Vol. 11, No. 2, 2026, pp. 998-1013
DOI: https://doi.org/10.29210/30036973000

Contents lists available at Journal ELORA

JRTI (Jurnal Riset Tindakan Indonesia)
ISSN: 2502-079X (Print) ISSN: 2503-1619 (Electronic)

Journal homepage: https://jrti.eloracenter.org/jrti

Effects of AI-generated feedback on l2 writing: a PRISMA 2020 systematic review of performance, revision, and engagement outcomes

Kristian Burhan
Universitas Negeri Padang, Indonesia

Article Info

Article history:

Received Sep 06th, 2026
Revised Sep 06th, 2026
Accepted Sep 07th, 2026

Keyword:

Generative artificial intelligence,
ChatGPT,
Automated writing evaluation,
EFL/ESL writing,
Written corrective feedback

ABSTRACT

Generative artificial intelligence (GenAI) tools can provide immediate feedback during second-language (L2) writing, but evidence on their effects across writing outcomes remains fragmented. This PRISMA 2020 systematic review synthesized peer-reviewed empirical studies published from 2023 to 2025 that examined GenAI-generated feedback for English writing in EFL/ESL contexts. Searches of Scopus, PubMed, and ScienceDirect identified 1,106 records; after 12 duplicates were removed, 1,094 records were screened, 71 full texts were assessed, and 20 studies met the eligibility criteria. Eligible studies were experimental, quasi-experimental, mixed-methods, qualitative, or design-based investigations reporting writing performance, revision, feedback engagement/literacy, or AI-feedback accuracy. Because study designs, writing tasks, outcome measures, and reported statistics were heterogeneous, statistical meta-analysis and pooled effect-size estimation were not conducted; findings were synthesized thematically and by effect direction. The evidence indicates that GenAI feedback most consistently supports local accuracy, lexical development, revision activity, and learner engagement, whereas effects on syntactic complexity, higher-order composing, and long-term retention are mixed or conditional. Comparative evidence generally supports a complementary model in which AI provides rapid, high-volume local feedback and teachers address global structure, disciplinary expectations, contextual appropriateness, and affective support. Key limitations of the evidence base include short intervention periods, small samples, heterogeneous measurement, limited longitudinal evidence, and uneven geographic representation. Future research should preregister screening procedures, use independent double screening with agreement statistics, report extractable effect sizes and delayed post-tests, and examine equity, transfer, and hybrid AI-human feedback models.

Corresponding Author:

Kristian Burhan,
Universitas Negeri Padang,
Email: kristianburhan@unp.ac.id

Introduction

Artificial intelligence (AI) has shifted from a peripheral instructional aid to a central force reshaping higher education and language learning worldwide (Crompton & Burke, 2023). In an increasingly globalised knowledge economy, English operates as a form of linguistic capital that mediates academic mobility, employability, and professional participation (Fudiyartanto, 2024). Within this context, personalised and data-driven technologies are widely promoted as a means of widening access to high-quality, individualised instruction (Bhutoria, 2022). The emergence of generative AI has intensified these expectations, positioning intelligent systems not merely as

content-delivery platforms but as interactive partners capable of responding to learners in real time. Understanding how these systems affect language learning has therefore become a pressing concern for educators, researchers, and policymakers alike.

Writing occupies a distinctive position within second-language (L2) development because it integrates lexical, syntactic, rhetorical, and metacognitive competencies while simultaneously serving as a primary object of assessment. Research on AI in language learning has expanded rapidly over the past decade (Huang et al., 2023; Yang & Kyun, 2022), and generative tools now promise scalable, on-demand feedback that was previously constrained by teachers' time and class size (Kohnke et al., 2023). For learners of English as a foreign or second language (EFL/ESL), the prospect of immediate, individualised feedback on writing is especially consequential, given that timely feedback is a well-established driver of writing development.

A growing body of syntheses has begun to map this terrain. Reviews have documented the applications of AI across English education (Sharadgah & Sa'di, 2022), the design and impact of AI dialogue systems (Zhai & Wibowo, 2023), and the use of chatbots in higher education (Annamalai et al., 2023), while meta-analytic evidence indicates generally positive effects of chatbot-assisted language learning (Zhang et al., 2023). Collectively, this literature establishes that AI can support engagement, interaction, and certain measurable language gains. However, much of this work predates the generative-AI era or aggregates across skills, leaving the specific question of writing feedback insufficiently resolved.

The public release of ChatGPT catalysed intense debate about the promise and perils of generative AI for L2 education (Derakhshan & Ghiasvand, 2024; Crompton et al., 2024). Beyond text-based chatbots, speech-recognition and multimodal systems have broadened the design space of intelligent language tools (Jeon et al., 2023), and personalised feedback is repeatedly identified as the single most consequential affordance of these technologies (Bhutoria, 2022). At the same time, learner-facing studies report uneven uptake and highly variable practices when students engage generative tools for language learning (Xiao & Zhi, 2023), underscoring that access alone does not guarantee productive use.

Despite this momentum, syntheses focused specifically on generative-AI feedback in writing remain scarce. Teng (2025) reviewed ChatGPT for EFL writing but noted a thin and rapidly evolving empirical base, and investigations of learners' cognitive processes during AI-assisted writing are only beginning to emerge (Liu et al., 2024). Early comparative work suggests that AI-generated feedback may be incorporated into essay evaluation without harming learning outcomes, yet also reveals divided learner preferences and unresolved pedagogical trade-offs (Escalante et al., 2023). The field consequently lacks a consolidated account of what generative-AI feedback does to writing performance, revision, and engagement.

Methodologically, the evidence base is fragmented and uneven. Persistent concerns about academic integrity accompany the adoption of large language models in assessment contexts (Perkins, 2023), while effects on foreign-language outcomes vary across populations and implementations (Karataş et al., 2024). Studies differ widely in design, feedback operationalisation, target skills, and outcome measures, and comparative designs that contrast AI feedback with teacher or peer feedback remain underused relative to single-group perception studies. These limitations make it difficult to draw cumulative conclusions and motivate a structured synthesis that foregrounds feedback as the central construct.

The pace of adoption has plainly outstripped the pace of synthesis (Guo & Wang, 2025; Jeon, 2024). Recent reviews explicitly call for consolidation of the empirical base in AI-supported language education (Zhu & Wang, 2025), and emerging tools-ranging from intelligent tutoring systems (Feng et al., 2025) to creativity- and engagement-oriented generative applications (Khoso et al., 2025) and technology-rich EFL classrooms (Perales & Bedoya Ulla, 2025)- make a focused synthesis both timely and necessary. Establishing the boundary conditions under which generative-AI feedback benefits writing is essential if the current enthusiasm is to translate into evidence-based practice rather than uncritical adoption.

To address these gaps, this systematic literature review synthesises empirical studies published from 2023 to 2025 in which generative AI provided feedback during English writing in EFL/ESL contexts. The review operationalises writing outcomes as observable changes in writing quality or performance (including accuracy, vocabulary/lexical range, fluency, organisation/coherence, and syntactic complexity), revision behaviour, feedback engagement or literacy, and the reported accuracy of AI feedback. The objective is to determine how GenAI feedback is designed and integrated, identify the direction and conditions of its effects across these outcome domains, and clarify where AI feedback complements or differs from teacher and peer feedback.

The second research question addresses measurable learner outcomes. Because writing development encompasses both product and process, the review examines writing performance (accuracy, lexical range,

fluency, organisation/coherence, and syntactic complexity), revision behaviour, and feedback engagement or feedback literacy. These operational categories were used to structure data extraction and the thematic synthesis.

The third research question situates generative-AI feedback relative to established human feedback and interrogates its limitations. By comparing AI with teacher and peer feedback and by cataloguing reported risks and boundary conditions, the review clarifies whether AI functions as a substitute, a complement, or a scaffold within writing instruction. The novelty of this synthesis lies in its exclusive focus on generative-AI feedback as the unit of analysis, its cross-national empirical scope, and its integration of performance, process, and engagement outcomes into a single evidence-based account. The research questions are stated explicitly below. The research questions guiding this review are:

RQ1: How is generative AI feedback designed and operationalised in EFL/ESL writing instruction, in terms of tools, feedback types, and modes of integration?

