevent overrepresentation. Additionally, 458 entries with default metadata values (e.g., text category "report", study year 1, School Type "Elementary school", and state "Bavaria") were excluded, as these improbable combinations suggested placeholder data or test inputs. Lastly, 90 entries with missing values in critical fields, such as feedback, were discarded, likely due to failed feedback generation. The resulting cleaned dataset comprised 23,650 unique submissions, providing a more reliable basis for analysis of authentic interactions with the AI feedback tool.

A key challenge was the lack of persistent user identifiers, which hindered tracking individual learners across submissions. To address this, we used a similarity-based

[1]https://www.edu.sot.tum.de/en/hctl/forschung/peer/

clustering approach to reconstruct feedback iteration cycles. Texts were vectorized using TF-IDF (excluding German stopwords), and pairwise cosine similarity scores were computed. Submissions with scores above 0.8 were grouped and assigned a shared identifier (text ID), yielding 18,741 unique texts. This method assumed that learners revising their work would make substantial but not complete changes between iterations. Clustering accuracy was spot-checked manually, and multiple thresholds were tested to ensure consistent grouping of text revisions. This approach allowed us to analyze temporal engagement and revision behavior despite the absence of explicit user tracking.

Analysis

To evaluate writing improvement following AI-generated feedback, we focused on multiple-submission cases where learners revised and resubmitted their texts. A total of 2,800 student texts were resubmitted at least once. Each original and revised submission, totaling 7,564 texts, was scored using an LLM-based multidimensional text evaluation approach validated in recent research by Seßler et al. (2025). The authors evaluated various LLMs against ratings from 37 human teachers on ten criteria for German student texts. The proprietary GPT o1 model showed the strongest alignment with human teacher ratings, particularly in language-related aspects (e.g., spelling $r = 0.814$, expression $r = 0.675$, overall $r = 0.742$), and demonstrated high internal consistency ($ICC = 0.80$). These findings support its use as a reliable tool for assessing the quality of student writing.

Building on this validated framework, we scored each text both before and after revision using a rubric adapted to our text types for pedagogical relevance. This enabled fine-grained tracking of writing changes beyond holistic scores. By using a model with benchmarked human alignment and scoring consistency, we ensured the validity of our automated assessment. We assessed pre- vs. post-feedback gains in overall and criterion-specific scores using two-tailed Wilcoxon signed-rank tests, as Shapiro–Wilk tests indicated non-normality for all criteria ($\alpha = 0.05$), reporting p-values and effect sizes (r). In addition, we fit ordinary least-squares regressions predicting score gain from revision rounds, initial score, and writer age/language background, using robust standard errors. The code used for all data preparation, text scoring, prompts, and analyses is available on GitLab.[2]

Results

Learner Engagement with AI Feedback in Real-World Contexts

Analyzing a year of interaction data, comprising 18,741 text submissions, from our open-access writing feedback tool reveals several patterns in how learners engage with AI-generated feedback in a self-directed context (see Figure 1 for an overview). The submissions were distributed across various educational settings, encompassing a range of text types, including reports, discussions, and stories. Among

