Vol. 10, No. 4, 2025, pp. 186-199
DOI: https://doi.org/10.29210/30036974000

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

Generative AI in second-language writing: a systematic review of learning, feedback, and literacy outcomes

Kristian Burhan
Universitas Negeri padang, Indonesia

Article Info

Article history:

Received Sep 28th, 2025
Revised Oct 07th, 2025
Accepted Dec 30th, 2025

Keyword:

Generative artificial intelligence,
ChatGPT,
English as a foreign language writing,
Automated writing evaluation,
Feedback literacy

ABSTRACT

The diffusion of generative artificial intelligence (AI) has challenged assumptions about how English L2 writing is taught, supported, and assessed. This systematic literature review synthesised peer-reviewed empirical evidence on generative AI in English L2 writing, focusing on pedagogical transformation, feedback and assessment, and academic literacy. Following PRISMA 2020, searches of Scopus, PubMed, and ScienceDirect identified 912 records. After duplicate removal and eligibility screening, 18 empirical studies published between 2023 and 2025 were included, although the search window covered 2020-2025. The eligible studies were appraised with the Mixed Methods Appraisal Tool (MMAT 2018), with 14 meeting at least 4 of 5 criteria under the manuscript's descriptive appraisal rubric and four scoring 3.0-3.5; none was excluded solely on quality. Owing to substantial heterogeneity in designs and outcomes, findings were synthesised narratively and no quantitative meta-analysis was attempted. Evidence indicates that generative AI can improve surface-level accuracy, revision activity, and selected language or self-regulatory outcomes, while effects on higher-order rhetorical and critical competencies remain inconsistent. AI is most defensible as a teacher-guided complement, with continuing concerns about academic integrity, over-reliance, assessment validity, and equity. The review identifies priorities for longitudinal, multi-model, and multi-site research.

Corresponding Author:

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

Introduction

The emergence of publicly accessible generative artificial intelligence (AI) has become one of the most consequential developments for higher education and language learning in a generation. Tools built on large language models (LLMs) can now produce fluent, context-sensitive text on demand, reshaping how knowledge is produced, communicated, and assessed across disciplines (Crompton et al., 2024; Gupta et al., 2024). Because English continues to serve as a major medium of international scholarship and professional communication, competence in English academic writing remains a high-stakes academic and professional skill (Chapelle, 2025; Tafazoli, 2024). Generative AI therefore intersects with a domain in which learning processes and assessed products are closely connected, raising urgent questions about appropriate pedagogical integration.

Within this landscape, the public release of ChatGPT in late 2022 catalysed unprecedented interest in the language classroom. Educators and researchers moved quickly from cautious speculation to empirical investigation of how conversational LLMs affect the teaching and learning of second-language (L2) writing (Barrot, 2024a; Kohnke et al., 2025). Writing is a particularly sensitive site of inquiry because it is both a learning

process and an assessed product, meaning that any tool capable of generating and revising text challenges established notions of authorship, effort, and evaluation (Derakhshan & Ghiasvand, 2024). This tension has positioned English L2 writing at the centre of debates about the promise and peril of generative AI in education.

A first wave of scholarship mapped the affordances and challenges of these tools, documenting their capacity to generate ideas, model genres, and provide instant corrective feedback, alongside concerns about accuracy and misuse (Pack & Maloney, 2023; Karataş et al., 2024). Early reviews began to consolidate this fragmented literature, noting the dominance of perception studies and the scarcity of rigorous intervention research (Teng, 2024; Jeon, 2025). Nonetheless, the field has expanded so rapidly that syntheses risk becoming outdated within months, and few have simultaneously addressed pedagogy, assessment, and literacy as an integrated whole.

Technological developments have compounded this pace of change. Generative capabilities are no longer confined to ChatGPT; competing systems such as Google's Gemini, Microsoft's Copilot, and other LLMs are now embedded in writing workflows, and specialised applications for automated writing evaluation (AWE) and automated essay scoring (AES) have proliferated (Pack et al., 2024; Tran, 2025). These advances promise to relieve teachers of time-consuming grading while offering learners individualised, round-the-clock support, but their validity and reliability for L2 assessment remain contested and unevenly evidenced.