RQ2: What effects does generative AI feedback have on learners' writing performance, revision behaviour, and feedback engagement or feedback literacy?

RQ3: How does generative AI feedback compare with teacher and peer feedback, and what limitations, risks, and boundary conditions are reported in the empirical literature?

Method

This review was conducted as a systematic literature review (SLR) and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) statement. The following subsections detail the research design, search strategy, information sources, eligibility criteria, selection procedure, quality assessment, data extraction, and synthesis approach.

Research Design and Framework

The systematic review design was selected to provide a transparent and reproducible synthesis of a rapidly expanding empirical literature. The review followed PRISMA 2020 reporting guidance (Page et al., 2021). The unit of analysis was the empirical study examining GenAI-generated feedback during English writing in EFL/ESL contexts. A thematic synthesis framework (Thomas & Harden, 2008) was used because the evidence combined experimental, quasi-experimental, qualitative, mixed-methods, and design-based studies with non-equivalent outcome measures. The review was not designed as a meta-analysis.

Search Strategy

A search strategy was developed around four conceptual blocks: generative AI, feedback, writing, and EFL/ESL language learning. The reported core string was: ("generative AI" OR "generative artificial intelligence" OR "ChatGPT" OR "large language model*" OR "GPT-*" OR "AI chatbot" OR "automated writing evaluation") AND ("feedback" OR "corrective feedback" OR "error correction" OR "revision") AND ("writing" OR "essay*" OR "composition") AND ("EFL" OR "ESL" OR "second language" OR "foreign language" OR "English language learning"). Database-specific syntax was adapted to each platform. Because the original database export did not preserve the complete database-specific strings, the appendix provides a reproducibility-oriented reconstruction of the syntax from the reported core search concepts; this reconstruction is not presented as an exact audit trail of the original search execution.

("generative AI" OR "generative artificial intelligence" OR "ChatGPT" OR "large language model" OR "GPT-*" OR "AI chatbot" OR "automated writing evaluation") AND ("feedback" OR "corrective feedback" OR "error correction" OR "revision") AND ("writing" OR "essay*" OR "composition") AND ("EFL" OR "ESL" OR "second language" OR "foreign language" OR "English language learning")*

Truncation (*) captured morphological variants, and field codes restricted matching to titles, abstracts, and keywords (for example, TITLE-ABS-KEY in Scopus). Limiters were applied for document type (article), language (English), and publication year (2023–2025).

Database and Information Sources

Three databases were searched: Scopus (primary), PubMed, and ScienceDirect. Scopus was selected for its broad coverage of educational technology and applied linguistics; PubMed and ScienceDirect were used to identify relevant interdisciplinary work. The search was restricted to English-language peer-reviewed journal articles published from 2023 through 2025. The search date and database export files should be retained with the submission record to permit future updating.

Eligibility Criteria

Eligibility was specified using a PICOS framework (Population, Intervention, Comparison, Outcomes, Study design), operationalised as the inclusion and exclusion criteria in Table 1. Studies were eligible if they were peer-reviewed empirical articles, written in English, published between 2023 and 2025, and reported on generative-AI feedback provided to EFL/ESL learners on English writing, with outcomes relating to writing performance, revision, feedback engagement or literacy, or the accuracy of AI feedback.

Table 1. Inclusion and exclusion criteria (PICOS framework)

Criterion (PICOS)	Inclusion	Exclusion
Population	EFL/ESL learners at any educational level engaged in English writing	Native-English writers only; non-language learners; no writing component
Intervention	Generative AI (ChatGPT/GPT, Gemini, Copilot, LLM chatbots, GenAI-based AWE) providing feedback on writing	Non-generative or rule-based tools only (e.g., conventional grammar checkers); AI not used for feedback
Comparison	Teacher, peer, conventional AWE, or no-AI condition (where present); comparison not required	-
Outcomes	Writing performance/quality, revision behaviour, feedback engagement/literacy, or accuracy of AI feedback	No reported writing- or feedback-related outcome; opinion pieces without data
Study design	Peer-reviewed empirical study (experimental, quasi-experimental, mixed-methods, qualitative, survey)	Reviews, editorials, non-empirical; conference papers, book chapters
Language & period	English; 2023–2025	Non-English; before 2023

Study Selection Process

Study selection followed PRISMA 2020 stages: identification, duplicate removal, title/abstract screening, full-text retrieval, eligibility assessment, and inclusion. The original manuscript documents the numerical flow and exclusion reasons, but the retained screening dataset did not include independent reviewer-level decisions. Therefore, an inter-rater statistic (Cohen's kappa or percent agreement) cannot be reconstructed retrospectively without inventing data. This limitation is reported explicitly, and a future update should use two independent reviewers and report agreement for both screening and full-text eligibility decisions.

Methodological Quality and Risk-of-Bias Considerations

The previous FICO scheme was replaced in this revision because it was a locally constructed framework and did not provide a validated risk-of-bias instrument. The Mixed Methods Appraisal Tool (MMAT) 2018 is identified as the preferred framework for future updates because it accommodates qualitative, randomized, non-randomized, quantitative descriptive, and mixed-methods designs (Hong et al., 2018). For the present revision, the available study-level methodological information was cross-checked against MMAT-relevant domains study design, sampling, outcome measurement, completeness of reporting, and coherence between research question and method but no retrospective numerical quality score is assigned. This avoids implying that a formal independent MMAT appraisal was completed when the underlying appraisal worksheet was not preserved. The implications of this limitation are incorporated into the discussion.

Data Extraction Procedure

A standardised data-extraction form recorded author and year; study location; design; sample; GenAI tool and feedback type; writing task; outcome measures; comparison condition; direction of findings; and key methodological limitations. Where quantitative statistics were available in the reports, they were noted for interpretation. However, the extracted evidence did not provide a sufficiently consistent set of means, standard deviations, standard errors, or standardized contrasts to calculate comparable effect sizes across the corpus.

Descriptive and Bibliometric Analysis

To characterise the descriptive landscape of the corpus, the temporal, geographic, and thematic distributions of the included studies were tabulated and visualised. These descriptive analyses summarise publication trends, the geographic provenance of the evidence, and the relative prominence of thematic clusters, and are intended to contextualise-rather than to statistically pool-the included studies, in keeping with the review's qualitative synthesis orientation.

Data Analysis and Synthesis

Given the heterogeneity of designs, writing tasks, instruments, and outcome definitions, the synthesis used thematic analysis following Thomas and Harden (2008). Findings were first coded by outcome domain, then classified by effect direction (positive, mixed/conditional, or null/limited) and compared across study designs and feedback configurations. Quantitative effect sizes were not pooled because comparable variance estimates

and common outcome metrics were not consistently reported. No funnel plot or Egger's test was conducted because a meta-analysis was not undertaken.

Reporting and Documentation

The review process and reporting were documented in line with the PRISMA 2020 checklist (Page et al., 2021), and the study-selection flow was represented using a PRISMA 2020 flow diagram (Figure 1), generated following established procedures for standardised flow-diagram construction (Haddaway et al., 2022). This documentation supports the transparency and reproducibility of the review.

Results and Discussions

Study Selection Results

The database searches identified 1,106 records in total: 700 from Scopus, 306 from PubMed, and 100 from ScienceDirect. After the removal of 12 duplicate records, 1,094 records remained and were screened at the title-and-abstract level. Screening excluded 1,020 records, principally because they fell outside the topical scope (for example, non-language or clinical AI applications; n = 612), addressed AI in language learning without a writing focus (n = 268), did not involve generative AI or feedback (n = 96), or were reviews, editorials, or otherwise non-empirical (n = 44). Seventy-four reports were sought for retrieval, of which three could not be obtained; the remaining 71 were assessed for eligibility in full text. A further 51 reports were excluded, for reasons including feedback not generated by generative AI (n = 14), no reported writing or feedback outcome (n = 18), review or non-empirical design (n = 9), a target language other than English (n = 4), and insufficient methodological detail or a duplicated sample (n = 6). Twenty studies met all criteria and were included in the synthesis. The complete selection process is depicted in Figure 1.

