



Castledown

 OPEN ACCESS

Technology in Language Teaching & Learning

ISSN 2652-1687

<https://www.castledown.com/journals/tltl/>

Technology in Language Teaching & Learning, 8, 104473 (2026)

<https://doi.org/10.29140/tltl.2026.104473>

AI-Mediated Feedback in L2 Writing: Associations with Writing Performance, Engagement, and Instructor Perspectives in an Armenian University Context



LUIZA MARABYAN^a  (Corresponding author)

LILIT AVETISYAN^b 

MARINE AVETISYAN^c 

ANNA KNYAZYAN^d 

^a*Yerevan State University, Republic of Armenia*
luizamarabyan@ysu.am

^b*European University of Armenia, Republic of Armenia*
lilavetis@gmail.com

^c*European University of Armenia, Republic of Armenia*
maravetis@yahoo.com

^d*Yerevan State University, Republic of Armenia*
annaknyazyan@ysu.am

Abstract

Written corrective feedback (WCF) is a cornerstone of second language (L2) writing instruction, and the emergence of AI-based tools has substantially transformed its delivery. However, a critical conceptual distinction exists between AI tools that provide corrective feedback on linguistic form, such as Grammarly and Cambridge Write & Improve, and large language model (LLM)-based tools, such as ChatGPT, which generate broader discourse-level feedback encompassing ideas, organisation, and cohesion. This mixed-methods study investigates how a sequenced combination of these AI tools is associated with the writing performance, perceptions, and engagement of 80 intermediate ESL learners at the European University of Armenia, alongside the perspectives of 10 ESL instructors. A four-week classroom intervention incorporated in-class, timed writing tasks, a learner questionnaire, structured written reflections, and semi-structured instructor interviews. Inter-rater reliability between the two independent raters was high ($ICC > .90$). Quantitative analysis using paired-sample t-tests

Copyright: © 2026 Luiza Marabyan, Lilit Avetisyan, Marine Avetisyan & Anna Knyazyan. This is an open access article distributed under the terms of the Creative Commons Attribution Non-Commercial 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

and a rubric adapted from Bitchener and Ferris (2012) revealed statistically significant pre-to-post improvements across grammatical accuracy, lexical variety, cohesion, and overall writing quality. In the absence of a control group and given the use of different essay prompts at pre- and post-test, these gains should be interpreted with caution rather than attributed causally to the AI intervention alone. Qualitative thematic analysis of reflections and interviews identified four learner themes: enhanced error awareness, structured revision support, difficulty evaluating feedback appropriateness, and emerging dependency, and five instructor themes centring on efficiency gains, risks of uncritical revision, uneven engagement, ethical concerns, and support for a hybrid pedagogical model. Drawing on feedback literacy (Carless & Boud, 2018), the Noticing Hypothesis (Schmidt, 1990), and Sociocultural Theory (Vygotsky, 1978), the study proposes a conceptual teacher-AI partnership framework in which automated tools complement, rather than replace, human pedagogical judgment, pending empirical validation through future research incorporating teacher-mediated designs. The study makes three contributions to the CALL literature: an empirical account of sequential multi-tool AWE integration examined across behavioural, cognitive, and affective engagement dimensions; evidence that feedback literacy development is proficiency-linked and requires explicit instructional scaffolding rather than tool exposure alone; and a theoretically grounded teacher-AI partnership framework positioned as a basis for future controlled experimental designs in EFL writing pedagogy.

Keywords: AI-mediated feedback, automated writing evaluation, L2 writing, written corrective feedback, learner engagement, feedback literacy, teacher-AI framework, dependency

Introduction

Written corrective feedback (WCF) occupies a central position in second language (L2) writing instruction, providing learners with targeted information about errors in grammatical accuracy, lexical choice, cohesion, and text organisation. Over the past two decades, WCF has attracted substantial empirical attention, evolving from early debates about its fundamental efficacy (Truscott, 1996) towards nuanced investigation of how, when, and under what conditions feedback best supports language development. Ferris (1999, 2004) demonstrated that well-implemented feedback can generate measurable improvements in linguistic accuracy and contribute to long-term interlanguage development. Subsequent research has examined the differential effects of direct and indirect feedback, focused and unfocused correction, and the inclusion of metalinguistic explanation, consistently indicating that WCF effectiveness is mediated by learner proficiency, task complexity, and instructional design (Bitchener & Knoch, 2008; Ferris, 2014).

The rapid advancement of artificial intelligence (AI) and natural language processing (NLP) technologies has created new possibilities for delivering writing feedback at scale. However, an important conceptual distinction must be maintained. Automated writing evaluation (AWE) tools such as Grammarly and Cambridge Write & Improve primarily function as corrective feedback systems targeting surface-level linguistic forms, including grammar, mechanics, and spelling, and thus align most directly with conventional WCF research. Large language model (LLM)-based tools such as ChatGPT generate a qualitatively different type of feedback, addressing ideas, argument structure, cohesion, and lexical reformulation, which extends beyond the traditional WCF construct (Barrot, 2023; Teng, 2024). This distinction is theoretically significant and has direct implications for study design, data interpretation, and pedagogical recommendations. The present study explicitly acknowledges this difference and examines both types of AI-mediated feedback, positioned within a broader framework of AI-mediated writing support.

Although the potential of AI tools in L2 writing instruction is widely recognised, important questions remain about their effects on learner engagement across behavioural, cognitive, and affective dimensions, and about the pedagogical and ethical challenges they present. AI feedback may encourage superficial revision, foster overreliance on automated suggestions, and raise concerns about authorship and academic integrity, particularly in higher education contexts (Barrot, 2023; Koltovskaia, 2020; Ranalli, 2021). Moreover, research in under-resourced or under-studied contexts, such as Armenian higher education, remains sparse, and the systematic comparison of multiple AI tools within a single study is limited in the existing literature. The present study is designed to address these gaps. It is guided by the following research questions:

1. How is a sequenced combination of AI-mediated feedback tools associated with L2 writing performance among intermediate ESL learners?
2. How do learners engage with AI-mediated feedback across behavioural, cognitive, and affective dimensions, and to what extent does perceived dependency emerge?
3. How do ESL instructors evaluate the pedagogical affordances and limitations of integrating AI tools into L2 writing instruction?

Positioned within feedback literacy frameworks (Carless & Boud, 2018), the Noticing Hypothesis (Schmidt, 1990), and Sociocultural Theory (Vygotsky, 1978), the study examines how the use of AI tools is reflected in relation to conventional WCF practice. Drawing on the findings, the study proposes a conceptual teacher-AI partnership framework in which AWE tools function as providers of immediate corrective feedback and LLM-based tools serve as resources for discourse-level revision support, with teachers retaining responsibility for context-sensitive guidance, ethical scaffolding, and the promotion of reflective writing practice. This framework represents a theoretically motivated proposal informed by the present findings, intended to guide future empirical investigation rather than a direct output of an intervention in which teacher mediation was embedded throughout.

Literature Review

Written Corrective Feedback in L2 Writing

Written corrective feedback has been widely identified as a central mechanism in facilitating L2 writing development. The foundational debate triggered by Truscott's (1996) claim that WCF is ineffective and potentially harmful prompted substantial empirical and theoretical research. Ferris (1999, 2004) systematically contested this position, demonstrating that appropriately designed feedback can enhance accuracy and support long-term language acquisition. Subsequent scholarship shifted attention from whether WCF works to how, when, and for whom it is most effective, considering factors such as learner proficiency, task difficulty, and the presence of metalinguistic cues (Bitchener & Knoch, 2008; Ferris, 2014).