Interest in AI-mediated learning extends well beyond language studies into the broader education sciences, including physical and sport-science education, where reviews have charted the adoption of intelligent and generative technologies for instruction and assessment (Wang & Wang, 2024; Zhong et al., 2025; Fazackerley et al., 2025). This cross-disciplinary momentum underscores that English academic literacy is a shared demand for scholars and students across fields, from applied linguistics to the health and movement sciences, and that insights about generative AI in English writing carry implications well beyond the language classroom.

Despite this activity, the empirical base remains fragmented. Studies are typically small in scale, geographically concentrated, and methodologically heterogeneous, and they seldom speak to one another across the pedagogical, evaluative, and literacy dimensions of writing (Zhang & Umeanowai, 2025; Teng, 2024). As a result, educators lack a consolidated, critically appraised account of what generative AI can and cannot reliably do for English L2 writers, and under what conditions.

A second limitation is theoretical and methodological. Much of the evidence rests on short interventions and self-reported perceptions rather than sustained, outcome-oriented designs, and constructs such as feedback literacy and academic integrity are frequently invoked but rarely synthesised across studies (Kurt & Kurt, 2024; Gozali et al., 2024; Perkins, 2023). A systematic review is therefore needed to distinguish findings supported by empirical evidence from provisional claims and to make the limits of the evidence base explicit.

These gaps make a timely, methodologically transparent synthesis necessary. This systematic literature review (SLR) consolidates and critically appraises empirical evidence on generative AI in English L2 writing. The review follows PRISMA 2020 reporting guidance, applies predefined eligibility criteria, appraises included studies with the Mixed Methods Appraisal Tool (MMAT), and uses thematic narrative synthesis because the studies report heterogeneous designs, outcomes, and measures. The review protocol was not prospectively registered, and this is acknowledged as a methodological limitation. The synthesis is organised around three inter-related dimensions: pedagogical transformation, feedback and assessment, and academic literacy. To achieve this aim, the review addresses three research questions:

RQ1. How is generative AI being integrated into English L2 writing pedagogy, and what learning processes and instructional practices does it reshape?

RQ2. How effective and reliable is generative AI as a source of writing feedback and automated evaluation relative to human or teacher feedback, and how does it influence learners' writing performance and feedback literacy?

RQ3. What affordances, challenges, and academic-integrity concerns do students and teachers report when using generative AI for English academic writing, and what implications follow for developing academic literacy?

By answering these questions through PRISMA-guided identification and screening, structured quality appraisal, and thematic narrative synthesis, the review offers an integrated account of a literature that has developed across pedagogical, evaluative, and literacy strands.

Method

Research Design and Framework

This study adopted a systematic literature review (SLR) design, a rigorous and replicable method for identifying, appraising, and synthesising the available evidence on a defined question (Tranfield et al., 2003). The SLR approach was selected because it minimises reviewer bias through transparent, pre-specified procedures and is well suited to consolidating a rapidly expanding and heterogeneous body of research (Liberati et al., 2009). Reporting was structured in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement (Page et al., 2021), which extends the earlier PRISMA guidance (Moher et al., 2009) and provides a standardised architecture for documenting identification, screening, eligibility, and inclusion.

Search Strategy

The research questions were first decomposed using the PICO framework (Population, Intervention/Interest, Comparison, Outcome) to ensure systematic identification of the core search concepts. The Population comprised EFL/ESL learners and English L2 writers in formal educational settings; the Intervention/Interest was the use of generative AI tools (ChatGPT and other LLMs) for writing instruction, feedback, or assessment; the Comparison, where present, was human or teacher feedback, traditional instruction, or no comparator; and the Outcome encompassed writing performance, feedback quality and literacy, learner perceptions, and academic-literacy development.

