ISSN 2663-8169 Vol. 8, Issue 7: 26-32, DOI: 10.25236/IJNDE.2026.080705

AI Large Language Models and Second Language Writing Development: A SLA Perspective

Hongxia Li[1], Xin Tuo[2]

[1]Xi'an Innovation College of Yan'an University, Xi'an, China
[2]Xi'an Innovation College of Yan'an University, Xi'an, China

Abstract: *This study examines how AI large language models (LLMs) affect second language (L2) writing development from a second language acquisition (SLA) perspective. Drawing on the Input, Noticing, Interaction, and Output Hypotheses and Sociocultural Theory, it investigates whether LLM-mediated instruction supports accuracy, lexical complexity, syntactic variety, coherence, and metalinguistic awareness. A mixed-methods quasi-experimental design involved 82 Chinese EFL undergraduates in an LLM-assisted experimental group (n = 42) and a teacher-feedback control group (n = 40) over 16 weeks. Writing tests were analyzed with t-tests and ANCOVA, while interviews and journals supplied qualitative evidence. Results showed significantly stronger gains in the experimental group (p < .001), especially in accuracy, lexical diversity, and coherence. Learners also reported stronger noticing, richer input, more pushed output, and lower anxiety, though over-reliance, generic feedback, and integrity concerns remained. The study concludes that LLMs can serve as effective SLA mediators within structured, teacher-guided pedagogy.*

Keywords: *Large Language Models, Second Language Writing, Second Language Acquisition, SLA Theories, EFL Writing, AI-Assisted Feedback, Writing Development, Sociocultural Theory*

1. Introduction

Second language (L2) writing is demanding because learners must coordinate grammar, vocabulary, discourse organization, and rhetoric in a non-native system [1]. Traditional instruction is often limited by large classes, delayed feedback, and insufficient individual support, leaving persistent problems in accuracy, lexical appropriacy, organization, and idea development (Zhang, 2021)[2]. These constraints have intensified interest in digital tools that can supplement teacher feedback.

Generative AI large language models (LLMs), including GPT-4 and ChatGPT, differ from earlier automated writing evaluation tools. Instead of mainly detecting errors or assigning scores, LLMs can offer interactive feedback, lexical paraphrasing, rhetorical scaffolding, and sustained dialogue (OpenAI, 2022)[3][4], aligning with SLA principles such as input, noticing, interaction, and output [5].

However, empirical research has not yet fully explained LLM effects through established SLA theories [6][7]. Many studies emphasize perceived usefulness or final writing scores rather than the mechanisms by which LLMs may support acquisition across planning, drafting, revision, and reflection. This study addresses that gap through a mixed-methods investigation with Chinese EFL undergraduates.

1.1. Research Questions

The study addresses four research questions:

How does LLM-assisted instruction affect L2 writing performance compared with traditional teacher feedback?

How do LLMs mediate writing development through input, noticing, interaction, output, and sociocultural scaffolding?

What benefits, challenges, and affective changes do learners report in LLM-supported writing?

What pedagogical conditions make LLMs effective SLA mediators?

ISSN 2663-8169 Vol. 8, Issue 7: 26-32, DOI: 10.25236/IJNDE.2026.080705

1.2. Theoretical Framework

The study draws on five SLA theories that together explain cognitive, interactional, and sociocultural dimensions of L2 writing development.

Krashen's [8] Input Hypothesis argues that acquisition requires comprehensible input slightly beyond learners' current competence. LLMs can operationalize this principle by adjusting vocabulary, syntax, and explanations to individual proficiency levels. Schmidt's [9] Noticing Hypothesis further suggests that learners must consciously attend to linguistic forms before input becomes intake; LLM feedback can promote this process by explaining errors and highlighting gaps between learner output and target norms.

Long's [10] Interaction Hypothesis emphasizes negotiation of meaning, while Swain's [11] Output Hypothesis stresses the value of pushed output. Unlike static AWE tools, LLMs allow learners to question feedback, request clarification, test alternatives, and revise independently. These dialogic revision cycles can support both interactional negotiation and syntactic processing.

Vygotsky's [12] Sociocultural Theory views learning as mediated by tools and interaction within the Zone of Proximal Development. In this sense, LLMs can function as digital scaffolds that help learners perform beyond their independent capacity while gradually internalizing writing strategies. Together, these theories frame LLMs as potential SLA mediators rather than merely editing tools.

