Proceedings of the 8th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2026.36194

Semantic Response Mechanisms of L2 Writing Revision Behavior under Generative AI Feedback

Yu Shi

Sydney School of Education and Social Work, The University of Sydney, Sydney, Australia
lingwadesu@gmail.com

Abstract. Generative AI is changing the way L2 writing feedback is produced and how learners revise their texts. Writing feedback is gradually moving from one way correction to human AI semantic interaction. Based on this change, the study focuses on non English major undergraduates in Chinese universities who have passed CET-4. It builds an experimental process that combines initial drafts, ChatGPT feedback, revised drafts, revision logs, and teacher ratings. Inputlog or Google Docs is used to record the revision process. SBERT is used to calculate semantic distance between initial drafts and revised drafts. Coh Metrix indicators, feedback uptake annotation, and mixed effects models are then used to examine how different feedback types affect writing quality improvement. The results show that grammar and vocabulary feedback are more easily accepted, but their explanatory power for score improvement is limited. Cohesion and argument feedback have lower uptake rates, yet they contribute more clearly to task response, paragraph organization, and coherence. The relationship between semantic distance and writing improvement is nonlinear. Moderate semantic change is most likely to produce effective revision. The findings indicate that the value of generative AI feedback does not depend on revision quantity. It depends on whether learners can make high quality semantic responses while preserving their original writing intentions.

Keywords: Generative AI feedback, L2 writing, revision behavior, semantic response, NLP learning analytics

1. Introduction

The integration of generative AI into L2 writing instruction has shifted feedback from teacher-centered final evaluation to process-oriented human–AI interaction. When learners receive AI suggestions, they do not only correct grammar and vocabulary but also reconsider argument organization, paragraph cohesion, and communicative intention. Research on automated writing evaluation has shown that immediate feedback can promote revision engagement, yet feedback uptake does not directly equal writing development, especially when suggestions involve content restructuring and may lead to selective acceptance or semantic drift [1]. Comparative studies of teacher feedback and automated feedback also indicate that feedback value depends on whether learners understand the problem and complete effective revision [2]. With Chinese university learners who have passed CET-4 as participants, a design combining ChatGPT feedback, revision

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.36194

logs, semantic similarity, textual indicators, and teacher ratings can further explain how AI feedback is transformed into observable semantic response mechanisms.

2. Literature review

2.1. AI feedback contexts

The development of automated writing evaluation has made L2 writing feedback more immediate and scalable, allowing learners to receive more frequent language support beyond teacher comments. Earlier systems mainly focused on grammar, spelling, and local expression, which helped reduce basic errors but offered limited support for argument depth, discourse coherence, and writer intention. Generative AI changes this feedback structure because it can not only identify problems but also provide rewritten expressions, organizational suggestions, and stylistic adjustments, extending feedback from error detection to meaning negotiation. Research on ChatGPT-assisted writing revision shows that learners engage with AI feedback behaviorally, cognitively, and affectively in different degrees [3]. A review of generative AI feedback studies also suggests that AI feedback effectiveness should be judged in relation to task type, feedback level, and learner processing [4]. The core issue in AI feedback contexts therefore lies in how learners distinguish surface-level language correction from higher-level content reconstruction.

2.2. Revision behavior traces

L2 writing revision is not a single process of textual correction but a dynamic behavior composed of deletion, replacement, addition, movement, and restructuring. Traditional final-draft scoring can show quality differences, yet it cannot explain how learners judge, pause, and choose after receiving feedback. Writing process research emphasizes that revision traces reveal learners' attention allocation and cognitive processing, especially because keystroke records, timestamps, and version comparison can capture revision frequency, pause duration, and repeated modification points [5]. Research on the development of L2 writers' revision behavior also shows that revision quality depends not only on the amount of change but also on how deeply learners identify textual problems [6]. In AI feedback environments, process data can map feedback types onto behavioral responses, making it possible to determine whether learners directly accept suggestions, partially rewrite them, extend their own expressions, or bypass AI advice and develop a new textual path.

