Jurnal Wahana Pendidikan, 13(2), 427-444, Agustus 2026
P-ISSN: 2355-2425 dan E-ISSN : 2715-6796

http://dx.doi.org/10.25157/jwp.v%vi%i.26521

Evidence on AI Tools in Second Language Writing: A Systematic Review of SLA Outcomes, Learner Experiences, and Pedagogical Dynamics

[1]Rifyal Kalam Mahardhika, [1]Margana Margana,[1]Rozanah Katrina Herda, [2] Bazilah Raihan Mat Shawal, [3]Shahzadi Hina Sain
[1] Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
[2] Universiti Malaysia Kelantan, Malaysia
[3]Beaconhouse National University, Pakistan
[1]Email: rifyalkalam.2025@student.uny.ac.id
[1]Email: margana@uny.ac.id
[1]Email: katrinaherda@uny.ac.id
[2]Email: bazilah@umk.edu.my
[3]Email: shahzadi.hina88@gmail.com

Abstract

Artificial intelligence (AI) tools have transformed second language (L2) writing instruction, yet empirical evidence spanning traditional Automated Writing Evaluation (AWE) and modern Generative AI (GenAI) remains fragmented. This systematic literature review synthesized 22 primary empirical studies published between 2018 and 2025 to evaluate the effects of AI tools on L2 writing performance and map associated pedagogical benefits, operational challenges, and ethical concerns within Second Language Acquisition (SLA) frameworks. Guided by PRISMA 2020 protocols, literature was gathered across five academic databases and analyzed using SLA performance constructs and thematic synthesis. Findings reveal that most included studies focusing on accuracy (11 out of 14) reported improvements in surface-level linguistic error reduction, whereas effects on syntactic complexity were variable across 8 studies depending on tool type (AWE vs. GenAI). GenAI also showed promising effects on coherence and structural organization when accompanied by explicit instructional scaffolding. While AI tools provide immediate scaffolding and reduce writing anxiety, operational challenges, such as declining engagement, feedback overload, and cognitive passivity, and ethical risks concerning academic integrity and loss of learner voice persist. The review concludes that AI tools are most productive when used as supportive writing companions within explicitly scaffolded and collaborative learning environments. Future research should employ longitudinal designs, include more diverse learner populations, and empirically examine critical AI literacy interventions and the transfer of AI-assisted gains to unassisted L2 writing. Educators should integrate AI critically while fostering learner agency, ethical awareness, and independent writing development.

Keywords: Artificial intelligence, Automated writing evaluation, Generative AI, L2 writing, Second language acquisition

Abstrak

Alat kecerdasan buatan (AI) telah mengubah pembelajaran menulis bahasa kedua (L2), tetapi bukti empiris yang mencakup Automated Writing Evaluation (AWE) tradisional dan Generative AI (GenAI) modern masih terfragmentasi. Tinjauan literatur sistematis ini mensintesis 22 studi empiris primer yang diterbitkan antara tahun 2018 dan 2025 untuk mengevaluasi pengaruh alat AI terhadap kemampuan menulis L2 serta memetakan manfaat pedagogis, tantangan operasional, dan persoalan etis yang terkait dalam kerangka Second Language Acquisition (SLA). Dengan berpedoman pada protokol PRISMA 2020, literatur dikumpulkan dari lima basis data akademik dan dianalisis menggunakan konstruk kemampuan menulis dalam SLA serta sintesis tematik. Hasil penelitian menunjukkan bahwa sebagian besar studi yang berfokus pada akurasi (11 dari 14 studi) melaporkan peningkatan dalam pengurangan kesalahan linguistik pada tingkat permukaan, sedangkan pengaruh terhadap kompleksitas sintaksis bervariasi dalam 8 studi, bergantung pada jenis alat yang digunakan (AWE atau GenAI).

GenAI juga menunjukkan pengaruh yang menjanjikan terhadap koherensi dan organisasi struktur teks ketika disertai dengan pemberian scaffolding pembelajaran secara eksplisit. Meskipun alat AI memberikan scaffolding secara langsung dan mengurangi kecemasan dalam menulis, berbagai tantangan operasional, seperti menurunnya keterlibatan, kelebihan umpan balik, dan kepasifan kognitif, serta risiko etis terkait integritas akademik dan hilangnya suara pembelajar masih tetap ada. Tinjauan ini menyimpulkan bahwa alat AI paling efektif ketika digunakan sebagai pendamping dalam kegiatan menulis di lingkungan pembelajaran yang didukung scaffolding secara eksplisit dan bersifat kolaboratif. Penelitian selanjutnya perlu menggunakan desain longitudinal, melibatkan populasi pembelajar yang lebih beragam, serta menguji secara empiris intervensi literasi AI kritis dan transfer peningkatan kemampuan yang diperoleh melalui bantuan AI ke dalam kegiatan menulis L2 tanpa bantuan AI. Pendidik perlu mengintegrasikan AI secara kritis sekaligus mengembangkan kemandirian pembelajar, kesadaran etis, dan kemampuan menulis secara mandiri.

Kata Kunci: AI generatif, Evaluasi penulisan otomatis, Kecerdasan buatan, Menulis bahasa kedua, Pemerolehan bahasa kedua

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

How to cite:

Mahardhika, R. K., Margana, Herda, R. K., Shawal, B. R. M., & Sain, S. H. (2026). The impact of AI tools on second language writing: A systematic literature review on SLA. *Jurnal Wahana Pendidikan*. 13 (2), 427-444

Article History:

Sent 17-08-2026, Revised 25-08-2026, Accepted 31-08-2026.

INTRODUCTION

The swift advancement of digital communication technology has revolutionized higher education, especially in language education programs that increasingly require students to demonstrate advanced writing proficiency, critical digital literacy, and global communication competence. In this setting, digital writing technologies have evolved from basic word processing software into sophisticated interactive systems that fundamentally alter how learners compose, revise, and communicate. For students in English as a Foreign or Second Language (EFL/ESL) programs, the incorporation of digital writing tools is increasingly relevant, as academic and professional communication now demands both linguistic precision and adaptive technology-mediated writing capabilities in global contexts.

Second language (L2) writing development is a complex, multi-dimensional cognitive and social endeavor central to Second Language Acquisition (SLA). In recent years, the rapid advancement of artificial intelligence (AI) tools, encompassing traditional Automated Writing Evaluation (AWE) systems, machine translation (MT), and generative AI (GenAI) powered by Large Language Models (LLMs), has fundamentally transformed L2 writing instruction and learning. In the context of SLA, AI writing tools refer to digital applications that utilize natural language processing (NLP) algorithms to provide automated feedback, stylistic reformulations, or text generation to assist writers. When integrated into language learning environments, these tools offer substantial pedagogical benefits: they deliver immediate, personalized feedback, scaffold drafting and revision, reduce cognitive load during composition, and foster learner self-regulation and autonomy in EFL/ESL writing contexts.