2. Literature Review

2.1. From AWE to LLMs: Evolution of AI in L2 Writing

AI support for L2 writing began with AWE systems such as e-rater, Criterion, and Pigai, which offered automatic scoring and basic error feedback (Educational Testing Service, 2018)[13]. Although useful for reducing teacher workload, these tools were often formulaic, surface-oriented, and unable to address rhetorical or content-level writing quality [14]. Their one-way feedback also limited the interaction valued in SLA research.

Transformer-based LLMs differ qualitatively from earlier AWE tools. They can sustain dialogue, explain errors, provide level-appropriate examples, support ideation and outlining, suggest collocations, and comment on coherence [4][15][16][17][18]. These functions expand AI writing support from correction toward interactive learning assistance.

2.2. Empirical Evidence on LLMs and L2 Writing

Recent studies generally report that LLM-based tools can improve grammatical accuracy, lexical diversity, organization, motivation, and self-efficacy, while reducing writing anxiety. Learners also report heightened awareness of recurring linguistic problems, a finding consistent with the Noticing Hypothesis. Yet findings remain mixed because designs, proficiency levels, and intervention protocols vary widely.

Evidence also suggests proficiency-mediated effects. Lower-proficiency learners often benefit from targeted form-focused correction, whereas advanced learners gain more from rhetorical and content-level feedback that develops argumentation, voice, and register [19][20]. LLM interaction protocols should therefore be matched to learner profiles.

Concerns persist. Over-reliance on generated text may weaken independent writing, and LLM feedback can be generic when tasks require disciplinary or rhetorical judgment. Academic integrity also requires disclosure norms, ethical guidance, and classroom routines that frame AI as learning support rather than shortcut production [21].

2.3. Research Gaps

Three gaps remain. Many studies lack explicit SLA grounding, product-oriented measures often overshadow process analysis, and mixed-methods evidence linking outcome gains with learner experience remains limited. This study addresses these gaps by combining quantitative writing data with qualitative evidence of developmental mechanisms.

ISSN 2663-8169 Vol. 8, Issue 7: 26-32, DOI: 10.25236/IJNDE.2026.080705

3. Methodology

3.1. Research Design

A mixed-methods quasi-experimental pretest-posttest control group design was used. Quantitative data came from analytic writing scores analyzed through t-tests and ANCOVA. Qualitative data came from interviews, weekly journals, and classroom observations, allowing the study to examine both outcome differences and the mechanisms behind them.

3.2. Participants

Participants were 82 non-English majors at a university in Northwest China (ages 18-22; 44 male, 38 female). The experimental group (n = 42) received LLM-assisted instruction, while the control group (n = 40) received traditional teacher feedback. All were intermediate EFL learners (CET-4: 390-470; IELTS equivalent: 4.5-5.5), and the same teacher taught both groups.

3.3. Instruments

3.3.1. Writing Tests

Pre- and post-tests required 300-350 word argumentative essays completed in 40 minutes on parallel topics. Essays were scored with a validated five-component rubric covering content, organization, vocabulary, grammar, and mechanics. Two trained raters scored all essays independently, producing strong reliability (ICC = 0.88, $p < .001$); disagreements were resolved through discussion.

3.3.2. LLM Tools and Protocol

The EG used GPT-4 and ChatGPT through a five-stage protocol: prompt analysis, outline and vocabulary support, drafting feedback, independent revision after LLM feedback, and reflective post-writing dialogue. The protocol was aligned with the SLA framework and piloted with non-participant students to ensure clarity and appropriate challenge.

3.3.3. Qualitative Instruments

Qualitative data were collected from semi-structured interviews with 12 EG participants and weekly learning journals. These sources documented perceived usefulness, challenges, affective responses, feedback episodes, strategy changes, and metalinguistic awareness. Data were transcribed and analyzed thematically.

3.4. Procedure

The intervention lasted 16 weeks, with two 90-minute sessions each week. EG students used LLMs for planning, vocabulary building, drafting support, revision, and reflection, but were required to revise independently before optional teacher consultation. The CG followed conventional instruction: teacher lecture, model text analysis, independent writing, and written teacher feedback returned one to two weeks later.

