Generative Artificial Intelligence and Automated Feedback in Second Language Writing: A Critical Review

A Critical Review

[1]*Yazdankulova Gulnigor Muzaffarovna.,*
1. Samarkand State institute of Foreign Languages
[2]*Jason Anderson*
2. University of Warwick, School of Education, Learning and Communication Sciences

Abstract

Background. Feedback is central to the development of second language (L2) writing, yet providing timely, individualised feedback to large cohorts remains a persistent challenge in higher education. Automated writing evaluation (AWE) has long promised to address this gap, and the emergence of generative artificial intelligence (GenAI) has dramatically expanded what automated feedback can do. **Aim.** This article critically reviews the evidence on automated and GenAI-generated feedback in L2 writing, evaluating its effectiveness, the nature of learner engagement, its use in automated scoring, and the pedagogical and ethical tensions it raises. **Methods.** Adopting a structured narrative-review design informed by systematic search principles, we synthesised peer-reviewed scholarship—prioritising 2020–2026 work from Scopus- and Web of Science-indexed journals—and organised the evidence thematically. **Results.** Meta-analytic evidence indicates moderate-to-large positive effects of AWE on writing quality, with larger effects for L2 and tertiary writers; GenAI tools can match human feedback in efficiency and certain textual dimensions while remaining weaker on higher-order, context-dependent aspects. Engagement with automated feedback is uneven and mediated by learner and contextual factors, and large language models show promising but imperfect reliability for automated essay scoring. **Conclusion.** Automated and GenAI feedback are best positioned as a complement to, rather than a replacement for, teacher feedback, and their value depends on principled integration that preserves learner agency. **Practical significance.** The review offers evidence-based

guidance for teachers, curriculum developers, institutions, and tool developers on integrating automated feedback while safeguarding higher-order writing development and academic integrity.

Keywords

second language writing; automated writing evaluation; generative artificial intelligence; ChatGPT; written corrective feedback; learner engagement; automated essay scoring; higher education; feedback literacy; assessment.

1. Introduction

Writing is among the most cognitively demanding competences in second language (L2) learning, and feedback is widely regarded as one of the most powerful levers for its development (Hattie & Timperley, 2007). Yet the provision of timely, individualised feedback is labour-intensive, and in the large, heterogeneous cohorts characteristic of higher education it is frequently delayed, uneven, or unsustainable. This tension has long motivated interest in automated writing evaluation (AWE)—software that analyses learner texts and returns feedback or scores—and, more recently, in generative artificial intelligence (GenAI) tools capable of producing fluent, context-sensitive feedback at scale (Barrot, 2023; Kohnke et al., 2023).

The public release of large language models (LLMs) in late 2022 marked a qualitative shift. Where earlier AWE systems relied on rule-based and statistical methods that excelled at surface-level error detection but struggled with higher-order concerns, GenAI tools can engage with content, organisation, and rhetorical purpose, and can adapt feedback to prompts and proficiency levels (Escalante et al., 2023; Mizumoto & Eguchi, 2023). This expansion of capability has been accompanied by a rapid proliferation of empirical and review scholarship, much of it enthusiastic but tool-focused and short-term, and insufficiently connected to established theory on feedback and L2 development (Liu et al., 2024).

This review addresses that gap by synthesising the evidence on automated and GenAI feedback in L2 writing and evaluating it critically against feedback theory

and second language acquisition (SLA) principles. The guiding questions are: (RQ1) What does current evidence indicate about the effectiveness of automated and GenAI feedback on L2 writing outcomes? (RQ2) How do learners engage with such feedback, and what mediates that engagement? (RQ3) What pedagogical and ethical tensions condition its principled integration? The novelty of the review lies in positioning GenAI feedback within the longer trajectory of AWE and feedback research, rather than treating it as an unprecedented phenomenon, and in foregrounding learner engagement and higher-order development as the decisive criteria of value.

