

Article

Do AI Writing Tools Improve L2 Writing Performance? A Meta-Analysis of Experimental Evidence, Moderators, and Methodological Quality

Jinming Du

Languages and Cultures Department, University of Otago, Dunedin 9016, New Zealand; duji6714@student.otago.ac.nz

How To Cite: Du, J. (2026). Do AI Writing Tools Improve L2 Writing Performance? A Meta-Analysis of Experimental Evidence, Moderators, and Methodological Quality. *Journal of Educational Technology and Innovation*, 8(3), 80–103. <https://doi.org/10.61414/awr9xk46>

Received: 21 May 2026

Revised: 24 August 2026

Accepted: 15 September 2026

Published: 30 September 2026

Abstract: Artificial intelligence (AI) writing tools—including large language model (LLM)-based systems, automated writing evaluation (AWE) platforms, and AI-powered writing assistants—have become an increasingly prominent feature of second language (L2) writing pedagogy. Yet whether these tools reliably improve L2 writing performance, and under what instructional conditions, remains an open question. This study reports a systematic meta-analysis of experimental and quasi-experimental evidence on the effects of AI writing tools on L2 writing performance. A multi-database search of Scopus, Web of Science, ERIC, and ProQuest Dissertations & Theses, supplemented by forward citation tracing and targeted hand-searching, identified 34 eligible focal effect sizes from studies published between 2021 and 2026. Effect sizes were expressed as Hedges’ g . A verified subset of 22 effects with directly extractable statistics was estimated using restricted maximum-likelihood (REML) with Knapp–Hartung adjustment; an expanded model ($k = 34$) incorporating conversion-based effects served as a robustness analysis. The verified model produced a pooled effect of $g = 0.714$, 95% CI [0.592, 0.836], with low observed heterogeneity ($I^2 = 24.3\%$). Exploratory moderator analyses indicated that multidimensional feedback designs, sustained interventions, particularly those lasting eight weeks or longer, were associated with larger effects, while single-session and lower-order-only feedback showed smaller gains. Methodological quality was generally adequate but uneven: many studies lacked delayed transfer measures and transparent effect-size reporting. These findings position AI writing tools as pedagogically contingent rather than inherently effective—their educational value depends critically on how feedback literacy, teacher mediation, and instructional design shape opportunities for noticing, revision, and transferable writing development.

Keywords: AI writing tools; second language writing; ChatGPT; generative AI; meta-analysis; automated writing evaluation.

1. Introduction

Artificial intelligence (AI) writing tools are now central to computer-assisted language learning (CALL), yet their effects on L2 writing development remain theoretically and empirically contested. Rather than treating them simply as automated correctors or text generators, this article conceptualises AI writing tools as CALL feedback systems that may reshape learner engagement, revision practices, teacher mediation, and human–technology agency. It examines whether controlled intervention evidence supports claims that these tools improve L2 writing performance. Their appeal is especially strong in L2 writing, where learners must coordinate linguistic, rhetorical, cognitive, and affective demands, including grammar, vocabulary, organisation, coherence, content development,



Copyright: © 2026 by the authors. This is an open access article under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Publisher’s Note: Scilight stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

audience awareness, and disciplinary voice (Hyland, 2019; Ferris, 2011). Because teachers often have limited capacity to provide timely and individualised feedback, AI writing tools are increasingly positioned as scalable feedback resources that can supplement teacher response, expand revision opportunities, and offer immediate language support (Godwin-Jones, 2022; Mahapatra, 2024; Wiboolyasarin et al., 2024).

1.1. AI Writing Tools and L2 Writing Development

AI writing tools may support L2 development through several routes: rapid corrective feedback on grammar and vocabulary helps learners address surface concerns (Alsaweed & Aljebreen, 2024; Ding et al., 2026; Du & Chen, 2026); generative AI extends this to higher-order concerns such as idea development and rhetorical refinement (Su et al., 2023; Wang & Ren, 2024; Wiboolyasarin et al., 2024); and iterative dialogue allows learners to ask follow-up questions and refine drafts progressively (Koltovskaia et al., 2024; Yan & Zhang, 2024). When feedback is used reflectively, AI-supported environments also show gains in motivation and self-regulated learning (Du et al., 2026; Jin et al., 2025; Liu et al., 2026).

At the same time, generative AI can produce fluent but inaccurate, generic, or rhetorically inappropriate text, and learners may accept suggestions without sufficient evaluation (Barrot, 2023; Cingillioglu, 2023; Lingard, 2023). L2 writers do not always possess the metacognitive or critical AI literacy needed to regulate AI feedback effectively (Koltovskaia et al., 2024; Ou et al., 2024; Wang & Wang, 2025; Yan & Zhang, 2024), so effectiveness likely depends on pedagogical design, feedback focus, learner level, task type, and teacher or peer mediation. At the broader educational-ecology level, recent evidence on acceptance of large language models in education also shows that perceived usefulness coexists with concerns about data privacy, over-reliance, bias, and unequal AI knowledge (Stroud & Du, 2025; Wan et al., 2026), reinforcing the need to treat AI writing support as a sociotechnical rather than purely technical intervention.

1.2. Fragmented Empirical Evidence and the Need for Quantitative Synthesis

A growing body of intervention studies has examined AI writing tools' effects on L2 writing performance: chatbot-based practice improved elementary EFL learners' writing (Kwon et al., 2023); LLM-supported strategy instruction improved EFL writing, self-regulated learning, and motivation (Liu et al., 2026); ChatGPT-based formative feedback was associated with improved ESL/EFL academic writing (Ghafouri et al., 2024; Mahapatra, 2024; Tsai et al., 2024); and AI-supported collaborative writing environments showed positive effects on specific writing aspects (Hu et al., 2025; Wiboolyasarin et al., 2024).

Despite this accumulation of evidence, the field remains fragmented. Studies use different labels—ChatGPT, LLM-supported instruction, AI chatbots, AWE systems, automated feedback systems—making cross-study comparisons difficult (Godwin-Jones, 2022; Shi et al., 2026), and measured outcomes vary from holistic writing scores to accuracy, complexity, grammar, and coherence across designs ranging from RCTs to small-scale classroom studies. These variations make it hard to determine from individual studies whether AI writing tools produce reliable improvements and under what conditions.

Existing reviews have mapped AI-assisted writing opportunities but are predominantly narrative, scoping, or perception-focused. The present study synthesises eligible experimental evidence quantitatively, reports study-disposition decisions transparently, and distinguishes the primary model from records retained for background only.

1.3. The Present Study

The present study conducts a meta-analysis of experimental, quasi-experimental, and intervention-based studies examining the effects of AI writing tools on L2 writing performance. The study is positioned as a CALL-focused quantitative synthesis rather than a broad review of all AI-in-writing scholarship. Its contribution is threefold: it estimates the overall effect of AI writing tools on L2 writing performance, examines moderator patterns that matter for CALL design, and evaluates the methodological maturity of this rapidly developing intervention literature. Three research questions guided the study:

RQ1. What is the overall effect of AI writing tools on L2 writing performance in controlled intervention studies?

RQ2. Do effects vary by AI architecture, feedback focus, outcome type, learner level, or intervention duration?

RQ3. What is the methodological quality of existing experimental evidence on AI writing tools and L2 writing performance, and what does the observed evidence map suggest for future CALL research design?

2. Literature Review

2.1. L2 Writing as a Cognitively and Pedagogically Demanding Activity

Second language writing is a complex cognitive, linguistic, and socio-rhetorical activity (Ferris, 2011; Hyland, 2019), and L2 learners face additional constraints from limited lexical resources, incomplete grammatical control, and restricted feedback access. Teachers frequently cannot provide detailed formative feedback to large classes, creating a strong rationale for technology-supported writing assistance. Research on L2 writing feedback has shifted toward viewing it as a process requiring learner engagement, interpretation, and revision (Hyland & Hyland, 2006): if AI feedback merely produces a polished text without learner understanding, it may improve the immediate product while contributing little to transferable competence, but when learners compare alternatives, diagnose weaknesses, and revise deliberately, AI tools may function as scaffolds for language development and metacognitive regulation (Koltovskaia et al., 2024; Yan & Zhang, 2024).

2.2. From Automated Writing Evaluation to Generative AI Writing Support

AI-supported writing assistance did not begin with ChatGPT. Earlier technologies such as AWE, automated written corrective feedback, and grammar checkers supported L2 writing through automated scoring, error detection, and revision prompts (Godwin-Jones, 2022; Nazari et al., 2021; Xu & Zhang, 2022). These constrained systems were strongest in lower-order concerns such as grammar and sentence-level accuracy, but were criticised for limited contextual sensitivity, formulaic feedback, and insufficient attention to higher-order concerns such as argumentation and disciplinary voice.

The expanded evidence map reinforces this pre-ChatGPT lineage: AWE and AI-writing-assistant studies (Fan, 2023; Nazari et al., 2021; Wei et al., 2023; Xu & Zhang, 2022) show that research has long addressed the same design questions now raised by LLMs: feedback timing, learner uptake, teacher mediation, and whether automated feedback supports durable learning.

Generative AI and LLMs have expanded AI-assisted writing's functional scope: such tools produce explanations, propose outlines, reformulate sentences, and engage in iterative dialogue, so empirical studies have examined ChatGPT and related systems as feedback providers, brainstorming partners, and self-regulated learning supports (Chen & Gong, 2025; Ghafouri et al., 2024; Liu et al., 2026; Mahapatra, 2024; Shi et al., 2026). This broader functionality also complicates evaluation because AI tools may simultaneously influence text quality, writing process, learner agency, and academic integrity.

The newer generative-AI literature broadens this trajectory from automated correction to interactive feedback: recent studies compare ChatGPT/GenAI feedback against teacher feedback, AWE feedback, restricted-access conditions, hybrid teacher-AI feedback, and AI-supported peer feedback (Chan et al., 2024; Teng et al., 2026; Escalante et al., 2023; Guo et al., 2024; Hong & Shin, 2025; Luo et al., 2025; Wang & Ren, 2024; Zou et al., 2026), shifting the field from asking whether automated feedback can identify errors to asking how learners orchestrate tool suggestions, teacher guidance, and their own rhetorical intentions.

2.3. Empirical Evidence on AI Tools and L2 Writing Performance

Experimental and quasi-experimental studies increasingly report positive effects: chatbot-based practice improved young learners' L2 writing (Kwon et al., 2023); LLM-supported strategy instruction enhanced EFL writing, self-regulated learning, and motivation (Liu et al., 2026); and ChatGPT-supported interventions improved overall writing quality, grammar, organisation, and coherence (Chen & Gong, 2025; Ghafouri et al., 2024; Hongxia & Razali, 2025; Mahapatra, 2024; Shi et al., 2026; Wiboolyasarin et al., 2024).

Across the expanded corpus, several partially overlapping research traditions emerge: Grammarly/AWE studies focus on accuracy, grammar, or complexity; ChatGPT/LLM studies examine holistic quality, revision quality, or self-regulated learning; hybrid studies ask whether AI feedback should replace or supplement teacher feedback. A positive grammar-accuracy effect in a Grammarly study is not equivalent to a positive argument-organisation effect in a ChatGPT study: effects are mediated by what kind of writing problem the tool is designed to address.

The evidence is also methodologically heterogeneous: studies differ in tool type, learner population, outcome measurement, and comparison condition, making it difficult to infer overall effectiveness from individual studies alone.

2.4. Potential Moderators in AI-Assisted L2 Writing Effects

The effects of AI writing tools are unlikely to be uniform because technologies differ in feedback architecture, interactional capacity, and learner control (Zimmerman, 2000). Earlier AWE and automated corrective feedback

systems typically provide constrained diagnostic or error-level feedback, whereas generative AI and LLM-based systems offer dialogic explanations, revision alternatives, and metalinguistic justifications—changing not only the feedback provided but how learners question, negotiate, and regulate it.

Several moderators therefore map directly onto CALL design principles: tool architecture distinguishes constrained AWE from dialogic LLM systems (Godwin-Jones, 2022; Nazari et al., 2021; Xu & Zhang, 2022); feedback source distinguishes AI-only, teacher-only, peer-supported, and hybrid teacher–AI designs; outcome focus distinguishes linguistic accuracy, syntactic complexity, coherence, rhetorical quality, revision quality, self-efficacy, and engagement; and duration distinguishes one-shot revision tasks from sustained writing instruction. Together, these dimensions capture learner–tool interaction, teacher mediation, feedback-cycle depth, and opportunities for revision and transfer. Feedback focus may also shape outcomes. Lower-order feedback on grammar and mechanics, higher-order feedback on content and organisation, and multidimensional feedback are likely to support different aspects of writing development (Ferris, 2011; Hyland, 2019). Outcome type is equally important: immediate holistic scores may be more responsive to AI-assisted revision than delayed measures of independent accuracy, complexity, or rhetorical organisation. Learner level and duration add further variation, as novice writers may benefit from scaffolding but remain more vulnerable to uncritical uptake, while longer interventions may allow feedback to be internalised rather than merely applied (Schmidt, 1990). These moderators link tool affordances, feedback-mediated noticing, self-regulated revision, and writing performance.

2.5. Risks, Methodological Gaps, and Rationale for the Present Meta-Analysis

Scholars caution that AI writing tools introduce pedagogical and ethical risks: generative AI may produce inaccurate content, learners may over-rely on suggestions, and extensive AI-mediated revision may obscure authorship (Barrot, 2023; Cingillioglu, 2023; Du & Kohnke, 2026; Lingard, 2023; Tsai et al., 2024). L2 writers' engagement varies substantially by proficiency and critical AI literacy (Koltovskaia et al., 2024; Ou et al., 2024; Wang & Wang, 2025; Yan, 2025; Yan & Zhang, 2024), so pedagogical value depends on how learners interpret, regulate, and learn from AI assistance.

The expanded corpus further shows that risks span both AWE and generative AI contexts. In AWE/Grammarly studies, risks include over-attention to surface form and uncritical acceptance of suggestions; in ChatGPT/GenAI studies, risks extend to rhetorical homogenisation, loss of authorial voice, and untraceable revision histories. The underlying question—whether AI improves the submitted text or the writer—cannot be answered without designs separating assisted performance from independent competence.

Existing reviews have mapped opportunities and challenges of AI tools in L2 writing, but many remain conceptual, narrative, or scoping. A quantitative synthesis is therefore useful for estimating the direction and magnitude of effects while exposing methodological limitations; the present study synthesises eligible experimental evidence and reports study-disposition decisions transparently, distinguishing the primary quantitative model from records retained only for background interpretation.

3. Methods

3.1. Review Design

This study adopted a systematic search and meta-analytic design to synthesise empirical evidence on the effectiveness of AI writing tools for L2 writing, estimating a pooled effect size across experimental and quasi-experimental studies. Given the limited number of eligible studies, the analysis was conducted as a random-effects meta-analysis with sensitivity and descriptive moderator analyses rather than a full multilevel model.

The focal construct was AI-supported L2 writing performance: digital systems using AI to support planning, drafting, feedback, revision, evaluation, or instruction, including generative AI systems such as ChatGPT and large language models (LLM)-supported systems, AI chatbots, AWE systems, automated written corrective feedback systems, and AI-powered digital writing assistants, reflecting the literature's overlapping traditions of AWE research, chatbot-assisted writing, generative AI feedback, and LLM-supported strategy instruction (Barrot, 2023; Godwin-Jones, 2022; Su et al., 2023; Warschauer et al., 2023).

The synthesis prioritised studies estimating changes in actual writing performance rather than perceptions alone, focusing on experimental, quasi-experimental, intervention, counter-balanced, and closely comparable pre-post designs. Studies reporting qualitative engagement, learner perceptions, AI literacy, feedback-seeking behaviour, or AI-detection performance were retained for contextual interpretation but excluded from the primary pooled effect unless they reported an extractable writing-performance effect.

3.2. Study Identification and Screening

A systematic database search (see Figure 1) was conducted in Scopus, Web of Science, ERIC, and ProQuest Dissertations & Theses, with the final search completed in December 2025. These databases were selected for their broad coverage of applied linguistics, CALL, educational technology, and education research. The initial search returned 1595 records (Scopus $n = 801$; Web of Science $n = 590$; ERIC $n = 120$; ProQuest $n = 76$; Call for Unpublished Work $n = 8$). The search string combined terms for AI writing tools, generative AI, automated writing evaluation, ChatGPT, L2 writing, ESL/EFL, intervention, experimental and quasi-experimental designs, and writing performance. The core Boolean structure was: (“AI writing tool*” OR “artificial intelligence” OR “generative AI” OR “ChatGPT” OR “large language model*” OR “automated writing evaluation” OR AWE OR “AI feedback” OR “AI writing assistant*”) AND (“second language writing” OR “L2 writing” OR ESL OR EFL OR “foreign language writing”) AND (intervention OR experiment* OR “quasi-experiment*” OR “control group” OR “writing performance” OR “writing quality” OR “writing score*”). After removing 982 duplicates, 613 unique records were screened by title and abstract. Of the 613 screened records, 541 were excluded at title-and-abstract screening because they were not focused on L2 writing, did not examine AI writing tools, were conceptual/review articles, or did not report an intervention or writing-performance outcome. The remaining 72 reports were sought for retrieval. A further 21 reports were excluded for not focusing on L2 writing or AI writing tools, 10 for lacking an intervention or comparison design, and 7 for not reporting extractable writing-performance data, leaving 34 reports from the database search. Reference searching of included publications identified 5 additional reports, and supplementary searching identified 29 further reports; all 34 of these passed full-text screening with no exclusions. Combining both pathways, 68 reports met full eligibility criteria; after removing 34 duplicate or overlapping datasets reported across multiple publications, 34 independent focal effects were retained for the quantitative synthesis.

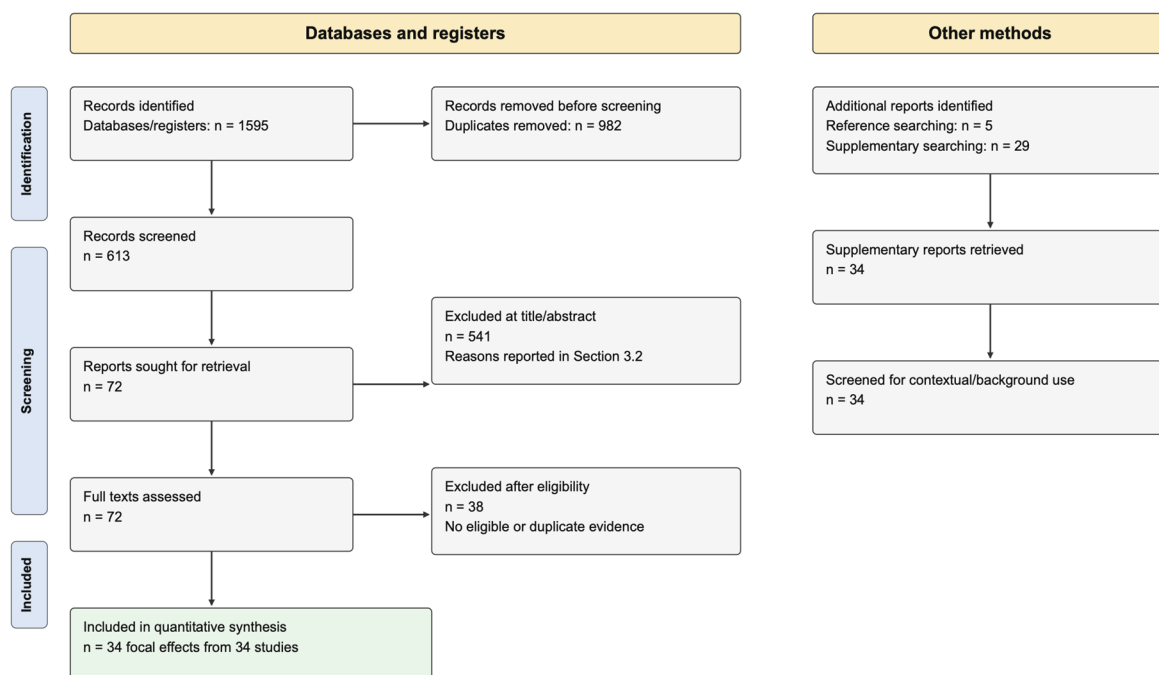


Figure 1. PRISMA flow diagram for study identification, screening, and inclusion.

3.3. Inclusion and Exclusion Criteria

Studies were included in the primary meta-analysis if they met four criteria. The study had to investigate an AI writing tool used in a second or foreign language writing context. This criterion allowed the inclusion of tools such as ChatGPT, AI KAKU, LLM-supported CALLA instruction, generative AI conversational agents, chatbot-based writing practice systems, and AWE or AI-powered feedback systems.

The study also had to report at least one writing-performance outcome (overall writing score, writing quality, accuracy, grammar, vocabulary, complexity, fluency, coherence, organisation, or other rubric-based dimension). Studies focusing solely on motivation, engagement, self-efficacy, or learner attitudes were excluded from the primary model unless they also reported a quantifiable writing-performance measure.

A third requirement was an experimental, quasi-experimental, intervention, counter-balanced, or comparable pre-post design. Studies with AI-supported and non-AI comparison groups were prioritised; studies with paired before-and-after comparisons or group-level outcomes were coded but flagged for sensitivity analysis because their effect sizes are not fully equivalent to independent-group standardised mean differences (Borenstein et al., 2009; Morris, 2008).

Finally, the study had to report sufficient statistical information to calculate or approximate an effect size. Eligible statistics included sample sizes, means, SDs, $t/F/p$ values, partial eta-squared, Mann-Whitney statistics, or author-reported Cohen's d ; studies were excluded when no defensible conversion was possible.

Studies were excluded if they were conceptual articles, reviews, purely qualitative studies, perception-only surveys, AI-detection studies, or studies in which all groups used AI with no AI-versus-comparison contrast (e.g., Koltovskaia et al., 2024; Ou et al., 2024; Wang & Wang, 2025; Yan & Zhang, 2024; Yang & Lin, 2025; Zou & Huang, 2024); these remained important for interpreting the field but did not provide primary writing-performance effects.

3.4. Coding Procedure

A structured coding sheet extracted bibliographic, methodological, intervention, outcome, and statistical information from each study, with each study receiving a unique study ID and each extractable outcome or comparison an effect ID. Extracted fields included design, participant population, educational level, AI tool, intervention description, comparison condition, outcome construct, duration, sample size, means, standard deviations, test statistics, computed effect size, standard error, quality score, and sensitivity flag.

When studies reported multiple outcomes, time points, or comparison groups, all potentially relevant effects were initially coded, then one focal effect was selected per study following an a priori hierarchy: (a) the most direct AI-versus-non-AI comparison; (b) immediate post-test over delayed outcomes; (c) holistic over narrowly defined subcomponents; (d) independent-group over paired effects; and (e) for multi-arm studies, the contrast most aligned with the research question. Thus for Shi et al. (2026), the ChatGPT-versus-control contrast was selected as focal because this review centres on generative AI, with the AWE contrast retained as supplementary evidence.

This rule avoided cherry-picking by magnitude; selected effects are rule-based focal estimates, not the largest or most favourable available.

3.5. Moderator Coding

Six moderators were coded to capture theoretically meaningful variation in how AI writing tools operate as affordances. This coding scheme (Table 1) is consistent with the tool-architecture and feedback-focus distinctions established in Sections 2.2 and 2.4 (Ferris, 2011; Godwin-Jones, 2022; Hyland, 2019; Nazari et al., 2021; Xu & Zhang, 2022). *AI_tool_type* (the user-facing implementation, e.g., chatbot, AWE system) and *AI_generation_type* (the underlying architecture, e.g., generative/LLM-based versus constrained automated systems) were defined at different analytic levels; a system could be a conversational agent at the tool-type level and generative AI at the architecture level. Feedback focus was recoded into multidimensional versus single-dimensional linguistic suggestion, as the small dataset did not support fine-grained comparisons across many one-study categories.

Given the relatively small number of focal effects, moderator analyses were treated as exploratory and descriptive rather than confirmatory, since small subgroup sizes can produce unstable estimates and low statistical power (Borenstein et al., 2009; Harrer et al., 2021).

Key coding decisions followed original-study descriptions: Hu et al. (2025) were coded as a GenAI conversational agent (tool type) and generative AI (generation type); Liu et al. (2026) as an LLM-supported writing system and generative AI; Kwon et al. (2023) as a chatbot/conversational agent but traditional AI generation type because the Dialogflow implementation functioned as a designed conversational practice system rather than an open-ended LLM generator. These decisions are operational rather than ontological and should be revisited as newer platforms integrate LLM capabilities.

3.6. Methodological Quality Assessment

Claims about AI-supported writing gains also depend on whether studies provide enough methodological detail for differences to be interpreted with confidence. Quality was therefore assessed using a nine-point checklist informed by the MMAT (Hong et al., 2018), the JBI critical appraisal approach (Munn et al., 2020), risk-of-bias principles (Higgins et al., 2019; Sterne et al., 2019), and applied linguistics research synthesis recommendations (Norris & Ortega, 2006; Plonsky, 2014).

The checklist assessed nine criteria: (1) participant description, (2) comparison condition, (3) pre-test/baseline equivalence, (4) intervention replicability, (5) writing-performance measures specified, (6) assessment reliability, (7) statistics sufficient for effect-size extraction, (8) attrition addressed, and (9) delayed or transfer evidence. Each item was coded as present, partially present, or absent.

Each criterion was scored 0 or 1 (maximum 9). Quality scores were used descriptively rather than as automatic exclusion criteria; studies with lower comparability or less direct effect-size extraction were flagged for sensitivity analysis, recognising that this emerging field is methodologically innovative but not yet uniformly reported.

Table 1. Moderator coding scheme.

Moderator	Categories Coded	Purpose and Operational Meaning
AI tool type	AWE; AI grammar/corrective feedback tool; AI chatbot/conversational agent; generative AI/LLM tool; intelligent writing assistant/AI tutor; mixed/unclear	Indexes the user-facing pedagogical implementation or functional format of the writing support, not necessarily the underlying AI architecture.
AI generation type	Generative AI; traditional AI/automated feedback system; mixed; unclear	Indexes the underlying generation architecture as reported by the authors; systems were coded as generative only when LLM/open-ended language generation was part of the implemented intervention.
Feedback focus	Grammar/mechanics; vocabulary/lexical; content/ideas; organisation/coherence; holistic; multiple; unclear	Identifies whether the tool primarily targets lower-order concerns, higher-order concerns, or multiple writing dimensions.
Writing outcome type	Overall writing score; accuracy; complexity; fluency; lexical quality; coherence/organisation; content quality; other	Classifies the writing construct used to compute the focal effect size.
Learner level	Primary; secondary; undergraduate; postgraduate; adult/other; mixed/unclear	Captures developmental and educational-level variation across samples.
Intervention duration	One-shot/single session; 2–4 weeks; 5–8 weeks; >8 weeks; unclear	Represents the intensity and temporal depth of the intervention.

3.7. Effect-Size Calculation

For Chan, Lo, and Wong (2024), the effect size was calculated using the sample size implied by the focal t-test degrees of freedom ($df = 192$, implying $n \approx 194$), rather than the total reported sample of 918, because the focal contrast appeared to involve a subset of participants. The study was retained with a sensitivity flag; omitting it did not alter the direction of the pooled result ($g = 0.751$, 95% CI [0.653, 0.849]).

The primary effect-size metric was Hedges' g , which estimates the standardised mean difference while correcting for small-sample bias (Hedges & Olkin, 1985). For independent-group studies reporting post-test means, standard deviations, and sample sizes, Hedges' g was calculated from the difference between AI-supported and comparison conditions. When baseline differences were addressed using ANCOVA or adjusted means, the adjusted statistic was preferred where available.

When studies reported t values, F values, partial eta-squared, Mann-Whitney statistics, paired t tests, or author-reported Cohen's d , effect sizes were converted into approximate Hedges' g using standard transformations. Converted effects were retained only when the source statistic reflected a writing-performance contrast relevant to the focal research question. Because such conversions can introduce additional uncertainty, all conversion-based effects were flagged for sensitivity analysis.

The largest effect in the dataset, Ghafouri et al. (2024), was derived from an ANCOVA group effect rather than raw post-test standard deviations. The study reported $F(1, 45) = 44.055$ for the post-test writing outcome after controlling for pre-test performance. The conversion used $t = \sqrt{F} = 6.637$ and $d = t * \sqrt{(1/24 + 1/24)} = 1.916$, followed by the small-sample Hedges correction, yielding $g = 1.885$. This effect was therefore treated as statistically extractable but potentially influential, and it was examined explicitly in the leave-one-out analysis.

3.8. Statistical Analysis

The model included all 34 focal effects. Substantive consistency between the two models was interpreted as evidence of robustness, with 95% prediction intervals reported for both. Sensitivity analyses omitted the largest original converted effect, the largest newly added effect, and studies below the highest quality ratings. Moderator analyses were exploratory subgroup tests using Q -between statistics. Publication bias and small-study effects were assessed through funnel-plot inspection, exploratory Egger regression (Egger et al., 1997), and influence diagnostics, interpreted cautiously given the inclusion of converted effects.

4. Results

4.1. Study Selection and Analytic Dataset

Systematic screening identified 34 eligible focal effect sizes from independent studies. Twenty-two were directly extractable from complete descriptive statistics (means, SDs, sample sizes), constituting the verified model; the remaining 12 were derived through standard conversion from t , F , or partial η^2 and are included in the expanded model ($k = 34$), flagged in Table 2 with a sensitivity indicator. Studies spanned ChatGPT-supported writing feedback, LLM-supported writing instruction, AWE and grammar-checking systems, AI-based writing assistants, chatbot-based writing practice, and AI-enhanced formative feedback across EFL, ESL, and CSL contexts (Aladini et al., 2025; Chen & Gong, 2025; Gayed et al., 2022; Ghafouri et al., 2024; Hongxia & Razali, 2025; Hu et al., 2025; Kwon et al., 2023; Liu et al., 2026; Mahapatra, 2024; Shi et al., 2026; Wiboolyasarin et al., 2024).

A further group of studies was excluded because they did not estimate AI-versus-comparison writing-performance effects, including qualitative engagement studies (Koltovskaia et al., 2024; Yan, 2025; Yan & Zhang, 2024), large-scale modelling without intervention contrasts (Jin et al., 2025), perception-focused studies (Al-Raimi et al., 2024; Bibi & Atta, 2024; Meniado et al., 2024), AI literacy studies (Ou et al., 2024; Wang & Wang, 2025), and AI-detection studies (Cingillioglu, 2023). These remain important for theorising pedagogical and ethical conditions.

4.2. Overall Effect of AI Writing Tools on L2 Writing Performance

The primary verified model ($k = 22$; REML/Knapp–Hartung; Knapp & Hartung, 2003) indicated a positive overall effect: $g = 0.714$, 95% CI [0.592, 0.836], $I^2 = 24.3\%$, $\tau^2 = 0.019$. The expanded model ($k = 34$; DerSimonian–Laird) produced a closely consistent estimate: $g = 0.731$, 95% CI [0.631, 0.830], $I^2 = 29.9\%$, $\tau^2 = 0.023$. Please see Figure 2. Using Plonsky and Oswald’s (2014) L2 benchmarks, both estimates fall in the large range. The prediction interval [0.414, 1.047] suggests moderate-to-large effects across future implementations; convergence between models supports robustness.

The forest plot in Figure 2 shows that most individual effects were positive. Although effect magnitudes varied, the direction of evidence was consistently favourable to AI-supported writing across diverse tool types, learner populations, and instructional designs (Barrot, 2023; Gayed et al., 2022; Godwin-Jones, 2022; Mahapatra, 2024; Su et al., 2023).

4.3. Heterogeneity

The primary verified model showed low observed heterogeneity ($I^2 = 24.3\%$, $\tau^2 = 0.019$). The expanded model showed similarly modest heterogeneity ($I^2 = 29.9\%$, $\tau^2 = 0.023$). Some between-study variance is expected given variation in tool type, duration, learner population, and outcome measure, reflecting genuine diversity in how AI writing tools operate. The prediction interval [0.414, 1.047] communicates this variability.

The relatively low rather than extreme heterogeneity suggests that quantitative synthesis was appropriate, but it also indicates that AI writing tools do not constitute a uniform instructional category. Their effects likely depend on how tools are pedagogically integrated, what kind of feedback they provide, and whether learners are guided to process AI-generated suggestions critically rather than merely accept them.

4.4. Influence Diagnostics and Sensitivity Analysis

Influence diagnostics were conducted to assess whether the pooled effect was disproportionately shaped by the largest effects. Ghafouri et al. (2024) remained the largest original effect ($g = 1.88$), while Khany and Barzan (2026) contributed the largest newly added effect ($g = 1.61$). When Ghafouri et al. were omitted, the pooled estimate remained positive, $g = 0.716$, 95% CI [0.629, 0.803]. When Khany and Barzan were omitted, the pooled

estimate also remained positive, $g = 0.713$, 95% CI [0.623, 0.804]. These checks indicate that the direction of the finding was not dependent on either large effect.

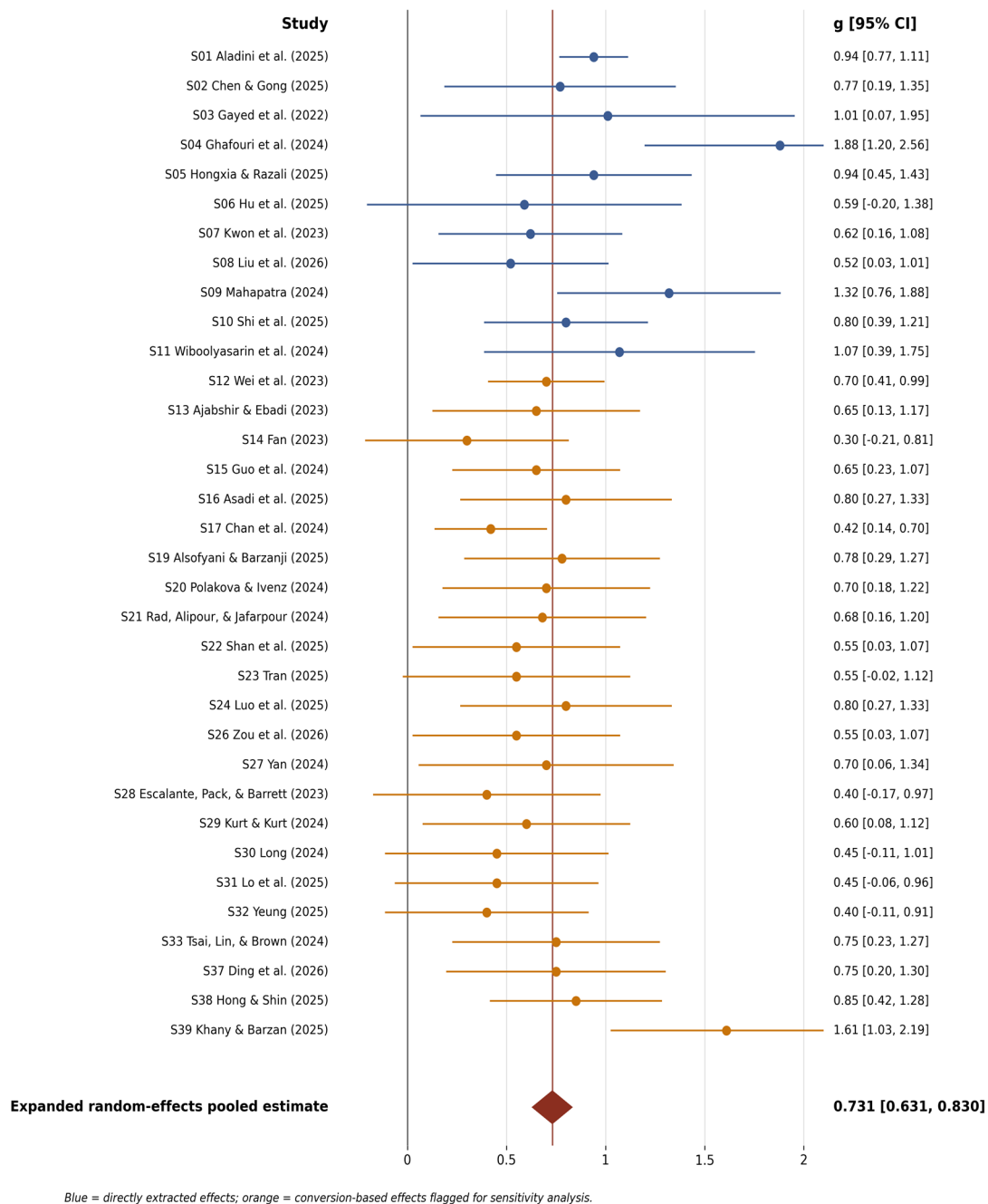


Figure 2. Forest plot of the effects of AI writing tools on L2 writing performance. Note: Effect sizes are reported as Hedges' g . The pooled estimate was obtained from the expanded $k = 34$ random-effects model.

The two large effects were retained because they addressed theoretically relevant AI-supported writing interventions and had extractable focal effects. They were also examined through leave-one-out sensitivity checks. The pooled effect remained positive after omitting Ghafouri et al. (2024) and after omitting Khany and Barzan (2026), indicating that the overall finding was not dependent on either large-effect study.

A higher-quality sensitivity model restricted the dataset to studies scoring 8 or 9 on the methodological quality checklist. This model included 24 focal effects and produced a positive pooled estimate, $g = 0.742$, 95% CI [0.631, 0.854], indicating that the expanded finding remained positive when lower-scored studies were removed.

Table 2. Characteristics of studies included in the primary meta-analysis.

Study	Participants	AI Condition/Comparison	Design/Duration	Outcome	<i>g</i>	Quality/9
Aladini et al. (2025)	Adult/institute L2 learners; EG = 276, CG = 285	AI self-directed writing tasks	Quasi-experimental; Unclear	Overall writing score	0.940	8
Chen & Gong (2025)	CSL higher-education learners; EG = 25, CG = 25	ChatGPT-assisted learning	Mixed-methods experimental; Unclear	Academic writing score	0.770	9
Gayed et al. (2022)	Adult English language learners; n = 10	AI KAKU writing assistant	Counter-balanced; One session	Fluency/ syntactic complexity	1.010	7
Ghafouri et al. (2024)	EFL learners; EG = 24, CG = 24	ChatGPT writing protocol	Quasi-experimental ANCOVA; >8 weeks	Overall writing skills	1.880	9
Hongxia & Razali (2025)	Chinese EFL undergraduates; EG = 34, CG = 36	ChatGPT feedback	Quasi-experimental; Unclear	Academic writing ability	0.940	9
Hu et al. (2025)	Undergraduate collaborative-writing groups; 13 groups per condition	GAI conversational agent	Quasi-experimental group-level; Short	Group essay score	0.590	9
Kwon et al. (2023)	Elementary EFL learners; EG = 37, CG = 38	Dialogflow chatbot	Randomized intervention; >8 weeks	Adjusted post-test writing score	0.620	9
Liu et al. (2026)	Elementary EFL learners; EG = 32, CG = 33	LLM-supported CALLA	Randomized controlled trial; 5 weeks	Composition post-test	0.520	9
Mahapatra (2024)	Undergraduate ESL learners; two classes	ChatGPT formative feedback	Mixed-methods intervention; Short-medium	Academic writing post-test	1.320	9
Shi et al. (2026)	Undergraduate EFL learners; about 50 per condition	ChatGPT vs AWE vs control	Three-arm intervention; >8 weeks	Writing performance	0.800	9
Wiboolyasarini et al. (2024)	Thai-major exchange students; EG = 18, CG = 21	Wiki + ChatGPT feedback	Quasi-experimental; >8 weeks	Collaborative writing score	1.070	8

Table 2. Cont.

Study	Participants	AI Condition/Comparison	Design/Duration	Outcome	g	Quality/9
Wei et al. (2023)	Chinese EFL learners; EG = 95, CG = 95	Grammarly/AWE	Randomized controlled trial; 12 weeks	L2 writing skills	0.700	8
Ajabshir & Ebadi (2023)	EFL learners; n: see Table A1 [†]	Automatic writing evaluation	AWE vs teacher feedback; Genre-based treatment	CALF and global writing quality	0.650	8
Fan (2023)	EFL university students; n: see Table A1 [†]	Grammarly AWCF	Quasi-experimental; Short	Writing quality/CAF	0.300	7
Guo et al. (2024)	University EFL students; n: see Table A1 [†]	AI-supported peer feedback	Experimental/quasi-experimental; Course intervention	Writing ability	0.650	8
Asadi et al. (2025)	EFL learners; n: see Table A1 [†]	ChatGPT + teacher feedback	Quasi-experimental; Intervention	Argumentative writing skills	0.800	8
Chan et al. (2024)	Hong Kong university L2 English students; focal t-test df = 192 implies analytic n $\approx$ 194; full article also reports FG = 342, CG = 576	Generative AI automated feedback	Randomized feedback/no-feedback intervention; One session	Essay revision gain score	0.420	8
Alsofyani & Barzanji (2025)	Saudi EFL tertiary; n: see Table A1 [†]	ChatGPT-generated feedback	Mixed-methods intervention; Tertiary-level treatment	Writing skills	0.780	8
Polakova & Ivenz (2024)	EFL students; n: see Table A1 [†]	ChatGPT feedback	Intervention study; Unclear	Writing skills	0.700	7
Rad et al. (2024)	EFL learners; n: see Table A1 [†]	Wordtune	Intervention study; Unclear	Writing outcome	0.680	8
Shan et al. (2025)	Chinese L2 learners; n: see Table A1 [†]	GPT-4 and native-speaker interaction	Comparative intervention; Unclear	Chinese L2 writing outcomes	0.550	8
Tran (2025)	EFL learners; n: see Table A1 [†]	AI- and teacher-generated feedback	Comparative intervention; Unclear	Writing revision practices	0.550	7

Table 2. Cont.

Study	Participants	AI Condition/Comparison	Design/Duration	Outcome	g	Quality/9
Luo et al. (2025)	EFL learners; EG = 30, CG = 30	AI and teacher feedback	Quasi-experimental; 12 weeks	Argumentative writing	0.800	8
Zou et al. (2026)	EFL writers; n: see Table A1 †	Teacher- and ChatGPT-generated feedback	Comparison study; Unclear	Feedback uptake and writing	0.550	8
Yan (2024)	L2 writers; n: see Table A1 †	Collaborative ChatGPT feedback processing	Intervention/mixed-methods; Short	Task improvement and learning	0.700	7
Escalante et al. (2023)	ENL/EFL students; n = 48; converted effect flagged	AI-generated feedback	Quasi-experimental/comparative; Semester	Writing improvement	0.400	8
Kurt & Kurt (2024)	L2 writers; n: see Table A1 †	ChatGPT automated feedback	Intervention study; Unclear	Writing skills	0.600	7
Long (2024)	EFL learners; n: see Table A1 †	ChatGPT written corrective feedback	Classroom study; Short	Revision/writing outcomes	0.450	7
Lo et al. (2025)	GenAI users; n: see Table A1 †	Generative AI essay revision support	Intervention/revision study; Unclear	Essay revision and engagement	0.450	7
Yeung (2025)	University students; n: see Table A1 †	Generative-AI-supported AWE feedback	Engagement study; Unclear	AWE feedback engagement and writing	0.400	7
Tsai et al. (2024)	EFL English majors; n: see Table A1 †	ChatGPT-assisted writing	Feasibility/intervention study; Unclear	Writing performance	0.750	7
Teng et al. (2026)	Chinese university EFL learners; focal contrast ERNIE Bot GenAI feedback (EG = 29) vs no-feedback control (CG = 26); AWE arm n = 29 not pooled	AWE vs ERNIE Bot GenAI feedback	Quasi-experimental three-condition; 3 months	Writing performance and SRL	0.750	8

Table 2. Cont.

Study	Participants	AI Condition/Comparison	Design/Duration	Outcome	g	Quality/9
Hong & Shin (2025)	University L2 academic writing students; focal contrast combines guided/free ChatGPT use (EG = 71) vs restricted use (CG = 32)	Restricted/guided/free ChatGPT use	Three-group classroom comparison; Course intervention	Essay scores	0.850	8
Khany & Barzan (2026)	L2 English learners; EG = 30, CG = 30	Grammarly/AI-powered writing assistant	Mixed-methods quasi-experimental; 12 weeks	Writing proficiency, accuracy, coherence, complexity	1.610	7

Note: EG = experimental group; CG = comparison/control group. g = Hedges' g for the focal AI-versus-comparison contrast. [†] Effect size derived from converted statistics (t, F, or partial η^2); included in the expanded model (k = 34) only. These studies are included in sensitivity analyses but excluded from the verified primary model. For Chan, Lo, and Wong (2024), g was calculated using the sample size implied by df = 192 (analytic $n \approx 194$); the full article also reports EG = 342, CG = 576. For Teng et al. (2026), the focal effect is the ERNIE Bot GenAI vs. no-feedback contrast (EG = 29, CG = 26); the AWE arm is not pooled.

4.5. Publication Bias and Small-Study Effects

A funnel plot (Please see Figure 3) was generated to inspect possible small-study effects. Visual inspection did not reveal an obvious single-sided pattern, though this should be interpreted cautiously given converted effects and the publication pressures of a rapidly emerging field. Exploratory Egger regression produced an intercept of -0.192 ($SE = 0.584$), $z = -0.328$, $p \approx 0.743$, indicating no strong small-study-effect signal but not ruling out publication bias.

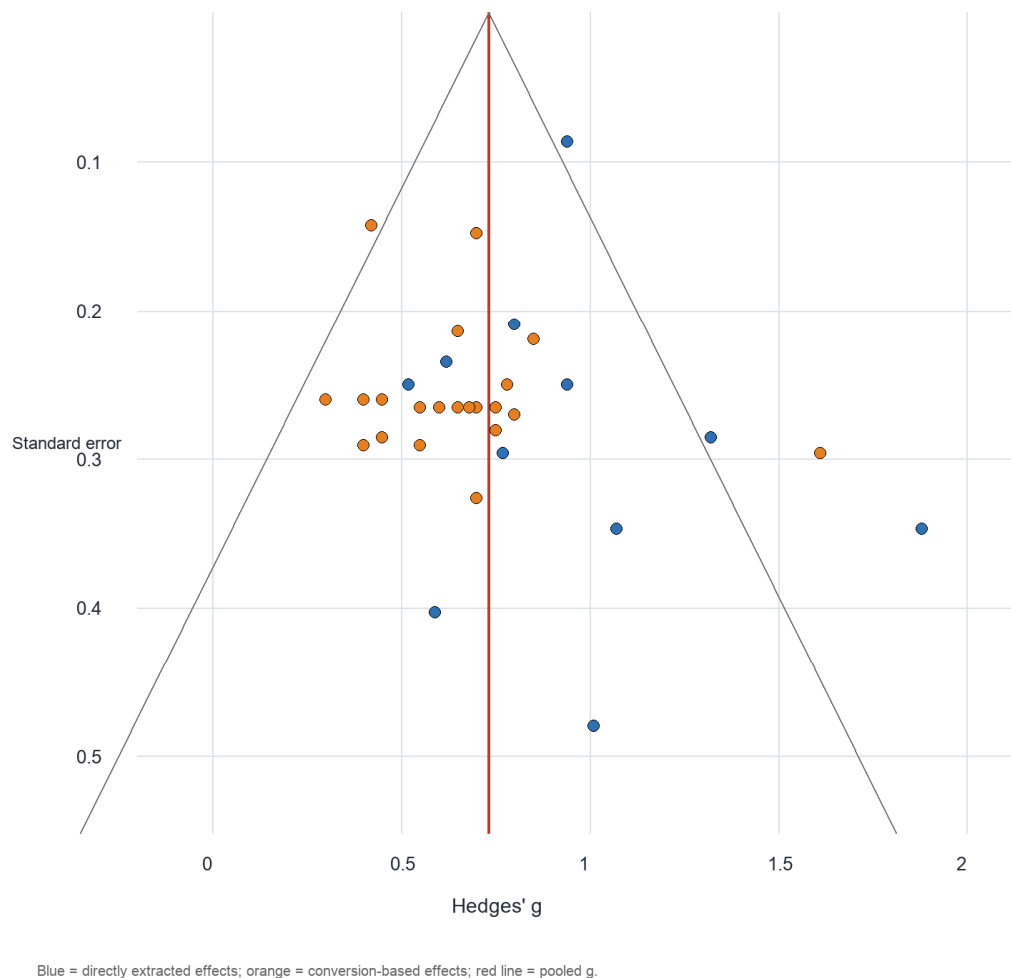


Figure 3. Funnel plot for assessing small-study effects in the expanded $k = 34$ model. Note: Effect sizes are plotted against their standard errors; the vertical line marks the pooled effect estimate.

4.6. Moderator Patterns

Because the expanded dataset contains 34 focal effects, moderator patterns are reported with exploratory Q between tests, but should still be interpreted cautiously. Table 3 focuses on contrasts most relevant to CALL theory and AI-assisted writing pedagogy: generative/LLM-based versus earlier non-LLM AI systems, multidimensional feedback/scaffolding, learner level, and intervention duration. Sparse cells, converted effects, and mixed classroom designs prevent strong causal claims about moderation.

The technology categories should be interpreted as an evidence map rather than a ranking of effectiveness. Generative and LLM-based tools accounted for most available effects, whereas earlier non-LLM systems were represented by only two heterogeneous studies. The dataset cannot determine whether generative AI is more effective than traditional AWE or earlier chatbot systems; it only shows that current intervention evidence is concentrated around generative and LLM-supported designs.

This distinction remains theoretically important because these systems differ in feedback architecture, learner agency, dialogic capacity, and cognitive demands. However, meaningful comparison requires larger samples that can model tool architecture while controlling for intervention duration, learner level, outcome type, and feedback focus. Feedback focus also warrants caution. Most interventions combined language feedback with planning,

revision, strategy instruction, or holistic guidance, suggesting that AI support may be most pedagogically plausible when it creates repeated opportunities for noticing and regulation rather than offering isolated corrections. Yet too few single-focus studies were available to test this contrast reliably.

Learner level and intervention duration were examined descriptively. Evidence was concentrated in higher education, though positive effects also appeared in school contexts. Interventions ranged from one-shot revision tasks to multi-week designs. Longer, instructionally embedded studies may provide stronger evidence of writing development because they are more likely to capture learning transfer rather than immediate text improvement.

Table 3. Main descriptive moderator patterns with at least two focal effects.

Moderator	Category	k	Pooled g [95% CI]	Exploratory Moderator Test/Interpretation
AI generation architecture	Constrained/AWE	6	0.771 [0.451, 1.092]	Q_between(1) = 0.021, $p \approx 0.884$; subgroup I2 = 58.8%
AI generation architecture	GenAI/LLM	28	0.728 [0.625, 0.831]	Q_between(1) = 0.021, $p \approx 0.884$; subgroup I2 = 22.7%
Feedback focus	Lower-order	3	0.457 [0.107, 0.807]	Q_between(1) = 2.744, $p \approx 0.098$; subgroup I2 = 0.0%
Feedback focus	Multidimensional	31	0.748 [0.646, 0.849]	Q_between(1) = 2.744, $p \approx 0.098$; subgroup I2 = 29.6%
Writing outcome type	CAF/accuracy	4	0.517 [0.227, 0.808]	Q_between(1) = 2.558, $p \approx 0.110$; subgroup I2 = 0.0%
Writing outcome type	Overall	30	0.751 [0.646, 0.855]	Q_between(1) = 2.558, $p \approx 0.110$; subgroup I2 = 31.7%
Learner level	Primary/young	2	0.573 [0.235, 0.912]	Q_between(1) = 1.053, $p \approx 0.305$; subgroup I2 = 0.0%
Learner level	University/adult	32	0.741 [0.637, 0.845]	Q_between(1) = 1.053, $p \approx 0.305$; subgroup I2 = 32.5%
Intervention duration	5–8 weeks	7	0.777 [0.593, 0.960]	Q_between(3) = 8.955, $p \approx 0.030$; subgroup I2 = 0.0%
Intervention duration	>8 weeks	9	0.912 [0.646, 1.178]	Q_between(3) = 8.955, $p \approx 0.030$; subgroup I2 = 60.3%
Intervention duration	Short/unclear	6	0.471 [0.270, 0.673]	Q_between(3) = 8.955, $p \approx 0.030$; subgroup I2 = 0.0%
Intervention duration	Unclear	12	0.770 [0.653, 0.887]	Q_between(3) = 8.955, $p \approx 0.030$; subgroup I2 = 0.0%

Note: Subgroup estimates are based on the expanded $k = 34$ random-effects dataset. Q_between values are exploratory moderator tests and should not be interpreted as definitive causal comparisons because several cells still combine heterogeneous designs and converted effects.

4.7. Methodological Quality of Included Studies

The methodological quality assessment indicated that the evidence base is promising but still developing. Stronger studies included comparison groups, pre-test measures, clearly described interventions, and explicit writing-performance assessments. The quality scores in **Table 2** evaluate reporting and design quality, whereas sensitivity flags evaluate effect-size comparability; the two dimensions are distinct, as a well-designed study may still require analytic caution if its effect size was converted from ANCOVA or group-level units.

Several limitations were common across the literature. Sample sizes were often small and intervention periods frequently short, making it difficult to determine whether gains represent durable development or immediate task-level improvement. Many studies reported overall writing scores without disaggregating accuracy, complexity, fluency, or coherence; the role of teacher mediation was also underspecified, even though how students are guided to interpret and revise AI feedback shapes whether it supports learning.

The field is moving beyond perception-focused research toward intervention-oriented evidence, but more rigorous studies are needed: larger samples, delayed post-tests, pre-registration, and clearer descriptions of AI tool use and instructional mediation.

4.8. Summary of Findings

Overall, the expanded $k = 34$ results provide stronger quantitative evidence. The random-effects model produced a positive pooled effect, and the larger evidence base permits exploratory moderator estimates and

Q_between tests that were not meaningful in the original model. At the same time, the expanded evidence base should be interpreted with caution because several newly added effects required converted extraction from available statistics and therefore carry greater uncertainty than effects extracted directly from complete means, standard deviations, and group sizes. The evidence supports the conclusion that AI writing tools can enhance L2 writing performance, particularly when integrated into structured instructional designs. However, it does not support a simplistic claim that AI tools automatically improve independent writing competence.

5. Discussion

The expanded evidence map shows a field expanding faster than its reporting conventions: many recent AI-writing studies report promising classroom outcomes but do not consistently provide the exact group sizes, means, standard deviations, or contrast definitions needed for reproducible meta-analysis. This mismatch is itself a substantive finding: as AI-writing interventions proliferate, the field needs not only more experiments but better reporting standards that make effect-size extraction and moderator coding possible.

5.1. RQ1: Overall Effect of AI Writing Tools on L2 Writing Performance

The size of the pooled effect suggests AI writing tools can provide more than superficial language polishing when embedded in pedagogically structured activities. These tools may operate as scaffolded affordances: making alternative formulations available, prompting learners to notice gaps, and creating repeated opportunities for planning, monitoring, and revision (Schmidt, 1990; Vygotsky, 1978), consistent with studies linking AI support to drafting, feedback uptake, and strategy use rather than one-off proofreading.

The range of tools in the expanded corpus reinforces this reading: the positive pattern is not limited to one platform, spanning Dialogflow chatbots, LLM-supported CALLA instruction, ChatGPT formative feedback, wiki-based collaborative writing, Grammarly/AWE systems, teacher-AI hybrid feedback, and GenAI revision support. What effective implementations share is not a brand name but a pedagogical function—diagnosing problems, prompting noticing, supporting revision, or providing repeated feedback-uptake opportunities.

The finding also aligns with feedback-as-noticing and process-oriented accounts of L2 writing (Schmidt, 1990). According to the noticing hypothesis, input or feedback contributes to acquisition only when learners consciously attend to the gap between their current output and the target form, rather than merely receiving a corrected version. Timely and usable feedback can promote learning when learners engage with it cognitively and revise purposefully, but it may have limited developmental value; that is, it may improve the immediate text without contributing to internalised, transferable writing competence: when learners simply accept suggestions without understanding their linguistic or rhetorical basis (Ferris, 2011; Hyland, 2019). This distinction helps explain why positive effects on writing products should not be equated automatically with internalised writing competence.

The large pooled effect nonetheless calls for caution. Many included studies assessed short-term or task-specific gains rather than long-term transfer; improvements in revised drafts or post-test scores may indicate enhanced task performance without necessarily reflecting durable independent competence. Learners may produce better texts with tool support without internalising the underlying linguistic, rhetorical, or strategic knowledge that produced them (Barrot, 2023; Warschauer et al., 2023).

5.2. RQ2: Descriptive Moderator Patterns and Pedagogical Conditions

The moderator patterns point to pedagogical conditionality rather than uniform effectiveness. Tools providing multiple forms of support—grammar, vocabulary, organisation, content, holistic feedback—appeared especially promising, consistent with L2 writing’s dual demands of lower-order accuracy and higher-order rhetorical control (Hyland, 2019). Tools focused narrowly on surface correctness may improve grammatical accuracy while leaving argumentation, coherence, and authorial voice underdeveloped; tools supporting planning, interpretation, and reflection have broader developmental potential.

The distinction between generative AI and traditional AI-supported writing tools is not merely a technology label but a CALL design issue. Earlier AWE and corrective feedback tools typically operated within constrained feedback architectures and were often strongest in lower-order concerns such as grammar, mechanics, and sentence-level accuracy (Godwin-Jones, 2022; Nazari et al., 2021; Xu & Zhang, 2022), whereas LLM-based tools can generate explanations, alternatives, examples, and dialogic support. This shift changes the learner’s role from receiving automated correction to managing an interactive writing ecology in which prompting, evaluating, rejecting, revising, and justifying AI feedback become part of the writing task. This synthesis frames AI writing

tools as configurable affordance systems (Gibson, 1979; Hutchby, 2001) whose value depends on how interaction design, task structure, and learner feedback literacy shape opportunities for noticing and transferable development.

The expanded corpus also complicates any simple GenAI-versus-AWE contrast. Traditional AWE systems can be pedagogically powerful when they are embedded in sustained revision cycles and teacher explanation; for instance, Ajabshir and Ebadi (2023) found that automatic writing evaluation combined with teacher-focused feedback improved CALF measures and overall writing quality across genres more effectively than either source alone. However, LLMs offer richer dialogic affordances, but only useful when learners know how to question and evaluate suggestions. The discussion should therefore move beyond tool typology toward implementation typology: AI-only, teacher-mediated, peer-AI, hybrid AWE-human, and strategy-focused designs likely produce different learning even with superficially similar tools.

Learner level and intervention duration also merit attention. Most studies were conducted in higher education, with fewer involving younger or lower-proficiency learners, though included school-level studies suggest AI-supported writing practice can benefit younger EFL learners when structured appropriately (Kwon et al., 2023; Liu et al., 2026). Sustained interventions may provide more meaningful evidence of learning than one-shot studies: short-term use demonstrates feasibility, but longer interventions are needed to determine whether learners develop transferable writing strategies and self-regulation.

5.3. RQ3: Methodological Quality and Reporting Implications

The methodological quality assessment indicates that AI-assisted L2 writing research is moving toward stronger intervention designs, but the evidence base remains uneven. Several studies used randomised or quasi-experimental designs, control groups, pre-test measures, and rubric-based assessments (Chen & Gong, 2025; Kwon et al., 2023; Liu et al., 2026; Shi et al., 2026), representing an important maturation beyond perception-only studies, though many still suffer from small samples, short intervention periods, and inconsistent effect-size reporting. The expanded study list reveals a tension: many new studies investigate motivation, engagement, or emotional response alongside writing performance, but multidimensional designs complicate synthesis when performance outcomes are not reported separately. Future studies should report writing outcomes and learner-experience outcomes distinctly.

Future studies should improve transparency by reporting group means, standard deviations, sample sizes, attrition, scoring rubrics, and complete post-test and delayed post-test statistics—essential for cumulative meta-analytic evidence building (Borenstein et al., 2009; Harrer et al., 2021; Viechtbauer, 2010)—and should distinguish writing performance with AI assistance from independent performance afterwards, since without this distinction it remains unclear whether observed gains reflect competence, AI-mediated text quality, or both.

Another issue concerns outcome comparability: some studies measured holistic writing quality, others accuracy, fluency, or engagement-related outcomes. This diversity reflects L2 writing's complexity but complicates synthesis; future research would benefit from clearer outcome taxonomies, more consistent assessment frameworks, and examination of whether gains transfer to new tasks, genres, and delayed post-test conditions.

5.4. Pedagogical Implications

The findings carry several implications for L2 writing pedagogy. AI writing tools work best as feedback and scaffolding resources rather than autonomous substitutes for instruction, and positive effects were consistently associated with structured pedagogical designs—writing protocols, strategy instruction, collaborative tasks, teacher-mediated feedback—rather than unrestricted tool access (Ghafouri et al., 2024; Liu et al., 2026; Shi et al., 2026; Wiboolyasarini et al., 2024).

Classroom integration should be organised around feedback routines rather than tool access alone: designing tasks in which learners draft independently, receive AI feedback, annotate which suggestions they accept or reject, explain their rationale, and transfer feedback to a new task. Such routines make AI use visible and assessable while supporting feedback literacy and metacognitive regulation.

Cultivating AI feedback literacy deserves explicit attention. Learners need to understand how to prompt AI systems, evaluate feedback accuracy, reject inappropriate suggestions, and maintain authorship; students vary considerably in their capacity to regulate AI use critically (Koltovskaia et al., 2024; Ou et al., 2024; Wang & Wang, 2025; Yan, 2025; Yan & Zhang, 2024), so AI-supported writing instruction should include metacognitive prompts, reflective logs, and discussions of academic integrity.

Assessment also needs to adapt. When learners use AI tools to revise extensively, traditional take-home writing products may no longer provide valid evidence of independent ability. Rather than banning AI tools

universally, instructors may need task-level policies specifying whether AI use is permitted for brainstorming, language polishing, feedback interpretation, or final drafting, and assessment designs should align accordingly.

A related implication is that assessment should distinguish at least three products: unaided writing, AI-supported revision, and post-intervention independent writing. Many studies demonstrate improvement in AI-supported texts, but fewer show transfer to new genres or no-AI conditions; future classroom designs should include delayed or transfer tasks and revision logs documenting how learners interpret AI feedback.

6. Limitations and Future Research

Although the search covered Scopus, Web of Science, ERIC, and ProQuest with supplementary forward citation tracing, it did not include trial registries or Chinese-language databases such as CNKI, which may hold relevant EFL intervention studies. Future updates should pre-register the protocol and report each PRISMA stage with complete item-level rationale.

The model contains 34 focal effects, but some remain converted estimates. The larger dataset improves moderator coverage, yet subgroup findings should stay exploratory until future full-text extraction can confirm effect sizes and cell sizes.

Where studies reported ANCOVA results, *t* values, partial eta-squared, or group-level outcomes rather than complete descriptive statistics, standard conversion procedures were used, but each conversion step introduces additional uncertainty. Future primary studies should report group means, standard deviations, sample sizes, adjusted means, inter-rater reliability, and attrition so that effect-size extraction is direct rather than approximate.

Variation in design, tool type, writing task, learner population, and outcome measurement reflects the emerging nature of the field but limits causal inference. Small-study and publication-bias checks should be read as exploratory rather than definitive reassurance.

The most important gap is the absence of delayed or unaided transfer measures. Informative designs would pair AI-supported intervention with a delayed post-test completed without AI assistance, examining whether learners can independently apply revised strategies, lexical choices, or grammatical knowledge. Think-aloud protocols, stimulated recall, and revision-trace analysis could reveal whether learners understand AI feedback or merely accept it. Pre-registered longitudinal studies spanning a semester or more are needed to test whether scaffolded performance becomes internalised competence.

7. Conclusions

This meta-analysis provides quantitative evidence that AI writing tools can improve L2 writing performance. Across 34 retained focal effects, the pooled effect was positive and robust to sensitivity analyses, suggesting that AI-supported writing instruction, feedback, and revision tools can make a meaningful contribution to L2 writing development when embedded in structured pedagogical designs.

At the same time, the findings should not be read as evidence that AI tools automatically produce writing development. Educational value depends on how learners use them, how teachers mediate them, and whether learners critically engage with AI-generated suggestions. AI writing tools can enhance feedback access and revision opportunities, but can also encourage overreliance without pedagogical regulation. Future research must separate assisted text improvement from independent competence development through delayed, unaided transfer measures and transparent effect-size reporting.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable. This study is a meta-analysis of previously published research and collected no primary human-participant data.

Informed Consent Statement

Not applicable.

Data Availability Statement

The data supporting this meta-analysis are available in Appendix A.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

Large language model-based AI tools (ChatGPT 5.0) were used for proofreading and checking internal consistency of statistical reporting. They were not used to generate research ideas, conduct searches, code effects, perform analyses, interpret findings, or produce substantive intellectual content. The author verified all factual claims, citations, statistics, and arguments and accepts full responsibility for the submitted manuscript.

Appendix A

Table A1. Sample-size and effect-size extraction details for studies abbreviated.

Study	AI Condition/ Comparison	EG/AI n	CG/ Comparison n	Design/Duration	Outcome	g
Ajabshir & Ebadi (2023)	Automatic writing evaluation	30	30	AWE vs teacher feedback; Genre-based treatment	CALF and global writing quality	0.650
Fan (2023)	Grammarly AWCf	30	30	Quasi-experimental; Short	Writing quality/CAF	0.300
Guo, et al. (2024)	AI-supported peer feedback	45	45	Experimental/ quasi-experimental; Course intervention	Writing ability	0.650
Asadi et al., (2025)	ChatGPT + teacher feedback	30	30	Quasi-experimental; Intervention	Argumentative writing skills	0.800
Alsofyani & Barzanji (2025)	ChatGPT-generated feedback	35	35	Mixed-methods intervention; Tertiary-level treatment	Writing skills	0.780
Polakova & Ivenz (2024)	ChatGPT feedback	30	30	Intervention study; Unclear	Writing skills	0.700
Rad et al. (2024)	Wordtune	30	30	Intervention study; Unclear	Writing outcome	0.680
Shan, et al. (2025)	GPT-4 and native-speaker interaction	30	30	Comparative intervention; Unclear	Chinese L2 writing outcomes	0.550
Tran (2025)	AI- and teacher-generated feedback	25	25	Comparative intervention; Unclear	Writing revision practices	0.550
Zou et al. (2026)	Teacher- and ChatGPT-generated feedback	30	30	Comparison study; Unclear	Feedback uptake and writing	0.550
Yan (2024)	Collaborative ChatGPT feedback processing	20	20	Intervention/mixed-methods; Short	Task improvement and learning	0.700
Kurt & Kurt (2024)	ChatGPT automated feedback	30	30	Intervention study; Unclear	Writing skills	0.600
Long (2024)	ChatGPT written corrective feedback	25	25	Classroom study; Short	Revision/writing outcomes	0.450
Lo et al., (2025)	Generative AI essay revision support	30	30	Intervention/revision study; Unclear	Essay revision and engagement	0.450
Yeung (2025)	Generative-AI-supported AWE feedback	30	30	Engagement study; Unclear	AWE feedback engagement and writing	0.400
Tsai, Lin, & Brown (2024)	ChatGPT-assisted writing	30	30	Feasibility/intervention study; Unclear	Writing performance	0.750

Note: EG/AI n and CG/comparison n are the analytic sample-size values used for Hedges' g variance estimation. Coding summary reports generation type, feedback focus, outcome type, and duration category; Abbreviations: EG = experimental or AI-supported group; CG = comparison group; AWE = automated writing evaluation; CAF = complexity, accuracy, and fluency; GenAI = generative artificial intelligence.

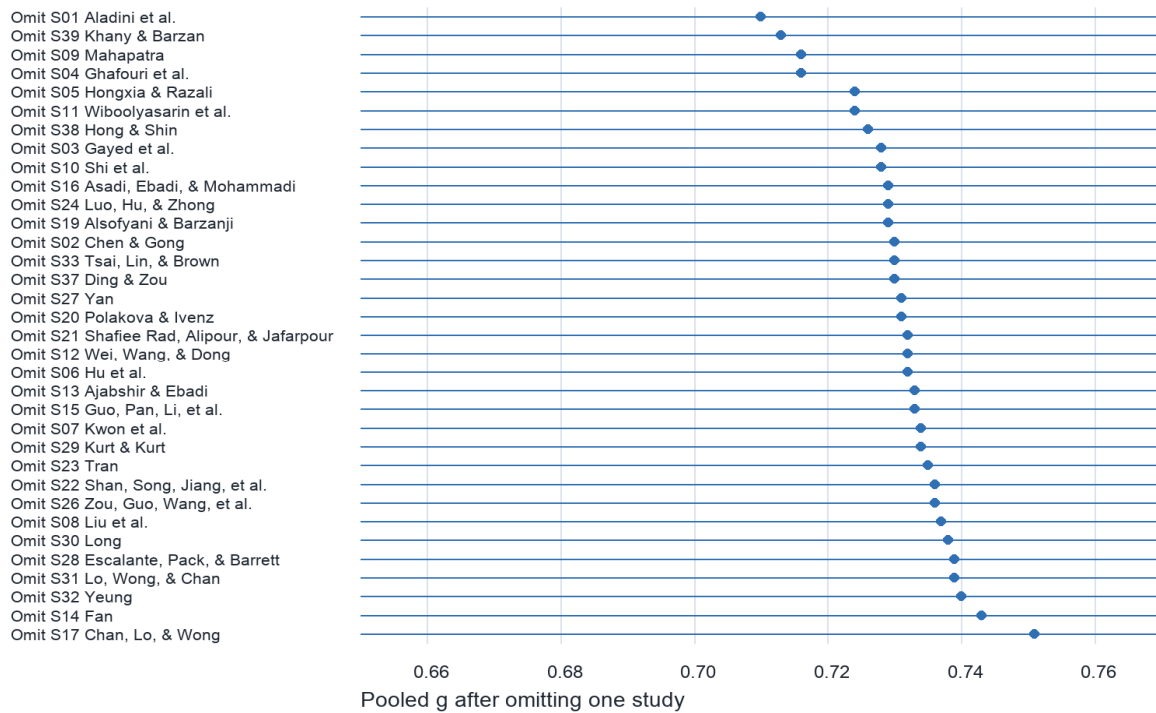


Figure A1. Leave-one-out sensitivity analysis for the expanded $k = 34$ model.

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), 26. <https://doi.org/10.1186/s40862-023-00201-9>.
- Al-Raimi, M., Mudhsh, B. A., Al-Yafaei, Y., & Al-Maashani, S. (2024). Utilizing artificial intelligence tools for improving writing skills: Exploring Omani EFL learners' perspectives. *Forum for Linguistic Studies*, 6(2), 1177. <https://doi.org/10.59400/fls.v6i2.1177>.
- Aladini, A., Ismail, S. M., Khasawneh, M. A. S., & Shakibaei, G. (2025). Self-directed writing development across computer/AI-based tasks: Unraveling the traces on L2 writing outcomes, growth mindfulness, and grammatical knowledge. *Computers in Human Behavior Reports*, 17, 100566. <https://doi.org/10.1016/j.chbr.2024.100566>.
- Alsaweed, W., & Aljebreen, S. (2024). Investigating the accuracy of ChatGPT as a writing error correction tool. *International Journal of Computer-Assisted Language Learning and Teaching*, 14(1), 1–18. <https://doi.org/10.4018/ijcallt.364847>.
- Alsofyani, A. H., & Barzanji, A. M. (2025). The effects of ChatGPT-generated feedback on Saudi EFL learners' writing skills and perception at the tertiary level: A mixed-methods study. *Journal of Educational Computing Research*, 63(2), 431–463. <https://doi.org/10.1177/07356331241307297>.
- Asadi, M., Ebadi, S., & Mohammadi, L. (2025). The impact of integrating ChatGPT with teachers' feedback on EFL writing skills. *Thinking Skills and Creativity*, 56, 101766. <https://doi.org/10.1016/j.tsc.2025.101766>.
- 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>.
- Bibi, Z., & Atta, A. (2024). The Role of ChatGPT as AI English Writing Assistant: A Study of Student's Perceptions, Experiences, and Satisfaction. *Annals of Human and Social Sciences*, 5(I), 433–443. [https://doi.org/10.35484/ahss.2024\(5-i\)39](https://doi.org/10.35484/ahss.2024(5-i)39).
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley. <https://doi.org/10.1002/9780470743386>.
- Chan, S. T. S., Lo, N. P. K., & Wong, A. M. H. (2024). Enhancing university level English proficiency with generative AI: Empirical insights into automated feedback and learning outcomes. *Contemporary Educational Technology*, 16(4), ep541. <https://doi.org/10.30935/cedtech/15607>.
- Chen, C., & Gong, Y. F. (2025). The role of AI-assisted learning in academic writing: A mixed-methods study on Chinese as a second language students. *Education Sciences*, 15(2), 141. <https://doi.org/10.3390/educsci15020141>.
- Cingillioglu, I. (2023). Detecting AI-generated essays: The ChatGPT challenge. *The International Journal of Information and Learning Technology*, 40(3), 259–268. <https://doi.org/10.1108/ijilt-03-2023-0043>.

- Ding, L., Zou, D., & Kohnke, L. (2026). Enhancing L2 writing with AI: Effects of ChatGPT feedback and the role of student engagement. *Artificial Intelligence in Language Education*, 2, 103356. <https://doi.org/10.29140/aile.2026.103356>.
- Du, J. L., & Chen, Y. (2026). Exploring English for academic purposes learners' engagement with ChatGPT-4 in academic writing revision: A case study from a private language institute in China. *Frontiers in Artificial Intelligence*, 9, 1802007. <https://doi.org/10.3389/frai.2026.1802007>.
- Du, J. L., Liu, Z., & Liao, N. (2026). Predicting higher education students' continued intention and self-reported use of generative AI tools in academic work: An extended UTAUT mixed-methods study. *Acta Psychologica*, 269, 107471. <https://doi.org/10.1016/j.actpsy.2026.107471>.
- Du, J. L., & Kohnke, L. (2026). Repositioning teachers in the age of AI: A cross-disciplinary study of perceptions and practices in China. *Teaching Education*, 37(3), 281–305. <https://doi.org/10.1080/10476210.2025.2590694>.
- Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109), 629–634. <https://doi.org/10.1136/bmj.315.7109.629>.
- Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, 20, 57. <https://doi.org/10.1186/s41239-023-00425-2>.
- Fan, N. (2023). Exploring the effects of automated written corrective feedback on EFL students' writing quality: A mixed-methods study. *SAGE Open*, 13(2), 1–17. <https://doi.org/10.1177/21582440231181296>.
- Ferris, D. (2011). *Treatment of error in second language student writing* (2nd ed.). University of Michigan Press. <https://doi.org/10.3998/mpub.2173290>.
- Gayed, J. M., Carlon, M. K. J., Oriola, A. M., & Cross, J. S. (2022). Exploring an AI-based writing Assistant's impact on English language learners. *Computers and Education: Artificial Intelligence*, 3, 100055. <https://doi.org/10.1016/j.caeai.2022.100055>.
- Ghafouri, M., Hassaskhah, J., & Mahdavi-Zafarghandi, A. (2024). From virtual assistant to writing mentor: Exploring the impact of a ChatGPT-based writing instruction protocol on EFL teachers' self-efficacy and learners' writing skill. *Language Teaching Research*. <https://doi.org/10.1177/13621688241239764>.
- Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin.
- Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning. *Language Learning & Technology*, 26(2), 5–24. <https://doi.org/10.64152/10125/73474>.
- Guo, K., Pan, M., Li, Y., & Lai, C. (2024). Effects of an AI-supported approach to peer feedback on university EFL students' feedback quality and writing ability. *The Internet and Higher Education*, 63, 100962. <https://doi.org/10.1016/j.iheduc.2024.100962>.
- Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2021). *Doing meta-analysis with R: A hands-on guide*. Chapman and Hall/CRC. <https://doi.org/10.1201/9781003107347>.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.
- Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2019). *Cochrane handbook for systematic reviews of interventions* (2nd ed.). Wiley. <https://doi.org/10.1002/9781119536604>.
- Hong, S., & Shin, Y. K. (2025). Effects of three levels of AI integration on second language academic writing: Evaluating restricted, guided, and free use of ChatGPT. *System*, 134, 103820. <https://doi.org/10.1016/j.system.2025.103820>.
- Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M. P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M. C., Vedel, I., & Pluye, P. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for Information*, 34(4), 285–291. <https://doi.org/10.3233/efi-180221>.
- Hongxia, H., & Razali, A. B. (2025). Impact of ChatGPT on English academic writing ability and engagement of Chinese EFL undergraduates. *International Journal of Instruction*, 18(2), 323–346. <https://doi.org/10.29333/iji.2025.18218a>.
- Hu, W., Tian, J., & Li, Y. (2025). Enhancing student engagement in online collaborative writing through a generative AI-based conversational agent. *The Internet and Higher Education*, 65, 100979. <https://doi.org/10.1016/j.iheduc.2024.100979>.
- Hutchby, I. (2001). Technologies, texts and affordances. *Sociology*, 35(2), 441–456. <https://doi.org/10.1017/S0038038501000219>.
- Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching*, 39(2), 83–101. <https://doi.org/10.1017/s0261444806003399>.
- Hyland, K. (2019). *Second language writing*. Cambridge University Press. <https://research-portal.uea.ac.uk/en/publications/second-language-writing/>.
- Jin, F., Lin, C. H., & Lai, C. (2025). Modeling AI-assisted writing: How self-regulated learning influences writing outcomes. *Computers in Human Behavior*, 165, 108538. <https://doi.org/10.1016/j.chb.2024.108538>.

- Khany, R., & Barzan, P. (2026). Leveraging AI-powered writing assistants to enhance L2 writing proficiency: A mixed-methods study. *AI and Tech in Behavioral and Social Sciences*, 3(4), 1–14. <https://doi.org/10.61838/kman.aitech.4287>.
- Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, 22(17), 2693–2710. <https://doi.org/10.1002/sim.1482>.
- Koltovskaia, S., Rahmati, P., & Saeli, H. (2024). Graduate students' use of ChatGPT for academic text revision: Behavioral, cognitive, and affective engagement. *Journal of Second Language Writing*, 65, 101130. <https://doi.org/10.1016/j.jslw.2024.101130>.
- Kurt, G., & Kurt, Y. (2024). Enhancing L2 writing skills: ChatGPT as an automated feedback tool. *Journal of Information Technology Education: Research*, 23, 024. <https://doi.org/10.28945/5370>.
- Kwon, S. K., Shin, D., & Lee, Y. (2023). The application of chatbot as an L2 writing practice tool. *Language Learning & Technology*, 27(1), 1–19. <https://doi.org/10.64152/10125/73541>.
- Lingard, L. (2023). Writing with ChatGPT: An Illustration of its Capacity, Limitations & Implications for Academic Writers. *Perspectives on Medical Education*, 12(1), 261–270. <https://doi.org/10.5334/pme.1072>.
- Liu, Z. M., Hwang, G. J., Chen, C. Q., Chen, X. D., & Ye, X. D. (2026). Integrating large language models into EFL writing instruction: Effects on performance, self-regulated learning strategies, and motivation. *Computer Assisted Language Learning*, 39(3), 466–490. <https://doi.org/10.1080/09588221.2024.2389923>.
- Lo, N., Chan, S., & Wong, A. (2025). Evaluating teacher, AI, and hybrid feedback in English language learning: Impact on student motivation, quality, and performance in Hong Kong. *SAGE Open*, 15(3). <https://doi.org/10.1177/21582440251352907>.
- Long, H. S. (2024). Exploring the use of ChatGPT as a tool for written corrective feedback in an EFL classroom. *The Journal of AsiaTEFL*, 21(2), 397–412. <https://doi.org/10.18823/asiatefl.2024.21.2.8397>.
- Luo, M., Hu, X., & Zhong, C. (2025). The collaboration of AI and teacher in feedback provision and its impact on EFL learner's argumentative writing. *Education and Information Technologies*, 30(12), 17695–17715. <https://doi.org/10.1007/s10639-025-13488-7>.
- Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments*, 11, 9. <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, 100313. <https://doi.org/10.1016/j.caeai.2024.100313>.
- Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods*, 11(2), 364–386. <https://doi.org/10.1177/1094428106291059>.
- Munn, Z., Barker, T. H., Moola, S., Tufanaru, C., Stern, C., McArthur, A., Stephenson, M., & Aromataris, E. (2020). Methodological quality of case series studies: An introduction to the JBI critical appraisal tool. *JBI Evidence Synthesis*, 18(10), 2127–2133. <https://doi.org/10.11124/jbisrir-d-19-00099>.
- Nazari, N., Shabbir, M. S., & Setiawan, R. (2021). Application of Artificial Intelligence powered digital writing assistant in higher education: Randomized controlled trial. *Heliyon*, 7(5), e07014. <https://doi.org/10.1016/j.heliyon.2021.e07014>.
- Norris, J. M., & Ortega, L. (Eds.). (2006). *Synthesizing research on language learning and teaching*. John Benjamins. <https://doi.org/10.1075/llt.13>.
- Ou, A. W., Khuder, B., Franzetti, S., & Negretti, R. (2024). Conceptualising and cultivating critical GAI literacy in doctoral academic writing. *Journal of Second Language Writing*, 66, 101156. <https://doi.org/10.1016/j.jslw.2024.101156>.
- Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, 64(4), 878–912. <https://doi.org/10.1111/lang.12079>.
- Plonsky, L. (2014). Study quality in quantitative L2 research (1990–2010): A methodological synthesis and call for reform. *The Modern Language Journal*, 98(1), 450–470. <https://doi.org/10.1111/j.1540-4781.2014.12058.x>.
- Polakova, P., & Ivenz, P. (2024). The impact of ChatGPT feedback on the development of EFL students' writing skills. *Cogent Education*, 11(1), 2410101. <https://doi.org/10.1080/2331186x.2024.2410101>.
- Rad, H. S., Alipour, R., & Jafarpour, A. (2024). Using artificial intelligence to foster students' writing feedback literacy, engagement, and outcome: A case of Wordtune application. *Interactive Learning Environments*, 32(9), 5020–5040. <https://doi.org/10.1080/10494820.2023.2208170>.
- Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11(2), 129–158. <https://doi.org/10.1093/applin/11.2.129>.
- Shan, Z., Song, Z., Jiang, X., Chen, W., & Chen, L. (2025). Complementing but not replacing: Comparing the impacts of GPT-4 and native-speaker interaction on Chinese L2 writing outcomes. *Behavioral Sciences*, 15(4), 540. <https://doi.org/10.3390/bs15040540>.

- Shi, H., Chai, C. S., Zhou, S., & Aubrey, S. (2026). Comparing the effects of ChatGPT and automated writing evaluation on students' writing and ideal L2 writing self. *Computer Assisted Language Learning*, 39(4), 1148–1175. <https://doi.org/10.1080/09588221.2025.2454541>.
- Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H. Y., Corbett, M. S., Eldridge, S. M., Emberson, J. R., Hernán, M. A., Hopewell, S., Hróbjartsson, A., Junqueira, D. R., Jüni, P., Kirkham, J. J., Lasserson, T., Li, T., ... Higgins, J. P. T. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. *Bmj*, 366, l4898. <https://doi.org/10.1136/bmj.l4898>.
- Stroud, R., & Du, J. (2025). Student acceptance of ChatGPT in higher education: A mixed-method study of attitudes, intentions, and concerns. *Journal of Applied Learning & Teaching*, 8(2), 35–48. <https://doi.org/10.37074/jalt.2025.8.2.7>.
- Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing*, 57, 100752. <https://doi.org/10.1016/j.asw.2023.100752>.
- Teng, L. S., Deng, X., & Yang, J. (2026). Can GenAI-empowered feedback promote L2 learners' self-regulation strategic behavior and writing performance? *System*, 138, 103970. <https://doi.org/10.1016/j.system.2026.103970>.
- Tran, T. T. T. (2025). Enhancing EFL writing revision practices: The impact of AI- and teacher-generated feedback and their sequences. *Education Sciences*, 15(2), 232. <https://doi.org/10.3390/educsci15020232>.
- Tsai, C. Y., Lin, Y. T., & Brown, I. K. (2024). Impacts of ChatGPT-assisted writing for EFL English majors: Feasibility and challenges. *Education and Information Technologies*, 29(17), 22427–22445. <https://doi.org/10.1007/s10639-024-12722-y>.
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1–48. <https://doi.org/10.18637/jss.v036.i03>.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
- Wan, X., Liu, Q., & Adebayo, S. A. (2026). Explore Parents' Views and Acceptance of Large Language Models Applied to Education. *Journal of Educational Technology and Innovation*, 8(2), 85–99. <https://doi.org/10.61414/wbgfyr12>.
- Wang, L., & Ren, B. (2024). Enhancing academic writing in a linguistics course with generative AI: An empirical study in a higher education institution in Hong Kong. *Education Sciences*, 14(12), 1329. <https://doi.org/10.3390/educsci14121329>.
- Wang, C., & Wang, Z. (2025). Investigating L2 writers' critical AI literacy in AI-assisted writing: An APSE model. *Journal of Second Language Writing*, 67, 101187. <https://doi.org/10.1016/j.jslw.2025.101187>.
- Warschauer, M., Tseng, W., Yim, S., Webster, T., Jacob, S., Du, Q., & Tate, T. (2023). The affordances and contradictions of AI-generated text for writers of English as a second or foreign language. *Journal of Second Language Writing*, 62, 101071. <https://doi.org/10.1016/j.jslw.2023.101071>.
- 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, 1249991. <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, 100228. <https://doi.org/10.1016/j.caeai.2024.100228>.
- Xu, J., & Zhang, S. (2022). Understanding AWE feedback and English writing of learners with different proficiency levels in an EFL classroom: A sociocultural perspective. *The Asia-Pacific Education Researcher*, 31(4), 357–367. <https://doi.org/10.1007/s40299-021-00577-7>.
- Yan, D. (2024). Comparing individual vs. collaborative processing of ChatGPT-generated feedback: Effects on L2 writing task improvement and learning. *Language Learning & Technology*, 28(1), 1–19. <https://doi.org/10.64152/10125/73597>.
- Yan, D. (2025). Feedback seeking abilities of L2 writers using ChatGPT: A mixed method multiple case study. *Kybernetes*, 54(7), 3757–3781. <https://doi.org/10.1108/k-09-2023-1933>.
- 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, 1086. <https://doi.org/10.1057/s41599-024-03543-y>.
- Yang, W., & Lin, C. (2025). Translanguaging with generative AI in EFL writing: Students' practices and perceptions. *Journal of Second Language Writing*, 67, 101181. <https://doi.org/10.1016/j.jslw.2025.101181>.
- Yeung, S. (2025). University students' engagement with generative AI-supported automated writing evaluation (AWE) feedback. *Journal of Second Language Writing*, 68, 101203. <https://doi.org/10.1016/j.jslw.2025.101203>.
- Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13–39). Academic Press. <https://doi.org/10.1016/b978-012109890-2/50031-7>.

- Zou, M., & Huang, L. (2024). The impact of ChatGPT on L2 writing and expected responses: Voice from doctoral students. *Education and Information Technologies*, 29, 13201–13219. <https://doi.org/10.1007/s10639-023-12397-x>.
- Zou, S., Guo, K., Wang, J., & Liu, Y. (2026). Investigating students' uptake of teacher- and ChatGPT-generated feedback in EFL writing: A comparison study. *Computer Assisted Language Learning*, 39(4), 1062–1091. <https://doi.org/10.1080/09588221.2024.2447279>.