3.5. Data Analysis

Quantitative data were analyzed in SPSS 26.0. Pretest equivalence was checked with independent samples t-tests, and ANCOVA compared posttest scores while controlling for pretest performance. Effect sizes were reported using Cohen's d and partial eta-squared (η^2). Qualitative data were analyzed using Braun and Clarke's [22] thematic procedure, then interpreted in relation to the SLA framework.

4. Results

4.1. Quantitative Results

4.1.1. Pretest Equivalence

Pretest scores showed no significant group difference (EG: M = 62.14, SD = 5.68; CG: M = 61.78, SD = 5.83; $t(80) = 0.29$, $p = .77$), confirming baseline equivalence (See Table 1).

ISSN 2663-8169 Vol. 8, Issue 7: 26-32, DOI: 10.25236/IJNDE.2026.080705

Table 1: Descriptive Statistics for Pretest and Posttest Writing Scores.

Group	N	Pretest M (SD)	Posttest M (SD)	Gain M (SD)
EG	42	62.14 (5.68)	79.88 (6.25)	17.74 (4.43)
CG	40	61.78 (5.83)	68.60 (5.79)	6.82 (3.68)

4.1.2. Posttest ANCOVA

ANCOVA showed a significant and large group effect, $F(1, 79) = 61.34$, $p < .001$, $\eta^2 = .44$. The adjusted EG posttest mean ($M = 79.62$) exceeded the CG mean ($M = 68.81$) by about 10.8 points, indicating a substantial intervention effect (See Table 2).

Table 2: ANCOVA Results for Posttest Writing Scores.

Source	Type III SS	df	F	p	η^2
Pretest (covariate)	1342.56	1	44.27	< .001	.36
Group	1863.79	1	61.34	< .001	.44
Error	2401.88	79	—	—	—

4.1.3. Subcomponent Analyses

The EG also outperformed the CG across all five writing subcomponents (Table 3). The strongest effects appeared in grammatical accuracy ($d = 1.72$), lexical diversity ($d = 1.46$), and organizational coherence ($d = 1.33$), matching the main affordances of LLM feedback: form explanation, vocabulary enrichment, and structural scaffolding (See Table 3).

Table 3: Between-Group Comparisons for Writing Subcomponents (Posttest).

Subscale	Group	M (SD)	t	p	Cohen's d
Content	EG	15.90 (2.11)	5.03	< .001	1.01
	CG	13.72 (2.28)			
Organization	EG	16.21 (2.23)	5.81	< .001	1.33
	CG	13.33 (2.15)			
Vocabulary	EG	15.84 (2.06)	6.42	< .001	1.46
	CG	12.88 (2.21)			
Grammar	EG	16.52 (2.14)	7.34	< .001	1.72
	CG	12.65 (2.30)			
Mechanics	EG	15.53 (2.09)	4.41	< .001	0.89
	CG	13.75 (2.26)			

4.2. Qualitative Results: Thematic Analysis

Thematic analysis generated four themes that help explain the quantitative gains and connect them to SLA mechanisms.

1) Theme 1: Enhanced Noticing and Metalinguistic Awareness

EG participants reported that LLM feedback changed how they attended to form because it explained errors rather than simply marking them:

"LLMs explain why errors occur, not just mark them. I now notice patterns I missed before." (Student 4)

"I catch mistakes while writing because the LLM trained me to notice grammar rules." (Student 9)

These comments support Schmidt's [9] Noticing Hypothesis. LLM explanations drew learners' attention to gaps between interlanguage forms and target norms, helping input become intake. Journals showed that learners began monitoring articles, verb aspect, and clause structure more independently, suggesting partial internalization of feedback-based strategies.

2) Theme 2: Personalized Comprehensible Input

Participants also perceived LLM input as more closely matched to individual proficiency than ordinary classroom input:

"LLMs give natural phrases at my level—not too hard, not too easy. I understand and absorb them."(Student 2)

ISSN 2663-8169 Vol. 8, Issue 7: 26-32, DOI: 10.25236/IJNDE.2026.080705

"I get many i+1 expressions that I can use in my own writing."(Student 7)

These responses align with Krashen's [8] Input Hypothesis. Learners experienced LLM output as understandable yet slightly challenging, and the low-pressure interaction reduced anxiety. This lowered affective filter encouraged experimentation with more complex grammar and vocabulary.

