Proceedings of CONF-MLA 2026 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2026.LD38380

Comparative Analysis of Application Modes and Feedback Effects of Large Language Models in L2 Writing

Jianghan Zhao

Faculty of Liberal Arts and Social Science, Education University of Hong Kong, Hong Kong, China
s1159661@s.eduhk.hk

Abstract. Traditionally, second language(L2)writing feedback has long depended on teachers' manual error correction. In practice, however, typical classroom settings often prevent such teacher-led feedback from being implemented effectively. This is why large language model (LLM)-based artificial intelligence (AI) feedback has rapidly emerged as a widely adopted alternative in recent years. Even so, existing relevant research does not produce consistent and reliable findings. At the same time, newly released AI tools on the market have made many teachers unsure about these tools' selection and practical classroom integration. This literature review divides LLM-generated feedback for L2 writing into four core types: holistic assessment, sentence-by-sentence error correction, structured revision suggestions, and multi-turn interactive dialogue.It analyzes six points: feedback type, learning effectiveness, learner autonomy, learner proficiency, and the strength of empirical research support. The results show that there is no specific feedback mode that can be used in every scenario, because each mode needs to be used with different learner groups. The review also finds some key limitations in existing review work. It provides actionable ways for both teachers and AI tool developers, gives suggestions to build unified criteria for LLM feedback, and notes that in second language writing instruction, there is a need to research how to combine human teacher feedback with LLMs' feedback flexibly.

Keywords: large language models, L2 writing, writing feedback

1. Introduction

1.1. Research background

Feedback is the most crucial part of second language writing. It is not just about correcting students' writing and it also decides how much progress students can make in writing quality. Teacher feedback is undoubtedly useful, but it does not always work well in real classrooms. Because of class sizes and the number of papers, teachers simply do not have enough time and energy to give every student timely and personalized feedback. So, when students receive feedback, it is often delayed and teachers only point out obvious mistakes and miss many key aspects of students' writing. It may interfere with the learning process of students of all levels. Weaker students are often not corrected on basic grammar mistakes. Advanced students receive little advice on rhetorical structure or logical flow. Such gaps are useless to students, slowing their progress in writing

© 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.LD38380

accuracy, text organization, and rhetorical expression. Before the rise of LLMs, Automated Writing Evaluation (AWE) systems like Criterion and Grammarly provided feedback based on rules or templates [1, 2]. But the advent of generative large language models has radically altered the landscape. Since the early 2020s, generative large language models such as GPT-4 and Claude have been transformative for the field of computer-assisted language learning. Big language models can write stuff, identify grammar errors and hold long conversations.

In practical terms, they can produce writing comments, give error analyzes, and customized revision prompts with relatively little time and labor cost. So LLMs are widely regarded as having the potential to compensate for the main shortcomings of manual teacher feedback. In recent years, many empirical studies have explored feedback on L2 writing supported by large language models. However, there are significant differences between the prompt word design and feedback output strategies used in existing research, and the research conclusions are relatively scattered, with some results even contradictory to each other. Some studies have found that large language models can generate rich feedback, whereas there are also studies that warn about AI-related problems such as generating hallucinations and giving vague and general suggestions. At present, there is still a lack of systematic integration research to classify various feedback modes of large language models and compare the feedback quality and actual learning effectiveness of different modes. It remains unclear which feedback mode best suits specific learner groups and classroom scenarios. This study addresses that gap.

1.2. Research questions and scope

The paper answers three core research questions. Firstly, when a Large Language Model supports L2 writing, what feedback modes can be used? Second, what different feedback quality and learning outcomes can be found in different modes? Finally, what teaching goals are proper for each feedback mode? To reach the three research questions above, this review classifies the different modes along two dimensions: granularity and interaction. The review searches peer-reviewed empirical studies published between 2020 and 2026. After searching and screening, 10 studies met all criteria and were included in the final comparative analysis. Non-empirical papers, opinion pieces, and technical reports lacking learner outcome data were all excluded from the sample.