2.3. Semantic response modeling

Semantic response modeling provides a computational path for explaining revision quality under AI feedback. Counting revision frequency alone may misread mechanical replacement as active revision, while relying only on teacher ratings cannot show the details of meaning change. SBERT sentence embeddings can measure semantic distance between initial and revised drafts, helping identify meaning preservation, moderate extension, and obvious deviation [7]. Coh-Metrix can describe changes in cohesion, readability, and syntactic complexity, allowing semantic revision to be connected with discourse-level textual features [8]. TAACO further strengthens the automatic analysis of textual cohesion and provides quantitative evidence for changes in paragraph organization and local connections [9]. When semantic distance, cohesion indicators, revision logs, and teacher ratings are jointly entered into statistical models, the combined effects of feedback type, semantic-change magnitude, and revision behavior on writing improvement can be examined.

DOI: 10.54254/2755-2721/2026.36194

3. Experimental methods

3.1. Data collection design

The sample includes 80 non English major undergraduates from Chinese universities. All participants have passed CET-4 with scores no lower than 425. CET-4 scores are used only as an entry criterion and a language proficiency control variable. Writing tasks are selected from publicly searchable CET-4 argumentative writing prompts from 2020 to 2024. The public examination information from the National Education Examinations Authority is also used to determine writing time, text length, and basic scoring dimensions. Each student completes two rounds of writing. Each round includes an initial draft, ChatGPT feedback, a revised draft, a revision log, and teacher ratings. A total of 160 initial draft, feedback, and revised draft text chains are obtained. ChatGPT feedback is generated through a unified prompt. The output is limited to five types of suggestions, namely grammar, vocabulary, cohesion, argument, and style. This reduces feedback bias caused by prompt variation. Inputlog or Google Docs version history is used to record deletion, addition, replacement, pause time, and revision sequence. The final dataset includes learner ID, CET-4 score, initial draft text, AI feedback text, revised draft text, revision operation, semantic indicators, and teacher ratings.

3.2. Semantic feature extraction

Semantic feature extraction follows the process of text alignment, feedback classification, semantic calculation, and quality indicator generation. Python is first used for text cleaning, sentence segmentation, and sentence level alignment between initial drafts and revised drafts. Names, student numbers, and irrelevant marks are removed. Two annotators then code AI feedback into five categories, including grammar, vocabulary, cohesion, argument, and style. Learner responses are annotated as uptake, partial uptake, rejection, and semantic drift. Semantic change is measured by calculating the cosine similarity between the SBERT vector of the initial sentence and that of the revised sentence [10]. A value closer to 1 indicates stronger meaning preservation. A lower value indicates a larger degree of semantic adjustment, as shown in Formula 1.

$$\mathrm{Sim}\left(d_i, d_i^{'}\right)=\frac{V(d_i)\cdot V\left(d_i^{'}\right)}{\|V(d_i)\|\|V\left(d_i^{'}\right)\|} \tag{1}$$

Here, d_i refers to the i_{th} sentence segment in the initial draft. $d_i^{'}$ refers to the corresponding revised sentence segment. $V(d_i)$ and $V\left(d_i^{'}\right)$ refer to the sentence vectors generated by SBERT. $\mathrm{Sim}\left(d_i, d_i^{'}\right)$ refers to the semantic similarity between the two sentence segments.

After semantic distance is obtained, the relationship between writing quality improvement, feedback uptake, semantic change, and CET-4 scores is further calculated. The mixed effects model treats individual students as random effects. It controls for differences in learners' baseline proficiency [11]. It also tests whether feedback type and semantic change can explain score improvement after revision, as shown in Formula 2.

$$\Delta \mathrm{Score}_{ij}=\beta_0+\beta_1 \mathrm{Distance}_{ij}+\beta_2 \mathrm{Uptake}_{ij}+\beta_3 \mathrm{Type}_{ij}+\beta_4 \mathrm{CET4}_i+u_i+\varepsilon_{ij} \tag{2}$$

DOI: 10.54254/2755-2721/2026.36194

Here, $\Delta Score_{ij}$ refers to the score improvement of learner i in writing task j. $Distance_{ij}$ refers to semantic distance. $ptake_{ij}$ refers to feedback uptake status. $Type_{ij}$ refers to feedback type. $CET4_i$ refers to the learner's CET-4 score. u_i refers to the individual random effect. ε_{ij} refers to the error term.

3.3. Modeling procedure

The analysis proceeds from raw writing chains to feature extraction and statistical modeling. First, 160 writing chains are organized by matching each AI feedback item with its revised sentence segment; unmatched feedback is classified as non-uptake, while partial uptake is assessed through semantic similarity and manual annotation. Second, essay-level indicators are calculated, including error reduction, semantic distance, revision frequency, pause duration, Coh-Metrix cohesion measures, and teacher rating differences. Inter-rater reliability is tested using ICC, with samples below 0.75 re-rated. Third, mixed-effects regression examines how feedback type, semantic distance, and uptake status predict score improvement, controlling for CET-4 scores. Finally, K-means or hierarchical clustering identifies learner response types based on uptake rate, semantic distance, high-level feedback uptake, and revision pauses.