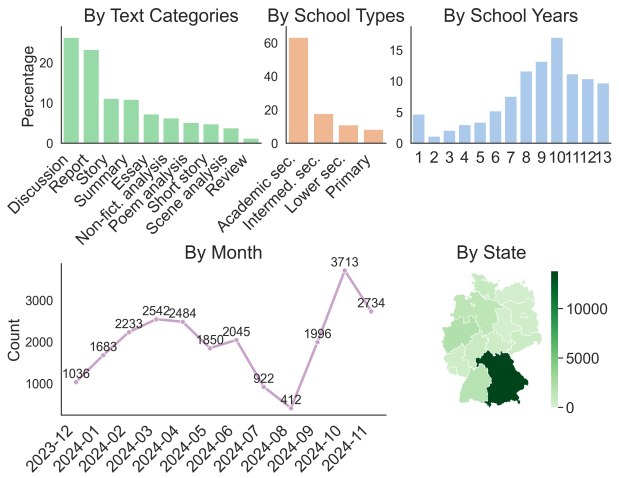

Figure 1: Interaction Data Overview

these, discussions and reports were the most frequently submitted text types, reflecting their prominence in school curricula. Submissions were drawn from various school forms (primary, lower secondary, intermediate secondary, and academic secondary), with the majority originating from academic secondary schools. Engagement also varied by grade: students in grades 8-13 accounted for the highest submissions, suggesting that middle and upper secondary students may be particularly receptive to AI feedback tools. A variation in dominant text types is observable across grade levels. Narrative texts dominate in lower grades, while reports in Grades 5–7 and analytical forms (e.g., discussions, essays) become more frequent from Grade 8 onward, reflecting a curricular shift toward analytical writing.

Temporal patterns indicate that learners most frequently engaged with the tool during the week, with lower activity observed on weekends. This suggests that usage is aligned with school schedules and potentially after-school homework routines rather than leisure-time writing. In line with these results, we observe a drop in engagement during July and August, concurrent with German summer holidays (see Figure 1). Geographical information indicates that, although users are spread across all German states, the vast majority are located in Bavaria, a region with high resources.

In terms of feedback iterations, a large proportion of learners interacted with the tool only once, writing one text. A notable subset of 15.5% engaged in multiple rounds of feedback and revision, of which 4.4% submitted a revised version of their text three to five times, indicating that some learners perceived value in refining their texts over successive submissions. The mean interval between submissions was 2.6 days, with a median of 0 days, over half of all revisions occurred on the same day as the initial submission, and more than 75% within a single day. To quantify textual changes across revisions, we computed the normalized Levenshtein edit distance (c.f. Schiller et al. 2024) between consecutive drafts. Across all revision rounds, the average normalized edit distance was 0.15, indicating moderate textual change between submissions. Most revision steps showed

[2]https://gitlab.lrz.de/hctl/ai-writing-feedback/-/tree/9f32f35d999fd54a34ab5d16a60bf133c69ac3ab/

consistent levels of modification (e.g., revision 1: 0.15; 2: 0.13), though some outliers were observed with unusually high distances (e.g., round 4: 0.29), suggesting more substantial rewrites. These findings suggest that while many learners made incremental edits, some engaged in more substantial textual restructuring during later revisions.

Writing Improvements Across Revisions

To analyze the effectiveness of the AI-generated feedback provided within our system, we applied automated LLM-based text scoring to a subset of texts within our data that were submitted multiple times, comprising 7,564 unique submissions and 2,800 distinct revision chains. Students' writing showed statistically significant improvements across most dimensions following AI feedback. Of the 2,800 texts resubmitted multiple times, 540 (19.3%) exhibited a positive increase in their overall judgment score from the first to the final draft (67.3% remained unchanged, 13% declined). Table 1 shows the mean score changes by rating criterion from the first to the last submission. Criterion-level score changes were tested using the Wilcoxon signed-rank test; all criteria except *Headline* showed statistically significant improvements from first to last draft ($p < .05$) with effect sizes (*r*) ranging from small to moderate. The holistic score (Overall judgment) increased by an average of 0.067 points, reflecting a modest but significant improvement in overall quality. Among content criteria, the largest gains occurred in the strength of the Conclusion (Δ = 0.138) and the Main part (Δ = 0.073), indicating many students' final drafts had stronger endings or syntheses and main part coherence. Language-related dimensions also improved significantly, with Spelling & punctuation rising by 0.068 points and Choice of words by 0.056 points. Together, these results suggest that iterative AI feedback use yields modest improvements in both content and language-related criteria.

To better understand how learners respond to AI-generated feedback, we analyzed whether the final recommendation section of the first feedback cycle explicitly referenced specific rating criteria and whether those criteria subsequently improved in the next revision. Using keyword matching for each dimension (e.g., Introduction, Spelling), we found that mentioning a criterion in the feedback was weakly associated with improvement in that area. For instance, when the "Conclusion" was mentioned (n = 1,067), 25.3% of those cases showed improved scores in that dimension in the second draft. Similarly, "Spelling & punctuation" (n = 1,021) and "Headline" (n = 103) saw implementation rates of 23.3% and 24.3%, respectively. Content-related criteria generally aligned more than language-related ones, though uptake was not uniform. Notably, feedback referencing "Linguistic style requirements" resulted in improvement only 10.7% of the time. These findings suggest that while students partially respond to targeted feedback, the uptake varies by the complexity of the criterion.