3) Theme 3: Pushed Output and Interactive Revision

Participants described LLM-supported revision as demanding and interactive rather than passive:

"I rewrite sentences many times with LLM feedback. It pushes me to be accurate and clear." (Student 6)

"Dialogue with LLMs helps me clarify ideas. Negotiation makes my writing stronger." (Student 11)

These accounts reflect Swain's [11] Output Hypothesis and Long's [10] Interaction Hypothesis. Students had to produce modified output, ask follow-up questions, and evaluate alternatives. From a sociocultural perspective, the LLM served as a temporary scaffold that enabled stronger performance while supporting gradual movement toward independent competence.

4) Theme 4: Affective Outcomes and Challenges

Participants also reported reduced writing anxiety, greater confidence, and stronger motivation to practice writing beyond class. These affective changes likely supported learning by making students more willing to test unfamiliar forms and revise repeatedly.

Challenges were also evident. About 70% of interviewees admitted a temptation to rely too heavily on LLM-generated text, though some deliberately drafted first to preserve independent effort. Advanced learners sometimes found rhetorical feedback too generic. Participants also expressed uncertainty about acceptable AI use in assessment, highlighting the need for clear institutional policies.

5. Discussion

5.1. LLMs as SLA Mediators: Theoretical Alignment

The findings show that LLMs can mediate L2 writing through reinforcing SLA mechanisms. Gains in accuracy, vocabulary, and coherence are consistent with comprehensible input [8], noticing [9], negotiation of meaning [10], pushed output [11], and ZPD scaffolding [12]. LLM effects appear to arise from cycles of input, explanation, interaction, revision, and reflection.

The strongest gains in grammar ($d = 1.72$) and vocabulary ($d = 1.46$) likely reflect the immediacy and specificity of LLM feedback. Mechanics showed a smaller but still meaningful effect ($d = 0.89$), possibly because learners already possessed basic convention knowledge before the intervention.

5.2. LLMs vs. Traditional Instruction: Key Advantages

Compared with traditional instruction, LLM-mediated support offered three advantages: immediate feedback, individualized responses to learner errors, and expanded target-language exposure. Because interaction was also non-judgmental, students reported less anxiety and more willingness to experiment with unfamiliar forms.

5.3. Challenges and Pedagogical Safeguards

The risks of over-reliance, generic feedback, and integrity problems require proactive pedagogy. Over-reliance can be reduced by requiring independent first drafts, learner explanations of revision choices, and reflective comparison between original and revised texts. Such routines preserve productive struggle while using LLMs for formative support.

Generic rhetorical feedback should be handled through teacher-AI complementarity. LLMs are well suited to frequent form-focused feedback, lexical precision, and sentence-level coherence, while teachers remain essential for argument quality, critical voice, disciplinary register, and audience awareness.

Academic integrity requires clear disclosure policies and LLM literacy education. Students need to understand what LLMs can and cannot do, how to evaluate outputs critically, and how to use AI for learning rather than substitute production [21].

ISSN 2663-8169 Vol. 8, Issue 7: 26-32, DOI: 10.25236/IJNDE.2026.080705

5.4. LLMs Across the Writing Process

LLM value also differed across writing stages. In pre-writing, LLMs supported idea generation, organization, and vocabulary planning. In drafting, they helped maintain accuracy and cohesion. In revision, feedback made gaps visible and motivated rewriting. Post-writing dialogue then promoted metacognitive awareness of progress and strategies.

5.5. Pedagogical Implications

Four principles follow: use LLMs in a human-in-the-loop model; teach LLM literacy, including prompting and output evaluation; embed AI use in process-oriented curricula that value revision and reflection; and establish evidence-based policies that protect integrity while permitting legitimate learning support.

6. Conclusion

This study shows that LLMs, when used within a theoretically grounded framework, can support L2 writing development. By combining input, noticing, interaction, pushed output, and ZPD scaffolding, they offer scalable assistance that complements teachers. The EG's stronger gains and qualitative evidence support cautiously optimistic conclusions.