2. Method of the Review

This article is a structured narrative review and reports no original participant data. Literature was identified through Scopus, Web of Science, and Google Scholar, complemented by hand-searching of leading journals including the Journal of Second Language Writing, Assessing Writing, System, Computer Assisted Language Learning, Language Teaching Research, Computers & Education, and the International Journal of Educational Technology in Higher Education. Search strings combined writing and feedback terms (e.g., "written corrective feedback", "automated writing evaluation", "feedback engagement") with AI terms ("generative AI", "ChatGPT", "large language model", "automated essay scoring"). Priority was given to peer-reviewed work from 2020–2026, while foundational earlier sources were retained. Findings were extracted and synthesised thematically, and divergent results were compared analytically. Limitations include the dominance of English-as-a-target-language scholarship and the rapid obsolescence of GenAI-specific evidence.

3. Conceptual Background: Feedback in L2 Writing

The role of feedback in L2 writing has been contested. Truscott (1996) famously argued that grammar correction is ineffective and potentially harmful, igniting a debate that generated decades of empirical work. Subsequent research, including controlled studies, provided evidence that written corrective feedback (WCF) can

contribute to language development under specific conditions (Bitchener & Knoch, 2010), and meta-analytic synthesis reported medium-to-large effects of WCF on written accuracy (Kang & Han, 2015). Contemporary scholarship has moved beyond the binary question of whether feedback works towards the more productive questions of how, for whom, and under what conditions it is effective (Bitchener & Ferris, 2012; Ferris, 2011).

Theoretically, feedback is understood as information that reduces the gap between current and desired performance, operating most powerfully when it addresses the task, the process, and self-regulation rather than the self (Hattie & Timperley, 2007). From an SLA perspective, feedback that prompts learners to notice gaps and to produce "pushed" output is consistent with the output hypothesis (Swain, 1985), while sociocultural accounts frame feedback as mediation within the learner's zone of proximal development (Vygotsky, 1978; Jiang et al., 2020). These frameworks supply the criteria against which automated and GenAI feedback should be judged: not merely whether errors are corrected, but whether learners notice, understand, and act on feedback in ways that advance development. This shifts attention from feedback provision to feedback engagement (Han & Hyland, 2015; Zhang & Hyland, 2018).

4. From Automated Writing Evaluation to Generative AI

Automated writing evaluation has a long history, evolving from word-processing tools that flagged surface errors to systems providing increasingly sophisticated feedback (Warschauer & Ware, 2006). Established AWE systems combined natural language processing with statistical scoring models, offering immediate, consistent, and scalable feedback. Reviews and classroom research documented both their promise—timeliness, opportunities for repeated revision, and support for autonomy—and their limitations, notably a tendency to privilege surface accuracy over content, organisation, and rhetorical effectiveness (Hockly, 2019; Li et al., 2015; Stevenson & Phakiti, 2014; Warschauer & Grimes, 2008). A recurrent

conclusion was that AWE is most effective when integrated with, rather than substituted for, teacher feedback (Li et al., 2015).

The advent of GenAI represents both continuity and rupture with this tradition. Like earlier AWE, LLM-based tools provide immediate, scalable feedback; unlike earlier systems, they can engage flexibly with meaning, generate explanations and exemplars, and adapt to prompts, positioning them closer to a dialogic partner than a static checker (Kohnke et al., 2023). Crucially, however, the gains in flexibility are accompanied by new risks—factual inaccuracy, fabricated content, and opacity—that earlier rule-based systems did not present to the same degree (Kasneci et al., 2023; Tlili et al., 2023). The literature is therefore best read as a single, evolving line of inquiry into automated feedback rather than as two unrelated bodies of work.

5. Effectiveness: What the Evidence Shows

Meta-analytic evidence on AWE is broadly positive. Zhai and Ma (2023), synthesising 26 studies with 2,468 participants, reported a large overall effect on writing quality ($g = 0.86$), with stronger effects for post-secondary than for secondary students and for L2/EFL than for native-English writers. Systematic review evidence indicates that AWE can affect not only surface-level but also deeper textual outcomes (Nunes et al., 2022). These findings position tertiary L2 writers—the focus of this review—as among the principal beneficiaries of automated feedback, while underscoring that effects are moderated by educational level, genre, and implementation.

