Language Teaching (2026), **59**, 122–127
doi:10.1017/S0261444825000199

RESEARCH IN PROGRESS

AMERICAN ASSOCIATION FOR APPLIED LINGUISTICS CONFERENCE, DENVER 2025 SELECTED POSTER PRESENTATIONS

Comparing AI and human feedback at higher education: Level appropriateness, quality and coverage

Yusuf Cengiz[1] and Nur Yigitoglu Aptoula[2]

[1]School of Foreign Languages, Boğaziçi University, Istanbul, Türkiye and [2]Department of Foreign Language Education, Boğaziçi University, Istanbul, Türkiye
Corresponding author: Yusuf Cengiz; Email: yusuf.cengiz@bogazici.edu.tr

(Received April 20 2025; accepted April 21 2025)

1. Introduction

Feedback is recognized as a vital component of second language (L2) writing due to its role in bridging the gap between current and desired performance levels (Hattie et al., 2021; Hattie & Timperley, 2007; Lee, 2017). Corrective feedback practices in L2 writing have been demonstrated to be effective in fostering a deeper understanding of language structures (Bitchener, 2008; Link et al., 2022; Van Beuningen, 2010) and lexical units, as well as improving the contextual (Kang, 2020) and organizational (Yu & Lee, 2014) aspects of written texts.

Although traditional feedback practices have been investigated for many years (Ngo et al., 2024), the use of automated tools for writing feedback has gained significant attraction with the introduction of Large Language Models (LLMs) (Steinert et al., 2024; Venter et al., 2024), such as ChatGPT, Gemini, and Copilot. These generative Artificial Intelligence (AI) tools have introduced a paradigm shift within the field of education (O'Dea, 2024) and their uses in corrective feedback in writing are being investigated in recent literature.

Research suggests that integrating AI tools in L2 writing processes can facilitate immediate and personalized (Kasneci et al., 2023; Venter et al., 2024) feedback opportunities, while also supporting the development learners' self-efficacy (Nazari et al., 2021), analytical thinking, and critical thinking skills (Han & Li, 2024; Li et al., 2024). However, issues have been raised regarding the consistency (Han & Li, 2024) and comprehensiveness (Liu & Chang, 2024) of AI feedback in L2 writing. These tools are susceptible to generating fabricated (Win Myint et al., 2024), oversimplified, or inaccurate responses (Al-khresheh, 2024). Although certain studies have compared human feedback to AI feedback to investigate the extent of its consistency, comprehensiveness, and correctness (Lin & Crosthwaite, 2024; Steiss et al., 2024), scant research has examined these dynamics within authentic classroom settings or made use of the prompt engineering opportunities while comparing AI feedback to human feedback.

A reproduction of the poster discussed is available in the supplementary material published alongside this article on Cambridge Core

© The Author(s), 2025. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike licence (http://creativecommons.org/licenses/by-nc-sa/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the same Creative Commons licence is used to distribute the re-used or adapted article and the original article is properly cited. The written permission of Cambridge University Press must be obtained prior to any commercial use.

https://doi.org/10.1017/S0261444825000199 Published online by Cambridge University Press

To investigate the appropriateness, quality, and coverage of AI feedback when compared to human feedback given in real-life classroom setting, this study aims to answer the following research questions:

1. What is the perceived quality of AI feedback compared to human feedback, as evaluated by teachers?

2. How does AI feedback compare to human feedback in term of coverage of content, grammar, vocabulary, organization, and mechanics?

2. Methodology

2.1. Context and participants

The study was conducted at the pre-departmental, intensive English program of a state university in Istanbul, Türkiye. Three English instructors and 56 students participated in the study. The students were 18-year-old Turkish learners of English. Based on CEFR (Council of Europe, 2020) standards, 10 of the students had A1-level English proficiency, 18 students had A2-level proficiency, while the remaining 28 students had B1-level proficiency. The teachers were aged between 26 and 30, and their years of experience ranged between 2 and 7.

2.2. Data collection

The data was collected throughout the 2024 fall term of the program. As part of their program, the students were required to write a paragraph, and at later stages an essay, on the given topic in a biweekly manner. The total number of paragraphs and essays produced by students was 117.

The instructors provided feedback on the assignments during their office hours. ChatGPT 4o (OpenAI., 2025) was used for AI feedback. The texts were submitted to ChatGPT 4o using a Python script. The AI prompt included the definition of the context, the task, feedback areas, a sample text, and some sample feedback for the text.