Each PICO element was translated into a cluster of keywords and synonyms, which were combined with Boolean operators. The representative search string was:

> *("generative artificial intelligence" OR "generative AI" OR "ChatGPT" OR "large language model*" OR "GPT-4" OR "LLM") AND ("English as a foreign language" OR "EFL" OR "ESL" OR "L2 writing" OR "second language writing" OR "academic writing" OR "writing instruction") AND ("feedback" OR "assessment" OR "writing" OR "academic literacy")*

Truncation and wildcard operators (e.g., model* to capture model and models) were used to maximise sensitivity. Field codes restricted the search to the title, abstract, and keyword fields (TITLE-ABS-KEY in Scopus, with the equivalent field tags applied in PubMed and ScienceDirect). Limiters were applied for language (English), document type (journal article), and publication window (2020–2025). The string was piloted and refined iteratively: an initial broad string returned an unmanageable number of tangential records, prompting the addition of the writing- and assessment-specific cluster to improve precision. The PICO framework here served solely to structure the search and is distinct from the quality-appraisal procedure described in Section Quality Assessment.

Database and Information Sources

Three databases were searched: Scopus (as the primary multidisciplinary source), PubMed (for education and health-adjacent coverage), and ScienceDirect (for full-text coverage of relevant journals). The final database search was executed in 2025. No supplementary hand-searching or citation-chasing was undertaken, so the evidence base reflects the retrieval characteristics and indexing coverage of the three databases.

Eligibility Criteria

Table 1. Inclusion and exclusion criteria

Criterion	Inclusion	Exclusion
Language	Studies reported in English	Non-English reports
Document type	Peer-reviewed empirical journal articles	Conference papers, book chapters, editorials, reviews, meta-analyses
Publication period	2020–2025	Published before 2020
Population	EFL/ESL learners or English L2 writers in formal education	Non-language populations; first-language-only writers
Intervention	Generative AI (ChatGPT/LLMs) as the focal tool for writing	Non-generative AI, or AI not applied to writing
Outcome	Writing performance, feedback, literacy, or perceptions	No writing-related outcome
Accessibility	Full text available	Abstract-only records

Study Selection Process

Study selection followed a multi-stage process consistent with PRISMA 2020. Records retrieved from the three databases were combined and duplicate records were removed. The remaining records were screened by title and abstract against the predefined eligibility criteria, followed by full-text assessment of potentially eligible

reports. At full text, a record was considered insufficiently detailed when the report did not provide enough information to establish the study design, participant population, generative-AI intervention, writing-related outcome, or principal result needed for extraction and synthesis. Decisions at each stage were documented, and reasons for full-text exclusion were logged. Ambiguous cases were resolved by re-examining the report against the predefined criteria. The original screening file did not retain item-level independent decisions from multiple reviewers; consequently, Cohen's kappa cannot be calculated retrospectively from the manuscript data, and no agreement statistic is reported to avoid fabricating a value. This limitation is explicitly acknowledged in the Discussion and should be addressed in future reviews through independent duplicate screening with a prespecified agreement procedure.

Quality Assessment

Quality appraisal was conducted independently of the PICO-based search formulation. Each included study was assessed using the Mixed Methods Appraisal Tool (MMAT), version 2018 (Hong et al., 2018), selected for its suitability in appraising the heterogeneous designs anticipated in this literature (quantitative descriptive, quantitative non-randomised, qualitative, and mixed-methods studies). For each study, the design category was first classified, and the corresponding five MMAT criteria were then applied and answered as Yes, Can't tell/Partial, or No. Responses were scored 1, 0.5, and 0, respectively. Studies meeting at least 60% of criteria (≥3 of 5) were considered to provide adequate methodological confidence. Consistent with MMAT developer guidance, no study was excluded on the basis of its quality score alone; lower-scoring studies were retained but their contribution to the synthesis was weighted accordingly. A summary of the appraisal is presented in Table 2.