Over the past decade, scientific inquiry into technology-assisted L2 writing has evolved rapidly. Early empirical research primarily examined classic AWE systems, focusing on how computer-generated automated written corrective feedback (AWCF) influences student revision patterns, writing quality, and measures of Complexity, Accuracy, Lexical Diversity, and Fluency (CALF) (Ajabshir & Ebadi, 2023; Ranalli, 2021; Wei et al., 2023; Zhang, 2020). Studies demonstrated that AWE feedback could positively guide learners from writing process to written product (Shen et al., 2023). However, the period between 2023 and 2025 marked a paradigm shift with the emergence of GenAI models (e.g., ChatGPT). Recent empirical studies have increasingly explored LLMs as interactive writing companions capable of reformulating essays (Chen, 2025), assessing writing accuracy (Mizumoto et al., 2024), supporting collaborative writing in digital environments (Meniado et al., 2024; Teng, 2024; Wiboolyasarin et al., 2024), and facilitating individual versus collaborative feedback processing (Yan D, 2024; Yan & Zhang, 2024). Scholars have also investigated how GenAI impacts deliberation during revision (Michel et al., 2025) and shapes overall academic writing skills (Mahapatra, 2024), often benchmarking GenAI feedback against traditional teacher feedback (Lin & Crosthwaite, 2024).

Despite this growing body of research, current scientific knowledge exhibits notable theoretical and empirical limitations. Previous research syntheses have provided valuable insights into automated feedback; for example, Karatay and Karatay (2024) conducted a research synthesis restricted to classic AWE systems, while Shi and Aryadoust (2024) systematically reviewed pre-2023 automated written feedback technologies. However, these earlier syntheses did not encompass the transformative shift to LLM-powered Generative AI (2023–2025), nor did they examine how different AI modalities (AWE vs. GenAI vs. MT) distinctively affect macro-structural coherence, critical AI literacy (e.g., the APSE framework), and learner agency under an integrated SLA lens. Existing technologies often prioritize immediate error correction over global textual coherence and deeper conceptual development. Furthermore, current literature highlights significant variations in how learners cognitively and socioculturally regulate their tool usage (Y. Wang, 2024), as well as ambiguities regarding when, why, and for how long learners actively engage with GenAI (Hwang et al., 2025). Crucially, there is a noticeable gap in critical digital literacy among L2 writers, creating risks of over-reliance, ethical compromises, and passive acceptance of AI-generated content (Darvin, 2025; Wang & Wang, 2025). A comprehensive review that bridges the empirical shift from classic AWE paradigms to GenAI models across 2018–2025 is urgently needed to resolve these fragmented insights.

To address these limitations, this study presents a Systematic Literature Review (SLR) analyzing empirical research published between 2018 and 2025. The primary objective is to evaluate how AI tools impact L2 writing performance and to map the broader pedagogical, cognitive, and ethical landscape of AI integration in SLA. Specifically, this review addresses two main research questions disaggregated by tool type and contextual variables:

1) How do different AI tool categories (AWE, Machine Translation, Generative AI) influence specific L2 writing performance constructs (accuracy, complexity, coherence, revision dynamics) across varying proficiency levels and intervention durations in empirical studies from 2018–2025?
2) What pedagogical benefits, operational challenges, and ethical concerns emerge from using AI tools in L2 writing contexts?

It is hypothesized that while AI tools consistently facilitate immediate linguistic revision and accuracy, their overall effect on higher-level writing dimensions (coherence and complexity) is heavily mediated by learner engagement, instructional scaffolding, and critical AI literacy. The novelty of this review lies in its unified framework that bridges traditional AWE and modern GenAI eras, offering a

holistic synthesis of linguistic outcomes alongside learner agency and ethics. Ultimately, this research provides significant theoretical and practical impacts: it refines SLA models of technology-mediated writing, guides educators in designing balanced AI-assisted writing curricula and equips academic institutions with evidence-based insights to address ethical AI literacy in language education.

RESEARCH METHODOLOGY

Research Design and Framework

We adopted a systematic literature review design adhering to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) statement (Page et al., 2021) to map and synthesize empirical evidence regarding AI tool integration in L2 writing. Prior to screening, a review protocol detailing eligibility criteria, database search queries, quality appraisal standards, and extraction procedures was established *a priori* and registered on the Open Science Framework (OSF). No major protocol deviations occurred during the review process. PRISMA provided a transparent, rigorous, and replicable operational framework across four phases: identification, screening, eligibility, and inclusion.

Data Sources and Search Strategy

Systematic searches were conducted across five academic databases: Web of Science, Scopus, ERIC, ScienceDirect, and Google Scholar. The search targeted literature published between January 2018 and December 2025, with final searches executed on December 15, 2025. Queries were restricted to peer-reviewed article title, abstract, and keyword fields. To manage the high volume and dynamic nature of Google Scholar, results were sorted by relevance and limited to the top 50 highly relevant records.

Boolean search strings were constructed across three core domains: AI technologies, target learner populations, and L2 writing-related outcomes. The AI Tool Domain captured studies involving artificial intelligence, Generative AI, ChatGPT, Automated Writing Evaluation (AWE), and machine translation (MT). The Target Population Domain identified studies involving second-language writers, EFL learners, ESL learners, and related L2 writing populations. The SLA & Writing Metrics Domain focused the search on studies addressing writing performance and learning-related constructs, including accuracy, complexity, coherence, revision, feedback, ethics, and engagement.

Table 1.

Database Search Domains and Query Parameters

Search Domain	Keywords & Boolean Terms	Yield per Database
AI Tool Domain	"artificial intelligence" OR "generative AI" OR "ChatGPT" OR "automated writing evaluation" OR "AWE" OR "machine translation"	Web of Science: 145 Scopus: 160
Target Population	"second language writing" OR "L2 writing" OR "EFL" OR "ESL" OR "L2 writers"	ScienceDirect: 90 ERIC: 75
SLA & Writing Metrics	"accuracy" OR "complexity" OR "coherence" OR "revision" OR "feedback" OR "ethics" OR "engagement"	Google Scholar: 50
		Total Identified: 520

Table 1 summarizes the search domains, principal keywords, and the number of records identified across the five databases. The distribution of records demonstrates that the search strategy was deliberately broad enough to capture both established AWE research and the more recent

expansion of GenAI research in L2 writing. Across all five databases, the search initially identified 520 records, which subsequently underwent duplicate removal and systematic screening.