A key distinction in the empirical literature concerns direct feedback, which provides explicit correction, and indirect feedback, which signals errors without supplying the corrected form (Bitchener & Knoch, 2008). Direct feedback has been found particularly effective for lower-proficiency learners and structurally complex targets, while indirect feedback promotes deeper cognitive processing and self-regulation. Focused feedback targeting specific error categories has been compared with unfocused feedback reflecting authentic writing conditions, revealing that each influences surface-level accuracy and higher-order revision processes in distinct ways (Ferris, 2014). Metalinguistic explanations facilitate rule internalisation, though their efficacy is contingent upon learner readiness (Ferris, 2014). Together, these findings confirm that WCF is a context-dependent, learner-sensitive practice rather than a uniformly effective intervention.

Automated Writing Evaluation and AI-Mediated Feedback

The development of automated writing evaluation (AWE) systems has expanded the possibilities for delivering feedback at scale. Rule-based and statistical NLP tools such as Grammarly, Cambridge Write & Improve, and ETS Criterion provide corrections and scoring on grammar, mechanics, and stylistic elements. Empirical research has demonstrated that AWE tools can enhance learner independence and accelerate surface-level revision, particularly with respect to grammar and mechanics (Barrot, 2021; Warschauer & Ware, 2006). Barrot's (2021) quasi-experimental study found that Grammarly-mediated automated corrective feedback significantly improved L2 writing accuracy, with gains attributed to noticing, adaptive metalinguistic explanation, and self-directed learning. Ranalli (2021) similarly found that engagement with Grammarly was influenced by individual factors, including trust in the tool and existing language knowledge, with less proficient learners at risk of uncritical acceptance of suggestions.

However, AWE systems have been widely critiqued for privileging surface-level correction while exhibiting limited sensitivity to contextual meaning, discourse structure, and higher-order thinking skills, thereby constraining self-regulated revision (Koltovskaia, 2020; Ranalli, 2021). Sari (2024) found that AWE tools significantly influenced EFL learners' writing self-efficacy and self-regulation, with learners reporting reduced writing anxiety alongside performance gains, though dependency on automated correction also emerged as a concern. Scholars caution against unmediated reliance on AWE and emphasise the need for pedagogically principled integration with teacher guidance (Al-Sofi, 2026; Barrot, 2021; Rahimi et al., 2025).

The emergence of large language models (LLMs) such as ChatGPT represents a qualitatively different development. Unlike AWE systems, LLMs generate contextualised, iterative feedback that extends beyond formal correctness to encompass ideas, argument structure, cohesion, and rhetorical strategy (Barrot, 2023; Teng, 2024). Mahapatra (2024) found significant positive effects of ChatGPT as a formative feedback tool on ESL students' academic writing, with learners reporting strong motivation and perceived utility. Teng (2024) found that while learners generally experienced ChatGPT as a supportive companion, concerns about over-reliance and the evaluation of conflicting suggestions were prevalent. Steiss et al. (2024) demonstrated that LLM-based tools outperformed traditional AWE on grammatical and syntactical corrections. Critically, Barrot (2023) cautions that while ChatGPT offers substantial affordances, it also introduces pitfalls including inaccurate feedback, context-insensitivity, and risks of academic dishonesty. Taken together, the literature positions both AWE and LLM-based tools as potentially valuable but insufficient as standalone feedback mechanisms. This insufficiency stems from what Stockwell and Wang (2025) describe as a 'conditional interdependence,' where the AI provides the generative capacity, but the 'pedagogically meaningful output' remains dependent on human task definition and evaluative guidance.

Learner Engagement with AI Feedback

Research on learner engagement with AI-mediated feedback has increasingly distinguished behavioural, cognitive, and affective dimensions. Behaviourally, learners may engage consistently with tools and complete revision cycles; cognitively, they may evaluate, compare, and selectively apply suggestions; and affectively, they may experience reduced anxiety, increased motivation, or, conversely, frustration and over-reliance (Koltovskaia, 2020; Ranalli, 2021). Koltovskaia (2020) found that lower-proficiency learners over-relied on Grammarly, uncritically accepting all suggestions including inaccurate ones, while higher-proficiency learners were more selective but also less trusting of the tool. Ranalli (2021) extended this by examining trust as a mediating factor, arguing that meaningful learning from AI feedback requires cognitive engagement beyond a surface-level proofreading orientation. Teng (2024)

found that learners who experienced ChatGPT as supportive were more likely to engage cognitively and reflectively with suggestions.

These findings underscore that effective AI-mediated revision depends not only on tool quality but on learners' feedback literacy, namely on their capacity to interpret, evaluate, and act upon feedback strategically (Carless & Boud, 2018). Learners who lack this capacity may engage behaviourally (completing revision cycles) without the cognitive engagement necessary for genuine language development.

Instructor Perspectives and Ethical Considerations

Research on instructor perspectives highlights both efficiency gains and persistent pedagogical concerns. Teachers report that AI tools reduce the burden of surface-level error correction, enabling focus on higher-order dimensions such as argumentation, coherence, and rhetorical strategy (Mahapatra, 2024; Teng, 2024). However, instructors consistently raise concerns about academic integrity and authorship, particularly when LLM-based tools reformulate entire sentences or paragraphs (Barrot, 2023). Concerns about the uneven quality of engagement with less proficient learners at greater risk of uncritical AI adoption and the absence of institutional policies governing AI use have also been highlighted in the literature (Ranalli, 2021; Rahimi et al., 2025). These concerns point to the need for explicit pedagogical scaffolding and institutional frameworks to govern responsible AI integration.

Theoretical Frameworks

These developments are most productively situated within feedback literacy theory (Carless & Boud, 2018), which is foregrounded as the primary interpretive lens in this study. Feedback literacy emphasises learners' capacity to interpret, evaluate, and act upon feedback – exactly the capacity that distinguishes surface-level from deep engagement with AI suggestions, and that the findings below illuminate most sharply. Schmidt's (1990) Noticing Hypothesis holds that acquisition is facilitated when learners consciously attend to salient linguistic forms; AI-mediated error highlighting may serve as an external noticing mechanism. From a Sociocultural perspective (Vygotsky, 1978), AI feedback can partially simulate scaffolded interaction within the learner's zone of proximal development, though it lacks authentic contingency, underscoring the continued importance of teacher mediation. These frameworks are applied where they most directly illuminate specific findings; their limitations in the context of an individual tool-use design are acknowledged.

Research Gaps and Study Rationale

Despite substantial progress in the literature, several gaps remain. Most empirical studies on AI-mediated feedback focus on a single tool, precluding comparison of AWE and LLM-based feedback within a single study. Luo et al. (2025), in a systematic review of AI in L2 writing education, similarly identified the comparison of multiple tools within a single study as a significant research gap, further underscoring the need for empirical investigations that examine AWE and LLM-based tools in combination. Few studies systematically examine all three dimensions of learner engagement (behavioural, cognitive, and affective) with AI feedback. Longitudinal studies are lacking, as are empirical investigations of teacher-AI hybrid models. Under-researched contexts, including Armenia, which may exhibit distinct pedagogical, cultural, and institutional dynamics, are rarely examined. It should be noted, however, that contextual specificity is claimed cautiously here: the Armenian institutional context is described as a rationale for the study's geographic focus, but no Armenian-specific factors are formally analysed in the present study, and the extent to which findings are context-specific remains an open empirical question for future comparative research. The present study addresses these gaps by empirically comparing the associations related to both AWE and LLM-based feedback tools, systematically

examining all three engagement dimensions, and exploring both learner and instructor perspectives in an Armenian university context.