Table 2. Quality appraisal summary using the Mixed Methods Appraisal Tool (MMAT, 2018)

Author(s), Year	Study design	Q1	Q2	Q3	Q4	Q5	Score	Category
Escalante et al. (2023)	Quant. non-rand.	Y	Y	Y	Y	Y	5.0	High
Liu et al. (2024)	Qualitative	Y	Y	Y	Y	P	4.5	High
Pack et al. (2024)	Quant. descriptive	Y	Y	Y	Y	Y	5.0	High
Guo et al. (2024)	Quant. non-rand.	Y	Y	Y	P	Y	4.5	High
Punar Özçelik & Yangın Ekşi (2024)	Quant. non-rand.	Y	P	Y	P	N	3.0	Medium
Tsai et al. (2024)	Quant. non-rand.	Y	Y	Y	Y	Y	5.0	High
Algaraady & Mahyoob (2023)	Quant. descriptive	Y	P	Y	P	Y	3.5	Medium
Nugroho et al. (2025)	Qualitative	Y	Y	Y	P	P	4.0	High
Guo et al. (2025)	Qualitative	Y	Y	Y	Y	P	4.5	High
Tran (2025)	Quant. non-rand.	Y	P	Y	P	Y	3.5	Medium
Alkamel & Alwagieh (2024)	Quant. descriptive	Y	P	Y	P	Y	3.5	Medium
Gozali et al. (2024)	Qualitative	Y	Y	Y	P	P	4.0	High
Meniado et al. (2024)	Mixed methods	Y	Y	Y	Y	P	4.5	High
Kurt & Kurt (2024)	Qualitative	Y	Y	Y	P	P	4.0	High
Seo (2024)	Quant. non-rand.	Y	Y	Y	P	Y	4.5	High
Abduljawad (2024)	Mixed methods	Y	Y	P	P	Y	4.0	High
Wang et al. (2025)	Mixed methods	Y	Y	Y	Y	P	4.5	High
Gao et al. (2025)	Quant. non-rand.	Y	Y	P	Y	P	4.0	High

Note. Y = Yes (1.0); P = Partial/Can't tell (0.5); N = No (0.0). Q1–Q5 correspond to the five MMAT criteria for the study's design category. Category: High ≥ 4.0; Medium 3.0–3.5; Low < 3.0.

Data Extraction Procedure

A structured data-extraction form was used to record, for each included study, the author(s) and year, the study context, the research design, the sample, the generative AI tool and intervention, the outcome measures, and the principal findings. Extraction focused on information reported in the primary sources to preserve fidelity to the original evidence.

Bibliometric and Descriptive Analysis

Descriptive analyses summarised the temporal distribution of the included studies, the geographic distribution of study contexts as reported by the authors, and the distribution of research designs. These descriptive characteristics are visualised in Figures 2–4 and contextualise the thematic synthesis.

Data Analysis and Synthesis

Given the methodological heterogeneity of the included studies, a thematic narrative synthesis was undertaken following Thomas and Harden (2008), complemented by the principles of reflexive thematic analysis described by Braun and Clarke (2006). Findings and outcome statements were coded inductively, codes were grouped into descriptive themes, and these were organised into the three analytical dimensions corresponding to the research questions: pedagogical transformation, feedback and assessment, and academic literacy and concerns. No meta-analysis or quantitative pooling was attempted because the studies differed substantially in design, intervention

dosage, comparators, outcome constructs, measurement instruments, and reporting formats. The synthesis therefore compares direction, consistency, and contextual conditions of findings rather than producing a pooled effect estimate. Quality-appraisal information was used to temper interpretation rather than to exclude studies.

Reporting and Documentation

The review was conducted and reported in accordance with the PRISMA 2020 statement (Page et al., 2021). The flow of records through identification, screening, eligibility, and inclusion is documented in Figure 1. Reporting was checked against the PRISMA 2020 checklist to improve transparency and reproducibility. Because the protocol was not prospectively registered and item-level dual-screening records were not retained, these methodological limitations are disclosed rather than inferred away.