1.3. Research significance and paper structure

In this paper, it theoretically builds a classification framework of mainstream LLM feedback modes, synthesize empirical findings across studies, and continues a computer assisted language learning (CALL) research on LLM-enabled L2 writing. Instead, it discusses the advantages and disadvantages and the most suitable contexts for each feedback model, thus providing L2 teachers with practical advice on how to choose and apply AI feedback tools in school teaching. This paper has a clear and well-organized structure. The second part demonstrates the methods of the review, defines some core concepts, and mentions a few relevant studies. Then the third part constructs a multi-dimensional evaluation framework to compare the four types of AI feedback modes. And then, the fourth part explores the differences across these four modes, points out the limitations of existing research, and proposes concrete methods. The final part summarizes the research findings and future research directions.

DOI: 10.54254/2755-2721/2026.LD38380

2. Literature review

2.1. Review methodology

This literature review takes a look at and analyzes all of the existing empirical studies on LLM-based feedback in the field of L2 writing [3]. It's all about the research published between 2020 and 2026, a period when mainstream LLM feedback tools became widely popular and commonly used in teaching. To set up its theoretical framework, the paper also cites some earlier studies on corrective feedback, but these classic works are not included in the core analysis of this review. It gathered cited sources from five databases (Google Scholar, Web of Science, Scopus, Education Resources Information Center (ERIC), and China National Knowledge Infrastructure (CNKI)). The paper's Boolean search query was ("large language model") AND ("L2 writing") AND ("feedback"). This review used three inclusion criteria: studies needed to address LLM feedback in an L2 writing context, provide empirical evidence regarding feedback quality or learning effectiveness, and appear in peer-reviewed journals. It left out any work that was not based on experiments, like opinion pieces, technical reports without actual experiment results, and any studies that were just about assessment instead of feedback.

2.2. Core theoretical concepts

In writing instruction for L2 learners, feedback refers to the responses learners get to their written texts, with the whole point being to push them to revise and, in the long run, become better writers [4]. A classic way of classifying corrective feedback sorts it into three buckets—direct, indirect, and metalinguistic—and this framework is still very much alive in research on AI feedback [5, 6]. Large language models are deep learning systems that can write, spot errors, and hold back-and-forth dialogue. Their feedback quality owes a lot to prompt design: for one essay, different prompts can produce feedback that varies hugely in detail and what it chooses to focus on [7, 8]. Still, LLMs have their flaws—occasional hallucinations, for instance, correcting things that did not really need correcting—so the sensible move is to treat them as an aid, not as a stand-in for a real teacher [9].

2.3. Overview of global related research

Since around 2022, there has been a notable surge in empirical work on LLM-generated feedback in L2 writing [3]. The literature falls into three broad lines of inquiry. One line pits LLM feedback against teacher feedback: the models turn out to be decent at flagging grammatical and lexical mistakes, but their suggestions on argumentation and rhetorical issues tend to fall short [10]. Another line centers on learners—and while students do seem to value how quickly AI can respond, many of them accept the suggested edits without really thinking them through, which raises concerns about durable learning [11, 12]. Recently, researchers have also divided some feedback modes into holistic assessment, sentence-by-sentence corrections, structured revision suggestions, and multi-turn interactive dialogue. But many of these studies only analyze one mode in a controlled experiment. What research really needs is a classification framework that can compare the advantages and disadvantages of all four different modes. And that is exactly what this review sets out to do.

DOI: 10.54254/2755-2721/2026.LD38380

3. Comparative analysis of LLM feedback modes

So, looking at the research in part 2, it has spotted four main ways large language models (LLMs) give feedback: they do a holistic evaluation, they give sentence-by-sentence corrections, they offer structured revision suggestions, and they engage in multi-turn dialogue. What really makes these differences is how detailed the feedback is, and how learners actually interact with the AI. So the paper then go through each of these modes one by one and see how they stack up against the six evaluative criteria the paper mentioned about earlier.