Methodology

Research Design

This study employed a concurrent mixed-methods design, integrating quantitative and qualitative data to provide a comprehensive account of the associations, perceptions, and pedagogical implications of AI-mediated writing support in L2 instruction. The concurrent triangulation approach enabled objective measurement of pre- to post-test differences in writing performance gains alongside in-depth exploration of learner and instructor perspectives, ensuring that numerical trends were contextualised by interpretive accounts of practice and experience. Both data strands were collected concurrently and triangulated at the interpretation stage.

Participants

Participants were 80 intermediate-level ESL learners enrolled in an academic writing programme at the European University of Armenia. Purposive sampling was employed, with participants selected according to the following inclusion criteria: aged 18 to 30 years; a minimum of three years of prior English instruction; and intermediate proficiency (B1–B2 on the Common European Framework of Reference for Languages), as confirmed by an institutional CEFR placement test administered prior to the study. Of the 80 participants, 73 were L1 Armenian speakers and 7 reported Russian as their first language. For the analysis of proficiency-linked engagement patterns, learners were classified, on the basis of their institutional CEFR placement result, into a lower-proficiency sub-group (CEFR B1) and a higher-proficiency sub-group (CEFR B2). The potential influence of learners' L1 backgrounds on error profiles and cognitive engagement with AI feedback could not be systematically examined given the unequal subgroup sizes; this linguistic heterogeneity is acknowledged as a limitation and is discussed further in the Limitations section. Participants were drawn from three intact classes, all taught by instructors included in the study. Additionally, 10 ESL instructors with a minimum of three years of experience in higher education writing instruction were recruited to provide professional perspectives on AI integration. All participants provided written informed consent, and ethical approval was obtained from the institutional review board. Participants were informed that data would be anonymised and that participation was voluntary.

Instruments and Conceptual Clarification

A critical conceptual distinction underpins the instrument design. Grammarly and Cambridge Write & Improve function primarily as AWE tools, providing corrective feedback on surface-level linguistic features (grammar, mechanics, spelling, and error categorisation) that aligns directly with traditional WCF research. ChatGPT, by contrast, is an LLM-based tool that generates broader discourse-level feedback on ideas, paragraph organisation, cohesion, and lexical reformulation – features that extend beyond the established WCF construct (Barrot, 2023). Both types of AI feedback are examined in this study as components of a broader AI-mediated writing support framework, with their distinct contributions to writing development considered separately throughout the analysis.

The study employed the following instruments. First, two timed, in-class argumentative writing tasks of 250 to 300 words each were administered under identical conditions: Essay 1 (pre-test), on the theme of technology in modern education, was completed without any AI support; Essay 2 (post-test), addressing the value of foreign language instruction in schools, was completed following the

four-week intervention. Both tasks were completed during the final 45 minutes of a scheduled class session under teacher supervision, ensuring standardised conditions. Second, the three AI feedback tools were Grammarly (grammar, mechanics, and style), Cambridge Write & Improve (error categorisation and scoring), and ChatGPT (discourse-level feedback on organisation, cohesion, and lexical development). Third, a post-intervention learner questionnaire used a five-point Likert scale to assess perceptions of AI feedback, motivation, confidence, and perceived dependency. Fourth, structured written reflections were collected after each AI-assisted revision cycle during the intervention, asking learners to describe their use of each tool, how they managed conflicting suggestions, and how they approached revision decisions. Fifth, semi-structured interviews conducted with the 10 instructors following the intervention explored pedagogical affordances, limitations, and ethical concerns associated with AI-mediated feedback.

Procedure

Pre-Test

Participants completed Essay 1 under timed, in-class conditions without access to any AI tools, establishing a standardised baseline measure of writing performance.

AI Tool Orientation

A dedicated one-hour training session familiarised all participants with the three AI platforms. The session included: a step-by-step demonstration of each tool using a sample essay; guidelines on interpreting different types of feedback; and explicit instruction on how to manage conflicting suggestions across tools. Learners were encouraged to compare recommendations before making independent revision decisions, and were informed that no suggestion needed to be accepted automatically.

Four-Week Intervention

Over four weeks, participants engaged in structured AI-mediated revision outside of class, completing at least two full revision cycles per essay per week. A prescribed sequence was established: Grammarly was consulted first for surface-level corrections; Write & Improve was used second to identify recurring error patterns and obtain structured scoring; and ChatGPT was consulted last for higher-order discourse feedback. When suggestions from different tools conflicted, participants were instructed to prioritise the revision that best preserved intended meaning and overall clarity, and to document their reasoning in their written reflection. This sequence was grounded on pedagogical principles: Grammarly was assigned first because its rule-based grammar corrections are most legible without prior revision work; Write & Improve was used second to enable structured pattern-awareness across revision cycles; and ChatGPT was assigned last because discourse-level feedback presupposes a degree of surface-level accuracy having been partially addressed. It must be acknowledged, however, that because all participants followed this same fixed sequence, the individual tools, their ordering, and cumulative exposure to revision across the four weeks are confounded. Consequently, the data cannot isolate the independent contribution of any single tool, and the sequential position of ChatGPT as the final instrument may have inflated learners' perceptions of its utility relative to earlier tools.

Post-Test and Data Collection

Following the intervention, Essay 2 was administered under conditions identical to the pre-test. Learner questionnaires were administered immediately after the post-test. Written reflections were collected after each revision cycle throughout the intervention period. Instructor interviews were conducted in the week immediately following the conclusion of the study and lasted approximately 45 minutes each.

Data Analysis

Quantitative analysis focused on writing performance and survey data. Essays were scored by two independent raters, both experienced ESL writing instructors who were not among the study participants and who were blind to the intervention condition of each essay. Scoring used a rubric adapted from Bitchener and Ferris (2012), assessing grammatical accuracy, lexical variety, cohesion, and overall writing quality on percentage-based scales. Lexical variety was operationalised as the range and appropriacy of vocabulary, the avoidance of unnecessary repetition, and the accuracy of word choice; cohesion was operationalised as logical paragraphing, the use of cohesive devices and transitions, and referential clarity. The full rubric, including band descriptors for each dimension, is provided in Appendix A. Inter-rater reliability was assessed using the Intraclass Correlation Coefficient (ICC), calculated from the independent pre-test ratings assigned by the two raters to all participant essays. The analysis was conducted using a two-way random-effects model with absolute agreement for single measurements (ICC(2,1)) in SPSS Statistics.

Table 1 *Inter-Rater Reliability of The Writing Rubric Based on Independent Pre-Test Ratings*

Rubric dimension	ICC (2, 1)	F	Interpretation
Grammatical Accuracy	.965	58.96	Excellent
Lexical Variety	.946	42.65	Excellent
Cohesion & Organization	.963	122.60	Excellent
Overall Writing Perform.	.966	56.72	Excellent

Note. ICC = Intraclass Correlation Coefficient. ICC (2,1) denotes a two-way random-effects model with absolute agreement for single measurements. Interpretation follows Koo and Li (2016): values < .50 indicate poor reliability, .50–.75 moderate reliability, .75–.90 good reliability, and > .90 excellent reliability.