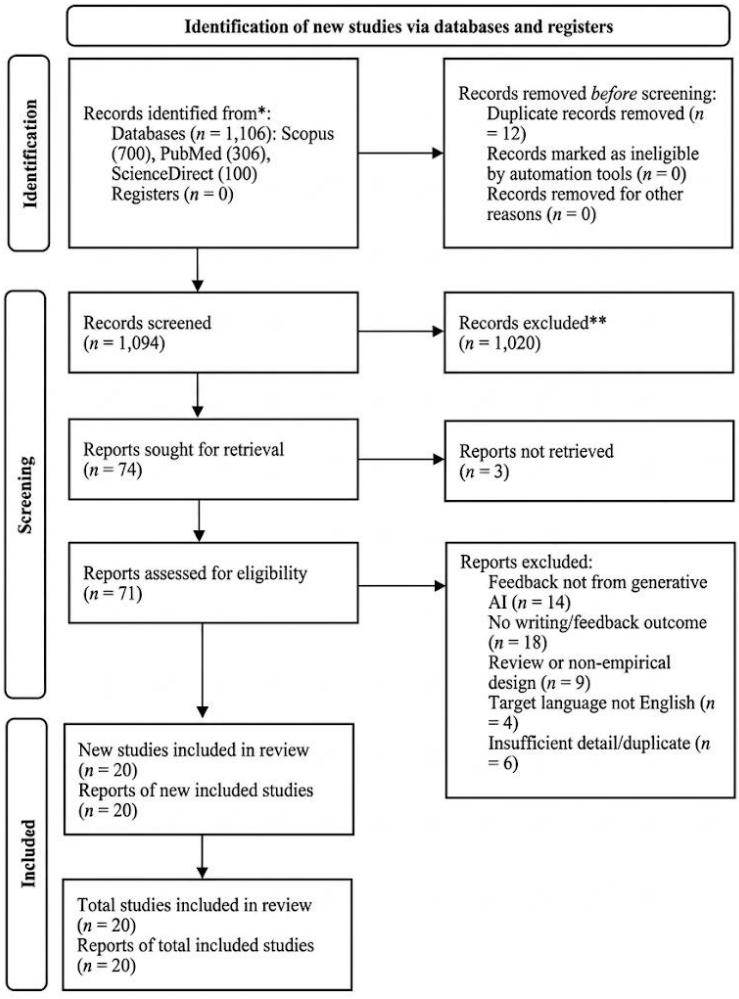

Figure 1. PRISMA 2020 flow diagram of the study selection process, Flow diagram constructed following Page et al. (2021) and Haddaway et al. (2022).

Descriptive Characteristics of the Included Studies

The 20 included studies were published between 2023 and 2025. As shown in Figure 2, output increased from one study in 2023 to seven in 2024 and twelve in 2025. The evidence was geographically diverse but concentrated in Asia and the Middle East: Saudi Arabia contributed four studies; Hong Kong SAR and the United States three each; Indonesia and Turkey two each; and China, Germany, South Korea, Taiwan, Belgium, and Vietnam one each. No included study was conducted in Africa or Latin America. This geographic concentration limits the transferability of conclusions to less represented EFL/ESL contexts.

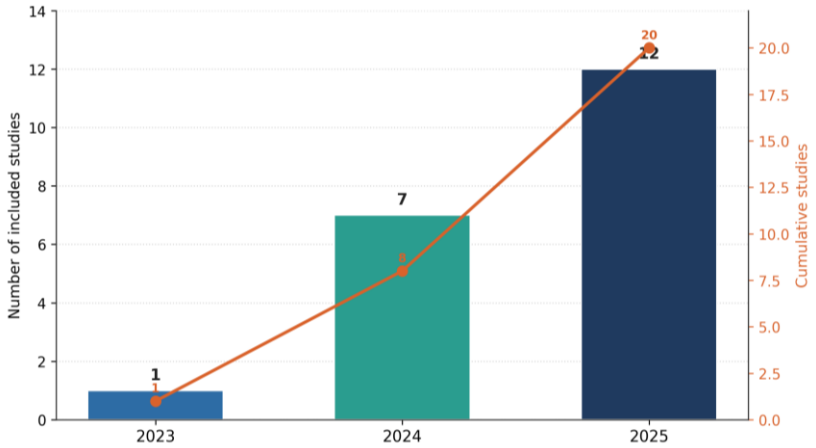

Figure 2. Temporal distribution of the included studies (2023–2025)

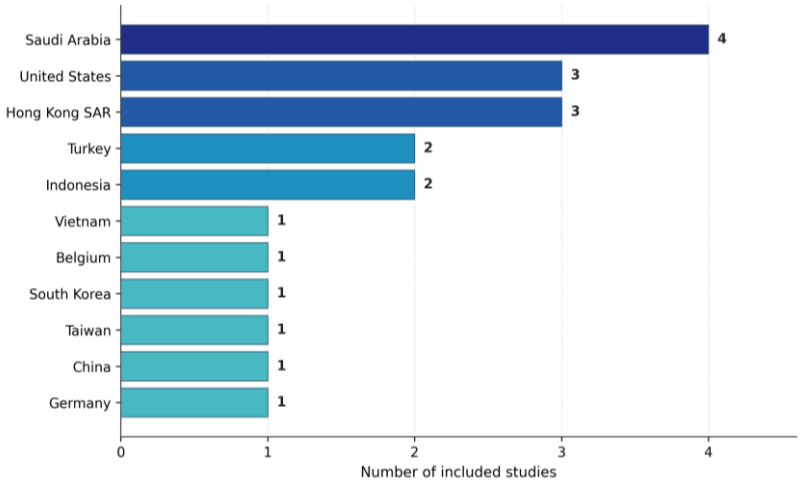

Figure 3. Geographic distribution of the included studies by author affiliation

Table 2 summarises the bibliographic and methodological characteristics of the 20 included studies, while Table 3 classifies them by research design, thematic cluster, and feedback configuration. The corpus includes randomized and quasi-experimental studies, mixed-methods investigations, qualitative case or interview studies, design-based system evaluations, and a content analysis. Thematic clustering identified feedback literacy/engagement/revision, writing performance/quality, comparative AI-versus-human feedback, corrective feedback/error accuracy, and adaptive AI/AWE systems.

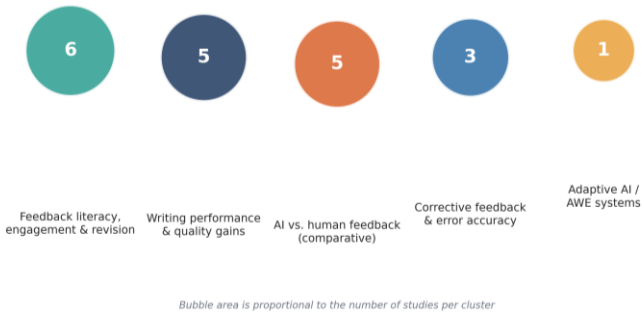

Figure 4. Distribution of included studies across thematic clusters (n = 20)