To explore who profits from the feedback, we examine whether the effectiveness of the feedback correlates with student grade or differs by school track. We found a small positive correlation between overall score and grade (r = 0.036, p = 0.058), indicating a modest improvement for older students. Notably, students from academic secondary schools exhibited the largest average gains in performance from first to final draft submissions (Δ = 0.083). In contrast, those from lower secondary(Δ = 0.034) and intermediate secondary schools (Δ = 0.037) showed the smallest gains. Intermediate improvements were observed for students from primary school (Δ = 0.061). These findings suggest slight but observable differences in feedback uptake across educational stages and school types.

Criterion	**Type**	**Δ 95% CI**	p-value	r
Overall judgment	C/L	0.067 [0.045, 0.090]	< 0.001	0.17
Headline	C	0.022 [–0.014, 0.058]	0.295	0.03
Introduction	C	0.049 [0.023, 0.076]	< 0.001	0.10
Main part	C	0.073 [0.049, 0.098]	< 0.001	0.16
Conclusion	C	0.138 [0.109, 0.168]	< 0.001	0.25
Choice of words	L	0.056 [0.029, 0.083]	< 0.001	0.11
Linguistic style req.	L	0.044 [0.018, 0.070]	0.001	0.09
Plot / Arg. logic	C	0.056 [0.029, 0.082]	< 0.001	0.12
Expr. & sent. struct.	L	0.045 [0.022, 0.068]	< 0.001	0.11
Spelling & punct.	L	0.068 [0.035, 0.100]	< 0.001	0.11

Table 1: Criterion-level score changes from first to last text submission.
Note: All p-values from Wilcoxon signed-rank tests. Δ = mean difference; C = content-related; L = language-related; r = effect size.

Greater Improvement Among Weaker Initial Drafts

Further, we investigate whether revision gains depend on initial text quality. Lower-quality drafts may offer more readily addressable issues, enabling greater improvements from AI feedback, while higher-quality drafts allow limited scope for gains and may even risk score declines due to less effective revisions. To assess this, we analyze the correlation between initial scores and score changes. Our data analysis indicates a moderate negative correlation (r=–0.29, $p < .001$) between initial holistic score and improvement, meaning students who started with lower scores tended to improve more. In contrast, those who started high often saw smaller gains or plateaued. This is also illustrated in Table 2. Learners who began with low first-draft scores (1–2 on overall judgement) exhibited the most considerable improvements, with an average increase of 0.388 in holistic score, 0.200 in content-related dimensions, and 0.192 in language-related dimensions. Those with medium initial performance (3 on overall judgement) saw more modest gains (Δ_{overall} = 0.207; Δ_{content} = 0.153;(Δ_{language} = 0.137), whereas high-performing students (overall judgement score of 4–5) showed virtually no change in holistic score (Δ_{overall} = –0.035) and minimal shifts in analytic criteria. These patterns indicate that iterative AI feedback disproportionately benefits students with weaker initial drafts.

Marginal Gains Across Revision Rounds

Another important aspect is how text quality evolves across successive rounds of feedback and revision. For a detailed

1st Score	n	Δ_{overall}	p_{overall}	Δ_{c}	p_{c}	Δ_{l}	p_{l}
Low (1–2)	225	0.387	< 0.001	0.196	< 0.001	0.191	< 0.001
Med. (3)	819	0.208	< 0.001	0.153	< 0.001	0.137	< 0.001
High (4–5)	1643	–0.035	0.013	0.016	0.096	–0.000	0.704

Table 2: Mean improvement by initial score group from first to last submission.
Note: Initial score refers to the first-draft overall score. All p-values are from Wilcoxon signed-rank tests comparing first and final drafts. c = content-related; l = language-related criteria.

iteration-by-iteration analysis, we focus on the first five revision rounds (i.e., six total submissions), as the number of texts still undergoing revision drops below 15% beyond this point, leading to unstable estimates. Figure 2 shows the average scores by submission (left panel) and average within-text score improvements from one draft to the next (right panel). Average scores increase from the first to sixth submission (e.g., overall from 3.67 to 3.82), suggesting higher quality in later drafts. However, these descriptive trends do not reflect within-text improvement, as different texts and students may appear across rounds.