At the same time, LLMs are not autonomous teachers. Their value depends on human design, ethical guidance, and classroom routines that prevent over-reliance. The most promising model is collaborative: LLMs provide immediate, individualized linguistic scaffolding, while teachers provide judgment, rhetorical expertise, and relational support.

6.1. Limitations

Several limitations should be noted. The sample came from one Chinese university, limiting generalizability. The 16-week duration cannot show long-term retention. The study focused on GPT-4 and ChatGPT, so other models may produce different outcomes. Finally, interaction logs were not systematically analyzed.

6.2. Future Research

Future research should use longitudinal and cross-cultural designs, include different proficiency levels, and examine pragmatic competence, academic voice, and writer identity. More work is also needed on teacher development for sustainable and ethical LLM integration.

References

[1] Hyland, K. (2019). The Routledge handbook of second language acquisition and writing. Routledge.
[2] Li, M. (2023). AI large language models in college EFL writing: Effects and challenges. Journal of Higher Education, 44(5), 112–118.
[3] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.
[4] Luo, Y., Zhang, L., & Wang, H. (2025). A systematic review of AI in second language writing education. Digital Studies in Language and Literature, 2(1), 309–332. https://doi.org/10.1515/dsll-2025-0001
[5] Ellis, R. (2023). The study of second language acquisition (3rd ed.). Oxford University Press.
[6] Chen, G. (2023). AI-assisted L2 writing: Practices, perceptions, and SLA implications. Journal of Second Language Writing, 41, 100445. https://doi.org/10.1016/j.jslw.2023.100445
[7] Guo, S., Zhang, Y., & Li, J. (2024). ChatGPT feedback and L2 writing revision quality. Language Learning & Technology, 28(1), 45–68.
[8] Krashen, S. D. (1985). Principles and practice in second language acquisition. Pergamon.
[9] Schmidt, R. (1990). The role of consciousness in second language learning. Applied Linguistics, 11(2), 129–158.
[10] Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C.

ISSN 2663-8169 Vol. 8, Issue 7: 26-32, DOI: 10.25236/IJNDE.2026.080705

Ritchie & T. K. Bhatia (Eds.), Handbook of second language acquisition (pp. 413–468). Academic Press.
[11] Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), Principles and practice in applied linguistics (pp. 125–144). Oxford University Press.
[12] Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Harvard University Press.
[13] Dalgish, S. M., & Smith, B. L. (2019). Automated writing evaluation: Research and practice. The Modern Language Journal, 103(Suppl. 1), 1–7. https://doi.org/10.1111/modl.12573
[14] Chen, H., & Cheng, L. (2023). Automated writing evaluation in L2 writing: A systematic review. Computer Assisted Language Learning, 36(3), 345–372. https://doi.org/10.1080/09588221.2022.2098761
[15] Darmawansah, I., Suciati, I., & Wardani, S. (2025). ChatGPT as a pre-writing scaffold for L2 argumentative writing. Digital Studies in Language and Literature, 2(1), 123–145. https://doi.org/10.1515/dsll-2025-0003
[16] Hsu, L., & Tai, J. (2024). Interactive AI feedback for L2 writing development: A mixed-methods study. Journal of Educational Technology & Society, 27(1), 78–95.
[17] Werdiningsih, I., Puspita, A., & Saputra, D. (2024). ChatGPT for conversational writing feedback: Learner perceptions and writing outcomes. TESOL Quarterly, 58(1), 201–228.
[18] Xiao, Y. (2024). LLM-assisted lexical development in EFL writing: A corpus-informed analysis. Journal of English for Academic Purposes, 68, 101350.
[19] Boudouaia, M., Li, Y., & Zhang, L. (2024). Proficiency differences in L2 writing development with AI feedback. Journal of Computer-Assisted Language Learning, 37(2), 189–212. https://doi.org/10.1080/09588221.2023.2234567
[20] Wang, Y., & Zhang, X. (2024). Proficiency-mediated LLM effects on L2 writing quality. System, 120, 103218.
[21] Huang, W., Jia, C., Hew, K. F., & Guo, J. (2024). Ethical use of LLMs in L2 writing education. Journal of Language, Identity, and Education, 23(2), 112–126. https://doi.org/10.1080/15348458.2023.2298765
[22] Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa