From Teacher to AI-Generated Multimodal Feedback – Investigating the Role of Feedback Agent in Academic Writing Instruction

Izabela Olszak – Jarosław Krajka

DOI: 10.18355/XL.2026.19.04.13

Abstract
This article compares the effectiveness of AI-generated written corrective feedback and teacher-provided multimodal feedback, combining written, voice, and facial cues in an academic writing context. While the integration of AI in educational settings continues to garner increasing attention, empirical investigations that contrast the depth and impact of human versus AI-mediated feedback remain scarce. Most existing studies focus on higher-level learners, leaving the needs of lower-proficiency academic English writers underexplored. This study seeks to evaluate not only the accuracy of each feedback modality but also its influence on learner outcomes, particularly in terms of writing development and perceived pedagogical value. The research involved 40 undergraduate students, divided into two groups: one receiving teacher-provided multimodal electronic feedback (written, voice, and facial) during structured workshops, and the other receiving AI-assisted feedback generated by ChatGPT after prior training by an instructor. The study first scrutinized current Large Language Models for capacity to deliver multimodal feedback, thereby addressing questions of how multimodal AI feedback can be and how it parallels or diverges from human multimodality. By assessing revision quality, learner perception and writing progress over time, this study advocates for a more conscious integration of teacher-generated and AI-generated feedback into a carefully designed writing pedagogy.
Key words: WCF, AI in L2 learning, Writing pedagogy, Multimodal feedback

1. Introduction

In recent years, Generative Artificial Intelligence (GenAI) has become an increasingly prominent tool in language education, opening up novel approaches to providing written corrective feedback (WCF) in response to second language (L2) writing. While GenAI tools promise speed, efficiency and scalability, questions remain about their pedagogical soundness, particularly compared with feedback from experienced human instructors. Current literature highlights that teachers generally outperform AI in terms of feedback accuracy and contextual relevance (Shabara et al., 2024; Er et al., 2024; Makwana, 2024, 2025; Henderson et al., 2025), at least in the perception of learners (Zeevy-Solovey, 2024). However, the practical implications of these findings within writing classroom settings are not yet fully understood.

To address these issues, the present study conducts a comparative analysis of AI-generated and teacher-provided multimodal feedback in an authentic academic setting. The study involved 40 undergraduate students enrolled in an academic writing course, divided into two equal groups. The 'teacher feedback' group (n=20) participated in structured writing workshops and received multimodal feedback from instructors, including written commentary, voice notes and facial cues. The 'AI feedback' group (n=20), on the other hand, received AI-generated feedback via ChatGPT, following training in prompting by the instructor. This design allows for an investigation of not only the relative accuracy of each feedback type but also its impact on student revision practices, writing improvement and perceptions of pedagogical value. By integrating both qualitative and quantitative data analysis procedures, this study contributes to the growing scholarship on the role of AI in writing instruction. It aims to inform best

practices in blending human expertise with AI innovation to support effective, inclusive, and scalable feedback systems for diverse L2 learners (Laurillard, 2024; Henderson et al., 2025).

2. Background to AI-Generated Feedback in the Writing Classroom

Teacher feedback, particularly when delivered multimodally—through written comments, voice explanations, and facial expressions—has been shown to foster deeper student engagement and more effective writing development when compared to Generative AI feedback (ElEbyary & Shabara, 2024; Er et al., 2024). At the same time, studies (e.g., Meyer et al., 2024; Zhang, 2025) indicate that when compared with revising with no feedback, using Large Language Models (LLMs) to deliver timely feedback can positively relate to students' cognitive and affective-motivational outcomes. While AI feedback is immediate and becoming increasingly sophisticated, it may lack the trustworthiness, nuanced judgment, interpersonal cues and emotions that characterize effective "human" pedagogy (Venter et al., 2025). Since certain studies (e.g., Alnemrat et al., 2025) underscore the superiority of AI feedback over teacher-generated one, additional qualities of AI feedback (such as multimodality or emotional load - Alsaiari et al., 2025; Mohammed & Khalid, 2025) need to be exploited to foster motivation to learn and achieve higher learning outcomes. As was proven by Tay (2024), it is not only the technology-rich environment that guarantees effective engagement with AI, but the personal (feedback value and self-efficacy) and the behavioral factors (acting upon feedback and transferring back to school) that make a difference. Thus, feedback needs to be meaningfully integrated into regular language pedagogy, which teacher-generated multimodal feedback is more inclined to do (Shi & Aryadoust, 2024). Pedagogical awareness and fostering critical thinking skills in using automated feedback (both teacher-generated and AI-generated) will enhance learners' writing skills more than exposure to AI-delivered feedback alone (Mekheimer, 2025).

To date, few empirical studies have directly examined how different feedback modalities of LLMs affect student learning, especially in diverse classroom contexts. Moreover, most research in this area has focused on advanced or professional L2 learners, with limited attention paid to undergraduate students still in the early stages of English for Academic Purposes writing competence development.

3. The Study

3.1. Aims of the Study

This study aims to critically examine the pedagogical value and practical implications of integrating AI-generated written corrective feedback (WCF) into writing instruction, with a specific focus on undergraduate learners in authentic classroom settings. While previous research has established that human instructors generally provide more accurate feedback than AI tools such as ChatGPT, there remains a gap in understanding how these two feedback modalities compare in terms of their influence on student learning, engagement and emotional response. Therefore, this study seeks not only to evaluate the accuracy and instructional effectiveness of AI-generated feedback versus teacher-delivered multimodal feedback (written, voice, and facial) but also to investigate how learners perceive, emotionally respond to, and benefit from each approach over time.

To set the stage for the comparative assessment of AI- and teacher-generated multimodal feedback, a pilot systematic investigation of the capacity of major contemporary large language models (ChatGPT, Mistral, Claude, Grok, and Gemini) to

generate multimodal feedback was undertaken. Ten anonymized L2 writing texts representing a range of proficiency levels and typical linguistic errors were selected for the pilot investigation. Feedback prompts were standardized to elicit textual, auditory, and visual feedback where supported.

In the second stage, the main study was conducted with two groups of undergraduate writing students – one group receiving teacher-generated multimodal feedback and the other receiving AI-generated feedback prompted by an instructor. This design allowed for the assessment of both quantitative development in students' writing and qualitative dimensions of their learning experiences, including engagement, emotional response and perceived pedagogical value.

3.2. Participants and the Teaching Context

Participants were 40 first-year undergraduate students enrolled in an English Philology Bachelor of Arts program at a medium-sized public university in eastern Poland. All participants were L2 English learners specializing in either teacher education or interpreter-translator pathways, reflecting a diverse yet academically motivated cohort with a strong focus on language proficiency. Students were enrolled in an academic writing course as part of their core curriculum. They were randomly assigned into two equal treatment groups: one that received teacher-provided multimodal feedback (written, voice, and facial cues) during structured writing workshops, and the other that received AI-generated written corrective feedback via ChatGPT, moderated by an instructor. The intervention was implemented over an 8-week period, during which both groups completed parallel writing tasks designed to assess the impact of each feedback modality on writing development, engagement, and learner perceptions. The purpose of providing each of the two groups with novel multimodal feedback treatment (teacher-generated vs. AI-generated) was to ensure that no students were put at a disadvantage by being deprived of exposure to innovative pedagogy and to remove the effect of novelty on learners' motivation.

3.3. Design, Procedure and Instrumentation

This study employed a quasi-experimental design involving two parallel groups of undergraduate English Philology students in their first year of BA studies, participating in an 8-week academic writing course treatment. The course focused on opinion and descriptive essays and was structured around selected units from *Successful Writing Upper Intermediate* by Virginia Evans (1997), supplemented with instructor-developed materials.

Participants in both groups were instructed to compose short argumentative essays on the topic "*Should students have homework every day?"* The writing task followed the standard tripartite structure of an essay: introduction, main body, and conclusion.

In each phase, students in the AI feedback group were encouraged to interact with an AI tool (ChatGPT) by submitting predefined prompts. This procedure was intended to replicate authentic writing practices while ensuring consistency across participants.