Results and Discussions

Study Selection Results

The database searches identified 912 records in total: 612 from Scopus, 200 from PubMed, and 100 from ScienceDirect. After removal of 14 duplicate records, 898 records were screened by title and abstract, of which 786 were excluded because they did not focus on generative AI (n = 402), did not address English L2/EFL/ESL writing (n = 320), or were non-empirical editorials and opinion pieces (n = 64). Of the 112 reports sought for retrieval, 9 could not be obtained in full text, leaving 103 reports assessed for eligibility. A further 85 reports were excluded because generative AI was not the focal intervention (n = 22), no writing or academic-literacy outcome was assessed (n = 28), the report was a review, meta-analysis, or conceptual paper (n = 19), or the report lacked sufficient methodological detail to establish the design, participants, focal intervention, writing-related outcome, or principal result (n = 16). Eighteen studies met all criteria and were included in the synthesis. The complete flow is depicted in Figure 1, following PRISMA 2020 (Page et al., 2021).

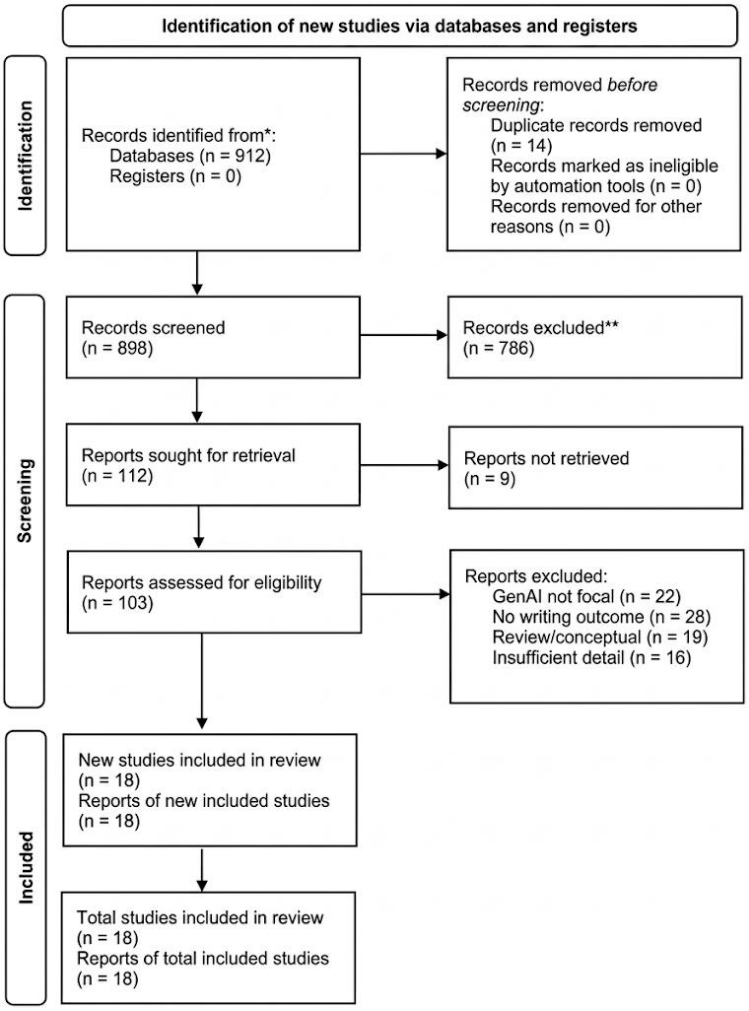

Figure 1. PRISMA 2020 flow diagram of the study selection process

Descriptive Characteristics of Included Studies

The 18 included studies were published between 2023 and 2025, with a marked concentration in 2024 (n = 11), reflecting the surge of empirical work following the public release of ChatGPT (Figure 2). Reported study contexts spanned at least eight countries, with China the most frequently represented; notably, eight studies did not specify the national context of their participants in the available metadata (Figure 3). Methodologically, the corpus was dominated by quantitative experimental and quasi-experimental designs (n = 7), followed by qualitative studies (n = 5), mixed-methods studies (n = 4), and quantitative descriptive or psychometric studies (n = 2) (Figure 4). Full characteristics of each study are reported in Table 3, and a thematic and methodological classification is provided in Table 4.