Evidence on GenAI feedback specifically is younger but accumulating. Comparative studies suggest that ChatGPT-generated feedback can match teacher feedback in efficiency and in certain dimensions such as readability and detail, while remaining weaker on aspects requiring cultural awareness and contextual judgement (Steiss et al., 2024). In a longitudinal quasi-experimental study, Escalante et al. (2023) found no significant difference in learning outcomes between learners receiving GenAI and human tutor feedback, with student preferences split—an indication that GenAI feedback is a viable supplement rather than a categorical

improvement. Experimental work outside L2 contexts likewise reports that LLM-generated feedback can increase text revision, motivation, and positive emotions (Meyer et al., 2024). Across these studies, the consistent pattern is parity-with-caveats rather than superiority: GenAI feedback is comparable to human feedback on some dimensions, weaker on higher-order and context-dependent ones, and valuable principally for its timeliness and scalability.

6. Learner Engagement and Perceptions

Effectiveness depends not on feedback provision but on feedback engagement—the cognitive, behavioural, and affective ways learners interact with feedback (Han & Hyland, 2015). Research on AWE engagement shows that learners differ markedly in how they process and act on automated feedback, with engagement mediated by proficiency, motivation, beliefs, and instructional framing (Zhang & Hyland, 2018). Case-study evidence on tools such as Grammarly indicates that learners may engage superficially, accepting suggestions without understanding, which limits developmental benefit (Koltovskaia, 2020). Ranalli (2018) similarly found that students' ability to make use of automated WCF is uneven and contingent on guidance.

These findings carry directly into the GenAI era. The flexibility and fluency of LLM feedback may heighten the risk of passive acceptance and over-reliance, as the apparent authority and ease of GenAI output can discourage the critical evaluation and effortful revision on which development depends (Barrot, 2023; Kasneci et al., 2023). Conversely, when GenAI is positioned as a collaborator that learners interrogate and negotiate with—rather than an oracle—engagement can become more active (Su et al., 2023; Yan, 2023). The decisive variable is therefore not the tool but the pedagogical framing that shapes how learners engage with it, which in turn implicates learners' feedback literacy.

7. Automated Essay Scoring with Large Language Models

Beyond formative feedback, LLMs have been examined for automated essay scoring (AES). Mizumoto and Eguchi (2023), scoring a corpus of 12,100 essays by non-native writers, found that GPT-based scoring achieved a useful level of reliability and accuracy and that incorporating linguistic features improved performance, suggesting LLMs can support—though not yet replace—human raters. Subsequent work indicates that while calibrated models can align reasonably with proficiency benchmarks, performance varies with task type and writers' first languages, raising fairness concerns across linguistic backgrounds. The construct validity of automated scoring systems remains an active concern (Myers & Wilson, 2023). The emerging consensus is that LLM-based AES is a promising assistive technology whose outputs require human oversight, particularly in high-stakes contexts.

8. Critical Tensions

Several tensions condition the principled integration of automated and GenAI feedback. The first is **over-reliance**: the ease and authority of automated feedback may erode the productive struggle and self-regulation central to writing development (Barrot, 2023; Kasneci et al., 2023). The second is **accuracy and trust**: unlike rule-based AWE, GenAI can produce fluent but incorrect or fabricated feedback, demanding critical evaluation that lower-proficiency learners may be ill-equipped to perform (Tlili et al., 2023). The third is **higher-order development**: automated systems remain comparatively strong on surface features and weaker on argumentation, coherence, and rhetorical purpose, risking a narrowing of writing instruction if used uncritically (Li et al., 2015; Steiss et al., 2024). The fourth is **equity**: differential access to advanced tools and the variable performance of models across first languages may entrench rather than reduce inequality (Mizumoto & Eguchi, 2023). The fifth is **academic integrity and the teacher's role**: as the boundary between feedback and text generation blurs, assessment validity and the redefinition of teacher and learner roles become pressing (Crompton & Burke, 2023;

Kohnke et al., 2023). Table 1 contrasts traditional AWE and GenAI feedback across these dimensions.

Table 1. Comparison of traditional AWE and GenAI feedback in L2 writing. Note. Synthesised from the sources reviewed in Sections 4–8.