The analyses, as presented in Table 1, demonstrated excellent agreement across all rubric dimensions: grammatical accuracy (ICC = .965), lexical variety (ICC = .946), cohesion and organization (ICC = .963), and overall writing performance (ICC = .966), supporting the reliability of the scoring procedure.

Discrepancies between raters were resolved through discussion. Paired-sample *t*-tests were conducted to examine pre- to post-test differences. To address the risk of Type I error inflation across four simultaneous paired *t*-tests, a Bonferroni-corrected significance threshold ($\alpha = .0125$) was applied; all four comparisons remained significant under this criterion (all $p < .001$). Survey data were analysed using descriptive statistics (means, standard deviations). A correlational analysis relating perceived usefulness, motivation, and dependency, proposed at the design stage, was not conducted: each construct was operationalised as a single five-point Likert item, and a bivariate correlation between single-item indicators would carry limited interpretive value and could not adequately represent these constructs. A validated, multi-item measure of each construct would be required to support meaningful correlational analysis in future replications (see Limitations).

Prior to conducting the paired-samples *t*-tests, the normality of the pre-post difference scores was assessed using the Shapiro-Wilk test. As Table 2 presents, the assumption of normality was met for lexical variety ($W = .984, p = .422$) and overall writing performance ($W = .981, p = .291$), but not for grammatical accuracy ($W = .953, p = .005$) or cohesion ($W = .891, p < .001$). Given the sample size

($N = 80$), and although normality assumptions were not fully satisfied for all variables, parametric procedures of paired-samples t-tests were retained because previous simulation and methodological research has demonstrated that t-tests are robust to violations of normality, particularly when sample sizes exceed 30 observations (Boneau, 1960; Lumley et al., 2002).

Table 2 *Shapiro-Wilk Tests of Normality for Pre-Post Difference Scores ($N = 80$)*

Rubric dimension	W	p	Interpretation
Grammatical Accuracy	.953	.005	Normality assumption not met
Lexical Variety	.984	.422	Normality assumption met
Cohesion & Organization	.891	.001	Normality assumption not met
Overall Writing Performance	.981	.291	Normality assumption met

Note. The Shapiro–Wilk test was applied to the difference scores (Pre-Post) for each rubric dimension. A non-significant result ($p > .05$) indicates that the normality assumption is satisfied.

Error presence was recorded per dimension per essay on a binary basis (present/absent), producing the proportion of essays containing each error type at each assessment point, as reflected in Table 5.

Qualitative data from learner reflections and instructor interviews were transcribed verbatim and subjected to thematic analysis following Braun and Clarke’s (2006) six-phase framework, using NVivo software. Coding followed a combined deductive–inductive scheme: an initial codebook was derived from the three engagement dimensions and from open coding of the first set of reflections, and codes were defined, exemplified, and refined across two coding cycles. The resulting codes were organised under the three engagement dimensions and the four learner themes; the complete coding frame, with code definitions, is provided in Appendix B. To support analytic transparency, the number of learners whose reflections expressed each theme was recorded in NVivo and is reported as indicative prevalence in the Results; a learner could contribute to more than one theme.

Two researchers independently generated initial codes and then collaboratively developed, reviewed, and refined themes through iterative discussion. Trustworthiness was enhanced through cross-coding, member-checking with a subset of participants (four learners and two instructors), and systematic review of discrepant cases. The member-checking sample is acknowledged as relatively small in relation to the breadth of qualitative claims; this is discussed in the Limitations section. Learner reflections were completed in English, consistent with the language of instruction; where minor surface-level errors affected readability, light editorial corrections were applied to quoted excerpts without altering meaning. It should be noted that the qualitative findings reported in the Results section draw on learner written reflections ($n = 80$) and instructor interviews ($n = 10$) as separate data sources; these are treated as complementary rather than directly comparable data.

Limitations of the Design

A control group was not included, primarily due to institutional constraints and the ethical imperative to ensure equitable access to AI-assisted learning for all enrolled students. Consequently, causal inference is limited; observed improvements cannot be definitively attributed to the AI intervention alone. Furthermore, pre- and post-test essays used different prompts (technology in education vs. foreign language instruction), meaning that prompt difficulty and topic familiarity constitute uncontrolled variables that may have contributed to performance differences independently of the intervention.

To partially address these constraints, the study employs pre- to post-test comparisons with triangulated survey and reflection data, reports effect sizes alongside significance values, and applies a validated inter-rater rubric to strengthen interpretive credibility. The four-week duration constrains conclusions regarding long-term interlanguage development. Additionally, familiarity with the three AI tools may have varied among participants despite the orientation session, potentially introducing differential engagement effects.

Results

This section presents quantitative and qualitative findings pertaining to writing performance, learner engagement, and instructor perspectives. Quantitative and qualitative data are reported separately before being triangulated in the Discussion.

Writing Performance

Pre- and Post-Test Comparisons

Table 3 presents the pre- and post-test writing performance scores across all four assessed dimensions, with paired-sample *t*-test results.

Table 3 *Pre- and Post-Test Comparisons of Writing Performance Across the Four Rubric Dimensions*

Writing dimension	Pre-test mean (SD)	Post-test mean (SD)	<i>t</i> (79)	<i>p</i>	Cohen's <i>d_z</i>	95% CI for <i>d</i>
Grammatical Accuracy	65.4 (8.2)	78.1 (7.5)	12.45	< .001	1.39	[1.08, 1.70]
Lexical Variety	61.7 (7.9)	70.5 (8.1)	9.87	< .001	1.10	[0.82, 1.38]
Cohesion & Organisation	68.2 (6.7)	75.8 (7.2)	8.23	< .001	0.92	[0.66, 1.18]
Overall Writing Score	65.1 (6.9)	74.6 (7.0)	11.12	< .001	1.24	[0.95, 1.53]

Note. Scores are expressed as percentages out of 100. SD = standard deviation. *p*-values are based on two-tailed paired-samples *t*-tests evaluated against a Bonferroni-corrected significance level of $\alpha = .0125$. Cohen's *d_z* represents the standardized mean difference for repeated measures. Confidence intervals were computed using the noncentral *t*-distribution.

Statistically significant improvements were observed across all four dimensions ($p < .001$ after Bonferroni correction), with large effect sizes across all measures. Grammatical accuracy demonstrated the largest absolute gain (12.7 percentage points, $d = 1.39$), indicating that the AI-mediated intervention was associated with particularly pronounced improvements in surface-level error types. Lexical variety and cohesion improved at moderate-to-large levels, reflecting some engagement with stylistic and structural dimensions of writing. The large effect sizes should be interpreted with considerable caution in the absence of a control group, as practice effects, task familiarity, and prompt differences may have contributed substantially to observed gains. No causal attribution to the AI intervention is warranted by this design.

Error Reduction by Category

Table 4 presents the percentage of essays containing each error type at pre-test and post-test, illustrating the differential associations between AI feedback and error reduction across categories.