To better capture revision effects, we analyze average improvements between drafts for the same text. The largest gains occur between the first and second submissions (Average $\Delta = 0.045$), with smaller and more variable gains in later rounds. Structural dimensions, such as the main part and conclusion, show strong early gains, while higher-order features, like linguistic style and argumentation, improve more steadily across rounds. Some criteria (e.g., headline, main part) show fluctuation or decline in later rounds, possibly due to revision fatigue or experimentation.

Overall, these findings suggest that early revisions are the most impactful, but some writing dimensions, particularly higher-order skills, can continue to benefit from further iteration. Importantly, we find no meaningful correlation between the amount of textual change and score improvement across early revisions (Pearson $r = -0.020$, Spearman $\rho = 0.046$, rounds 1–5), indicating that larger edits do not necessarily yield better results.

To analyze whether students with more revisions achieve higher overall improvements, we examined the mean score improvement from the first to the last draft in relation to the total number of revisions. Due to the aforementioned sample size constraints, we only included texts in our sample with equal to or fewer than five revision rounds (i.e., six total submissions). As shown in Table 3, mean improvements in students' writing quality vary systematically with the number of revision rounds. Texts revised only once (n = 1,864) exhibited a modest overall gain of 0.049 points, with content-related criteria improving by 0.058 and language-related criteria by 0.036. Twice-revised texts (n = 527) demonstrated larger gains (Δ_{overall} = 0.076; Δ_{content} = 0.076; Δ_{language} = 0.068), and three rounds of revision (n = 195) yielded even greater improvements (Δ_{overall} = 0.154; (Δ_{content} = 0.080; Δ_{language} = 0.081). The largest

Rev.	n	Δ_{overall}	p_{overall}	Δ_{c}	p_{c}	Δ_{l}	p_{l}
1	1864	0.049	0.001	0.058	< 0.001	0.036	0.004
2	527	0.076	0.004	0.076	< 0.001	0.068	0.002
3	195	0.154	< 0.001	0.080	0.014	0.081	0.014
4	81	0.198	0.003	0.143	0.010	0.259	< 0.001
5	52	0.135	0.090	0.238	0.004	0.154	0.030

Table 3: Mean score improvements by number of revision rounds from first to last submission.
Note: Only texts with ≤ 5 submissions were included. All p-values are from Wilcoxon signed-rank tests comparing first and last drafts. c = content-related criteria; l = language-related criteria.

average gains occurred after four revisions (n = 81), particularly in language (Δ_{language} = 0.259) and content (Δ_{content} = 0.143), with an overall mean increase of 0.198.

To assess whether score gains were explained by confounding factors, we ran an OLS regression including text type, school type, federal state, study year, initial score, number of revisions (capped at 5), time elapsed between drafts, and normalized edit distance as predictors. The model explained limited variance ($R^2 = 0.113$), with only initial score ($\beta = -0.209$, $p < 0.001$) and number of revisions ($\beta = 0.036$, $p = 0.004$) showing significant effects. All other predictors, including school type, state, text type, time elapsed, and magnitude of textual change, had no significant impact on improvement. These results suggest that score gains are not systematically driven by structural, demographic, or surface-level revision factors.

Discussion

This study offers novel insights into the real-world usage of an open-access, generative AI writing feedback tool, providing the first large-scale analysis (n = 23,650) of student engagement in voluntary, self-directed contexts. Learners from various school types and grade levels, particularly those in academic secondary schools, utilized the tool. However, consistent with prior work on engagement with feedback (Jansen, Horbach, and Meyer 2025; Meyer, Jansen, and Fleckenstein 2025), only a small subset (15.5%) engaged in multiple rounds of revision, reinforcing that availability of feedback alone does not ensure uptake, particularly in unsupervised settings where motivation and self-regulation play a key role (Panadero 2017).