Eligibility Criteria and Language Selection

Explicit eligibility criteria were established before screening to isolate primary empirical research conducted in L2 writing contexts. The criteria were designed to ensure that the final evidence base directly addressed the effects, experiences, or pedagogical implications of AI-assisted L2 writing rather than merely discussing AI theoretically. As shown in Table 2, studies were included when they involved authentic AI writing technologies, examined L2/EFL/ESL writing, and provided primary quantitative or qualitative evidence. Studies were excluded when they were secondary literature, focused on L1 writing or non-writing applications, lacked primary empirical data, or did not provide sufficient methodological information. Non-English publications were also excluded because of resource constraints associated with consistently interpreting fine-grained SLA linguistic measures across languages. This language restriction is acknowledged as a potential source of publication bias.

Table 2.

Eligibility Criteria for Literature Selection

Inclusion Criteria	Exclusion Criteria
Primary empirical studies published in peer-reviewed journals (2018–2025).	Review articles, syntheses, meta-analyses, conceptual frameworks, dissertations, and book chapters.
Evaluation of authentic AI tools (AWE, GenAI/ChatGPT, MT) in L2/EFL/ESL writing settings.	Studies focusing on L1 writing, general non-writing language software, or traditional non-AI digital tools.
Measurement of L2 writing performance gains (CALF, revision) or qualitative learner/teacher experiences.	Non-empirical papers or studies lacking primary qualitative/quantitative data metrics.
Rigorous empirical design (experimental, quasi-experimental, mixed-methods, qualitative case studies).	Non-English publications or studies with incomplete methodologies.

The criteria in Table 2 served two complementary purposes. First, the inclusion criteria ensured that the review was restricted to empirical evidence directly relevant to AI-mediated L2 writing. Second, the exclusion criteria reduced conceptual and methodological heterogeneity by removing studies that did not provide primary evidence or did not directly address the review's population, intervention, or outcomes. Consequently, the final sample was designed to support meaningful comparisons across AWE, MT, and GenAI while retaining both quantitative performance outcomes and qualitative evidence concerning learner experiences, engagement, and ethical issues.

Selection Procedure and Inter-Rater Agreement

The screening process was executed independently by two reviewers across all phases to minimize selection bias. Database searches initially yielded 520 records. After removing 180 duplicate entries via reference management software, 340 unique records underwent title and abstract screening. Title and abstract screening eliminated 282 records failing to meet population, intervention, or design criteria. Subsequently, 58 full-text articles were retrieved and independently assessed for eligibility. Disagreements between the two reviewers were resolved through structured consensus meetings. Inter-rater reliability evaluated via Cohen's kappa demonstrated strong agreement at title/abstract screening (\kappa = 0.88) and full-text eligibility (\kappa = 0.84). Thirty-six full-text articles were excluded with documented reasons: non-primary empirical design/reviews (n = 14), non-L2 or non-writing focus (n = 12), and insufficient performance metrics (n = 10). Ultimately, 22 primary empirical studies met all criteria and were included in the final synthesis.

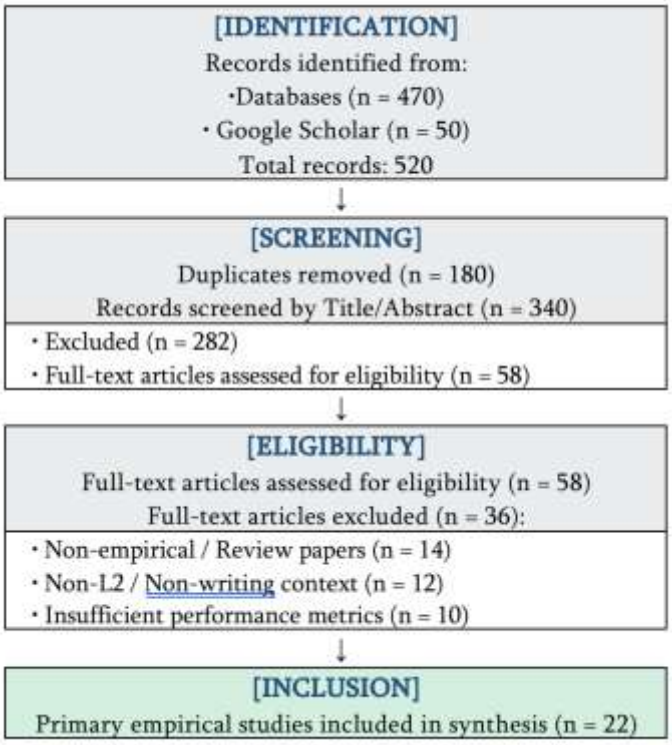

Figure 1. PRISMA 2020 Flowchart of the Literature Selection Process.

Figure 1 illustrates the complete study-selection pathway following the PRISMA 2020 framework. The identification phase began with 520 records retrieved from five academic databases. After 180 duplicate records were removed, 340 unique records remained for title and abstract screening. At this stage, 282 records were excluded because they did not satisfy the predefined population, intervention, or study-design requirements. The remaining 58 records were subjected to full-text assessment. Of these, 36 articles were excluded after detailed examination: 14 were reviews or other non-primary empirical studies, 12 did not address L2 writing or the target population, and 10 did not provide sufficient performance-related or relevant empirical evidence. The final 22 studies therefore represent the empirical evidence base that satisfied all inclusion criteria and formed the basis for the subsequent quality appraisal and thematic and narrative synthesis. This staged process ensured that study selection was transparent, reproducible, and consistent with the eligibility criteria established a priori.

Quality Appraisal and Risk of Bias

All 22 included primary studies underwent formal quality appraisal using the Mixed Methods Appraisal Tool (MMAT, 2018 version), which accommodates quantitative randomized controlled trials, non-randomized designs, qualitative studies, and mixed-methods research. Two reviewers independently scored each study across MMAT criteria (ranging from 0% to 100%). Studies were categorized as high quality (80–100%; n = 15) or moderate quality (60–75%; n = 7). Methodological limitations, such as short intervention durations, lack of control groups, or homogeneous university samples, were explicitly factored into the narrative synthesis to ensure claims reflect study rigor.

Characteristics of Primary Included Studies

To ensure transparency, the core parameters of the 22 included primary empirical studies are detailed in Table 3.

Table 3.