Table 4 *Percentage of Essays Containing Each Error Type at Pre-Test and Post-Test*

Error type	Pre-test (%)	Post-test (%)	Reduction (%)
Subject-Verb Agreement	21	8	61.9
Article Use	18	7	61.1
Verb Tense	15	6	60.0
Sentence Structure	12	6	50.0
Lexical Choice	10	5	50.0

Note. Percentages reflect the proportion of essays in which each error type was recorded as present at each assessment point. All reductions were observed across the full sample ($n = 80$).

The largest reductions occurred in grammar-related categories: subject-verb agreement (61.9%), article use (61.1%), and verb tense (60.0%). Sentence structure and lexical choice showed more moderate reductions (50.0% each), indicating that while the intervention was associated with substantial reductions in mechanical accuracy errors, particularly through Grammarly and Write & Improve, errors related to syntactic complexity and word selection proved more resistant to automated correction.

To move beyond the descriptive proportions and to determine whether the observed reductions in individual error categories were statistically significant, exact two-sided McNemar tests were conducted using the paired binary error classifications (present vs. absent) for all 80 participants. The results are demonstrated in Table 5.

Table 5 *McNemar Test Results for Changes in Writing Error Categories Between Pre-Test and Post-Test*

Error category	Present → Absent	Absent → Present	McNemar p	Interpretation
Subject-Verb Agreement	10	0	.002	Significant
Article Use	9	0	.004	Significant
Verb Tense	7	0	.016	Significant*
Sentence Structure	5	0	.063	Not significant
Lexical Choice	4	0	.125	Not significant

* Significant at $\alpha = .05$; not significant after Bonferroni correction ($\alpha = .01$)

Significant reductions were observed for subject-verb agreement ($p = .002$) and article use ($p = .004$). Verb tense also decreased ($p = .016$), although this result did not remain significant following Bonferroni adjustment for multiple comparisons. In contrast, reductions in sentence structure ($p = .063$) and lexical choice ($p = .125$) were not statistically significant. Overall, the inferential analyses support the descriptive finding that AI-mediated feedback was associated with greater improvements in rule-based grammatical errors than in higher-order syntactic and lexical categories.

Differential Patterns of Tool Use

Qualitative analysis of learner reflections revealed distinct patterns of tool use. Table 6 summarises the primary feedback focus and observed patterns for each tool.

Table 6 *Learners' Reported Patterns of AI Writing Tool Use*

AI tool	Primary feedback focus	Learner-reported pattern
Grammarly	Grammar, mechanics, style	Most consistently used for immediate grammar correction; perceived as most effective for surface-level error reduction
Write & Improve	Error categorisation, scoring, feedback type classification	Supported awareness of recurring error patterns; helped identify structural weaknesses across drafts
ChatGPT	Organisation, cohesion, idea development, lexical reformulation	Most frequently consulted for rephrasing and paragraph organisation; learners described it as addressing text-level concerns

These patterns are consistent with the conceptual distinction between AWE and LLM-based feedback. However, because all learners used the tools in the same fixed sequence, the data cannot attribute effects to individual tools, and the observation that learners used each tool as instructed reflects protocol compliance rather than independently discovered differentiation. Interpretation of these patterns is developed further in the Discussion.

Learner Engagement and Perceptions

Survey Findings

Table 7 presents learner responses to the post-intervention questionnaire (five-point Likert scale; 1 = strongly disagree, 5 = strongly agree).

Table 7 *Descriptive Statistics for Learners' Perceptions of AI-Assisted Writing Feedback*

Survey item	Mean	SD
AI feedback improved my grammatical accuracy	4.3	0.7
AI feedback helped me revise and develop my ideas	3.9	0.8
AI feedback increased my motivation to revise	4.1	0.6
AI feedback was easy to understand and apply	4.4	0.5
I felt dependent on AI suggestions during revision	3.2	0.9

Note. Responses recorded on a five-point Likert scale (1 = strongly disagree, 5 = strongly agree). SD = standard deviation. n = 80.

Learners rated AI feedback as highly accessible ($M = 4.4$, $SD = 0.5$) and reported positive associations with grammatical accuracy ($M = 4.3$, $SD = 0.7$) and motivation ($M = 4.1$, $SD = 0.6$). Perceptions of support for idea development were more moderate ($M = 3.9$, $SD = 0.8$), consistent with the view that learners' higher-order revision processes were less effectively scaffolded by automated tools alone. A mean dependency score of 3.2 ($SD = 0.9$) suggests that a proportion of learners perceived themselves as reliant on AI suggestions; however, this item is a single Likert indicator, and conclusions about dependency should be drawn cautiously pending a more robust multi-item measure in future studies.

Three Dimensions of Learner Engagement

Thematic analysis of 80 written reflections provided evidence across all three dimensions of engagement: behavioural, cognitive, and affective. Table 8 summarises the key indicators and representative learner reflections for each dimension.

Table 8 *Dimensions of Learner Engagement with AI-Assisted Writing and Representative Reflections*

Engagement dimension	Key indicators	Representative learner reflection
Behavioural	Consistent tool use; multiple revision cycles; adherence to the prescribed tool sequence	<i>"I used all three tools at least twice each week, just as the teacher explained."</i>
Cognitive	Evaluation of suggestions; comparison across tools; selective uptake; self-correction	<i>"When two tools gave different advice, I stopped and thought about which one made my meaning clearer."</i>
Affective	Reported confidence, motivation, anxiety reduction; emotional responses to feedback	<i>"Using the AI made me less afraid of making mistakes. I could fix them privately before showing anyone."</i>

Behaviourally, the majority of learners adhered to the prescribed tool sequence and completed multiple revision cycles. However, behavioural compliance did not uniformly indicate meaningful learning: consistent participation does not itself confirm cognitive engagement. Cognitive engagement was more variable. Higher-proficiency learners demonstrated evaluative and selective engagement, comparing suggestions across tools and exercising independent judgment: "When two tools gave different advice, I stopped and thought about which one made my meaning clearer." In contrast, several lower-proficiency learners described accepting AI suggestions without evaluation: "I relied on the AI to fix the grammar automatically, so I didn't always think about why the correction was made." This proficiency-linked pattern is consistent with prior findings in the literature and underscores the importance of explicit scaffolding to promote critical engagement.

Affective engagement was broadly positive. Learners consistently reported reduced anxiety and increased confidence: "Using the AI made me less afraid of making mistakes. I could fix them privately before showing anyone." However, affective investment also manifested as dependency: "I felt more confident writing after checking everything with AI. But now I notice I wait for the tool to confirm my choices instead of trusting myself."

Four Qualitative Themes from Learner Reflections

Four primary themes emerged from thematic analysis of the 80 learner reflections, as provided in Table 9. Enhanced error awareness was the most widely reported theme, identified in 52 of the 80 reflections (65%): the majority of learners described AI-mediated highlighting as making recurrent patterns more visible and easier to internalise. "I didn't realise how often I repeated the same mistakes with articles until the AI pointed them out every time. After a few corrections, I started noticing the pattern myself." Structured revision support, particularly from ChatGPT, was reported by 48 learners (60%) and helped them approach paragraph organisation more systematically. "The suggestions helped me understand how to reorganise my paragraphs. When I saw examples of better transitions, it became easier to rewrite the text more logically."

Table 9 *Indicative Prevalence of Learner Themes Identified Through Thematic Analysis (N = 80)*

Learner theme	Participants (n)	Prevalence (%)
Enhanced error awareness	52	65
Structured revision support	48	60
Difficulty evaluating feedback	29	36
Emerging dependency	18	23

Difficulty evaluating feedback appropriateness was raised by 29 learners (36%). “Sometimes the AI gave recommendations that didn’t match my meaning, and I wasn’t sure whether to accept them or not. I needed more guidance to decide.” Finally, emerging dependency was acknowledged by 18 learners (23%). “I relied on the AI to fix my grammar automatically, so I didn’t always think about why the correction was made.” Across the four themes, prevalence followed a clear ordering, from enhanced error awareness (the most frequently reported) to emerging dependency (the least frequently reported). Consistent with Braun and Clarke’s (2006) approach, these frequencies are reported as indicative of prevalence within the present sample of 80 reflections rather than as quantitative measures intended for statistical generalisation; a theme was counted as present for a learner when it was substantively expressed in that learner’s reflections, and a learner could contribute to more than one theme.

Instructor Perspectives

Semi-structured interviews with 10 ESL instructors identified five principal themes. These perspectives are treated as complementary to the learner data rather than as corroborating or independent validation.

Theme 1 Efficiency gains: All 10 instructors acknowledged that AI tools reduced time spent on surface-level error marking, freeing capacity for higher-order feedback. Representative: “AI definitely saves time on correcting basic grammar mistakes. Instead of marking every article or verb tense error, I can focus on argumentation and coherence. But I still have to check whether students understand why the correction was made.”

Theme 2 Risks of uncritical revision: Concerns about uncritical revision were prominent across interviews. “Many students accept every suggestion without questioning it. Their final drafts look polished, but when I ask them why they made certain changes, they can’t explain.” This theme aligns with the proficiency-linked cognitive engagement patterns observed in learner reflections and points to a shared priority for future scaffolding design.

Theme 3 Uneven engagement quality: Instructors observed wide variation between learners: “A few students used the feedback very effectively. They compared alternatives, asked questions, and reflected. Others simply clicked through all suggestions.” This echoes the behavioural–cognitive distinction in learner engagement data reported above.

Theme 4 Ethical concerns (authorship and integrity): “It’s getting harder to know what part of the writing is the student’s and what is the tool’s. If AI rewrites entire sentences, it becomes a question of authorship.” These concerns were raised by all 10 instructors and centred on LLM-generated reformulations rather than AWE corrections.

Theme 5 Qualified support for a structured hybrid model: Despite these concerns, all instructors expressed support for a model in which AI handles first-round error correction while teachers address

higher-order concerns: “AI is great for first-round corrections. After that, I can work on higher-order issues like structure, clarity, and argumentation. The combination works better than either one alone.” It should be noted that instructors were not embedded in the AI-mediated feedback cycle during the intervention itself; their perspectives therefore represent professional judgements rather than direct experience of the partnership model as implemented.

Discussion

The findings of this study provide a nuanced account of the potential and limitations of AI-mediated writing support in L2 instruction. Across quantitative and qualitative data strands, a consistent pattern emerges: AWE tools are associated with accessible and measurable surface-level error reduction; LLM-based feedback extends to discourse-level concerns; and both types of AI support are insufficient as standalone feedback mechanisms for promoting deep cognitive engagement, higher-order revision, and autonomous writing development. These observations must be qualified by the design’s principal constraints, which include the absence of a control group, the use of non-equivalent essay prompts, and the confounded tool sequence, and together preclude causal claims.

Writing Performance and Differential Tool Patterns

The statistically significant improvements across all writing dimensions, with large effect sizes, are consistent with prior research demonstrating that timely, explicit, structured feedback can promote error noticing and correction (Bitchener & Knoch, 2008; Barrot, 2021). However, effect sizes of this magnitude ($d = 0.92$ – 1.39 over four weeks) in a design without a control group should prompt scrutiny. Practice effects arising from writing argumentative essays twice in the same genre and prompt familiarity effects are plausible alternative explanations that cannot be ruled out. The effect sizes should therefore be reported as upper-bound estimates rather than evidence of a uniquely AI-driven effect. The largest gains in grammatical accuracy align with evidence that AWE tools are most effective when targeting high-frequency, rule-governed error types (Barrot, 2021; Warschauer & Ware, 2006). The more modest improvements in lexical variety and cohesion suggest that discourse-level features are less consistently addressed by automated systems, although learner reflections on ChatGPT’s utility for organisation and lexical reformulation indicate that LLM-based feedback may partially address these dimensions. These observations support a tentative differentiated view of AI-mediated feedback: learners perceived AWE and LLM tools as serving complementary rather than redundant functions, with the pedagogically sequenced use of both tools perceived as providing support across multiple dimensions of the writing process.

Learner Engagement Across Three Dimensions

The three-dimensional engagement framework (behavioural, cognitive, affective) provides a more differentiated account of learner interaction with AI feedback than has been reported in prior studies focused primarily on surface-level revision outcomes (Koltovskaia, 2020; Ranalli, 2021). Behavioural engagement was high. Cognitive engagement was more variable and strongly mediated by proficiency: higher-proficiency learners evaluated suggestions selectively and exercised independent judgment, while lower-proficiency learners often accepted all recommendations without evaluation: a pattern consistent with Koltovskaia’s (2020) findings, now observed in a larger sample. Affective engagement was broadly positive, with learners reporting reduced anxiety and enhanced confidence; however, affective investment also manifested as dependency. The dependency finding rests on a single Likert item ($M = 3.2$, $SD = 0.9$), and should be treated as indicative rather than conclusive; a validated multi-item instrument would be required to characterise dependency more reliably.

Instructor Perspectives

Instructor perspectives complement the learner data, reinforcing the view that AI feedback is most educationally valuable as a preliminary mechanism, with teacher-led instruction addressing dimensions that automated tools cannot reliably support (Mahapatra, 2024; Teng, 2024). This is consistent with Zhang et al. (2025), who found in a controlled comparative study that combining GenAI feedback with human tutor input produced significantly greater gains in organisation, critical thinking, and sentence variety than AI feedback alone, directly supporting the case for a structured hybrid model over AI-only approaches. The ethical concerns raised by instructors, particularly regarding authorship and the challenge of distinguishing learner-generated from AI-generated text, echo broader debates in higher education (Barrot, 2023) and remain largely unresolved at the institutional level. Because instructors were not directly embedded in the feedback loop during the intervention, their perspectives reflect professional judgements about AI tools generally rather than direct experience of the specific three-tool sequence used by their students.

Theoretical Implications

Foregrounding feedback literacy (Carless & Boud, 2018) as the primary theoretical lens offers the sharpest interpretive purchase on the present findings. Learners who demonstrated strategic, evaluative engagement with AI feedback can be understood as exercising developing feedback literacy, while those who engaged behaviourally but not cognitively require explicit instruction in how to interpret, interrogate, and selectively apply automated suggestions. The proficiency-linked engagement pattern suggests that feedback literacy development is likely to be uneven, with lower-proficiency learners requiring more deliberate scaffolding. This is consistent with recent empirical work showing that GenAI-supported feedback practices can enhance feedback literacy when embedded within structured pedagogical scaffolding, including preparatory instruction and reflective reinforcement, but show limited impact on learners' capacity to make independent judgements about feedback quality (Lu et al., 2026). Interpreted through Schmidt's (1990) Noticing Hypothesis, AI-mediated error highlighting appears to function as an external mechanism for directing learner attention to salient linguistic forms, consistent with observed error reductions. However, the dependency patterns identified suggest that noticing facilitated by AI does not automatically translate into autonomous internalisation, particularly without metacognitive scaffolding. From a Sociocultural Theory perspective (Vygotsky, 1978), AI feedback partially simulates scaffolded interaction within the zone of proximal development, yet its inability to respond contingently to individual communicative intent underscores the irreplaceable role of the teacher as mediator. Sociocultural Theory's emphasis on authentic contingent interaction, however, sits somewhat awkwardly with the individual, asynchronous tool-use design of this study; its application here is therefore illustrative rather than central.