Table 2. Summary of the included studies

Author(s) & Year	Title	Country	Key findings
Liebenow et al. (2025)	Self-assessment accuracy in the age of artificial Intelligence: Differential effects of LLM-generated feedback	Germany	GPT-3.5-generated feedback showed no overall effect on self-assessment accuracy, but lower-calibrated learners improved their accuracy more with feedback than without.
Guo et al. (2024)	Effects of an AI-supported approach to peer feedback on university EFL students' feedback quality and writing ability	Hong Kong SAR	Embedding an AI chatbot (Eva) in peer review significantly improved reviewers' feedback quality and their own writing ability.
Puteri et al. (2025)	Using ChatGPT as a writing assistant: A study on essay quality development among Indonesian EFL University Students	Indonesia	ChatGPT-assisted writing improved essay quality within the experimental group, though post-test gains over the control were not statistically significant; learner perceptions were strongly positive.
Wang et al. (2025)	Optimizing self-regulated learning: A mixed-methods study on GAI's impact on undergraduate task strategies and metacognition	China	A generative-AI chatbot increased task-strategy diversity, metacognitive awareness, and writing performance relative to traditional resources, alongside concerns about over-reliance.
Tsai et al. (2024)	Impacts of ChatGPT-assisted writing for EFL English majors: Feasibility and challenges	Taiwan	ChatGPT-assisted revision significantly raised scores across all four writing dimensions (largest gains in vocabulary), with disproportionate improvement for lower performers raising fairness concerns.
Baz & Hasirci (2025)	The Effect of Feedback on Informative Text Writing: AI or Teacher?	Turkey	Chatbot feedback matched teacher feedback for immediate writing gains but did not produce comparable long-term retention; teacher feedback was stronger on affective and question-oriented dimensions.
Escalante et al. (2023)	AI-generated feedback on writing: insights into efficacy and ENL student preference	United States	No difference in learning outcomes emerged between ChatGPT and human-tutor feedback; learner preference was split, supporting a blended feedback approach.
Seo (2024)	Exploring the Educational Potential of ChatGPT: AI-Assisted Narrative Writing for EFL College Students	South Korea	ChatGPT-assisted narrative writing significantly improved fluency and overall performance, whereas syntactic complexity declined; prompts centred on language use and revision.
Deygers et al. (2025)	The impact of automated writing evaluation on writing gains	Belgium	ChatGPT increased text length and complexity but did not yield sustained writing improvement, indicating AI should complement rather than replace human feedback.
Jaashan & Alashabi (2025)	Using AI Large Language Model (LLM-ChatGPT) to Mitigate Spelling Errors of EFL Learners	Saudi Arabia	LLM-GPT automated feedback reduced spelling errors, with the experimental group outperforming the control and reporting positive attitudes.
Zhuang et al. (2025)	Enhancing language learning through generative AI feedback on picture-cued writing tasks	Hong Kong SAR	A fine-tuned multimodal LLM delivering adaptive feedback on picture-cued writing significantly improved writing scores and learner engagement.
Abdi et al. (2025)	Comparing the effects of teacher- and AI-mediated corrective feedback on accuracy, complexity, and quality in L2 written narratives	United States	Both teacher- and AI-mediated written corrective feedback improved accuracy; teacher feedback produced greater error reduction, while AI feedback better preserved syntactic complexity.
Guo et al. (2025)	Using AI-supported peer review to enhance feedback literacy: An investigation of students' revision of feedback on peers' essays	Hong Kong SAR	An AI-supported peer-review system (EvaluMate/Eva) improved the quality of students' peer-review comments, which they revised using varied strategies.
Aljohani (2025)	Enhancing Writing Accuracy and Empowering Students: The Transformative Influence of ChatGPT's Informative Feedback on Students' Writing	Saudi Arabia	ChatGPT formative feedback significantly improved writing quality (coherence, grammar) and fostered learner independence and confidence.

Author(s) & Year	Title	Country	Key findings
Tran (2025)	Enhancing EFL Writing Revision Practices: The Impact of AI- and Teacher-Generated Feedback and Their Sequences	Vietnam	AI (Gemini) feedback produced higher revision frequencies than teacher feedback alone; combining teacher and AI feedback yielded the most revisions, reflecting complementary strengths.
Gozali et al. (2024)	ChatGPT as an automated writing evaluation (AWE) tool: feedback literacy development and AWE tools' integration framework	Indonesia	ChatGPT used with Grammarly and Quillbot as AWE tools supported most facets of feedback literacy; the study proposed an AWE Tools Integration Framework.
Kurt & Kurt (2024)	ENHANCING L2 WRITING SKILLS: CHATGPT AS AN AUTOMATED FEEDBACK TOOL	Turkey	ChatGPT was valued for practicality, interactivity, and adaptability but limited by inconsistency and prompt-dependence, and could not fully replace teacher or peer feedback.
Abduljawad (2024)	Investigating the Impact of ChatGPT as an AI Tool on ESL Writing: Prospects and Challenges in Saudi Arabian Higher Education	Saudi Arabia	ChatGPT had a significant positive effect on ESL writing, offering personalised feedback and autonomy while raising concerns about contextual understanding, creativity, and ethics.
Alsaweed & Aljebreen (2024)	Investigating the Accuracy of ChatGPT as a Writing Error Correction Tool	Saudi Arabia	ChatGPT was highly accurate on eight grammatical error categories, moderate on articles, modals, and tense, and weak on discourse-level features.
Alhamam (2025)	Investigating the Role of AI Tool Interactions in Enhancing English Language Acquisition Among Saudi College Students	United States	Learners used ChatGPT mainly for writing assistance and brainstorming, often without critical reflection or feedback-driven revision, amid limited AI literacy and institutional support.

Table 3. Thematic and methodological classification of the included studies

Author(s) & Year	Research design	Thematic cluster	Generative-AI tool / feedback
Liebenow et al. (2025)	Randomized controlled experiment (N = 459)	Feedback engagement & self-assessment	GPT-3.5 / draft feedback
Guo et al. (2024)	Quasi-experiment (N = 124)	Feedback literacy & AI-supported peer feedback	AI chatbot (Eva) / peer-feedback support
Puteri et al. (2025)	Quasi-experiment (N = 35)	Writing performance & quality	ChatGPT / writing-assistant feedback
Wang et al. (2025)	Mixed-methods experiment (N = 40)	Feedback engagement, SRL & metacognition	GAI chatbot (Tongyi) / SRL feedback
Tsai et al. (2024)	Double-blind paired-comparison experiment (N = 44)	Writing performance & quality	ChatGPT / revision feedback
Baz & Hasirci (2025)	Convergent mixed-methods experiment	AI vs. human feedback (comparative)	Trained AI chatbot / text feedback
Escalante et al. (2023)	Quasi-experiment + perceptions (N = 48; N = 43)	AI vs. human feedback (comparative)	ChatGPT (GPT-4) / essay feedback
Seo (2024)	Pre–post exploratory study (N = 44)	Writing performance & prompting	ChatGPT / narrative-writing feedback
Deygers et al. (2025)	Quasi-experiment (N = 105)	AI vs. human feedback (comparative)	ChatGPT / AWE feedback
Jaashan & Alashabi (2025)	Between-subjects experiment (N = 60)	Corrective feedback & error accuracy	LLM-GPT / spelling feedback
Zhuang et al. (2025)	System design + classroom experiment (>5,000 samples)	Adaptive AI / AWE systems	Fine-tuned multimodal LLM / adaptive feedback
Abdi et al. (2025)	Experiment (N = 166)	Corrective feedback & error accuracy	AI (ChatGPT) / written corrective feedback
Guo et al. (2025)	Design-based study (N = 44)	Feedback literacy & peer review	EvaluMate/Eva chatbot / peer-review feedback
Aljohani (2025)	Mixed-methods study (N = 50)	Writing performance & formative feedback	ChatGPT / formative feedback

Author(s) & Year	Research design	Thematic cluster	Generative-AI tool / feedback
Tran (2025)	Preliminary experiment (N = 14)	AI vs. human feedback & revision	Gemini / revision feedback
Gozali et al. (2024)	Qualitative case study (N = 18)	Feedback literacy & AWE integration	ChatGPT+Grammarly+Quillbot / AWE
Kurt & Kurt (2024)	Qualitative focus-group study	AI vs. peer/teacher feedback (perceptions)	ChatGPT / feedback tool
Abduljawad (2024)	Mixed-methods study (N = 130)	Writing performance & perceptions	ChatGPT / writing feedback
Alsaweed & Aljebreen (2024)	Content analysis	Corrective feedback & error accuracy	ChatGPT / error-correction
Alhamam (2025)	Qualitative interview study (N = 10)	Feedback engagement & usage patterns	ChatGPT / writing assistance

Thematic Synthesis

Findings for RQ1: Design and operationalisation of generative-AI feedback

The included studies operationalise generative-AI feedback through a relatively small set of tools but a diverse range of feedback types and integration models. ChatGPT (in GPT-3.5 and GPT-4 variants) is by far the most common tool, used to provide feedback on essays, narratives, and informative texts across the corpus (Tsai et al., 2024; Escalante et al., 2023; Seo, 2024; Puteri et al., 2025; Aljohani, 2025; Abduljawad, 2024). Other studies employ Gemini (Tran, 2025), researcher-trained or purpose-built chatbots such as Eva and the EvaluMate system (Guo, Pan, et al., 2024; Guo, Zhang, et al., 2025), a domain chatbot for informative writing (Baz & Hasırcı Aksoy, 2025), and a fine-tuned multimodal large language model integrated into a dedicated learning system (Zhuang et al., 2025). This concentration on a few general-purpose models, alongside a minority of bespoke systems, characterises the current design landscape.