Characteristics of Included Primary Empirical Studies (N = 22)

Study	Setting & Sample	Target AI Tool	Design & Duration	SLA Outcome Measures	Main Findings
Ajabshir & Ebadi (2023)	Iran; 60 EFL Undergraduates	AWE (Grammarly) vs. Teacher	Quasi-experimental; 8 weeks	CALF measures across genres	AWE significantly improved accuracy in argumentative writing; minor impact on complexity.
Chen (2025)	China; 45 EFL Undergraduates	GenAI (ChatGPT reformulations)	Mixed-methods; 4 weeks	Structural revision & coherence	Essay reformulations enhanced macro-structural reorganization and rhetorical cohesion.
Darvin (2025)	Canada; 18 ESL Graduate Students	GenAI (ChatGPT)	Qualitative case study; 1 semester	Critical digital literacy & voice	Highlighted risks of voice homogenization and the need for critical AI literacy frameworks.
Hwang et al. (2025)	Korea; 82 EFL Undergraduates	GenAI (ChatGPT)	Longitudinal mixed-methods; 12 weeks	Engagement trajectories	Usage declined over time; cognitive overload led to surface-level edits without explicit prompts.
Lin & Crosthwaite (2024)	Hong Kong; 52 EFL Undergraduates	GenAI (GPT-4) vs. Teacher	Quasi-experimental; 6 weeks	Corrective feedback precision	GPT-4 feedback matched teacher accuracy for local errors but lagged in deep contextual nuance.
Li & Ke (2025)	China; 64 EFL Undergraduates	AWE & MT	Mixed-methods; 6 weeks	Revision depth & processing	High automated feedback volume caused cognitive passivity and uncritical error acceptance.
Mahapatra (2024)	India; 75 ESL Undergraduates	GenAI (ChatGPT)	Mixed-methods intervention; 4 weeks	CALF & academic writing	Gains observed in productive lexical diversity and complex clause linkages.
Meniado et al. (2024)	Thailand/Vietnam; 120 EFL Students	GenAI (ChatGPT)	Survey & Qualitative; 1 semester	Affective factors & perceptions	Reduced writing anxiety; acted as a non-judgmental

					interactive companion.
Michel et al. (2025)	Germany; 38 GFL Undergraduates	GenAI (ChatGPT)	Experimental (Individual vs. Collaborative); 3 weeks	Deliberation & revision process	Collaborative peer processing of AI prompts yielded deeper cognitive deliberation.
Mizumoto et al. (2024)	Japan; 40 EFL Undergraduates	GenAI (ChatGPT)	Quantitative evaluation; 4 weeks	Diagnostic accuracy assessment	ChatGPT showed high reliability in diagnostic grammar/orthographic error detection.
Ranalli (2021)	USA; 32 ESL Undergraduates	AWE (Grammarly)	Mixed-methods; 5 weeks	Learner engagement & trust	Passive uptake occurred when learners overly trusted automated surface corrections.
Shen et al. (2023)	China; 94 EFL Undergraduates	AWE (Pigai)	Longitudinal design; 10 weeks	Process-to-product revision	AWCF positively guided process-to-product revisions, though engagement fluctuated.
Teng (2024)	Hong Kong; 28 EFL Undergraduates	GenAI (ChatGPT)	Qualitative multiple-case; 8 weeks	Affective filter & feedback perception	Lowered affective filter; acted as a supportive dialogic partner during drafting.
Wang & Wang (2025)	China; 42 EFL Undergraduates	GenAI (ChatGPT)	Case study; 6 weeks	Critical AI literacy (APSE model)	Uncritical learners accepted hallucinated outputs; APSE training improved agency.
Wang (2024)	China; 50 EFL Undergraduates	MT (DeepL) & GenAI	Mixed-methods; 8 weeks	Self-regulated learning (SRL)	High reliance on MT/GenAI weakened internal monitoring and self-regulation.
Wei et al. (2023)	China; 110 EFL Undergraduates	AWE (Pigai)	Randomized Controlled Trial; 12 weeks	Grammatical accuracy gains	RCT confirmed significant surface error reduction in the experimental group.
Wiboolyasarin et al. (2024)	Thailand; 48 EFL Undergraduates	GenAI & Wiki platform	Mixed-methods intervention; 6 weeks	Collaborative writing proficiency	Wiki-integrated AI feedback improved structural co-construction and cohesion.
Xiaosa & Ping (2025)	China; 36 EFL Undergraduates	AWE (Pigai)	Longitudinal case study; 16 weeks	Engagement retention	Found engagement decline over 16 weeks without ongoing instructor scaffolding.
Yan D. (2024)	Hong Kong; 54 EFL Undergraduates	GenAI (ChatGPT)	Experimental; 4 weeks	Peer deliberation vs. Individual	Collaborative processing of AI feedback triggered higher meta-talk than individual processing.
Yan & Zhang (2024)	China; 8 EFL Undergraduates	GenAI (ChatGPT)	Mixed-methods case study; 6 weeks	Multidimensional engagement	Identified complex behavioral, cognitive, and affective engagement profiles.

Zhang (2020)	China; 86 EFL Undergraduates	AWE (Grammarly)	Quasi-experimental; 8 weeks	Revision patterns	Students predominantly made surface-level mechanical edits rather than structural revisions.
Zheng et al. (2024)	China; 62 EFL Undergraduates	GenAI (ChatGPT)	Quasi-experimental; 6 weeks	Coherence and rhetorical stance	Scaffolding prompts improved essay macro-structure and topical coherence.

Data Synthesis Techniques

Data synthesis proceeded in two complementary stages corresponding to the research questions.

To address RQ1 (impacts on SLA performance constructs), quantitative metrics and process outcomes were grouped into four analytical domains: grammatical accuracy, syntactic/lexical complexity, macro-coherence, and revision dynamics. Findings were disaggregated by AI technology type (AWE, MT, GenAI) and contextual variables. Syntheses by Karatay and Karatay (2024) and Shi and Aryadoust (2024) were utilized strictly as comparative background literature and were not counted among the 22 primary empirical studies.

To address RQ2 (benefits, challenges, and ethics), thematic synthesis followed Braun and Clarke's qualitative framework. Two reviewers independently conducted a six-phase thematic analysis: familiarization with data, initial code generation, searching for themes, reviewing candidate themes, defining/naming themes, and producing the report. Double-coded text extracts yielded high inter-coder agreement (91%), with discrepancies resolved via consensus.

Table 4.