(1) **Introduction**: Students were provided with a model introduction:

"Many people believe that homework is essential for learning. However, in my opinion, students should not have homework every day. Daily homework can cause stress, reduce free time, and may not always lead to better academic results." They were then instructed to request specific types of feedback from the AI, using prompts such as:

- *Check my grammar and spelling*
- *Suggest stronger examples*

- *Help me rewrite the introduction to make it more engaging*
- *Find better linking words for my paragraphs*
- *Paraphrase the introduction by applying one of the following techniques: (a) reference an unusual or striking idea, (b) address the reader directly, or (c) start with a quotation or thought-provoking statement*

(2) **Main body**: Participants were presented with three model body paragraphs: (1) Argument in favor of daily homework; (2) Counterargument concerning negative effects of daily homework; (3) Balanced view proposing a middle-ground solution. For this section, students were instructed to interact with ChatGPT through the following prompts:

- *Check my grammar and spelling*
- *Suggest stronger examples*
- *Help me rewrite the main body to make it more engaging*
- *Find better linking words for my paragraphs*
- *Use more advanced vocabulary and grammar structures*
- *Apply discussion clock ideas – present the arguments from educational, social, and psychological standpoints*

(3) **Conclusion**: A model conclusion was also provided:

"In conclusion, while daily homework can reinforce classroom learning and build essential skills, it also risks causing unnecessary stress and reducing students' enthusiasm for learning. A balanced approach, where homework is assigned thoughtfully and not every day, seems to be the most effective way to support both academic growth and personal well-being. By focusing on meaningful assignments and allowing time for rest, schools can help students achieve success while maintaining a healthy and motivated attitude toward education."

Participants were asked to request further feedback from the AI chatbot by means of the following prompts:

- *Check my grammar and spelling*
- *Suggest stronger examples*
- *Help me rewrite the conclusion to make it more engaging*
- *Find better linking words for my paragraph.*
- *Paraphrase the conclusion by applying one of the following techniques: (a) reference an unusual or striking idea, (b) address the reader directly, or (c) start with a quotation or thought-provoking statement*

In Week 1, all students completed a pre-test opinion essay, after which the two feedback conditions were implemented. Over the following weeks, students engaged in a series of iterative writing workshops, biweekly feedback sessions and guided reflection activities. Each participant from both groups completed three major writing tasks (Essays 1–3), with feedback integrated at each stage to facilitate revision and track improvement. Instructional topics included the structure of academic essays, effective argumentation, the role of feedback in revision and comparison of AI vs. human feedback. Students' perceptions and emotional responses were captured through two

structured surveys (Weeks 4 and 6), while Week 7 featured focus group discussions aimed at exploring learner engagement, confidence, and preference in relation to the feedback received.

Data were collected through a multi-method approach. Writing performance was assessed using standardized rubrics covering *(1) text type requirements, (2) metaorganization, (3) microorganization, (4) content/congruence, and (5) repetition/redundancy*. Surveys and focus groups provided qualitative and quantitative insights into student perceptions, while instructor observations offered contextual data on student engagement and interaction with the feedback. The final phase of the study included a comparative analysis of pre- and post-test writing performance and a research integration session where students reflected on the pedagogical value and implications of AI feedback in academic writing instruction.

To ensure comparability between the two feedback conditions, both teacher-provided multimodal feedback and AI-generated feedback were guided by a common set of structured prompts targeting core features of the essays, such as coherence, grammar and argument development. In the multimodal feedback group, the teacher's focus on multimodal feedback was systematically aligned with these prompts, emphasizing key aspects specified in the writing rubric. Similarly, the AI feedback was generated in response to identical prompts with instructions to elaborate on the same core elements. This approach aimed to minimize variability in feedback content and focus, ensuring that both modalities targeted similar dimensions of students' writing.