Proposed Conceptual Framework: Teacher-AI Partnership

Informed by the findings of this study and grounded in the literature on teacher-AI collaboration in language education (Stockwell & Wang, 2025; Tran, 2026), the following conceptual framework is proposed as a structure for future pedagogical design and empirical investigation. The framework is deliberately human-centred: consistent with recent calls in this journal to position AI as a pedagogical partner that complements, rather than replaces, teacher feedback, peer collaboration, and the development of learner feedback literacy (Tran, 2026), it treats automated tools as a first layer of support whose educational value is realised only through subsequent teacher mediation. It is important to note that the present study did not implement this framework directly. Specifically, Phase 3 (Teacher Mediation, described below) was not a formal component of the four-week intervention, during which teachers were not systematically embedded in the AI-mediated feedback cycle. The framework,

therefore, represents a theoretically motivated proposal rather than an empirically tested model, and empirical validation through future research, particularly research that directly involves teacher mediation alongside AI tools, is an essential next step.

Phase 1 AI Corrective Feedback: AWE tools (Grammarly, Write & Improve) are used for surface-level error correction, with structured training on how to evaluate and selectively apply suggestions.

Phase 2 AI Discourse Feedback: LLM-based tools (ChatGPT) are used for higher-order discourse feedback on organisation, cohesion, and lexical reformulation, with explicit instruction distinguishing this from corrective WCF.

Phase 3 Teacher Mediation: Instructors engage learners in guided reflection, peer discussion, and conferencing to address higher-order concerns, evaluate AI suggestions critically, and develop metacognitive awareness.

Revision logs and written reflections are embedded throughout all phases to promote self-regulation and mitigate dependency. Institutional guidelines on authorship, attribution, and responsible AI use are positioned as foundational prerequisites. This framework is proposed as a theoretically informed structure pending direct empirical testing with designs that incorporate formal teacher participation in the AI-mediated feedback process.

Conclusion

This study has investigated AI-mediated writing support in L2 instruction, examining the associations between a sequenced combination of AWE tools (Grammarly and Write & Improve) and an LLM-based tool (ChatGPT) and writing performance, learner engagement, and instructor perspectives in an Armenian higher education context. A central contribution of the study is its explicit acknowledgement of the conceptual distinction between corrective AWE feedback and broader LLM-based discourse feedback.

The findings demonstrate statistically significant pre-to-post improvements with large effect sizes in grammatical accuracy, lexical variety, cohesion, and overall writing quality among intermediate ESL learners. These findings should not be interpreted as evidence of a causal AI effect, given the absence of a control group, non-equivalent prompts, and the confounded tool sequence. Analysis across behavioural, cognitive, and affective engagement dimensions revealed that engagement quality was strongly mediated by proficiency. Moderate levels of perceived dependency, based on a single survey item, indicate a risk of overreliance that requires explicit pedagogical scaffolding. Instructor perspectives complemented the learner data while highlighting concerns about academic integrity, authorship, and the need for structured guidance to promote critical engagement.

The study's contributions include its empirical comparison of AWE and LLM-based tools within a single protocol, its systematic examination of all three engagement dimensions, and its theoretically grounded, conceptual teacher-AI partnership framework. Acknowledged limitations, such as the absence of a control group, the short intervention window, non-equivalent prompts, and the confounded tool sequence, define the boundaries of interpretation and point directly to the methodological priorities for the next generation of studies in this area.

The central insight that emerges from this study is not that AI feedback helps or harms L2 writing development, but that the observed associations are shaped by learner proficiency, the quality of pedagogical scaffolding, and the presence or absence of teacher mediation. As AWE and LLM-based tools

become ubiquitous in language learning environments, the field urgently needs controlled experimental designs that embed teacher participation alongside automated feedback, designs capable of testing whether deliberate feedback literacy instruction can convert behavioural tool engagement into durable cognitive and linguistic gains. This study provides the empirical baseline and the theoretical scaffolding for that next step.

Limitations

The following limitations are acknowledged and should be considered when interpreting the findings.

1. **Absence of a control group.** No control or comparison condition was included. Pre-to-post improvements cannot be causally attributed to the AI intervention; practice effects from writing argumentative essays twice, increased writing self-efficacy over four weeks, and regular classroom participation are plausible alternative explanations.
2. **Non-equivalent essay prompts.** Pre- and post-test essays used different prompts, meaning that prompt difficulty, topic familiarity, and genre alignment are uncontrolled variables. Pre- and post-test scores are therefore not directly comparable, limiting the validity of gain-score interpretation.
3. **Confounded tool sequence.** All participants followed the same Grammarly → Write & Improve → ChatGPT sequence, meaning that tool identity, sequential position, and cumulative revision experience are inextricably confounded. No causal or comparative claim about individual tools can be sustained by the data.
4. **Short duration.** The four-week intervention precludes conclusions about long-term interlanguage development, retention of gains, or the durability of reduced dependency over time.
5. **Single institutional context.** All participants were drawn from one institution. The generalisability of findings to other cultural, institutional, or curricular contexts — including other Armenian universities or non-Armenian settings — cannot be assumed.
6. **L1 heterogeneity.** The sample comprised primarily L1 Armenian speakers with a minority of Russian-L1 learners. Subgroup sizes were unequal and too small for systematic comparison. The potential influence of L1 background on error types and cognitive engagement with AI corrections therefore remains unexamined.
7. **Single-item dependency measure.** Perceived dependency was assessed via a single five-point Likert item. A validated multi-item scale would be required to characterise dependency more reliably and to support correlational analyses with other learner variables.
8. **Self-report measures.** Learner engagement data derived from questionnaires and written reflections are subject to social desirability bias and retrospective reconstruction. Behavioural engagement data from tool logs would provide a more objective complement to self-reported engagement.
9. **Member-checking sample size.** Member-checking was conducted with four learners and two instructors — a relatively small validation base relative to the breadth of qualitative claims made. This limits the strength of trustworthiness claims based on member-checking alone.
10. **Teacher non-participation in the feedback cycle.** Teachers were not systematically embedded in the AI-mediated feedback process during the intervention. The proposed teacher-AI partnership framework therefore cannot be understood as an empirical finding of this study; rather, it is a conceptual proposal for future research.

Data Availability Statement

The anonymised data supporting the findings of this study (writing scores, error-coding records, questionnaire responses, and de-identified learner reflections and interview transcripts) are available from

the corresponding author on reasonable request, subject to the conditions of the institutional ethics approval governing participant confidentiality.

Use of Generative AI

The authors declare the use of artificial intelligence, specifically Grammarly and QuillBot, solely for spelling, grammar, and language refinement. These tools did not influence the study design, data analysis, interpretation of results, or conclusions.