Analytic Matrix Mapping Empirical Studies to SLR Research Questions

Analytic Domain	Target SLA Construct / Theme	Primary Target AI Tools	Representative Literature
RQ1: Writing Performance	• Accuracy (grammatical errors) • Complexity (syntactic/lexical) • Coherence & Rhetoric • Revision & Feedback Engagement	AWE Systems (Grammarly, Pigai), Generative AI (ChatGPT, GPT-4), MT (DeepL)	Ajabshir & Ebadi (2023); Chen (2025); Karatay & Karatay (2024); Lin & Crosthwaite (2024); Mizumoto et al. (2024); Shen et al. (2023); Wei et al. (2023); Xiaosa & Ping (2025); Zhang (2020)
RQ2: Perceptions & Ethics	• Cognitive & sociocultural benefits • Learner engagement hurdles • Critical AI literacy (APSE model) • Academic integrity & agency	Generative AI (ChatGPT), Wiki-based AI tools, Automated Feedback platforms	Darvin (2025); Hwang et al. (2025); Mahapatra (2024); Meniado et al. (2024); Michel et al. (2025); Ranalli (2021); Shi & Aryadoust (2024); Teng (2024); Wang & Wang (2025); Wang (2024); Wiboolyasarin et al. (2024); Yan (2024); Yan & Zhang (2024)

The matrix in Table 4 demonstrates how the 22 primary empirical studies were organized according to the two research questions. For RQ1, studies were mapped according to measurable L2

writing outcomes, including accuracy, complexity, coherence, and revision processes, while also being differentiated by AI technology. For RQ2, studies were grouped according to pedagogical benefits, learner engagement, operational challenges, critical AI literacy, and ethical or agency-related concerns. Secondary syntheses, including Karatay & Karatay (2024) and Shi and Aryadoust (2024), were used only as contextual background and were therefore excluded from the representative primary-study categories in the table.

RESULTS AND DISCUSSION

Impacts of AI Tools on L2 Writing Performance

The synthesis of the 22 primary empirical studies demonstrates that AI tools influence L2 writing performance in diverse ways across different tool architectures and instructional settings.

Table 5.
Summary of AI Tool Impacts on L2 Writing Performance Constructs

Performance Construct	Primary Empirical Outcomes	Supporting Primary Studies (N)	SLA Theoretical Alignment
Linguistic Accuracy	11 of 14 studies reported positive reductions in surface errors; AWE and GenAI match human diagnostic precision for local mechanics.	n = 14 (e.g., Ajabshir & Ebadi, 2023; Mizumoto et al., 2024; Wei et al., 2023)	Noticing Hypothesis (cognitive gap detection)
Syntactic & Lexical Complexity	Mixed outcomes across 8 studies; traditional AWE showed minimal effect, while GenAI reformulations fostered sophisticated vocabulary.	n = 8 (e.g., Chen, 2025; Mahapatra, 2024)	Input Enhancement & Modelled Output
Coherence & Text Structure	Promising gains in 5 studies; LLM reformulations assisted in macro-structural reorganization when accompanied by teacher prompts.	n = 5 (e.g., Chen, 2025; Zheng et al., 2024)	Process Writing Approach & Scaffolding
Revision Dynamics	Shift from surface edits to structural revisions observed in 9 studies; collaborative AI interaction increased engagement.	n = 9 (e.g., Michel et al., 2025; Shen et al., 2023; Yan D., 2024)	Cognitive Process Model of Writing

Linguistic Accuracy and Syntactic/Lexical Complexity

Across the primary evidence base, AI integration consistently demonstrated positive impacts on surface-level linguistic accuracy. Of the 14 primary studies evaluating accuracy metrics, 11 reported positive error reduction. Automated Written Corrective Feedback (AWCF) generated by legacy AWE platforms (e.g., Grammarly, Pigai) and LLM-powered GenAI acts as a cognitive trigger for error detection (Wei et al., 2023). Empirical testing confirms that tools like ChatGPT exhibit high diagnostic agreement with human raters in identifying mechanical, orthographic, and local grammatical errors (Mizumoto et al., 2024). Within Schmidt's (1990) Noticing Hypothesis, AI feedback provides immediate input enhancement, allowing L2 learners to "notice the gap" between their interlanguage output and target language norms.

When comparing feedback modalities, teacher feedback retained superior efficacy in identifying deep contextual errors. However, recent empirical comparisons indicate that LLM feedback equals teacher feedback in local error precision while providing immediate response times (Lin & Crosthwaite, 2024). Accuracy gains were also genre-dependent: structured feedback yielded greater

accuracy improvements in argumentative compositions than in descriptive tasks (Ajabshir & Ebadi, 2023).

Regarding complexity, evidence across 8 relevant studies revealed divergence based on AI architecture. Rule-based AWE systems produced minimal gains in syntactic complexity or lexical variety, as their underlying algorithms prioritize error elimination over stylistic expansion. Conversely, LLM-driven generative AI tools provide holistic reformulations that model sophisticated academic syntax and vocabulary (Chen, 2025; Mahapatra, 2024). ESL learners exposed to ChatGPT reformulations demonstrated measurable increases in productive vocabulary diversity and complex clause structures in post-test measures (Mahapatra, 2024).

Coherence, Organization, and Textual Quality

Evaluating macro-level text quality, including structural coherence, topical unity, and rhetorical organization, represents a developing area of inquiry supported by 5 primary studies in this review. Historically, classic AWE platforms were unable to evaluate paragraph-level rhetoric, offering generic feedback on transition markers. Evidence shows that GenAI models help address this limitation by providing contextualized structural suggestions (Chen, 2025; Zhang, 2020).

When learners analyze ChatGPT reformulations of their own drafts, they internalize macro-structural patterns, leading to better-organized essays (Chen, 2025). However, because these conclusions stem from a limited number of studies, claims regarding macro-coherence must remain qualified. Comparative findings emphasize that while GenAI offers useful structural outlines, human instruction remains essential for developing complex argumentative depth and domain-specific nuance (Lin & Crosthwaite, 2024). Thus, AI platforms act as complementary tools for structural scaffolding rather than complete substitutes for pedagogical instruction.

Revision Behaviors and Engagement Dynamics

The impact of AI feedback on L2 writing is mediated by how learners process and operationalize suggestions during revision. Synthesizing findings across 9 primary studies reveals an engagement continuum spanning behavioral, cognitive, and affective dimensions (Xiaosa & Ping, 2025; Yan & Zhang, 2024). In individual, unguided contexts, learners frequently default to passive uptake, executing surface-level mechanical edits without deep cognitive processing (Ranalli, 2021; Zhang, 2020).

Learner engagement patterns varied depending on the social configuration of tool implementation. Emerging evidence from 3 studies indicates that integrating GenAI into collaborative environments (e.g., peer-review or wiki tasks) increases cognitive processing (Michel et al., 2025; Wiboolyasarin et al., 2024; Yan, 2024). Comparative testing demonstrated that collaborative processing of ChatGPT feedback triggered higher levels of peer deliberation, meta-talk, and deep structural revision than individual processing (Yan, 2024). Furthermore, longitudinal studies (Shen et al., 2023; Xiaosa & Ping, 2025) indicate that engagement with AI tools wanes over time without ongoing instructional scaffolding, risking a return to uncritical copy-pasting behavior.