4. Results and Findings

4.1 Comparative Evaluation of Multimodality of AI-Generated Feedback

The comparative analysis of interactions with five LLMs revealed clear distinctions in their multimodal capacities across three feedback modalities: text, audio, and visual, as elaborated below:

- **Text feedback.** All models demonstrated the ability to generate high-quality text-based feedback, which remains the most established and consistently available modality across current LLM architectures.
- **Audio feedback.** Three of the models—ChatGPT (GPT-4o), Mistral Voxtral, and Grok 4—were capable of producing audio feedback. In the case of ChatGPT, audio required integration with external text-to-speech (TTS) synthesis. Mistral Voxtral demonstrated particularly strong audio generation capacities but lacked corresponding visual feedback features. Grok 4 and Gemini 2.5 supported fully integrated audio outputs without external reliance. Claude 3.5 Sonnet did not provide native audio support.
- **Visual feedback.** Visual generation capacities varied substantially. Grok 4 and Gemini 2.5 were the only models to provide full multimodal outputs (text, audio, and visuals). Gemini 2.5, in particular, exhibited advanced multimodal reasoning with robust integration across text, image, and even video. ChatGPT offered limited diagrammatic and illustrative support, whereas Claude 3.5 Sonnet demonstrated more effective text-image reasoning, though without audio integration. Mistral Voxtral lacked visual generation capacity altogether.
- **Overall multimodal capacity.** When considered holistically, Grok 4 and Gemini 2.5 displayed the most advanced multimodal integration, supporting seamless generation of text, audio, and visuals. ChatGPT (GPT-4o) showed

strong text-image reasoning, though audio remained dependent on external synthesis, while Claude 3.5 Sonnet and Mistral Voxtral offered partial but non-complete multimodal capabilities.

More specific evaluation results in terms of particular criteria for academic writing assessment are as follows:

1. **Grammar and spelling checks**: the AI tools show relative consistency, reliably detecting basic grammatical correctness. This suggests that current GenAI tools as writing assistants have reached a high level of competence in grammar validation.
2. **Variation in example suggestions**: Each tool prioritizes different types of examples (empirical studies, policy references, cognitive psychology, personal narratives). This indicates that AI tools are influenced by distinct internal datasets and contextual strategies, highlighting the importance of selecting AI tools aligned with the disciplinary focus of the research.
3. **Diverse introductory approaches**: AI tools employ different strategies — rhetorical questions, statistics, narrative hooks, surprising facts, or bold statements. One can be inclined to believe that AI-generated content can cater to a variety of rhetorical styles and audience engagement goals, suggesting a strong potential for AI in creative academic writing.
4. **Linking word variability**: Differences in suggested linking words show that AI tools incorporate unique syntactic preferences. This leads to the conclusion that no single AI tool dominates in optimal flow structuring and multiple tools should be used in combination for richer discourse coherence.
5. **Paraphrasing diversity**: The use of different paraphrasing techniques by the tools suggests a high adaptability of AI in re-framing arguments, offering academic writers varied rhetorical perspectives to enhance originality and reader engagement.

A summary of feedback characteristics of selected Generative AI tools can be found in Table 1 in Appendix 1.

4.2. Multimodal Teacher- vs. AI-Generated Feedback

An initial comparative analysis between the two groups was conducted based on the pre-test opinion essays written by students during the first week of the quasi-experimental treatment. The primary purpose of the analysis was to determine whether any notable differences in writing proficiency existed between the two groups prior to the implementation of the treatment. Establishing baseline equivalence in writing ability was essential to ensure the internal validity of the experimental design. Once confirming that both groups of 20 English philology students demonstrated comparable levels of proficiency at the outset, the researchers could attribute subsequent differences in post-treatment performance to the instructional intervention rather than to pre-existing disparities.

The analysis focused on eight key dimensions of academic writing: (1) compliance with genre requirements, (2) meta-organization (overall structure of the essay), (3) micro-organization (cohesion within paragraphs), (4) content accuracy and congruence with the task, (5) avoidance of repetition and redundancy, (6) overall effect on the reader, (7) layout and formatting, and (8) keeping the expected length (250–350 words). Each criterion was evaluated on a six-point scale (0–5). Specific descriptors were developed for each dimension to ensure objectivity of judgment. For each of the eight assessment

categories, the arithmetic mean (M) to capture individual student scores and standard deviation (SD) to reflect the degree of variability within each group were calculated for both the treatment and control groups.