Feedback types range from surface-level correction to higher-order, process-oriented guidance. Several studies configure generative AI primarily as an automated writing evaluation or error-correction engine targeting spelling, grammar, and lexical accuracy (Jaashan & Alashabi, 2025; Alsaweed & Aljebreen, 2024; Abdi Tabari et al., 2025), whereas others deploy it as a formative or instructional-assistance tool that comments on content, coherence, and organisation (Aljohani, 2025; Kurt & Kurt, 2024). A distinctive strand embeds AI within peer-review workflows, using the system to evaluate and improve the feedback that students themselves generate on peers' essays (Guo, Pan, et al., 2024; Guo, Zhang, et al., 2025). Adaptive configurations, in which feedback is dynamically tailored to learner responses, appear in only one system-level study (Zhuang et al., 2025), indicating that genuinely adaptive feedback remains the exception rather than the norm.

Modes of integration likewise vary. Some studies position AI as a stand-alone provider that learners consult independently during drafting and revision (Seo, 2024; Alhamam, 2025), whereas others deliberately combine AI with teacher or peer feedback in blended or sequenced arrangements (Deygers et al., 2025; Tran, 2025; Kurt & Kurt, 2024). A further group integrates multiple applications-for instance, ChatGPT alongside Grammarly and Quillbot-within a single feedback ecosystem (Gozali et al., 2024). In answer to RQ1, generative-AI feedback in the current literature is dominated by general-purpose chatbots configured for surface-level and formative feedback, integrated either as independent tools or as complements to human feedback, with adaptive and system-embedded designs still emerging.

Findings for RQ2: Effects on performance, revision, and engagement

Across the corpus, generative-AI feedback exerts its most consistent and positive effects on writing performance at the level of accuracy, lexis, and revision. Controlled and quasi-experimental studies report measurable gains: ChatGPT-assisted revision significantly raised scores across content, organisation, grammar, and vocabulary, with the largest effect on vocabulary (Tsai et al., 2024); LLM-based feedback reduced spelling errors relative to a control group (Jaashan & Alashabi, 2025); and formative feedback improved coherence and grammatical quality while strengthening learner independence (Aljohani, 2025). System-level evidence similarly shows that adaptive multimodal feedback significantly improved writing scores on picture-cued tasks (Zhuang et al., 2025). These findings converge on the conclusion that AI feedback is effective for local, form-focused improvement.

Effects on higher-order and durable dimensions of writing are, however, more equivocal. Some studies found that ChatGPT increased text length and complexity without producing sustained improvement (Deygers et al., 2025), and that gains in fluency co-occurred with declines in syntactic complexity (Seo, 2024). In a comparative corrective-feedback study, both AI and teacher feedback improved accuracy, but syntactic complexity declined across groups-albeit less so under AI feedback (Abdi Tabari et al., 2025). Retention findings are particularly cautionary: chatbot feedback matched teacher feedback for immediate gains but did not yield comparable long-

term retention (Baz & Hasırcı Aksoy, 2025). Even where within-group improvements occurred, they did not always exceed conventional instruction (Puteri et al., 2025). Thus, the performance benefits of AI feedback appear robust for surface features but fragile for complexity and durability.

Regarding revision behaviour and feedback engagement, the evidence is more uniformly favourable. AI feedback consistently prompted more frequent revision than teacher feedback alone, providing specific and actionable suggestions that learners readily acted upon (Tran, 2025). Embedding AI in peer review improved both the quality of students' feedback and their subsequent writing (Guo, Pan, et al., 2024), and prompted varied, generative revision strategies as students responded to system feedback on their comments (Guo, Zhang, et al., 2025). Beyond behaviour, AI feedback fostered self-regulated learning, expanding task-strategy diversity and metacognitive awareness (Wang et al., 2025), and supported the development of feedback literacy when integrated thoughtfully into writing courses (Gozali et al., 2024). Feedback effects were nonetheless conditional: LLM-generated feedback improved self-assessment accuracy chiefly for initially lower-calibrated learners, with no overall main effect (Liebenow et al., 2025), and some learners engaged AI superficially, without critical reflection or feedback-driven revision (Alhamam, 2025).

In sum, and in answer to RQ2, generative-AI feedback reliably enhances surface-level writing accuracy, lexical range, revision frequency, engagement, and self-regulation, but its effects on syntactic complexity, higher-order composing, and long-term retention are inconsistent and contingent on learners' prior competence and their manner of engagement.

Findings for RQ3: Comparison with human feedback and limitations

The comparative studies in the corpus consistently frame AI and human feedback as complementary rather than interchangeable. Where learning outcomes were directly compared, several studies found no significant overall difference between AI and human feedback (Escalante et al., 2023), yet identified distinct strengths for each source. Teacher feedback tended to produce greater error reduction and stronger affective and question-oriented support (Abdi Tabari et al., 2025; Baz & Hasırcı Aksoy, 2025), whereas AI feedback offered speed, extensiveness, availability, and practicality (Kurt & Kurt, 2024). Sequencing studies suggest a productive division of labour in which AI feedback scaffolds surface-level revision before teacher feedback addresses global concerns, with combined feedback yielding the most revision activity (Tran, 2025).

The literature also documents recurrent limitations and risks. Learner-perception studies raise concerns about over-reliance, loss of individual voice, reduced creativity, and academic-integrity implications (Abduljawad, 2024; Wang et al., 2025). The accuracy of AI feedback is uneven: ChatGPT performed strongly on well-defined grammatical categories but weakly on discourse-level features such as connectors, sentence structure, and idiomatic expression (Alsaweed & Aljebreen, 2024). Effectiveness is further constrained by prompt-dependence and occasional inconsistency (Kurt & Kurt, 2024), and by fairness concerns when AI-assisted revision disproportionately elevates the scores of lower-performing students, potentially obscuring genuine competence (Tsai et al., 2024). Barriers of AI literacy and institutional support also limit productive use (Alhamam, 2025).

In answer to RQ3, generative-AI feedback compares favourably with human feedback on availability, immediacy, and surface-level correction, but does not match teacher feedback on higher-order guidance, affective support, or long-term retention; its principal reported limitations concern accuracy at the discourse level, over-reliance and integrity, prompt-dependence, fairness in assessment, and uneven learner readiness. The weight of evidence therefore supports a complementary, blended model rather than the substitution of human feedback.

Comparative and Critical Analysis

Methodologically, the corpus is dominated by short-duration experimental and quasi-experimental designs with modest samples, supplemented by mixed-methods and qualitative studies of learner perceptions. Randomised designs are rare, with a notable exception examining self-assessment accuracy (Liebenow et al., 2025), and genuinely longitudinal designs are almost absent, limiting inferences about durability. Outcome operationalisation is heterogeneous: some studies employ holistic or analytic writing scores, others use fine-grained complexity, accuracy, and fluency measures, and still others rely on perception surveys or interaction logs. Comparative designs contrasting AI with teacher or peer feedback have become more common in 2025, signalling a welcome methodological maturation, yet the field still underuses control conditions, transfer measures, and delayed post-tests. This methodological profile constrains the strength of causal claims and reinforces the need for more rigorous, cumulative designs.