Dimension	**Traditional AWE**	**GenAI feedback**
Mechanism	Rule-based / statistical NLP	Large language models
Strength	Consistent surface-level accuracy	Flexible, content- and context-sensitive
Higher-order feedback	Limited	Improved but uneven
Principal risk	Narrow, surface focus	Inaccuracy / fabrication; over-reliance
Best role	Supplement to teacher feedback	Supplement; collaborator under guidance

9. Discussion

Read as a whole, the evidence supports three conclusions. First, automated feedback—both traditional AWE and GenAI—has measurable benefits for L2 writing quality, with meta-analytic effects that are particularly favourable for tertiary L2 writers (Zhai & Ma, 2023). Second, GenAI feedback is best characterised as comparable to human feedback on efficiency and selected textual dimensions, but weaker on higher-order, context-dependent aspects, making it a complement rather than a replacement for teacher feedback (Escalante et al., 2023; Steiss et al., 2024). Third, and most importantly, the value of automated feedback is realised through engagement, not provision: outcomes depend on whether learners notice, understand, critically evaluate, and act upon feedback, which is shaped by pedagogical framing and feedback literacy rather than by the tool itself (Han & Hyland, 2015; Zhang & Hyland, 2018).

These findings both extend and complicate the historical AWE literature. They extend it by showing that GenAI overcomes some of the surface-level narrowness of earlier systems; they complicate it by introducing accuracy and over-reliance risks that earlier rule-based tools did not present (Kasneci et al., 2023; Tlili et al., 2023). The pedagogical implication is that GenAI should be embedded within feedback practices that cultivate learner agency—teaching students to interrogate, verify, and selectively act on automated feedback. The theoretical implication is the need to articulate human–AI feedback interaction within SLA and feedback theory, rather than importing claims of effectiveness uncritically. The technological implication is that tools should be designed to scaffold revision and explanation rather than to supply finished text, and the policy implication is the necessity of guidelines, teacher development, and equitable access.

10. Practical Recommendations

For teachers: integrate automated and GenAI feedback as a complement to—not a substitute for—teacher feedback; explicitly teach learners to evaluate, verify, and act on automated feedback; and reserve human attention for higher-order, context-dependent concerns.

For curriculum developers: build feedback literacy and critical AI literacy into writing curricula; design revision-focused tasks that require learners to engage with, rather than passively accept, automated feedback.

For institutions and policymakers: provide clear guidance on the ethical use of GenAI in writing instruction and assessment; ensure equitable access to tools; and invest in teacher professional development.

For tool developers: design feedback systems that scaffold learner revision, expose uncertainty, and support explanation rather than text substitution, in collaboration with writing specialists.

For students: treat automated feedback as a starting point for revision rather than an authoritative answer, verifying suggestions and prioritising effortful, reflective engagement.

11. Future Research Directions

Priorities include: longitudinal and classroom-based studies of GenAI feedback's effects on higher-order writing development and durable learning; research on learner engagement and feedback literacy in GenAI-rich settings; investigations of fairness across first languages and proficiency levels; the development and validation of LLM-based scoring with adequate human oversight; and theoretically grounded work articulating human–AI feedback interaction within SLA. Mixed-methods and design-based research are especially suited to these aims.

12. Conclusion

This review has synthesised the evidence on automated and GenAI feedback in L2 writing. The principal finding is that automated feedback offers genuine, measurable benefits—especially for tertiary L2 writers—but is best positioned as a complement to teacher feedback, with its value realised through active learner engagement rather than mere provision. The scientific contribution lies in situating GenAI feedback within the longer trajectory of AWE and feedback research and in foregrounding engagement and higher-order development as the decisive criteria of value. The practical contribution is a set of evidence-based recommendations for teachers, developers, institutions, and learners. The review's limitations include its reliance on English-language scholarship and the rapid evolution of GenAI evidence, constraints that themselves motivate the research agenda outlined above. The overarching perspective is one of principled integration: harnessing the timeliness and scalability of automated feedback while safeguarding the critical engagement and effortful revision on which L2 writing development ultimately depends.

Additional information

Conflict of interest. [The authors declare no conflict of interest.] **Funding.** [This research received no external funding.]