Table 1. Results of the pre-test opinion essays

Criterion	**Teacher Feedback Group (M / SD)**	**AI Feedback Group (M / SD)**
Genre Requirements	3.70 / 0.39	3.70 / 0.36
Metaorganization	3.62 / 0.45	3.78 / 0.50
Microorganization	3.78 / 0.40	3.75 / 0.42
Content / Congruence	3.62 / 0.43	3.80 / 0.39
Repetition / Redundancy	3.55 / 0.57	3.52 / 0.53
Effect on the Reader	3.62 / 0.48	3.62 / 0.49
Layout	3.75 / 0.40	3.70 / 0.46
Length (250–350 words)	3.70 / 0.47	3.70 / 0.43

The descriptive statistics reveal no substantial disparities between the teacher feedback and the AI feedback groups across any of the assessment dimensions. The average scores in each category were closely aligned, with differences never exceeding 0.18 points. Both groups demonstrated intermediate levels of proficiency, with the strongest performance in *Genre Requirements* and *Microorganization*, and relatively lower consistency in controlling redundancy. The identical scores in *Genre Requirements* and *Length* (3.70) indicate equal mastery of genre conventions and ability to adhere to word count constraints. The control group showed marginally higher scores in *Metaorganization* (3.78 vs. 3.62) and *Content/Congruence* (3.80 vs. 3.62), suggesting slightly better logical sequencing and topical relevance. However, these differences are minimal and fall within the range of normal variance. Scores for *Effect on the Reader, Layout* and *Microorganization* were comparable, implying equivalent rhetorical effectiveness and structural awareness among students. The *Repetition/Redundancy* category exhibited the highest variability in both groups, reflecting a shared area for development regarding lexical diversity and economy of expression.

Overall, the results support the equivalence of the two groups at the outset of the study. No meaningful discrepancies were observed in initial writing performance, thus establishing a solid foundation for assessing the impact of subsequent instructional interventions. The initial homogeneity in writing proficiency strengthens the internal validity of the study and ensures that any observed improvements post-treatment can be more reliably attributed to the feedback method applied (human expertise vs. AI innovation).

The analysis of the final opinion essays produced by students in the teacher feedback group and the AI feedback group after the writing intervention shows the following results:

Table 2. Results of the post-test opinion essays

Criterion	Teacher Feedback Group mean	AI Feedback Group mean	Difference (Treatment-Control Groups)
Genre Requirements	4.95	4.68	+0.27
Metaorganization	4.88	4.38	+0.50
Microorganization	4.86	4.53	+0.33
Content / Congruence	4.78	4.51	+0.27
Repetition/Redundancy	4.80	4.51	+0.29
Effect on the Reader	4.83	4.59	+0.24
Layout	4.78	4.61	+0.17
Length (250–350 words)	4.90	4.66	+0.24
Overall Mean	4.85	4.56	+0.29

The teacher group outperformed the AI feedback group across all evaluated criteria. The most pronounced improvement was seen in *Metaorganization (+0.50)*, indicating enhanced structural organization and logical flow in essays after treatment. Other areas, such as *Microorganization* and *Repetition/Redundancy,* also showed meaningful gains.

While the post-intervention essays demonstrated improvements in several writing dimensions, the present study did not include direct measures of how students integrated received feedback into their revisions. Therefore, the extent to which improvements can be attributed to active feedback utilization remains to be investigated. Future research should incorporate qualitative analyses, such as revision logs, think-aloud protocols, or student interviews, to explicitly assess the degree of feedback incorporation and its relationship with writing development.

The implementation of teachers' multimodal electronic feedback and AI-assisted tools such as ChatGPT has proven influential in students' growth as writers. Feedback from both groups provides a nuanced picture of how the selected feedback modalities contribute to or constrain learning. The following analysis synthesizes their responses, highlighting specific writing features impacted by feedback, the comparative strengths and weaknesses of feedback types, and students' evolving perspectives on academic writing.

Regarding observed improvements in academic writing, the following can be highlighted:

1. **Lexical enrichment and grammatical development**

Students consistently cited vocabulary expansion as a direct result of AI feedback. As one participant stated, *"AI offers a way to improve your vocabulary and add new words to your essay."* Another commented, *"We stopped using over-generalization, started to use richer vocabulary and clearer arguments,"* pointing to both lexical depth and semantic precision.

A vast majority of students from the treatment group underlined that their grammatical competence improved. They highlighted that AI *"gave us suggestions that included advanced constructions and grammar,"* which refined their essays beyond basic correctness. Teacher feedback supported this process further by offering individualized guidance: *"Our teacher's feedback has been very beneficial… It was personalized and helped each of us notice our specific mistakes."*

2. Clarity and structural coherence

The impact of feedback on structural clarity was widely reported. *"ChatGPT has helped with making the essay more concise and clear by shortening it and getting rid of extra words,"* as one student noted. Another emphasized improved paragraphing: *"It helped differentiate between the paragraphs, arguments, etc."*

Several students acknowledged a clear shift toward structured development of ideas, stating that they now *"pay more attention to using multiple structured paragraphs with clear development of ideas."*

3. Argumentation and academic awareness

AI support was particularly helpful in overcoming cognitive or topical blocks: *"It helps when it comes to topics that you have problems with."* However, deeper development of argumentation was more strongly linked to teacher feedback. *"Teachers can help in explaining the grammar and suggesting better ways to write because they are professionals,"* one participant noted, adding that *"they can also give personal tips that help students to improve step by step."*

Students also demonstrated growing academic awareness. *"We also learned that academic writing vastly differed from what we were taught in high school,"* indicating that feedback not only corrected but also reoriented their conceptual understanding of academic writing conventions.

More specific responses on particular types of feedback can be found in Table 3 below.

Table 3. Students' responses on the type of feedback given

Type of feedback	Participants' opinions (authentic feedback)
Direct	*- Very helpful because the mistake is clearly pointed out.* *- Easiest to remember and helps me avoid repeating the same errors.* *- Most productive, as it focuses on specific mistakes.* *- Clear and not confusing; tells me exactly what to fix.* *- Helps me improve faster because I don't waste time guessing.* *- Best option for me, as it's legible and leaves no doubt about what's right/wrong.* *- Useful when combined with metalinguistic feedback, as it also provides explanation.* *- Helps me memorize examples where I made mistakes.*
Indirect	*- Encourages me to think more on my own.* *- Makes me rethink and reflect on my answer.* *- Stimulates the brain and helps shape ideas about the correct answer.* *- Leads to deeper thought about what's right.* *- Allows me to actually understand mistakes while correcting them.* *- Gives me a chance to figure out the right answer by myself.* *- Not always helpful, as it can be confusing when I can't spot the mistake after looking at my work for a long time.*

Metalinguistic	*- Most effective when combined with direct feedback: I get the right form and an explanation.* *- Provides key information and rules so I avoid making the same mistake again.* *- Explains why something is wrong, not just what is wrong.* *- Helps me understand grammar rules, vocabulary use, or structures better.* *- Forces me to think and correct my own work, which is the best way to learn.* *- Applies to more examples in the future, making learning transferable.* *- I prefer feedback that gives the corrected form and the explanation – something between metalinguistic and direct.*

5. Discussion

In terms of the comparative analysis of feedback modalities, the results indicate that teacher feedback emerged as the most valued and impactful modality. It was praised for being *"authentic as it comes from someone who can understand the student."* Moreover, students appreciated its alignment with academic standards: *"It also showed us how we will be graded in future exams, which will help us in the future."* However, students did recognize its limitations. *"Teachers are real people, so they take a lot longer to go through essays,"* noted some participants, and others observed that *"some teachers are incapable of explaining things clearly and concisely."* These comments suggest that while teacher feedback is pedagogically rich, its effectiveness can be constrained by logistical and communicative issues.

AI-assisted feedback, on the other hand, was praised for immediacy and clarity. As one student remarked, *"AI gave us suggestions highlighting that our constructions were grammatically correct but could be revised to make our writing refined."* Others noted its structural support as follows: *"AI pointed out that while our sentences were grammatically correct, they could be refined for clarity and style."* Despite these strengths, students acknowledged that AI *"lacks a personal touch while making basic mistakes"* and *"is not always that precise."* These comments reinforce the idea that AI can offer surface-level refinement but lacks the contextual sensitivity of human feedback. Finally, some study participants highlighted peer feedback as well and described it as emotionally supportive and *"less harsh,"* but ultimately less professional. As one respondent put it, *"Peer feedback is usually more friendly… but it is not professional."* Students valued it as a complementary source, not a replacement. *"Peer feedback can serve as a good second opinion, even if it is not coming from a professional."*

The results of the study highly support the need for implementing an integrated feedback model in the development of students' writing competence. Many participants advocated for a blended approach that combines the strengths of all three feedback sources. One participant articulated this clearly: *"The best way to learn may be to combine all three of these methods, AI for quickness, peers for emotional support, and teachers for professional advice."* Such an integrated feedback model allows for: (1) AI to offer immediate micro-level corrections, (2) peer feedback to provide emotional engagement and informal review, and (3) teacher feedback to deliver macro-level, pedagogically grounded critique.

Additionally, a notable meta-theme was the development of confidence and self-awareness as academic writers in the process of the quasi-treatment employed. Students expressed a sense of ownership over their learning, stating that *"We have become more confident writers as we are aware of our flaws and try to avoid them."* Others echoed this growth, noting that feedback helped them *"organize paragraphs more effectively"* and *"understand [their] weaknesses."*

6. Final Conclusions

Taking into account both parts of the study, the current findings corroborate the claims made by Laurillard (2024) and Henderson et al. (2025) of the need to create a rich feedback environment for writing instruction which brings together multimodality, learner-sensitivity and emotional richness of teacher-generated feedback with accuracy and immediacy of AI-generated one. While the current limitations of Generative Artificial Intelligence tools for the provision of multimodal feedback indicated in the comparative phase of the present research are more than likely to be overcome soon, the predicted availability of enhanced feedback opportunities only in paid versions makes the sole use of AI for feedback delivery rather dubious. Therefore, writing instructors should be aware of the opportunities, limitations and effective prompting strategies to put LLMs into feedback generation and delivery, while at the same time be careful enough to merge teacher-generated multimodal feedback with automatic correction of certain selected areas of writing competence for the most effective writing skills development.

AI Use Declaration

The authors declare that no generative artificial intelligence (AI) tools were used in the preparation of this manuscript.

Bibliographic references

Alnemrat, A., Aldamen, H., Almashour, M., Al-Deaibes, M., & AlSharefeen, R. (2025). AI vs. teacher feedback on EFL argumentative writing: A quantitative study. Frontiers in Education, 10, 1614673. https://doi.org/10.3389/feduc.2025.1614673

Alsaiari, O., Baghaei, N., Lahza, H., Lodge, J. M., Boden, M., & Khosravi, H. (2025). Emotionally enriched AI-generated feedback: Supporting student well-being without compromising learning. Computers & Education, 239, 105363. https://doi.org/10.1016/j.compedu.2025.105363

ElEbyary, K., & Shabara, R. (2024). Chat-GPT corrective feedback: Does it do what it says on the tin? Teaching English with Technology, 24(3), 68–89. https://doi.org/10.56297/vaca6841//BFFO7057/MYEH4562

Er, E., Akçapınar, G., Bayazıt, A., Noroozi, O., & Banihashem, S. K. (2024). Assessing student perceptions and use of instructor versus AI-generated feedback. British Journal of Educational Technology, 56(3), 1074-1091. https://doi.org/10.1111/bjet.13558

Henderson, M., Bearman, M., Chung, J., Fawns, T., Buckingham Shum, S., Matthews, K. E., & de Mello Heredia, J. (2025). Comparing Generative AI and teacher feedback: student perceptions of usefulness and trustworthiness. Assessment & Evaluation in Higher Education. Advance online publication. https://doi.org/10.1080/02602938.2025.2502582

Laurillard, D. (2024). How AI-generated feedback could assist student learning. HEPI Blog, 22 October 2024. https://www.hepi.ac.uk/2024/10/22/how-ai-generated-feedback-could-assist-student-learning/

Makwana, R. (2024). Comparative analysis of teacher-provided and ChatGPT-generated written corrective feedback in ESL writing. System, 122, 103117. https://www.sciencedirect.com/science/article/pii/S0346251X24003117

Makwana, R. (2025). The effectiveness of AI-powered feedback compared to teacher-led feedback on ESL students' academic writing development. Journal of English Language Teaching, 67(1), 102–118. https://journals.eltai.in/jelt/article/view/JELT670102
Mekheimer, M. (2025). Generative AI-assisted feedback and EFL writing: a study on proficiency, revision frequency and writing quality. Discover Education, 4, 170. https://doi.org/10.1007/s44217-025-00602-7
Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & 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. https://doi.org/10.1016/j.caeai.2023.100199
Mohammed, S. J., & Khalid, M. W. (2025). Under the world of AI-generated feedback on writing: mirroring motivation, foreign language peace of mind, trait emotional intelligence, and writing development. Language Testing in Asia, 15(7). https://doi.org/10.1186/s40468-025-00343-2
Shabara, R., ElEbyary, K., & Boraie, D. (2024). Teachers or ChatGPT: The issue of accuracy and consistency in L2 assessment. Teaching English with Technology, 24(2), 71–92. https://doi.org/10.56297/vaca6841/LRDX3699/XSEZ5215
Shi, H., & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. ReCALL, 36(2), 187–209. https://doi.org/10.1017/S0958344023000265
Tay, H. Y. (2024). AI feedback: Moving beyond the hype to integrating it into a coherent feedback pedagogy. Routledge Open Research, 3, 26. https://doi.org/10.12688/routledgeopenres.18443.1
Venter, J., Coetzee, S. A., & Schmulian, A. (2025). Exploring the use of artificial intelligence (AI) in the delivery of effective feedback. Assessment & Evaluation in Higher Education, 50(4), 516–536. https://doi.org/10.1080/02602938.2024.2415649
Zeevy-Solovey, M. (2024). EFL students' perceptions and preferences regarding AI (ChatGPT), peer, and teacher feedback. Technology in Language Teaching & Learning, 6(2), 1482–1498. https://www.castledown.com/journals/tltl/article/view/tltl.v6n2.1482
Zhang, K. (2025). Enhancing critical writing through AI feedback: A randomized control study. Behavioral Sciences, 15(5), 600. https://doi.org/10.3390/bs15050600

Words: 5407
Characters: 38 719 (21 standard pages)

Izabela Olszak
John Paul II Catholic University of
Lublin
Poland
izabela.olszak@kul.pl
ORCID: 0000-0002-8504-7814

Jarosław Krajka
Maria Curie-Sklodowska University,
Lublin
Poland
jaroslaw.krajka@mail.umcs.pl
ORCID: 0000-0002-4172-9960

Appendix 1. Comparison of selected GenAI tools for feedback characteristics

LLM Model	Strengths	Weaknesses
ChatGPT / GPT-4o	Balanced, polished, adaptive. Good depth, explanation, and clarity.	Sometimes too "safe" or neutral; less bold in voice.
Mistral Voxtral	Sharp, elegant, persuasive. Better stylistic flair.	Could be less cautious (occasional overstatement).
Claude 3.5 Sonnet	Thoughtful, moderate, well-balanced tone.	May be conservative or less dramatic.
Grok 4	Direct, punchy, clear. Good for persuasive essays.	Might gloss over nuance or subtlety.
Gemini 2.5	Accessible, reader-friendly, supportive voice.	Sometimes less formal, maybe less "academic" sounding.

Appendix 2. Capabilities of selected Large Language Models for multimodal feedback delivery

No.	LLM Model	Text feedback	Audio feedback	Visual feedback	Notes on multimodal capacity
1.	**ChatGPT (GPT-4o)**	yes	yes, via TTS	yes, but limited diagrams	advanced text-image integration; audio requiring external synthesis
2.	**Mistral Voxtral**	yes	yes	no	strong audio capabilities; limited visual support
3.	**Claude 3.5 Sonnet**	yes	no	yes	effective text-image reasoning; no native audio
4.	**Grok 4**	yes	yes	yes	fully multimodal; text, audio, and visual outputs supported
5.	**Gemini 2.5**	yes	yes	yes	advanced multimodal reasoning; strong integration of text, image, and video