3.1. Evaluation framework and indicators

So, to really make a fair comparison across different ways that LLM-based feedback can work, the review put together an evaluation framework. It got the idea from previous research in CALL and L2 feedback. The paper is looking at four feedback modes and judging them along six dimensions. The first dimension is granularity. This just means which linguistic unit the feedback operates on, the whole text or just a single word. The second one is feedback type. That includes direct error correction, indirect prompts, and metalinguistic comments. The third aspect is learning effectiveness, and it aims to record the student's progress, such as writing accuracy. The fourth dimension is learner autonomy. It is about figuring out whether students can think independently and make their own judgments. The fifth dimension is learners' language proficiency. The paper splits learners into beginner, intermediate, and advanced groups. The last dimension evaluates the strength of empirical support, and it elaborates on the credibility of the model. In conclusion, these six dimensions help this review get a full picture.

3.2. Holistic evaluation feedback

So, with the holistic feedback mode, the AI focuses on an entire essay and then gives users a general evaluation. It will not point out every single little mistake in the text. Instead, it just comments on how accurate the article topic is, how the paragraphs are organized, and the overall quality of the language. Simultaneously, it points out the highlights of the article and the prior errors that need to be corrected. However, opinions remain divided on this feedback type. For students, they feel it can rapidly help them know their writing level, yet it has no capacity for generating specific revision steps. Advanced learners benefit from these broad suggestions, while weaker learners feel it is too difficult to grasp the comments, so they cannot revise the essay. So far, research on this feedback mode consists mostly of small-scale case studies, lacking some comparative experiments and empirical evidence in real classes. In practice, it is mostly suited for advanced students' writing instruction. It is effective in classes that focus on global organization and rhetorical features such as essay structure and rhetoric features like essay structure and rhetoric, but it is less suitable for fixing tiny grammar mistakes in individual sentences.

3.3. Sentence-by-sentence error correction

Sentence-by-sentence error correction goes through the text line by line, finding and fixing grammatical slips, wrong word choices, and correcting issues along the way. This mode rests on a simple assumption—that getting the details right at the sentence level should come first. The large language model will mark the problems in sentences, provide the modified expression, and give a brief description of each error. Empirical studies show that the error correction mode can effectively

DOI: 10.54254/2755-2721/2026.LD38380

help learners pay attention to surface language errors and has a high adoption rate in modifying grammar and vocabulary problems. However, extensive sentence-level corrections may weaken learners' autonomous learning ability, and students tend to mechanically accept the revisions given by AI without deep thinking. In addition, the large language model occasionally produces incorrect corrections or excessive error correction, which distorts the original intention of learners. This mode is most suitable for junior and intermediate-level learners whose main difficulty is language accuracy. It is also suitable for practical writing tasks, which focus on the appropriateness of grammar and vocabulary.

3.4. Structured revision suggestions

Structured revision suggestions require a large language model to output feedback according to clear, fixed dimensions, such as content, cohesion, argument, grammar, and vocabulary. Unlike simple sentence-level error correction, this mode divides feedback into clear sections. Empirical evidence shows that structured recommendations give consideration to overall guidance and practical operability. Compared with holistic evaluation feedback, the sub-dimensional output can provide learners with clearer revision directions, and learners can find targeted suggestions according to different revision goals. However, this mode still belongs to one-way communication: feedback is pre-generated, and learners cannot ask for further explanation. This mode is suitable for most language learners, as well as for routine writing assignments in the L2 classroom.

3.5. Multi-turn interactive dialogue feedback

Multi-turn interactive dialogue feedback regards the large language model as a conversational writing tutor. This mode no longer provides one-time comments but supports two-way reciprocating communication. Learners can ask for further explanation, question the views of AI, or request to generate other modified versions. This mode is a conversation with the AI: the learner and the model go back and forth, with the AI encouraging the students to keep raising questions. The evidence suggests that this mode truly pushes learners to think on their own; they come to understand why their expression is problematic, instead of just copying the AI's generated content. But the flaws are hard to ignore: it is time-consuming, and it demands a lot of computing power. Moreover, published studies show that there is almost no research on how this mode plays out in real classrooms over the long term. In short, it is best suited for intermediate or advanced learners who want to practice academic writing.

3.6. Cross-mode comparison and synthesis

Based on the above six indicators, the characteristics of the four feedback modes can be summarized as shown in Table 1.