2.3. Data analysis

Using the error analysis categories developed by Ferris (2006) and Ene and Upton (2014), all human and AI feedback instances were coded by the researchers based on their area of focus, that is, grammar, vocabulary, organization, content, or mechanics.

To answer **RQ1**, each AI feedback instances were investigated to determine whether (i) they were correct, and (ii) they were appropriate and necessary for the level of the students. In order to answer **RQ2**, the remaining correct AI feedback instances were juxtaposed with human feedback to determine overlapping and distant feedback instances according to each feedback area.

3. Results

3.1. AI feedback quality and level appropriateness

In total, ChatGPT provided 1986 feedback instances for all the texts. Table 1 provides the feedback areas of correct, incorrect, and unnecessary instances.

In 1750 of these instances, or 88.11% of all instances, ChatGPT correctly identified the issues with the text and provided level-appropriate and useful feedback. In 126 feedback instances (6.35%), ChatGPT marked a correct student production as incorrect or provided incorrect feedback to an erroneous use, while 110 instances (5.54%) included unnecessary feedback which was correct but was either beyond the students' current proficiency levels, or was not necessary to implement into the texts.

https://doi.org/10.1017/S0261444825000199 Published online by Cambridge University Press

Table 1. Feedback areas for correct, incorrect, and unnecessary AI feedback instances

						Total	
Feedback type	Grammar	Vocabulary	Organization	Content	Mechanics	Total	%
Correct	1008	291	68	76	307	1750	88.11
Incorrect	53	15	21	31	6	126	6.35
Unnecessary	15	25	40	29	1	110	5.54

Table 2. Matches between human and correct AI feedback based on feedback areas

	Human feedback		AI feedback		Number of matches	
Feedback area	Total	%	Total	%	Total	%
Grammar	1005	57.40	1008	57.60	654	65.07
Vocabulary	313	17.88	291	16.63	160	51.12
Content	116	6.62	76	4.34	27	23.28
Organization	75	4.28	68	3.89	17	22.67
Mechanics	183	10.45	307	17.54	124	67.76
Total	1751*	100*	1750	100	985	56.29

Note: Matching percentage is calculated by dividing the total number of matches by the total number of human feedback and multiplying by 100 for each feedback area.
*Human feedback includes 59 (3.37%) instances of personal remarks of the teachers, which could not be classified under these categories and thus were not included in the table or analysis.

3.2. Comparing AI and human feedback

The teachers in this study provided 1751 instances of feedback for the whole dataset, while the number of correct ChatGPT feedback instances was 1750. The analysis has shown that ChatGPT matched 985 (56.29%) human feedback instances. Table 2 provides the statistics regarding these matches and feedback areas.

The highest alignment between AI and human feedback was found in mechanics-related issues – such as punctuation, capitalization, and spelling – with AI covering 67.76% of human feedback. Grammar feedback showed a 65.07% match, while vocabulary issues aligned at 51.12%. Among the subcategories, the highest overlaps were observed in capitalization (100%), word forms (86.96%), spelling (84.76%), subject-verb agreement (84.06%), and idiomatic usage (81.82%).

Conversely, the lowest alignment was observed in content (23.28%) and organization (22.67%). Apart from these categories, the lowest matches occurred in the use of pronouns (35.29%), punctuation (35.82%), and word choice (37.80%).

3.3. AI feedback without corresponding human feedback

While the overall alignment between human and AI feedback was 56.29%, the remaining 858 instances – representing 49.08% of all accurate AI feedback – addressed errors not identified by the teachers. Table 3 presents the total number of these AI feedback instances where there were no corresponding human feedback AI feedback instances by category.

4. Discussion and implications

Even though concerns have been raised in terms of the quality, accuracy (Steiss et al., 2024), and consistency (Lin & Crosthwaite, 2024) of AI-generated feedback in L2 writing, our study revealed that ChatGPT 4o can produce highly accurate feedback, as 88.11% of all AI feedback instances

https://doi.org/10.1017/S0261444825000199 Published online by Cambridge University Press

Table 3. Correct AI feedback without corresponding human feedback

Feedback area	The number of feedback instances without teacher focus	Percentage (%)
Grammar	443	51.64
Vocabulary	149	17.36
Content	41	4.78
Organization	37	4.31
Mechanics	188	21.91
Total	858	100

correctly addressed the issues and only 11.98% were classified as inaccurate or unnecessary. Our findings align with those of the literature, showing that AI tools can effectively provide corrective feedback (Jamshed et al., 2024), particularly when context relevant prompts are utilized (Venter et al., 2024).