Among students who did revise, we observed statistically significant improvements across both content and language dimensions, with the largest gains among initially lower-performing writers. This pattern aligns with prior studies, which show that generative AI feedback can serve as an effective scaffold (Meyer et al. 2024), particularly for students. Most improvements occurred during first revisions, though further gains, especially in higher-order skills such as argumentation and linguistic style, continued across later revisions, albeit more variably. This suggests a design implication for feedback systems: while prompting iterative revision is useful, tools should also monitor when progress plateaus

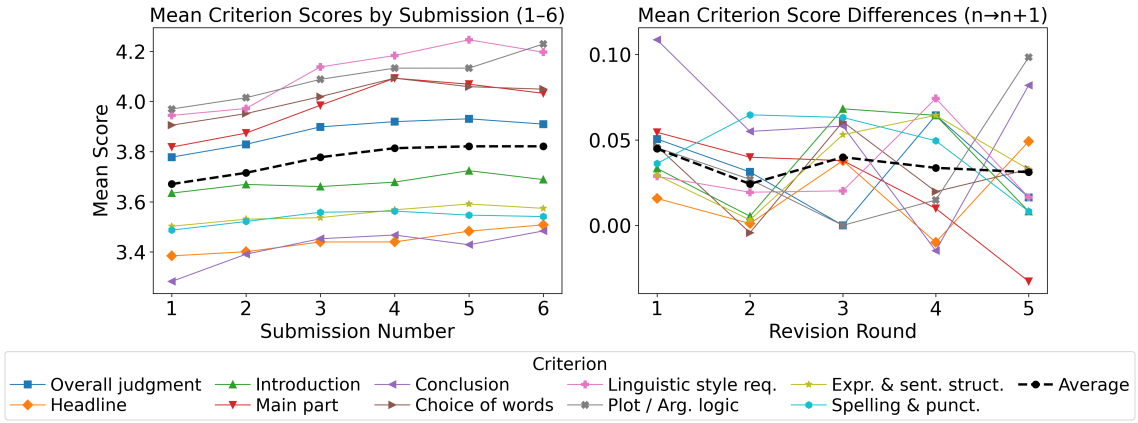

Figure 2: Text Criteria Scores by Submission (left) and Score Gains by Revision Rounds (right)

or high quality is achieved, to help students focus effort without causing unnecessary or demotivating revisions.

Given the scale of our dataset (7,564 revised submissions), we used a validated LLM-based multidimensional scoring method (Seßler et al. 2025). While human ratings are ideal, automated scoring offers a practical, pedagogically aligned alternative at this scale. Additionally, our use of privacy-preserving methods (e.g., text similarity clustering) demonstrates that meaningful insights into revision behavior can be obtained without persistent user tracking, thereby supporting the ethical use of large-scale learning analytics.

From an equity perspective, our data raises critical questions. The current user base appears skewed toward relatively privileged, academically tracked, older students, despite the tool's open-access design. This echoes broader concerns in educational technology about the "inverse equity" effect, where the students who stand to benefit the most are the least likely to engage (Bulathwela et al. 2024). Reaching underrepresented learners, including those with limited digital literacy or access, will require targeted outreach, simplified interfaces, and possibly alternative feedback modalities. For example, younger learners or those with reading difficulties may benefit from speech-based interactions or dialogic feedback formats that reduce the cognitive demands of processing long, text-based suggestions.

Limitations. While promising, these findings must be interpreted with caution. As a non-experimental study, causal claims cannot be made, and improvements may be attributed to other confounding factors, such as teacher input, peer feedback, or learner motivation. Although the evaluation model employed here has been validated in prior work for its alignment with teacher ratings, using LLMs for both feedback generation and assessment introduces the potential for systemic bias or feedback–evaluation entanglement. Additionally, due to anonymized data, revision chains were reconstructed via text similarity, which may introduce classification errors. We were unable to directly assess learner engagement based on the feedback. Future research using behavioral data (e.g., clickstream logs) could offer a stronger basis for interpreting uptake and engagement patterns (Schiller et al. 2024). Finally, although the tool is publicly accessible, its user base is skewed toward students from well-resourced, academically oriented schools and is restricted to Germany. Lacking data on socioeconomic status, digital access, or learner background limits the evaluation of its impact on educational equity. While gains among lower-performing students are promising, more inclusive sampling and outreach are needed to assess and support equity claims.