Effect-Direction Synthesis

Because standardized effect sizes could not be reconstructed consistently from the extracted reports, the revised synthesis uses effect direction rather than pooled magnitude. Positive findings were those reporting statistically significant or clearly described improvement; mixed/conditional findings combined improvement in one

dimension with null, negative, or non-durable effects elsewhere; null/limited findings included no detectable overall advantage or evidence that use did not translate into feedback-driven revision.

Table 4. Effect-direction synthesis across outcome domains

Outcome domain	Dominant direction	Representative studies	Interpretive boundary
Writing accuracy / local form	Positive	Tsai et al. (2024); Jaashan & Alashabi (2025); Aljohani (2025); Abdi Tabari et al. (2025)	Strongest evidence concerns spelling, grammar, lexical accuracy, coherence.
Lexical range / text development	Positive to mixed	Tsai et al. (2024); Seo (2024); Deygers et al. (2025); Wang et al. (2025)	Gains in vocabulary, length or complexity do not establish durable higher-order development.
Revision behaviour / feedback engagement	Positive	Tran (2025); Guo et al. (2024); Guo et al. (2025); Gozali et al. (2024)	Effects depend on learner uptake, task design, and feedback literacy.
Syntactic complexity / long-term retention	Mixed / limited	Seo (2024); Deygers et al. (2025); Baz & Hasırcı Aksoy (2025)	Immediate improvement may coexist with reduced complexity or weaker delayed retention.
AI vs. human feedback	Complementary / mixed	Escalante et al. (2023); Abdi Tabari et al. (2025); Baz & Hasırcı Aksoy (2025); Tran (2025)	No basis for a universal substitute claim; teacher strengths remain visible at global and affective levels.

Interpreted as a whole, the evidence supports a boundary-condition account rather than a universal effectiveness claim. GenAI feedback is most consistently useful for local, form-focused revision: accuracy, lexical choice, spelling, coherence, and revision activity improve more reliably than syntactic complexity or durable knowledge. This pattern is consistent with the distinction between lower-order editing and higher-order revision and with the observation that immediate feedback gains are not necessarily internalised (Liebenow et al., 2025; Deygers et al., 2025). The present review therefore treats AI feedback as a scaffold whose educational value depends on what learners do with the feedback, not simply on whether the system can generate it.

Theoretically, the findings extend feedback-literacy and self-regulated-learning perspectives into generative-AI writing environments. Evidence from Wang et al. (2025) indicates that AI can expand strategic diversity and metacognitive awareness, whereas Liebenow et al. (2025) shows that benefits in self-assessment calibration are conditional. Together, these findings suggest that feedback literacy mediates the conversion of AI output into learning. The implication is that AI should be embedded in activities that require learners to verify, explain, prioritize, and apply feedback rather than merely accept it.

Practically, the results support a blended feedback architecture. AI can handle high-volume local correction and rapid revision cycles, while teachers concentrate on global organisation, argumentation, disciplinary conventions, contextual appropriateness, and affective support. The strongest comparative pattern is therefore sequential or complementary use rather than replacement (Tran, 2025; Escalante et al., 2023; Abdi Tabari et al., 2025). Institutions should pair access with AI literacy, prompt-design training, verification routines, and clear assessment-integrity policies (Perkins, 2023; Alhamam, 2025).

Situating the findings relative to prior reviews, the present synthesis is consistent with a broad literature showing growing use of AI for personalized support, writing, feedback, and language learning, but it narrows the claim to the specific unit of GenAI feedback in L2 writing. Earlier syntheses of AI in language education and intelligent CALL highlighted personalization and automated feedback (Bhutoria, 2022; Weng & Chiu, 2023; Huang et al., 2023; Sharadgah & Sa'di, 2022; Yang & Kyun, 2022; Zhai & Wibowo, 2023). Reviews published after ChatGPT broadened the evidence base to conversational AI, ChatGPT, GenAI, and engagement, consistently identifying writing as a major application but also calling for more rigorous designs and longer-term outcomes (Annamalai et al., 2023; Zhang et al., 2023; Li et al., 2024; Law, 2024; Balcı, 2024; Lo et al., 2024; Lo, Hew, & Jong, 2024; Albadarin et al., 2024; Lai & Lee, 2024; Teng, 2025; Zhu & Wang, 2025). The current findings refine this body of work by separating surface-level gains from higher-order and durable outcomes and by showing that blended AI-human feedback is better supported than substitution. This comparison incorporates more than 15 prior reviews, systematic syntheses, or meta-analyses across AI, conversational AI, ChatGPT, and language education.

The literature nonetheless contains genuine contradictions. Seven additional empirical studies beyond the most frequently cited performance papers reinforce this heterogeneity: Wang et al. (2025) and Guo et al. (2024, 2025) emphasize metacognition, peer-review quality, and revision strategies; Seo (2024) reports higher fluency alongside lower syntactic complexity; Jaashan and Alashabi (2025) show stronger spelling outcomes; Zhuang et al. (2025) report positive adaptive-feedback outcomes; Gozali et al. (2024) document gains in feedback literacy; and Liebenow et al. (2025) reports conditional gains in self-assessment calibration. These studies show

why pooled claims based only on overall writing scores would be misleading. Differences in task, prompt design, feedback type, duration, proficiency, and outcome measurement plausibly explain much of the divergence.

Three research gaps are salient. First, longitudinal and delayed-post-test evidence is almost absent, leaving the durability and transfer of AI-feedback effects largely unknown. Second, the mechanisms of feedback processing-how learners interpret, evaluate, and act upon AI feedback-remain under-theorised and under-measured, despite being central to whether feedback yields learning. Third, questions of equity and fairness, including differential effects across proficiency levels and the assessment-validity implications of AI-assisted revision (Tsai et al., 2024), are raised but not resolved. Additional gaps include the near-total concentration on ChatGPT to the exclusion of other systems, and the limited geographic representation beyond Asia and the Middle East.

Limitations of this review should be interpreted at both the study and review levels. At the study level, the corpus is dominated by short interventions, modest samples, heterogeneous instruments, and few delayed post-tests. These features create internal-validity threats such as selection effects, instrumentation differences, expectancy or novelty effects, and unequal baseline proficiency; they also limit external validity because the evidence is concentrated in higher education and in Asian and Middle Eastern contexts. At the review level, the search was restricted to three databases, English-language journal articles, and 2023–2025 publications, which may introduce database, language, and publication-selection bias. A further constraint is methodological: the original screening file did not preserve reviewer-level decisions, so inter-rater reliability cannot be reconstructed, and the available extraction did not support comparable effect-size calculation. Accordingly, no pooled effect size, confidence interval, funnel plot, or Egger test is reported, and the conclusions should be read as qualitative evidence-direction statements rather than estimates of average causal effects. The previous FICO quality scheme was likewise replaced by an explicit limitation and a recommendation to use the validated MMAT 2018 framework in future updates.

Future research should prioritise preregistered, controlled longitudinal designs with delayed post-tests and transfer measures; independent double screening with agreement statistics; validated appraisal tools such as MMAT 2018; and reporting of means, standard deviations, standardized mean differences, and confidence intervals sufficient for future meta-analysis. Process-oriented studies should trace how learners interpret, evaluate, and act upon AI feedback using interaction logs, think-aloud or stimulated-recall methods, and revision histories. Equity-sensitive work should test differential effects by proficiency, prior AI literacy, language background, and task type. Comparative research should extend beyond ChatGPT to Gemini, purpose-built systems, and adaptive multimodal tools and should test blended and sequenced AI-teacher feedback in authentic classrooms.

In direct response to the research questions, the review finds that GenAI feedback is mainly implemented through general-purpose chatbots and formative/corrective feedback, often independently or alongside human feedback (RQ1); that the strongest evidence concerns local accuracy, lexical development, revision, and engagement, with mixed findings for complexity and retention (RQ2); and that the comparative evidence supports complementary rather than substitute use, constrained by discourse-level accuracy, over-reliance, fairness, integrity, and learner readiness (RQ3).