Table 1. Comparison of four LLM feedback modes across six evaluation dimensions

Mode	Granularity	Type	Learning Effectiveness	Autonomy	Suitable Level	Evidence Strength
Holistic Evaluation Feedback	Full-text level	prompts, explanations	Improves global text awareness	one-way judgment,	Intermediate -advanced learners	Mostly small-scale

DOI: 10.54254/2755-2721/2026.LD38380

Table 1. (continued)

Sentence-by-Sentence Error Correction	Sentence and vocabulary level	Mainly direct corrective feedback	High adoption rate for grammar-vocabulary errors	Learners accept AI revisions	Beginner-intermediate learners	small-scale experiments
Structured Revision Suggestions	text-paragraph-sentence	Mixed direct-indirect feedback	Balances global guidance and concrete revision directions	clear suggestions	All proficiency levels	Moderate amount
Multi-turn Interactive Dialogue Feedback	Adjustable, determined by learner-AI conversation	Corrective and meta-linguistic explanations	Outcome depends heavily on learner participation	Learners may question	Intermediate-advanced learners	Limited empirical evidence

So, this study demonstrates that no single way of giving feedback works best all the time. In a real classroom, teachers cannot just pick one type of feedback and expect it to be perfect for every situation. Teachers need to weigh out the good and bad points and choose the right kind of feedback. For instance, if you give feedback that looks at the whole picture (holistic evaluation), it can apparently help students understand the main writing goals. But it will not tell them exactly what to do next. Correcting errors sentence by sentence is great for making sure students' language is accurate. However, students might just become passive recipients, losing their ability to think independently. When it suggests revisions for the structure, those are usually clear and easy to understand. This means students can just follow the instructions to modify the article, but it doesn't allow for a longer chat with the AI. A multi-turn interactive dialogue can really help students think for themselves, but it takes a lot more computing power and time.

4. Discussion

4.1. Trade-offs among the four feedback modes

Having compared the four core feedback modes, this review can answer the three research questions raised in the previous text. This research examines two core dimensions: feedback granularity and the mode of interaction between learners and AI tools. So, it sorts out LLMs' four typical feedback modes in L2 writing, and different modes provide obvious differences in feedback quality and learning effectiveness. Holistic evaluation only provides the strengths and weaknesses of students' papers, but there is a lack of the next revision step. Sentence-by-sentence error correction can easily deal with surface-level errors, but it is easy for learners to foster a sense of dependence on AI tools. Structured revision suggestions reach a balance between instructions and operations because they are clear and user-friendly, but the drawback is a lack of flexibility, failing to support continuous interaction with AI. Multi-turn interactive dialogues are the best at getting students to think actively, but they require considerable time. In conclusion, four modes fit different learner groups. There is

DOI: 10.54254/2755-2721/2026.LD38380

no omnipotent feedback that can suit every scenario. So, it is sensible to choose feedback mode flexibly according to specific teaching goals and students' linguistic proficiencies.

4.2. Limitations of existing research

Current empirical research in this field still has several obvious limitations. Firstly, most research chooses a narrow range of students' samples, so the conclusion cannot be applied in real classroom situations, where students have different backgrounds and uneven linguistic levels. Second, there is not a unified standard to judge the feedback quality generated by LLMs; in other words, current research uses different metrics. Besides, LLMs' feedback itself has technical flaws: sometimes it may create mistakes that do not exist, and even over-correct sentences that do not need any changes. The most important issue is that few studies can track students' skill changes over time in real classroom settings, and nearly all current experiments are conducted over a short period. So, it is unclear how students' writing abilities can develop in the long term when they keep using AI feedback.

4.3. Practical implications

For teachers teaching L2 writing, the first step before using any AI feedback tool is to clarify the core goals of the lesson. Sentence-by-sentence error correction is a proper and stable choice, especially for low-level learners who have weak basic skills, need to consolidate grammar, and accumulate vocabulary. In contrast, advanced learners focus more on improving rhetorical quality and logical thinking, so they may benefit from holistic evaluation or multi-turn interactive dialogue feedback. While AI tool developers should adopt a sensible direction—create an adjustable type of feedback to let teachers change feedback granularity according to real settings. Ultimately, LLMs are just teaching aids and can never replace real teachers. The effective pedagogical approach is to combine AI feedback with teacher manual feedback. At the same time, teachers need to retain full control over grading and guiding the teaching [13].