References

- Al-Sofi, B. (2026). The effectiveness of AI-generated feedback in enhancing students' writing proficiency: A study of Google Gemini. *Technology in Language Teaching & Learning*, 8, 103403. <https://doi.org/10.29140/tltl.2026.103403>
- Barrot, J. S. (2021). Using automated written corrective feedback in the writing classrooms: Effects on L2 writing accuracy. *Computer Assisted Language Learning*, 36(4), 584–607. <https://doi.org/10.1080/09588221.2021.1936071>
- 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>
- Bitchener, J., & Ferris, D. R. (2012). *Written corrective feedback in second language acquisition and writing*. Routledge.
- Bitchener, J., & Knoch, U. (2008). The value of written corrective feedback for migrant and international students. *Language Teaching Research*, 12(3), 409–431. <https://doi.org/10.1177/1362168808089924>
- Boneau, C. A. (1960). The effects of violations of assumptions underlying the t test. *Psychological Bulletin*, 57(1), 49–64. <https://doi.org/10.1037/h0041412>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, 43(8), 1315–1325. <https://doi.org/10.1080/02602938.2018.1463354>
- Ferris, D. R. (1999). The case for grammar correction in L2 writing classes: A response to Truscott (1996). *Journal of Second Language Writing*, 8(1), 1–11. [https://doi.org/10.1016/S1060-3743\(99\)80111-6](https://doi.org/10.1016/S1060-3743(99)80111-6)
- Ferris, D. R. (2004). The 'grammar correction' debate in L2 writing: Where are we, and where do we go from here? *Journal of Second Language Writing*, 13(1), 49–62. <https://doi.org/10.1016/j.jslw.2004.04.005>
- Ferris, D. R. (2014). Responding to student writing. In M. Celce-Murcia, D. M. Brinton, & M. A. Snow (Eds.), *Teaching English as a second or foreign language* (4th ed., pp. 222–237). National Geographic Learning.
- 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>
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Lu, Q., Yao, Y., Xiao, L., Zhu, X., & Yin, H. (2026). Exploring the impact of a GenAI-supported feedback practice on student feedback literacy in L2 writing. *Computer Assisted Language Learning*. Advance online publication. <https://doi.org/10.1080/09588221.2025.2605541>

- Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, 23, 151–169. <https://doi.org/10.1146/annurev.publhealth.23.100901.140546>
- Luo, Y., Xia, Y., & Lu, X. (2025). A systematic review of the role of artificial intelligence in second language writing education. *Digital Studies in Language and Literature*, 2(2), 302–325. <https://doi.org/10.1515/dsll-2025-0001>
- Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments*, 11(1), 9. <https://doi.org/10.1186/s40561-024-00295-9>
- Rahimi, M., Fathi, J., & Zou, D. (2025). Exploring the impact of automated written corrective feedback on the academic writing skills of EFL learners: An activity theory perspective. *Education and Information Technologies*, 30, 2691–2735. <https://doi.org/10.1007/s10639-024-12896-5>
- Ranalli, J. (2021). L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing*, 52, 100816. <https://doi.org/10.1016/j.jslw.2021.100816>
- Sari, E. (2024). The impact of automated writing evaluation on English as a foreign language learners' writing self-efficacy, self-regulation, anxiety, and performance. *Journal of Computer Assisted Learning*, 40(5), 2065–2080. <https://doi.org/10.1111/jcal.13004>
- Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11(2), 129–158. <https://doi.org/10.1093/applin/11.2.129>
- Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Tseng, W., Warschauer, M., & Kennedy, 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>
- Stockwell, G., & Wang, Y. (2025). Framing human-AI collaboration in language education. *Technology in Language Teaching & Learning*, 7(4), 103824. <https://doi.org/10.29140/tltl.v7n4.103824>
- Teng, M. F. (2024). “ChatGPT is the companion, not enemies”: EFL learners' perceptions and experiences in using ChatGPT for feedback in writing. *Computers & Education: Artificial Intelligence*, 6, 100215. <https://doi.org/10.1016/j.caeai.2024.100215>
- Tran, H. (2026). How does AI work as a “game changer”? Self-determination and influencing factors in EFL students' use of AI for academic writing. *Technology in Language Teaching & Learning*, 8, 103723. <https://doi.org/10.29140/tltl.2026.103723>
- 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>
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
- Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, 10(2), 157–180. <https://doi.org/10.1191/1362168806lr188oa>
- Zhang, Z., Aubrey, S., Huang, X., & Chiu, T. K. F. (2025). The role of generative AI and hybrid feedback in improving L2 writing skills: A comparative study. *Innovation in Language Learning and Teaching*. Advance online publication. <https://doi.org/10.1080/17501229.2025.2503890>

Appendix A

Writing Assessment Rubric (adapted from Bitchener & Ferris, 2012)

Each dimension was scored on a 0–100 percentage scale; the bands below summarise the descriptors applied during scoring.

Dimension	Excellent (85–100)	Good (70–84)	Adequate (55–69)	Weak (40–54)	Poor (<40)
Grammatical accuracy	Virtually free of grammatical error; complex structures used accurately.	Occasional minor errors that do not impede meaning; generally accurate control.	Noticeable errors in some structures; meaning generally clear but control inconsistent.	Frequent errors that sometimes obscure meaning; limited control of basic structures.	Pervasive errors that frequently impede comprehension.
Lexical variety	Wide, precise and varied vocabulary; natural word choice and collocation.	Good range with occasional imprecision or repetition; mostly appropriate choice.	Adequate but limited range; some repetition and inexact word choice.	Narrow vocabulary; frequent repetition and inappropriate word choice.	Very restricted vocabulary that limits the expression of ideas.
Cohesion & organisation	Logically organised with seamless paragraphing and varied cohesive devices.	Clear organisation with effective, if at times mechanical, cohesion.	Discernible organisation; cohesive devices present but used inconsistently.	Limited organisation; cohesion faulty or overused, affecting flow.	Little discernible organisation; lack of cohesion impedes reading.
Overall writing quality	Fully addresses the task with well-developed, coherent and fluent writing.	Addresses the task competently with generally coherent, developed writing.	Addresses the task adequately; development and coherence uneven.	Partially addresses the task; underdeveloped and frequently unclear.	Fails to address the task; largely undeveloped or incoherent.

Appendix B

Coding Frame

Codes were organised under the three engagement dimensions and the four learner themes reported in the Results.

Code/category	Definition
Behavioural engagement (dimension)	Consistent use of the tools and completion of multiple revision cycles, including adherence to the prescribed tool sequence.
Cognitive engagement (dimension)	Evaluating, comparing and selectively applying AI suggestions, and exercising independent judgement about feedback.
Affective engagement (dimension)	Emotional responses to AI use, including reduced anxiety and increased confidence and, conversely, dependency.
Enhanced error awareness (theme)	Reflections in which learners reported that AI highlighting made recurrent error patterns more visible and easier to notice and internalise.
Structured revision support (theme)	Reflections describing AI (particularly ChatGPT) as helping learners organise paragraphs and approach revision more systematically.
Difficulty evaluating feedback appropriateness (theme)	Reflections expressing uncertainty about whether to accept AI suggestions that did not match intended meaning, and a wish for more guidance.
Emerging dependency (theme)	Reflections indicating reliance on AI to correct work without understanding the reason, or reduced trust in the learner's own judgement.