Future work. Future studies should prioritize rigorous evaluation designs and adaptive feedback strategies that account for learner profiles and engagement patterns. Longitudinal studies will be necessary to determine whether short-term improvements in text quality translate into durable writing competence. New approaches to improving LLM-generated feedback, such as feedback-in-the-loop optimization, have recently been proposed, where feedback is iteratively refined based on simulated student revisions Nair et al. (2024). Such techniques may have the potential to enhance the educational impact of feedback by optimizing its quality, and also its effectiveness in prompting meaningful revision; however, they warrant real-world evaluations. Furthermore, ensuring equitable access across sociocultural contexts and educational systems remains a central challenge.

Conclusion. Despite these limitations, the results offer encouraging evidence for the pedagogical potential of generative AI systems in educational settings. When designed to deliver constructive, curriculum-aligned, and timely feedback, such tools can support learners in engaging with complex writing tasks through iterative self-assessment. The high levels of voluntary engagement and measurable improvements in writing quality suggest that generative feedback tools may serve as a valuable supplement to traditional instruction, particularly in contexts where teacher feedback is scarce or delayed. This study contributes to the growing body of research on socially beneficial AI by demonstrating that carefully deployed generative feedback systems can foster meaningful learner engagement and incremental writing gains beyond formal instruction. These findings support developing inclusive, pedagogically grounded AI tools that broaden access to individualized learning support at scale.

Acknowledgements

We sincerely thank Kathrin Seßler for her foundational work on the first versions of the Feedback Tool PEER.

References

Attali, Y. 2004. Exploring the feedback and revision features of Criterion. *Journal of Second Language Writing*, 14(3): 191–205.

Banihashem, S. K.; Kerman, N. T.; Noroozi, O.; Moon, J.; and Drachsler, H. 2024. Feedback sources in essay writing: peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education*, 21(1): 23.

Brügelmann, H.; and Richter, S. 1994. *Wie wir recht schreiben lernen. 10 Jahre Kinder auf dem Weg zur Schrift*. Lengwil: Libelle.

Bulathwela, S.; Pérez-Ortiz, M.; Holloway, C.; Cukurova, M.; and Shawe-Taylor, J. 2024. Artificial intelligence alone will not democratise education: On educational inequality, techno-solutionism and inclusive tools. *Sustainability*, 16(2): 781.

Dai, W.; Lin, J.; Jin, H.; Li, T.; Tsai, Y.-S.; Gašević, D.; and Chen, G. 2023. Can large language models provide feedback to students? A case study on ChatGPT. In *2023 IEEE international conference on advanced learning technologies (ICALT)*, 323–325. IEEE.

Escalante, J.; Pack, A.; and Barrett, A. 2023. AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, 20(1): 57.

Fleckenstein, J.; Jansen, T.; Meyer, J.; Trueb, R.; Raubach, E. E.; and Keller, S. D. 2024. How am I going? Behavioral engagement mediates the effect of individual feedback on writing performance. *Learning and Instruction*, 93: 101977.

Fleckenstein, J.; Liebenow, L. W.; and Meyer, J. 2023. Automated feedback and writing: a multi-level meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence*, 6: 1162454.

Freedman, S. W.; Hull, G. A.; Higgs, J. M.; and Booten, K. P. 2016. Teaching writing in a digital and global age: Toward access, learning, and development for all. *Handbook of research on teaching*, 5: 1389–1450.

Graham, S.; Hebert, M.; and Harris, K. R. 2015. Formative assessment and writing: A meta-analysis. *The elementary school journal*, 115(4): 523–547.