Providing feedback within a classroom setting can be challenging, exhausting, and time-consuming for L2 teachers (Dikli & Bleyle, 2014; Lee, 2019; Wilson et al., 2014). Given that ChatGPT 4o aligned with 56.29% of human feedback instances and 858 other instances that the teachers had not addressed, covering 49.08% of all correct AI feedback instances, it can be suggested that ChatGPT 4o provides a valuable feedback mechanism through which it can meaningfully complement and support human feedback practices (Han & Li, 2024; Lin & Crosthwaite, 2024; Xue, 2024). The data further indicate that ChatGPT 4o could provide extensive feedback on mechanics and grammar. By utilizing AI to identify surface-level errors, the efforts of the language teachers could be directed towards more attention-demanding language related issues, such as organization and content (Link et al., 2022; Steinert et al., 2024), where AI feedback showed limited alignment in this study.

This study has certain limitations. Student evaluations of ChatGPT 4o's feedback are needed to assess comprehensibility and perceived usefulness of the feedback. Additionally, potential biases may have influenced the classification of "unnecessary feedback" instances, as the decision was based on the teachers' evaluation of the feedback according to their own teaching preferences. Moreover, despite prompt optimization, ChatGPT remains susceptible to generating different responses even with the same texts and prompts (Lin & Crosthwaite, 2024).

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0261444825000199.

References

Al-khresheh, M. H. (2024). Bridging technology and pedagogy from a global lens: Teachers' perspectives on integrating ChatGPT in English language teaching. *Computers and Education: Artificial Intelligence*, *6*, 100218. https://doi.org/10.1016/j.caeai.2024.100218

Bitchener, J. (2008). Evidence in support of written corrective feedback. *Journal of Second Language Writing*, *17*(2), 102–118. https://doi.org/10.1016/j.jslw.2007.11.004

Council of Europe. (2020). *Common European Framework of Reference for Languages: Companion Volume*. Cambridge University Press.

Dikli, S., & Bleyle, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing*, *22*, 1–17. https://doi.org/10.1016/j.asw.2014.03.006

Ene, E., & Upton, T. A. (2014). Learner uptake of teacher electronic feedback in ESL composition. *System*, *46*, 80–95. https://doi.org/10.1016/j.system.2014.07.011

Ferris, D. (2006). Does error feedback help student writers? New evidence on the short- and long-term effects of written error correction. In F. Hyland, & K. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 81–104). Cambridge University Press. https://doi.org/10.1017/CBO9781139524742.007

https://doi.org/10.1017/S0261444825000199 Published online by Cambridge University Press

Han, J., & Li, M. (2024). Exploring ChatGPT-supported teacher feedback in the EFL context. *System*, *126*, 103502. https://doi.org/10.1016/j.system.2024.103502

Hattie, J., Crivelli, J., Van Gompel, K., West-Smith, P., & Wike, K. (2021). Feedback that leads to improvement in student essays: Testing the hypothesis that "Where to Next" feedback is most powerful. *Frontiers in Education*, 6. https://doi.org/10.3389/feduc.2021.645758

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Jamshed, M., Ahmed, A. S. M. M., Sarfaraj, M., & Warda, W. U. (2024). The impact of Chatgpt on English language learners' writing skills: An assessment of AI feedback on mobile. *International Journal of Interactive Mobile Technologies (Ijim) 18*(19), 18–36. https://doi.org/10.3991/ijim.v18i19.50361

Kang, E. Y. (2020). Using model texts as a form of feedback in L2 writing. *System*, *89*, 102196. https://doi.org/10.1016/j.system.2019.102196

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., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Lee, I. (2017). *Classroom writing assessment and feedback in L2 school contexts*. Springer. https://doi.org/10.1007/978-981-10-3924-9

Lee, I. (2019). Teacher written corrective feedback: Less is more. *Language Teaching*, *52*(4), 524–536. https://doi.org/10.1017/S0261444819000247

Li, X., Jiang, S., Hu, Y., Feng, X., Chen, W., & Ouyang, F. (2024). Investigating the impact of structured knowledge feedback on collaborative academic writing. *Education and Information Technologies*, *29*(14), 19005–19033. https://doi.org/10.1007/s10639-024-12560-y