4. Results

4.1. Feedback uptake patterns

The illustrative pre experiment results show that 1176 items of AI feedback are extracted from 160 writing chains. Grammar and vocabulary feedback account for the largest proportions, with 412 and 296 items respectively. Their uptake rates reach 78.4% and 71.6%. This indicates that learners who have passed CET-4 can identify local language problems more quickly. They also tend to accept low risk revisions. Cohesion and argument feedback include 184 and 156 items respectively. Their uptake rates decrease to 54.9% and 48.7%. However, their average score improvements reach 0.61 and 0.74. These values are higher than the 0.32 improvement linked to grammar feedback. Table 1 shows that high level feedback is more difficult to accept, yet it has a clearer relationship with improvement in task response, coherence, and paragraph organization. The uptake rate of style feedback is 51.6%. Its average score improvement is 0.42. This suggests that expression optimization is useful to some extent. Its effect remains limited if it does not lead to changes in argument structure. Overall, feedback uptake should not be judged only by quantity. Semantic level and revision depth are key variables for evaluating the effectiveness of AI feedback, as shown in Table 1.

Table 1. Feedback type, uptake rate, and writing score improvement

Feedback type	Feedback count	Uptake rate	Average semantic distance	Average score improvement after revision
Grammar feedback	412	78.4%	0.08	0.32
Vocabulary feedback	296	71.6%	0.12	0.38
Cohesion feedback	184	54.9%	0.21	0.61
Argument feedback	156	48.7%	0.27	0.74

DOI: 10.54254/2755-2721/2026.36194

Table 1. (continued)

Style feedback	128	51.6%	0.18	0.42

4.2. Semantic transfer pathways

The semantic transfer results show that the relationship between semantic distance and writing score improvement is not linear. Samples with semantic distance below 0.10 are mainly concentrated in local grammar and vocabulary replacement. Their average score improvement is about 0.29. This indicates that revision remains at the surface level when meaning change is limited. Samples with semantic distance between 0.18 and 0.30 show the highest score improvement, with an average value of 0.68. They usually correspond to argument enrichment, stronger paragraph cohesion, and expression compression. When semantic distance exceeds 0.38, score improvement falls back to about 0.34. Some texts show deviation from the original argument, excessive reliance on AI rewriting, or paragraph level logical breaks. Figure 1 uses hexagonal density distribution and a smoothed response curve to show score changes across different semantic distance ranges. It shows that effective revision is concentrated in the moderate semantic change zone rather than in the extremely low or extremely high change zones, as shown in Figure 1. This result corresponds to Table 1. High level feedback is more likely to become writing quality improvement only when it causes moderate semantic adjustment while preserving the original writing intention.