Hahn, M. G.; Navarro, S. M. B.; Valentín, L. D. L. F.; and Burgos, D. 2021. A systematic review of the effects of automatic scoring and automatic feedback in educational settings. *Ieee Access*, 9: 108190–108198.

Hattie, J.; and Timperley, H. 2007. The power of feedback. *Review of educational research*, 77(1): 81–112.

Holmes, W.; Miao, F.; et al. 2023. *Guidance for generative AI in education and research*. Unesco Publishing.

Jansen, T.; Horbach, A.; and Meyer, J. 2025. Feedback from Generative AI: Correlates of Student Engagement in Text Revision from 655 Classes from Primary and Secondary School. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, LAK '25, 831–836. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0701-8.

Kasneci, E.; Sessler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; Krusche, S.; Kutyniok, G.; Michaeli, T.; Nerdel, C.; Pfeffer, J.; Poquet, O.; Sailer, M.; Schmidt, A.; Seidel, T.; Stadler, M.; Weller, J.; Kuhn, J.; and Kasneci, G. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103: 102274.

Kinder, A.; Briese, F. J.; Jacobs, M.; Dern, N.; Glodny, N.; Jacobs, S.; and Leßmann, S. 2025. Effects of adaptive feedback generated by a large language model: A case study in teacher education. *Computers and Education: Artificial Intelligence*, 8: 100349.

Lo, N.; Wong, A.; and Chan, S. 2025. The impact of generative AI on essay revisions and student engagement. *Computers and Education Open*, 100249.

Meyer, J.; Jansen, T.; and Fleckenstein, J. 2025. Nonengagement and unsuccessful engagement with feedback in lower secondary education: The role of student characteristics. *Contemporary Educational Psychology*, 81: 102363.

Meyer, J.; Jansen, T.; Schiller, R.; Liebenow, L. W.; Steinbach, M.; Horbach, A.; and Fleckenstein, J. 2024. Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, 6: 100199.

Nair, I.; Tan, J.; Su, X.; Gere, A.; Wang, X.; and Wang, L. 2024. Closing the Loop: Learning to Generate Writing Feedback via Language Model Simulated Student Revisions. *arXiv preprint arXiv:2410.08058*.

Panadero, E. 2017. A review of self-regulated learning: Six models and four directions for research. *Frontiers in psychology*, 8: 422.

Ramesh, D.; and Sanampudi, S. K. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3): 2495–2527.

Rashkin, H.; Clark, E.; Huot, F.; and Lapata, M. 2025. Help Me Write a Story: Evaluating LLMs' Ability to Generate Writing Feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 25827–25847.

Schiller, R.; Fleckenstein, J.; Mertens, U.; Horbach, A.; and Meyer, J. 2024. Understanding the effectiveness of automated feedback: Using process data to uncover the role of behavioral engagement. *Computers & Education*, 223: 105163.

Seßler, K.; Kepir, O.; and Kasneci, E. 2024. Enhancing Student Motivation Through LLM-Powered Learning Environments: A Comparative Study. In *European Conference on Technology Enhanced Learning*, 156–162. Springer.

Seßler, K.; Fürstenberg, M.; Bühler, B.; and Kasneci, E. 2025. Can AI grade your essays? A comparative analysis

of large language models and teacher ratings in multidimensional essay scoring. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, LAK '25, 462–472. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0701-8.

Seßler, K.; Xiang, T.; Bogenrieder, L.; and Kasneci, E. 2023. PEER: Empowering Writing with Large Language Models. In Viberg, O.; Jivet, I.; Muñoz-Merino, P. J.; Perifanou, M.; and Papathoma, T., eds., *Responsive and Sustainable Educational Futures*, 755–761. Cham: Springer Nature Switzerland. ISBN 978-3-031-42682-7.

Steiss, J.; Tate, T.; Graham, S.; Cruz, J.; Hebert, M.; Wang, J.; Moon, Y.; Tseng, W.; Warschauer, M.; and Olson, C. B. 2024. Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, 91: 101894.

Wiley, D. A. 2006. Learning objects in public and higher education. In *Innovations in instructional technology*, 1–9. Routledge.