Benefits, Challenges, and Ethical Concerns in AI-Assisted L2 Writing

Table 6.
Matrix of Perceived Benefits, Operational Challenges, and Ethical Issues

Category	Primary Synthesized Findings	Supporting Primary Studies	Pedagogical & SLA Implications
Pedagogical Benefits	Immediate individualized feedback; non-judgmental environment; reduced writing anxiety.	Meniado et al. (2024); Teng (2024)	Lowers Affective Filter; expands Zone of Proximal Development
Operational Challenges	Temporal engagement decline; feedback overload; cognitive passivity; uncritical copying.	Hwang et al. (2025); Li & Ke (2025); Wang (2024)	Disruption of Self-Regulated Learning (SRL) cycles
Ethical & Literacy Concerns	Plagiarism risks; potential loss of authentic learner voice; uncritical acceptance of AI errors.	Darvin (2025); Wang & Wang (2025)	Development of Critical AI Literacy frameworks (APSE model)

Table 6 provides an overview of the major themes emerging from the synthesis of evidence addressing RQ2. The findings indicate that AI-assisted L2 writing produces a simultaneous combination of pedagogical opportunities and potential risks. On the positive side, AI provides immediate and individualized support and can reduce affective barriers to writing. However, these benefits are accompanied by operational challenges, particularly declining engagement, feedback overload, and cognitive passivity. At the ethical level, the literature raises concerns about academic integrity, uncritical acceptance of AI-generated content, and possible erosion of authentic learner voice. Thus, the evidence suggests that the pedagogical value of AI depends not simply on access to the technology but on how learners are guided to interpret, evaluate, and use its outputs.

Coherence, Organization, and Textual Quality

Evaluating macro-level text quality, including structural coherence, topical unity, and rhetorical organization, represents a developing area of inquiry supported by five primary studies in this review. Compared with accuracy, the evidence base for coherence is considerably smaller and should therefore be interpreted cautiously. Classic AWE platforms have traditionally focused more strongly on linguistic and mechanical features than on paragraph-level rhetoric and complex argumentative organization. By contrast, recent evidence suggests that GenAI models can provide contextualized structural suggestions and reformulations that support macro-level organization (Chen, 2025; Zhang, 2020).

When learners analyze ChatGPT reformulations of their own drafts, they may become more aware of macro-structural patterns and subsequently produce better-organized essays (Chen, 2025). Similarly, scaffolding prompts can help learners use GenAI feedback to improve topical coherence and rhetorical organization (Zhang, 2020). However, these findings do not establish that GenAI independently produces sustained improvements in higher-order writing quality. Comparative evidence indicates that teacher feedback remains important for contextual interpretation, argumentative depth, and domain-specific nuance (Lin & Crosthwaite, 2024). Therefore, GenAI should be viewed as a potential structural scaffold rather than a substitute for human pedagogical guidance.

Pedagogical and Cognitive Benefits

The primary benefit identified across qualitative and mixed-methods empirical studies is the provision of immediate, scalable scaffolding. L2 learners perceive AI platforms like ChatGPT as non-judgmental, accessible writing companions (Teng, 2024). This environment reduces foreign language writing anxiety, lowering Krashen's Affective Filter and encouraging exploratory output generation (Meniado et al., 2024; Teng, 2024).

From a Sociocultural perspective, AI platforms can be conceptualized through an analytical metaphor as a "pseudo-More Knowledgeable Other" (pseudo-MKO) that operates within the learner's Zone of Proximal Development (ZPD). Unlike a human MKO, an AI tool lacks human intentionality, pedagogical empathy, and contextual understanding, and it carries a risk of generating factual errors ("hallucinations"). Nevertheless, as an interactive partner, GenAI allows L2 writers to request explanations, clarify grammar rules, and experiment with stylistic options (Meniado et al., 2024), providing scalable dialogic practice outside the classroom.

Operational Challenges: Cognitive Overload and Behavioral Passivity

Despite these benefits, empirical findings highlight operational challenges. Longitudinal tracking reveals a temporal decline in student engagement over extended interventions (Hwang et al., 2025). As initial novelty wears off, usage becomes more pragmatic and selective. Furthermore, when AI systems generate extensive feedback, learners can experience cognitive overload, leading them to ignore complex stylistic advice in favor of quick mechanical fixes (Hwang et al., 2025; Y. Wang, 2024; Xiaosa & Ping, 2025).

Additionally, automated correction can reduce productive cognitive effort. When AI platforms automatically rewrite text without requiring active problem-solving, learners may demonstrate cognitive passivity (Wang, 2024; Xiaosa & Ping, 2025). High reliance on automated translation or generative output is associated with reduced depth of processing and weakened internal Monitoring processes (Wang, 2024). Unscaffolded learners risk developing passive dependency rather than dynamic self-regulation (Zimmerman, 2000).

Ethical Concerns, Critical AI Literacy, and Learner Agency

Integrating AI into L2 writing raises ethical questions surrounding academic integrity, authorship, and learner agency (Darvin, 2025). Unregulated AI adoption creates ambiguity regarding legitimate assistance versus unauthorized text generation. Qualitative findings in selective studies (e.g., Darvin, 2025) suggest that over-reliance on generative outputs may homogenize learner writing, as AI algorithms impose standardized, westernized academic registers that dilute authentic voice and personal style.

To address these risks, recent research emphasizes developing critical AI literacy ((Darvin, 2025; Wang & Wang, 2025). Frameworks like the APSE model (Access, Purpose, Assessment, Ethics) provide structured guidelines for evaluating L2 student engagement with AI (Wang & Wang, 2025). Empirical evaluations indicate that learners lacking critical AI literacy tend to treat AI outputs as authoritative, accepting incorrect reformulations without reflection. Cultivating critical agency enables L2 writers to position AI as an advisory tool, ensuring human agency remains central to the writing process.

Critical Evaluation of Primary Literature: Methodological Constraints and Bias

A critical evaluation of the 22 primary empirical studies reveals substantial methodological heterogeneity, sample constraints, and control limitations that must qualify general claims regarding AI efficacy in SLA.

Sample Constraints and Population Homogeneity

Most included studies examined undergraduate EFL learners in Asian university contexts, with a substantial concentration of participants at intermediate proficiency levels. Advanced L2 writers, young learners, adult professional L2 communicators, and learners in immersive ESL environments remain comparatively underrepresented. Consequently, findings concerning writing performance, cognitive processing, engagement, and affective responses may not generalize directly to populations with different educational backgrounds, proficiency levels, or communicative needs.

Variability in Experimental Controls and AI Usage Modalities

The primary literature exhibits substantial variation in AI implementation protocols. Studies differed between unguided or relatively unstructured AI use, in which learners independently submitted drafts or generated text, and pedagogically scaffolded use, in which instructors provided guidance on prompting, evaluation, revision, or dialogic interaction. This distinction is theoretically important because observed outcomes may reflect not only the technological capabilities of AI tools but also the quality and intensity of instructional scaffolding surrounding their use. Studies with limited control over prompting protocols or concurrent instructor intervention therefore make it difficult to isolate the independent contribution of the AI technology itself.

Measurement and Longitudinal Limitations

The duration and outcome measures of the included studies also varied substantially. Many studies measured immediate or relatively short-term outcomes, such as revision uptake, error reduction, or post-intervention writing performance, whereas fewer studies examined whether AI-assisted gains were retained or transferred to subsequent unassisted writing. Long-term SLA outcomes, including sustained grammatical internalization, independent revision ability, and durable self-regulation, therefore remain insufficiently investigated. Future longitudinal research should distinguish between immediate improvement in AI-assisted products and genuine development of learners' independent L2 writing competence.

Theoretical Synthesis and Proposed Conceptual Framework

The synthesized findings from empirical studies spanning 2018–2025 reflect a paradigm shift in AI-assisted L2 writing, moving from predominantly rule-based automated evaluation toward increasingly interactive and generative forms of AI-mediated writing support. Earlier AWE research primarily emphasized automated diagnostic feedback and revision of linguistic errors, whereas recent GenAI research demonstrates that conversational AI can also influence cognitive engagement, affective experiences, collaborative interaction, and learner agency ((Darvin, 2025; Teng, 2024; Y. Wang, 2024).

To advance theoretical interpretation beyond a simple categorization of AI tools, this review proposes the Integrated AI-Mediated Second Language Acquisition (I-AM-SLA) model. The framework brings together cognitive, sociocultural, and ethical/agency dimensions in a dynamic relationship rather than treating AI feedback as an isolated technological intervention.

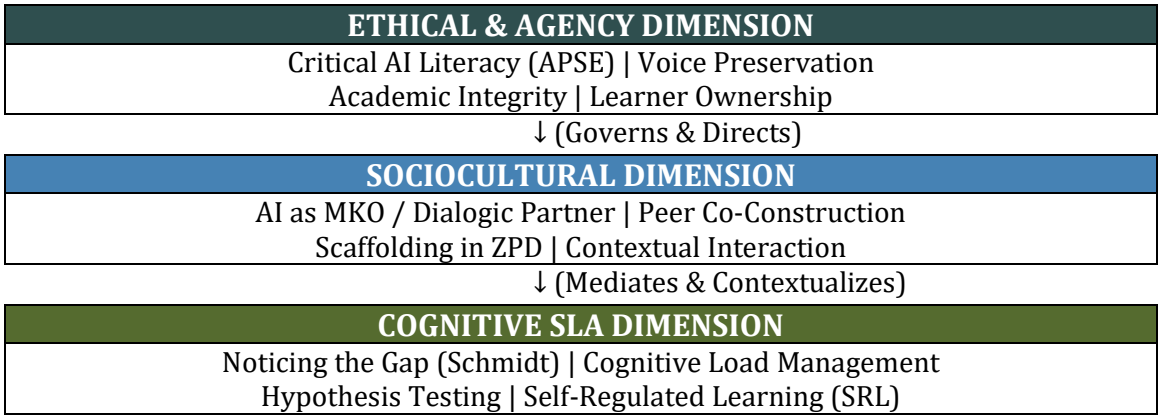

Figure 2. The Integrated AI-Mediated Second Language Acquisition (I-AM-SLA) Framework.

1) Cognitive SLA Dimension (Foundation): Grounded in Schmidt's (1990) Noticing Hypothesis and Sweller's (1988) Cognitive Load Theory, AI provides real-time input enhancement. Automated corrections highlight discrepancies between interlanguage output and target norms, facilitating "noticing" while reducing cognitive load during initial composition.

2) Sociocultural Dimension (Mediator): Positioned within Vygotskian (1978) SLA, AI functions not as an authoritative evaluator, but as a conversational More Knowledgeable Other (MKO). Collaborative peer interaction around AI feedback elevates processing from passive acceptance to active negotiation of meaning, expanding the Zone of Proximal Development (ZPD).
3) Ethical & Agency Dimension (Governing Tier): Encompassed by Critical AI Literacy (e.g., Wang & Wang's APSE framework), this tier governs interaction quality. Without active critical evaluation, the learner risks cognitive passivity and loss of authentic voice. Learner agency ensures that AI outputs are vetted, adapted, and critically synthesized rather than copied, preserving learner autonomy and authentic L2 identity.

Ultimately, AI tools yield the greatest gains in L2 writing performance when embedded within socio-collaborative learning environments and guided by explicit instructional scaffolding (Michel et al., 2025; Wiboolyasarin et al., 2024).

CONCLUSION

This systematic literature review synthesized 22 primary empirical studies published between 2018 and 2025 to examine the influence of AI tools on L2 writing performance and to identify their broader pedagogical, operational, and ethical implications. In relation to RQ1, the evidence indicates that AI-assisted writing most consistently supports linguistic accuracy, with 11 of 14 studies reporting reductions in surface-level errors. Effects on syntactic and lexical complexity were more variable across eight studies and appeared to depend on the type of AI technology and the manner in which it was pedagogically implemented. Evidence concerning coherence and rhetorical organization was promising but more limited, with five studies indicating that GenAI could support macro-structural improvement when accompanied by explicit instructional scaffolding. Revision outcomes similarly demonstrate that AI does not automatically produce deeper learning: learners may engage in surface-level correction when using AI individually and without guidance, whereas collaborative processing and teacher scaffolding can promote deeper deliberation and structural revision.

In relation to RQ2, the review identifies a dual pattern of opportunities and risks. AI tools can provide immediate individualized feedback, create a relatively non-judgmental writing environment, and reduce writing anxiety, thereby expanding opportunities for L2 writing practice. At the same time, declining engagement, feedback overload, cognitive passivity, uncritical copying, and weakened self-regulation may emerge when learners rely excessively on automated assistance. Ethical concerns regarding academic integrity, learner authorship, and the preservation of authentic voice further demonstrate that technological access alone is insufficient for effective AI integration. The proposed Integrated AI-Mediated Second Language Acquisition (I-AM-SLA) framework consequently positions cognitive processing, sociocultural interaction, and ethical learner agency as interdependent dimensions of effective AI-mediated L2 writing.

Several implications emerge for future research and academic practice. First, future studies should employ longitudinal designs to determine whether improvements observed in AI-assisted writing transfer to unassisted L2 writing and remain stable over time. Second, researchers should empirically evaluate critical AI literacy interventions, including frameworks such as APSE, by examining their effects on learners' evaluative judgment, prompting practices, and ethical decision-making. Third, future research should broaden participant populations beyond tertiary EFL learners to include young learners, adult professional L2 writers, advanced learners, learners in immersive ESL contexts, and additional target languages. Finally, instructors are encouraged to investigate AI-supported socio-collaborative practices, such as peer review, collaborative revision, and wiki-based writing, as promising alternatives to isolated AI use. Such approaches should nevertheless be empirically tested across diverse educational settings. Overall, the evidence supports neither complete prohibition nor unguided adoption of AI in L2 writing; rather, it supports carefully scaffolded integration in which AI functions as a writing support while learners retain responsibility for evaluating, revising, and authoring their texts.

ACKNOWLEDGMENT

The authors express their deepest gratitude to the Directorate of Research and Community Service at Universitas Negeri Yogyakarta for providing the research infrastructure and access to academic databases that made this systematic review possible.

REFERENCES

Ajabshir, Z. F., & Ebadi, S. (2023). The effects of automatic writing evaluation and teacher-focused feedback on CALF measures and overall quality of L2 writing across different genres. *Asian-Pacific Journal of Second and Foreign Language Education*, *8*(1). https://doi.org/10.1186/s40862-023-00201-9

Chen, Y. (2025). Evaluating the potential of ChatGPT-reformulated essays as written feedback in L2 writing. *Computers and Education: Artificial Intelligence*, *9*. https://doi.org/10.1016/j.caeai.2025.100500

Darvin, R. (2025). The need for critical digital literacies in generative AI-mediated L2 writing. *Journal of Second Language Writing*, *67*. https://doi.org/10.1016/j.jslw.2025.101186

Hwang, H., Chang, X., & Sun, J. (2025). Generative AI is useful for second language writing, but when, why, and for how long do learners use it? *Journal of Second Language Writing*, *69*. https://doi.org/10.1016/j.jslw.2025.101230

Karatay, Y., & Karatay, L. (2024). Automated writing evaluation use in second language classrooms: A research synthesis. In *System* (Vol. 123). Elsevier Ltd. https://doi.org/10.1016/j.system.2024.103332

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

Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: a mixed methods intervention study. *Smart Learning Environments*, *11*(1). https://doi.org/10.1186/s40561-024-00295-9

Meniado, J. C., Huyen, D. T. T., Panyadilokpong, N., & Lertkomolwit, P. (2024). Using ChatGPT for second language writing: Experiences and perceptions of EFL learners in Thailand and Vietnam. *Computers and Education: Artificial Intelligence*, *7*. https://doi.org/10.1016/j.caeai.2024.100313

Michel, M., Bazhutkina, I., Abel, N., & Strobl, C. (2025). Collaborative writing based on generative AI models: Revision and deliberation processes in German as a foreign language. *Journal of Second Language Writing*, *67*. https://doi.org/10.1016/j.jslw.2025.101185

Mizumoto, A., Shintani, N., Sasaki, M., & Teng, M. F. (2024). Testing the viability of ChatGPT as a companion in L2 writing accuracy assessment. *Research Methods in Applied Linguistics*, *3*(2). https://doi.org/10.1016/j.rmal.2024.100116

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Journal of Clinical Epidemiology*, *134*, 178–189. https://doi.org/10.1016/j.jclinepi.2021.03.001

Ranalli, J. (2021). L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing*, *52*. https://doi.org/10.1016/j.jslw.2021.100816

Schimdt, R. W. (1990). The Role of Consciousness in Second Language Learning1. *Applied Linguistics*, *11*(2), 129–158. https://doi.org/10.1093/applin/11.2.129

Shen, C., Shi, P., Guo, J., Xu, S., & Tian, J. (2023). From process to product: writing engagement and performance of EFL learners under computer-generated feedback instruction. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1258286

Shi, H., & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. *ReCALL*, *36*(2), 187–209. https://doi.org/10.1017/S0958344023000265

Sweller, J. (1988). Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Teng, M. F. (2024). "ChatGPT is the companion, not enemies": EFL learners' perceptions and experiences in using ChatGPT for feedback in writing. *Computers and Education: Artificial Intelligence*, *7*. https://doi.org/10.1016/j.caeai.2024.100270

Vygotsky, L. S. (1980). *Mind in Society* (M. Cole, V. Jolm-Steiner, S. Scribner, & E. Souberman, Eds.). Harvard University Press. https://doi.org/10.2307/j.ctvjf9vz4

Wang, C., & Wang, Z. (2025). Investigating L2 writers' critical AI literacy in AI-assisted writing: An APSE model. *Journal of Second Language Writing*, *67*. https://doi.org/10.1016/j.jslw.2025.101187

Wang, Y. (2024). Cognitive and sociocultural dynamics of self-regulated use of machine translation and generative AI tools in academic EFL writing. *System*, *126*. https://doi.org/10.1016/j.system.2024.103505

Wei, P., Wang, X., & Dong, H. (2023). The impact of automated writing evaluation on second language writing skills of Chinese EFL learners: a randomized controlled trial. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1249991

Wiboolyasarin, W., Wiboolyasarin, K., Suwanwihok, K., Jinowat, N., & Muenjanchoey, R. (2024). Synergizing collaborative writing and AI feedback: An investigation into enhancing L2 writing proficiency in wiki-based environments. *Computers and Education: Artificial Intelligence*, *6*. https://doi.org/10.1016/j.caeai.2024.100228

Xiaosa, L., & Ping, K. (2025). How L2 student writers engage with automated feedback: A longitudinal perspective. *Assessing Writing*, *64*. https://doi.org/10.1016/j.asw.2025.100919

Yan D. (2024). Comparing individual vs. collaborative processing of ChatGPT-generated feedback: Effects on L2 writing task improvement and learning. *Language Learning & Technology*, *2024*(1), 1–19. https://hdl.handle.net/10125/73597

Yan, D., & Zhang, S. (2024). L2 writer engagement with automated written corrective feedback provided by ChatGPT: A mixed-method multiple case study. *Humanities and Social Sciences Communications*, *11*(1). https://doi.org/10.1057/s41599-024-03543-y

Zhang, Z. (Victor). (2020). Engaging with automated writing evaluation (AWE) feedback on L2 writing: Student perceptions and revisions. *Assessing Writing*, *43*. https://doi.org/10.1016/j.asw.2019.100439

Zimmerman, B. J. (2000). Attaining Self-Regulation. In *Handbook of Self-Regulation* (pp. 13–39). Elsevier. https://doi.org/10.1016/B978-012109890-2/50031-7