5. Conclusion

This review sorts out LLMs' four feedback modes in L2 writing and answers three core research questions raised in the introductory section. Each mode has its own strengths and weaknesses, suiting specific learner groups and teaching objectives instead of working equally well for every instructional scenario. Its real value mainly shows in two aspects: it builds a clear classification framework for LLMs' writing feedback and brings together the scattered research results in this fast-growing field, forming a complete evidence chain that integrates disjointed empirical observations from prior studies. Research clearly shows that LLMs are practical auxiliary tools for teachers, rather than replacing teachers in L2 writing instruction. When picking the right way to give feedback in class, teachers really need to think about their students' writing level and what they are trying to achieve with each writing assignment. Of course, this paper is not perfect. It only looks at 10 studies that fit some pretty strict rules, and it still does not have enough long-term experiments that track how well students learn over time in a classroom setting. In the future, the research should probably come up with a consistent way to measure the quality of comments from LLMs. It also needs to include more people from different languages and educational backgrounds. It is great to create a flexible feedback system that can adjust to different teaching needs, and it should definitely test out the "teacher + large model" approach in actual, real-world classrooms to see if it really works. All of

DOI: 10.54254/2755-2721/2026.LD38380

this will help teachers figure out exactly when and how to bring AI feedback into their teaching. Ultimately, that will make teaching more effective and improve how students learn a second language.

References

[1] Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing, 27*, 1–18. https://doi.org/10.1016/j.jslw.2014.10.004

[2] Warschauer, M., & Grimes, D. (2008). Automated writing evaluation in the classroom. *Pedagogies: An International Journal, 3*(1), 22–36. https://doi.org/10.1080/15544800701771580

[3] Craven, L., & Fredrick, D. R. (2026). LLM-generated feedback in L2 writing: A scoping review. *Education Sciences, 16*(8), 1196. https://doi.org/10.3390/educsci16081196

[4] Hyland, F. (2016). Challenges faced by second language doctoral student writers in Hong Kong and their writing strategies. *Australian Review of Applied Linguistics, 39*(2), 158–180. https://doi.org/10.1075/aral.39.2.04hyl

[5] Ellis, R. (2017). Oral corrective feedback in L2 classrooms: What we know so far. In *Corrective feedback in second language teaching and learning* (pp. 3–18). Routledge. https://doi.org/10.4324/9781315621432-2

[6] Bitchener, J., Young, S., & Cameron, D. (2005). The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing, 14*(3), 191–205. https://doi.org/10.1016/j.jslw.2005.08.001

[7] Li, M. (2024). Leveraging ChatGPT for L2 writing feedback and assessment. *International Journal of Computer-Assisted Language Learning and Teaching, 14*(1). https://doi.org/10.4018/ijcallt.360382

[8] Luo, N., Wang, Y., & Zhang, Z. V. (2026). Impact of prompt sophistication on ChatGPT's output for automated written corrective feedback. *ReCALL, 38*(2), 240–254. https://doi.org/10.1017/S095834402510044X

[9] Zhang, Y. (2024). Incorporating ChatGPT as an automated written corrective feedback tool into L2 writing class. *Journal of Language Teaching, 4*(4), 22–34. https://doi.org/10.54475/jlt.2024.024

[10] Heng, B., & Nawas, A. (2026). ChatGPT in EFL academic writing instruction in Asian higher education: A scoping review. *Review of Education, 14*(2), e70213. https://doi.org/10.1002/rev3.70213

[11] Ma, X., Zhang, C., & Lee, I. (2026). L2 learners' engagement with AI-generated feedback on writing. *Assessing Writing, 68*, 101020. https://doi.org/10.1016/j.asw.2026.101020

[12] Zou, S., Guo, K., Wang, J. (2026). Investigating students' uptake of teacher- and ChatGPT-generated feedback in EFL writing: A comparison study. *Computer Assisted Language Learning, 39*(4), 1062–1091. https://doi.org/10.1080/09588221.2024.2447279

[13] Li, M., Han, J., & Taha, G. (2026). *Generative AI for second language writing feedback, assessment, and instruction*. Cambridge University Press. https://doi.org/10.1017/9781009657334