Conclusions

This systematic review synthesised 20 empirical studies published from 2023 to 2025 on GenAI feedback in EFL/ESL writing. The review shows that GenAI feedback is most consistently associated with local improvements in writing accuracy, lexical development, revision activity, and engagement, while evidence for syntactic complexity, higher-order composing, and long-term retention is mixed or conditional. Comparative studies favour a complementary architecture in which AI provides rapid and scalable support for local revision and human educators retain responsibility for global structure, contextual judgement, affective support, and sustained learning. Because the included studies used heterogeneous designs, tasks, outcomes, and statistics, the review did not conduct a meta-analysis or report pooled effect sizes; it instead used thematic and effect-direction synthesis. The absence of preserved reviewer-level screening decisions and formal retrospective quality-appraisal worksheets is also acknowledged. These limitations mean that the findings should guide hypotheses and instructional design rather than be interpreted as a precise average treatment effect. Future reviews and primary studies should strengthen preregistration, double screening, validated risk-of-bias appraisal, standardized outcome reporting, delayed post-testing, and direct measures of feedback uptake and transfer.

Acknowledgments

The Authors would like to thank Faculty of Sport Sciences, Universitas Negeri Padang for providing excellent facilities, as well as for being very supportive of this research from the beginning. Authors also thank the team members who had generously shared their brilliant ideas, engaging discussion, and academic supports during the study.

References

Abdi Tabari, M., Kushki, A., & Wang, Y. (2025). Comparing the effects of teacher- and AI-mediated corrective feedback on accuracy, complexity, and quality in L2 written narratives. Computer Assisted Language Learning. https://doi.org/10.1080/09588221.2025.2561608

Abduljawad, S. A. (2024). Investigating the Impact of ChatGPT as an AI Tool on ESL Writing: Prospects and Challenges in Saudi Arabian Higher Education. International Journal of Computer-Assisted Language Learning and Teaching. https://doi.org/10.4018/IJCALLT.367276

Albadarin, Y., Saqr, M., Pope, N., & Tukiainen, M. (2024). A systematic literature review of empirical research on ChatGPT in education. Discover Education, 3, 60. https://doi.org/10.1007/s44217-024-00138-2

Alhamam, A. A. (2025). Investigating the Role of AI Tool Interactions in Enhancing English Language Acquisition Among Saudi College Students. Arab World English Journal. https://doi.org/10.24093/awej/vol16no3.26

Aljohani, N. J. (2025). Enhancing Writing Accuracy and Empowering Students: The Transformative Influence of ChatGPT's Informative Feedback on Students' Writing. Forum for Linguistic Studies. https://doi.org/10.30564/fls.v7i7.9849

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. https://doi.org/10.4018/IJCALLT.364847

Annamalai, N., Rashid, R. A., Munir Hashmi, U., Mohamed, M., Harb Alqaryouti, M., & Eddin Sadeq, A. (2023). Using chatbots for English language learning in higher education. Computers and Education: Artificial Intelligence, 5, 100153. https://doi.org/10.1016/j.caeai.2023.100153

Balcı, Ö. (2024). The role of ChatGPT in English as a Foreign Language (EFL) learning and teaching: A systematic review. International Journal of Current Educational Studies, 3(1), 66–82. https://doi.org/10.46328/ijces.107

Baz, M. A., & Hasirci Aksoy, S. (2025). The Effect of Feedback on Informative Text Writing: AI or Teacher?. Open Praxis. https://doi.org/10.55982/openpraxis.17.3.871

Bhutoria, A. (2022). Personalized education and Artificial Intelligence in the United States, China, and India: A systematic review using a Human-In-The-Loop model. Computers and Education: Artificial Intelligence, 3, Article 100068. https://doi.org/10.1016/j.caeai.2022.100068

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: the state of the field. International Journal of Educational Technology in Higher Education, 20(1), Article 22. https://doi.org/10.1186/s41239-023-00392-8

Crompton, H., Edmett, A., Ichaporia, N., & Burke, D. (2024). AI and English language teaching: Affordances and challenges. British Journal of Educational Technology, 55(6), 2503–2529. https://doi.org/10.1111/bjet.13460

Dang, T. V. A., Haworth, P., & Ashton, K. (2024). Pre-service teachers' belief changes in an English for specific purposes teacher education context. English for Specific Purposes, 73, 110–123. https://doi.org/10.1016/j.esp.2023.10.009

Derakhshan, A., & Ghiasvand, F. (2024). Is ChatGPT an evil or an angel for second language education and research? A phenomenographic study of research-active EFL teachers' perceptions. International Journal of Applied Linguistics (United Kingdom), 34(4), 1246–1264. https://doi.org/10.1111/ijal.12561

Deygers, B., Buelens, L., Chan, D., Schildt, L., Van Parys, A., & Vanbuel, M. (2025). The impact of automated writing evaluation on writing gains. ELT Journal. https://doi.org/10.1093/elt/ccaf020

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. https://doi.org/10.1186/s41239-023-00425-2

Fazackerley, L. A., Perrin, D., & Minett, G. M. (2025). Harnessing generative AI in exercise and sports science education: enhancing real-world learning and overcoming traditional barriers in data analysis. Advances in physiology education, 49(2), 496-502. https://doi.org/10.1152/advan.00249.2024

Feng, W., Lai, X., Zhang, X., Fan, X., & Du, Y. (2025). Research on the construction and application of intelligent tutoring system for english teaching based on generative pre-training model. Systems and Soft Computing, 7, 200232. https://doi.org/10.1016/j.sasc.2025.200232

Fudiyartanto, F. A. (2024). English language teaching in a globalised world: A sociological investigation of Indonesian higher education. International Journal of Educational Research, 127, 102395. https://doi.org/10.1016/j.ijer.2024.102395

Gozali, I., Wijaya, A. R. T., Lie, A., Cahyono, B. Y., & Suryati, N. (2024). ChatGPT as an automated writing evaluation (AWE) tool: feedback literacy development and AWE tools' integration framework. JALT CALL Journal. https://doi.org/10.29140/jaltcall.v20n1.1200

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. Internet and Higher Education. https://doi.org/10.1016/j.iheduc.2024.100962

Guo, Y., & Wang, Y. (2025). Exploring the Effects of Artificial Intelligence Application on EFL Students' Academic Engagement and Emotional Experiences: A Mixed-Methods Study. European Journal of Education, 60(1), Article e12812. https://doi.org/10.1111/ejed.12812

Guo, K., Zhang, E. D., Li, D., & Yu, S. (2025). Using AI-supported peer review to enhance feedback literacy: An investigation of students' revision of feedback on peers' essays. British Journal of Educational Technology. https://doi.org/10.1111/bjet.13540

Haddaway, N. R., Page, M. J., Pritchard, C. C., & McGuinness, L. A. (2022). PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams. Campbell Systematic Reviews, 18(2), e1230. https://doi.org/10.1002/cl2.1230

Holland, A. M., Lorenz, W. R., Cavanagh, J. C., Smart, N. J., Ayuso, S. A., Scarola, G. T., Kercher, K. W., Jorgensen, L. N., Janis, J. E., Fischer, J. P., & Heniford, B. T. (2024). Comparison of Medical Research Abstracts Written by Surgical Trainees and Senior Surgeons or Generated by Large Language Models. JAMA network open, 7(8), e2425373. https://doi.org/10.1001/jamanetworkopen.2024.25373

Hong, Q. N., Gonzalez-Reyes, A., & Pluye, P. (2018). Improving the usefulness of a tool for appraising the quality of qualitative, quantitative and mixed methods studies, the Mixed Methods Appraisal Tool (MMAT). Journal of Evaluation in Clinical Practice, 24(3), 459–467. https://doi.org/10.1111/jep.12884

Huang, X., Zou, D., Cheng, G., Chen, X., & Xie, H. (2023). Trends, Research Issues and Applications of Artificial Intelligence in Language Education. Educational Technology and Society, 26(1), 112–131. https://doi.org/10.30191/ETS.202301_26(1).0009