Lin, S., & Crosthwaite, P. (2024). The grass is not always greener: Teacher vs. GPT-assisted written corrective feedback. *System*, *127*, 103529. https://doi.org/10.1016/j.system.2024.103529

Link, S., Mehrzad, M., & Rahimi, M. (2022). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning*, *35*(4), 605–634. https://doi.org/10.1080/09588221.2020.1743323

Liu, Y., & Chang, P. (2024). Exploring EFL teachers' emotional experiences and adaptive expertise in the context of AI advancements: A positive psychology perspective. *System*, *126*, 103463. https://doi.org/10.1016/j.system.2024.103463

Nazari, N., Shabbir, M. S., & Setiawan, R. (2021). Application of Artificial Intelligence powered digital writing assistant in higher education: Randomized controlled trial. *Heliyon*, *7*(5), e07014. doi:https://doi.org/10.1016/j.heliyon.2021.e07014

Ngo, T. T.-N., Chen, H. H.-J., & Lai, K. K.-W. (2024). The effectiveness of automated writing evaluation in EFL/ESL writing: A three-level meta-analysis. *Interactive Learning Environments*, *32*(2), 727–744. https://doi.org/10.1080/10494820.2022.2096642

O'Dea, X. (2024). Generative AI: Is it a paradigm shift for higher education? *Studies in Higher Education*, *49*(5), 811–816. https://doi.org/10.1080/03075079.2024.2332944

OpenAI. (2025). *ChatGPT 4o* (February 2025 version). https://chat.openai.com/chat

Steinert, S., Avila, K. E., Ruzika, S., Kuhn, J., & Küchemann, S. (2024). Harnessing large language models to develop research-based learning assistants for formative feedback. *Smart Learning Environments*, *11*(1), 62. https://doi.org/10.1186/s40561-024-00354-1

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, *91*, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Van Beuningen, C. (2010). Corrective feedback in L2 writing: Theoretical perspectives, empirical insights, and future directions. *International Journal of English Studies*, *10*(2), 1. https://doi.org/10.6018/ijes/2010/2/119171

Venter, J., Coetzee, S. A., & Schmulian, A. (2024). Exploring the use of artificial intelligence (AI) in the delivery of effective feedback. *Assessment and Evaluation in Higher Education*, 1–21. https://doi.org/10.6018/ijes/2010/2/119171

Wilson, J., Olinghouse, N. G., & Andrada, G. N. (2014). Does automated feedback improve writing quality? *Learning Disabilities: A Contemporary Journal*, *12*(1), 93–118. https://eric.ed.gov/?id=EJ1039856

Win Myint, P. Y., Lo, S. L., & Zhang, Y. (2024). Harnessing the power of AI-instructor collaborative grading approach: Topic-based effective grading for semi open-ended multipart questions. *Computers and Education: Artificial Intelligence*, *7*, 100339. https://doi.org/10.1016/j.caeai.2024.100339

Xue, Y. (2024). Towards automated writing evaluation: A comprehensive review with bibliometric, scientometric, and meta-analytic approaches. *Education and Information Technologies*, *29*(15), 19553–19594. https://doi.org/10.1007/s10639-024-12596-0

Yu, S., & Lee, I. (2014). An analysis of Chinese EFL students' use of first and second language in peer feedback of L2 writing. *System*, *47*, 28–38. https://doi.org/10.1016/j.system.2014.08.007

https://doi.org/10.1017/S0261444825000199 Published online by Cambridge University Press

Yusuf Cengiz is a Ph.D. candidate in English Language Teaching at Boğaziçi University, Istanbul, Türkiye, where he also works as an English instructor. His research focuses on second language (L2) writing, corrective feedback, and use of AI.

Nur Yiğitoğlu Aptoula is an associate professor in the Department of Foreign Language Education at Boğaziçi University, Istanbul, Türkiye. Her current research focuses on L2 writing and L2 teacher education. She is one of the recipients of the 2023 Turkish Academy of Sciences (TÜBA) Outstanding Young Scientists Awards (GEBİP).

Cite this article: Cengiz, Y., & Yigitoglu Aptoula, N. (2026). Comparing AI and human feedback at higher education: Level appropriateness, quality and coverage. *Language Teaching*, 59(1), 122–127. https://doi.org/10.1017/S0261444825000199

https://doi.org/10.1017/S0261444825000199 Published online by Cambridge University Press