References

1. Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. Assessing Writing, 57, 100745. https://doi.org/10.1016/j.asw.2023.100745
2. Bitchener, J., & Ferris, D. R. (2012). Written corrective feedback in second language acquisition and writing. Routledge.
3. Bitchener, J., & Knoch, U. (2010). The contribution of written corrective feedback to language development: A ten-month investigation. Applied Linguistics, 31(2), 193–214. https://doi.org/10.1093/applin/amp016
4. Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. International Journal of Educational Technology in Higher Education, 20, 22. https://doi.org/10.1186/s41239-023-00392-8
5. Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. International Journal of Educational Technology in Higher Education, 20, 57. https://doi.org/10.1186/s41239-023-00425-2
6. Ferris, D. R. (2011). Treatment of error in second language student writing (2nd ed.). University of Michigan Press.
7. Han, Y., & Hyland, F. (2015). Exploring learner engagement with written corrective feedback in a Chinese tertiary EFL classroom. Journal of Second Language Writing, 30, 31–44. https://doi.org/10.1016/j.jslw.2015.08.002
8. Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81–112. https://doi.org/10.3102/003465430298487

9. Hockly, N. (2019). Automated writing evaluation. ELT Journal, 73(1), 82–88. https://doi.org/10.1093/elt/ccy044

10. Hyland, K. (2019). Second language writing (2nd ed.). Cambridge University Press.

11. Hyland, K., & Hyland, F. (Eds.). (2019). Feedback in second language writing: Contexts and issues (2nd ed.). Cambridge University Press.

12. Jiang, L., Yu, S., & Wang, C. (2020). Second language writing instructors' feedback practice in response to automated writing evaluation: A sociocultural perspective. System, 93, 102302. https://doi.org/10.1016/j.system.2020.102302

13. Kang, E., & Han, Z. (2015). The efficacy of written corrective feedback in improving L2 written accuracy: A meta-analysis. The Modern Language Journal, 99(1), 1–18. https://doi.org/10.1111/modl.12189

14. 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

15. Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. RELC Journal, 54(2), 537–550. https://doi.org/10.1177/00336882231162868

16. Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. Assessing Writing, 44, 100450. https://doi.org/10.1016/j.asw.2020.100450

17. Li, Z., 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

18. Liu, M., Zhang, L. J., & Biebricher, C. (2024). Investigating students' cognitive processes in generative AI-assisted digital multimodal composing and traditional writing. Computers & Education, 211, 104977. https://doi.org/10.1016/j.compedu.2023.104977

19. 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

20. Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. Research Methods in Applied Linguistics, 2(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

21. Myers, M. C., & Wilson, J. (2023). Evaluating the construct validity of an automated writing evaluation system with a randomization algorithm. International Journal of Artificial Intelligence in Education, 33(3), 609–634. https://doi.org/10.1007/s40593-022-00301-6

22. Nunes, A., Cordeiro, C., Limpo, T., & Castro, S. L. (2022). Effectiveness of automated writing evaluation systems in school settings: A systematic review. Journal of Computer Assisted Learning, 38(2), 599–620. https://doi.org/10.1111/jcal.12635

23. Ranalli, J. (2018). Automated written corrective feedback: How well can students make use of it? Computer Assisted Language Learning, 31(7), 653–674. https://doi.org/10.1080/09588221.2018.1428994

24. 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

25. Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. Assessing Writing, 19, 51–65. https://doi.org/10.1016/j.asw.2013.11.007

26. Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. Assessing Writing, 57, 100752. https://doi.org/10.1016/j.asw.2023.100752

27. Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. M. Gass & C. G. Madden (Eds.), Input in second language acquisition (pp. 235–253). Newbury House.

28. Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. Smart Learning Environments, 10, 15. https://doi.org/10.1186/s40561-023-00237-x

29. Truscott, J. (1996). The case against grammar correction in L2 writing classes. Language Learning, 46(2), 327–369. https://doi.org/10.1111/j.1467-1770.1996.tb01238.x

30. Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Harvard University Press.

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

32. Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. Language Teaching Research, 10(2), 157–180. https://doi.org/10.1191/1362168806lr190oa

33. Yan, D. (2023). Impact of ChatGPT on learners in an L2 writing practicum: An exploratory investigation. Education and Information Technologies, 28(11), 13943–13967. https://doi.org/10.1007/s10639-023-11742-4

34. Zhai, N., & Ma, X. (2023). The effectiveness of automated writing evaluation on writing quality: A meta-analysis. Journal of Educational Computing Research, 61(4), 875–900. https://doi.org/10.1177/07356331221127300

35. Zhang, Z. (V.), & Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. Assessing Writing, 36, 90–102. https://doi.org/10.1016/j.asw.2018.02.004