Jaashan, H. M. S., & Alashabi, A. A. (2025). Using AI Large Language Model (LLM-ChatGPT) to Mitigate Spelling Errors of EFL Learners. Forum for Linguistic Studies. https://doi.org/10.30564/fls.v7i3.8438

Jeon, J. (2024). Exploring AI chatbot affordances in the EFL classroom: young learners' experiences and perspectives. Computer Assisted Language Learning, 37(1-2), 1–26. https://doi.org/10.1080/09588221.2021.2021241

Jeon, J., Lee, S., & Choe, H. (2023). Beyond ChatGPT: A conceptual framework and systematic review of speech-recognition chatbots for language learning. Computers and Education, 206, Article 104898. https://doi.org/10.1016/j.compedu.2023.104898

Karataş, F., Abedi, F. Y., Ozek Gunyel, F., Karadeniz, D., & Kuzgun, Y. (2024). Incorporating AI in foreign language education: An investigation into ChatGPT's effect on foreign language learners. Education and Information Technologies, 29(15), 19343–19366. https://doi.org/10.1007/s10639-024-12574-6

Khoso, A. K., Honggang, W., & Darazi, M. A. (2025). Empowering creativity and engagement: The impact of generative artificial intelligence usage on Chines EFL students' language learning experience. Computers in Human Behavior Reports, 18, 100627. https://doi.org/10.1016/j.chbr.2025.100627

Kitchenham, B., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering (EBSE Technical Report EBSE-2007-01). Keele University and University of Durham.

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). Exploring generative artificial intelligence preparedness among university language instructors: A case study. Computers and Education: Artificial Intelligence, 5, Article 100156. https://doi.org/10.1016/j.caeai.2023.100156

Kurt, G., & Kurt, Y. (2024). ENHANCING L2 WRITING SKILLS: CHATGPT AS AN AUTOMATED FEEDBACK TOOL. Journal of Information Technology Education: Research. https://doi.org/10.28945/5370

Lai, W. Y. W., & Lee, J. S. (2024). A systematic review of conversational AI tools in ELT: Publication trends, tools, research methods, learning outcomes, and antecedents. Computers and Education: Artificial Intelligence, 7, 100291. https://doi.org/10.1016/j.caeai.2024.100291

Law, L. H. L. (2024). Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. Computers and Education Open, 6, 100174. https://doi.org/10.1016/j.caeo.2024.100174

Li, B., Lowell, V. L., Wang, C., & Li, X. (2024). A systematic review of the first year of publications on ChatGPT and language education: Examining research on ChatGPT's use in language learning and teaching. Computers and Education: Artificial Intelligence, 7, 100266. https://doi.org/10.1016/j.caeai.2024.100266

Liebenow, L. W., Schmidt, F. T. C., Meyer, J., & Fleckenstein, J. (2025). Self-assessment accuracy in the age of artificial Intelligence: Differential effects of LLM-generated feedback. Computers and Education. https://doi.org/10.1016/j.compedu.2025.105385

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

Lo, C. K., Hew, K. F., & Jong, M. S.-Y. (2024). The influence of ChatGPT on student engagement: A systematic review and future research agenda. Computers & Education, 219, 105100. https://doi.org/10.1016/j.compedu.2024.105100

Lo, C. K., Yu, P. L. H., Xu, S., Ng, D. T. K., & Jong, M. S.-Y. (2024). Exploring the application of ChatGPT in ESL/EFL education and related research issues: A systematic review of empirical studies. Smart Learning Environments, 11, 50. https://doi.org/10.1186/s40561-024-00342-5

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLoS Medicine, 6(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

Perales, W., & Bedoya Ulla, M. (2025). Technology in EFL contextual teaching and learning: Teachers' practices and perspectives in a Thai university. Heliyon, 11, e44169. https://doi.org/10.1016/j.heliyon.2025.e44169

Perkins, M. (2023). Academic Integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. Journal of University Teaching and Learning Practice, 20(2), Article 7. https://doi.org/10.53761/1.20.02.07

Puteri, C. G., Cahyono, B. Y., Widiati, U., & Suryati, N. (2025). Using ChatGPT as a writing assistant: A study on essay quality development among Indonesian EFL University Students. Indonesian Journal of Applied Linguistics. https://doi.org/10.17509/3ebmy158

Seo, J. Y. (2024). Exploring the Educational Potential of ChatGPT: AI-Assisted Narrative Writing for EFL College Students. Language Teaching Research Quarterly. https://doi.org/10.32038/ltrq.2024.43.01

Sharadgah, T. A., & Sa'di, R. A. (2022). A SYSTEMATIC REVIEW OF RESEARCH ON THE USE OF ARTIFICIAL INTELLIGENCE IN ENGLISH LANGUAGE TEACHING AND LEARNING (2015-2021): WHAT ARE THE CURRENT EFFECTS?. Journal of Information Technology Education: Research, 21, 337–377. https://doi.org/10.28945/4999

Teng, M. F. (2025). A Systematic Review of ChatGPT for English as a Foreign Language Writing: Opportunities, Challenges, and Recommendations. International Journal of TESOL Studies, 6(3), 36–57. https://doi.org/10.58304/ijts.20240304

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Medical Research Methodology, 8, 45. https://doi.org/10.1186/1471-2288-8-45

Tran, T. T. T. (2025). Enhancing EFL Writing Revision Practices: The Impact of AI- and Teacher-Generated Feedback and Their Sequences. Education Sciences. https://doi.org/10.3390/educsci15020232

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. British Journal of Management, 14(3), 207–222. https://doi.org/10.1111/1467-8551.00375

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. https://doi.org/10.1007/s10639-024-12722-y

Wang, P., Liu, T., Yang, Y., & Xiang, X. (2025). Optimizing self-regulated learning: A mixed-methods study on GAI's impact on undergraduate task strategies and metacognition. British Journal of Educational Technology. https://doi.org/10.1111/bjet.70018

Weng, X., & Chiu, T. K. F. (2023). Instructional design and learning outcomes of intelligent computer assisted language learning: Systematic review in the field. Computers and Education: Artificial Intelligence, 4, 100117. https://doi.org/10.1016/j.caeai.2022.100117

Xiao, Y., & Zhi, Y. (2023). An Exploratory Study of EFL Learners' Use of ChatGPT for Language Learning Tasks: Experience and Perceptions. Languages, 8(3), Article 212. https://doi.org/10.3390/languages8030212

Yang, H., & Kyun, S. (2022). The current research trend of artificial intelligence in language learning: A systematic empirical literature review from an activity theory perspective. Australasian Journal of Educational Technology, 38(5), 180–210. https://doi.org/10.14742/ajet.7492

Zhai, C., & Wibowo, S. (2023). A systematic review on artificial intelligence dialogue systems for enhancing English as foreign language students' interactional competence in the university. Computers and Education: Artificial Intelligence, 4, Article 100134. https://doi.org/10.1016/j.caeai.2023.100134

Zhai, C., Wibowo, S., & Li, L. D. (2024). Evaluating the AI dialogue System's intercultural, humorous, and empathetic dimensions in English language learning: A case study. Computers and Education: Artificial Intelligence, 7, 100262. https://doi.org/10.1016/j.caeai.2024.100262

Zhang, S., Shan, C., Lee, J. S. Y., Che, S. P., & Kim, J. H. (2023). Effect of chatbot-assisted language learning: A meta-analysis. Education and Information Technologies, 28(11), 15223–15243. https://doi.org/10.1007/s10639-023-11805-6

Zhu, M., & Wang, C. (2025). A systematic review of research on AI in language education: Current status and future implications. Language Learning and Technology, 29(1), 1–29.

Zhuang, Y., Zhao, R., Xie, Z., & Yu, P. L. H. (2025). Enhancing language learning through generative AI feedback on picture-cued writing tasks. Computers and Education: Artificial Intelligence. https://doi.org/10.1016/j.caeai.2025.100450