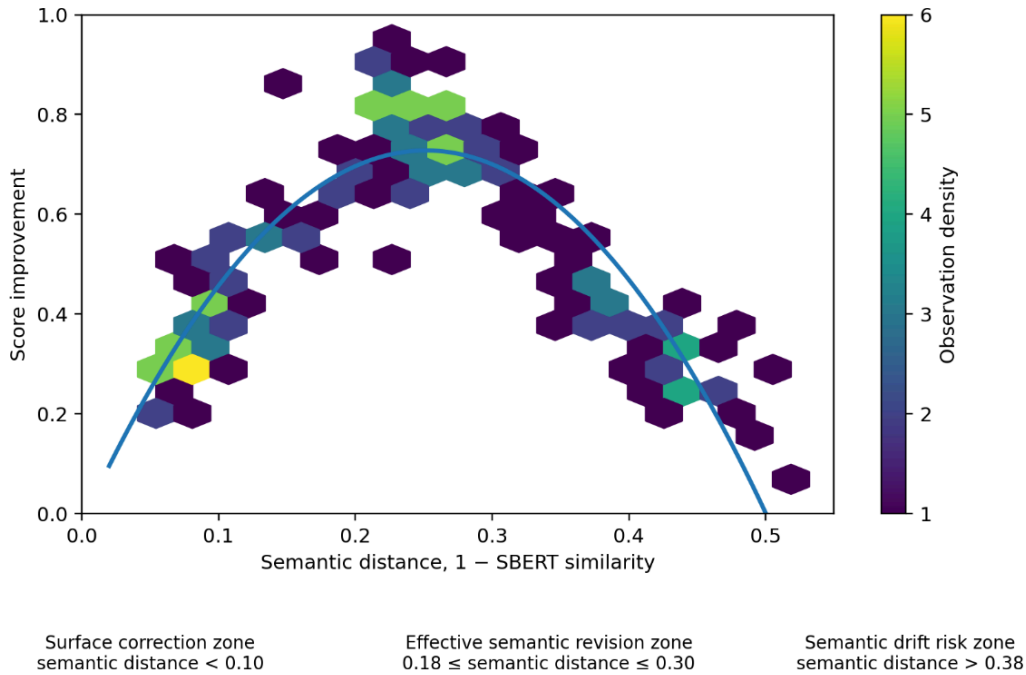


Figure 1. Nonlinear response distribution between semantic distance and writing score improvement

5. Discussion

The results show that learners who have passed CET-4 are able to process basic language feedback, so grammar and vocabulary suggestions are more easily transformed into direct revisions. However, these revisions mostly remain local and have limited influence on overall writing quality. Cohesion and argument feedback require learners to reconsider paragraph relations, idea development, and

DOI: 10.54254/2755-2721/2026.36194

communicative purpose. They are more difficult to accept, but effective transformation of such feedback is more likely to improve scores. The nonlinear pattern of semantic distance further shows that AI assisted revision does not become better when the amount of change increases. Very low semantic change often leads to surface replacement. Very high semantic change may cause deviation from the original meaning. Moderate semantic adjustment reflects learners' ability to understand, select, and reformulate AI suggestions.

6. Conclusion

The study builds an analytical framework for examining L2 writing revision behavior under generative AI feedback. It combines writing texts, AI feedback, revision logs, semantic indicators, and teacher ratings. Using CET-4 passers as participants controls differences in basic English proficiency and fits the Chinese university English teaching context. The results show clear differences in revision quality across feedback types. High level semantic feedback explains writing improvement more effectively than surface language feedback. The combination of SBERT semantic distance, Coh Metrix textual indicators, and mixed effects modeling provides an operational path for identifying learners' semantic response mechanisms. This framework can support AI assisted writing instruction, feedback design, and the development of learners' revision competence.

References

[1] Crosthwaite, P., & Sun, S. (2026). Generative AI and L2 written feedback studies: A scoping review. *RELC Journal, 57*(1), 207–219.
[2] Muñoz, B. C. M., Nassaji, H., & Bello Carrillo, F. I. (2025). ChatGPT-generated versus human direct corrective feedback on L2 writing. *System, 134*, 103805.
[3] Lin, S., & Crosthwaite, P. (2024). The grass is not always greener: Teacher vs. GPT-assisted written corrective feedback. *System, 127*, 103529.
[4] Han, J., & Li, M. (2024). Exploring ChatGPT-supported teacher feedback in the EFL context. *System, 126*, 103502.
[5] Karatay, Y., & Karatay, L. (2024). Automated writing evaluation use in second language classrooms: A research synthesis. *System, 123*, 103332.
[6] Shi, H., & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. *ReCALL, 36*(2), 187–209.
[7] Shi, H., et al. (2025). Comparing the effects of ChatGPT and automated writing evaluation on students' writing and ideal L2 writing self. *Computer Assisted Language Learning*, 1–28.
[8] Chen, Z., et al. (2026). L2 students' barriers in engaging with form and content-focused AI-generated feedback in revising their compositions. *Computer Assisted Language Learning, 39*(3), 715–735.
[9] Zou, S., et al. (2024). Investigating students' uptake of teacher-and ChatGPT-generated feedback in EFL writing: A comparison study. *Computer Assisted Language Learning*, 1–30.
[10] Koltovskaia, S., Rahmati, P., & Saeli, H. (2024). Graduate students' use of ChatGPT for academic text revision: Behavioral, cognitive, and affective engagement. *Journal of Second Language Writing, 65*, 101130.
[11] Yao, Y., et al. (2025). Secondary school English teachers' application of artificial intelligence-guided chatbot in the provision of feedback on student writing: An activity theory perspective. *Journal of Second Language Writing, 67*, 101179.