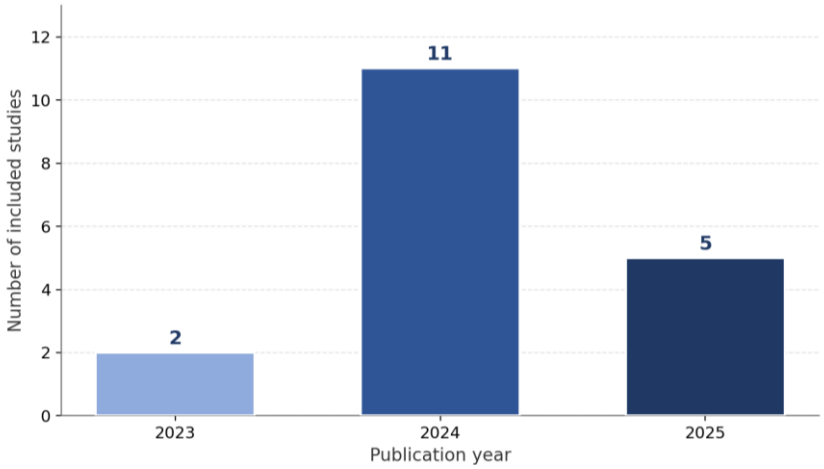

Figure 2. Annual distribution of the 18 included studies (2023–2025)

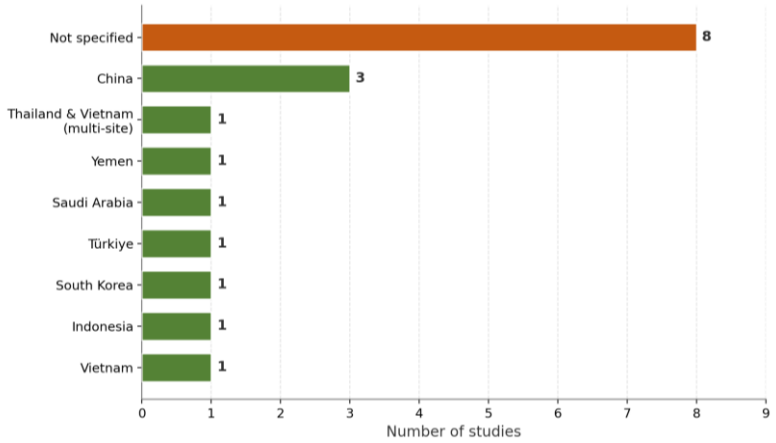

Figure 3. Geographic distribution of reported study contexts (N = 18)

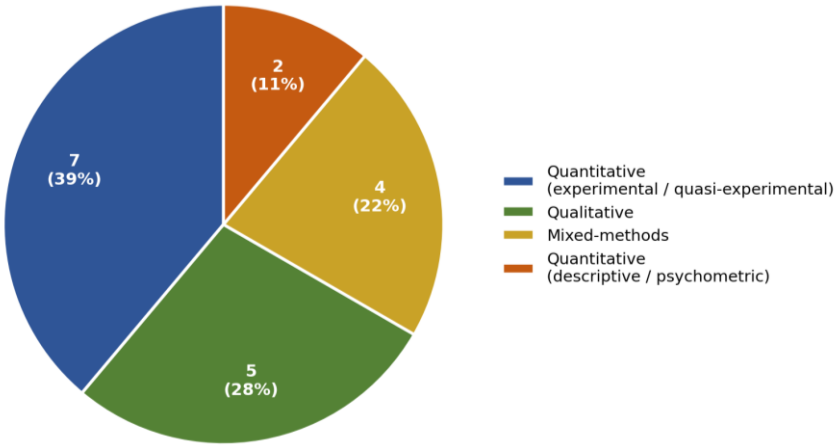

Figure 4. Distribution of research designs across included studies (N = 18)

Table 3. Characteristics and key findings of the 18 included studies

Author(s), Year	Title	Context	Design & sample	Key findings
Escalante et al. (2023)	AI-generated feedback on writing: insights into efficacy and ENL student preference	Not specified	Quasi-experimental; 48 + 43 ENL learners	No difference in learning outcomes between GPT-4 and human-tutor feedback; learner preference split; a blended approach is recommended.
Liu et al. (2024)	Investigating students' cognitive processes in generative AI-assisted digital multimodal composing and traditional writing	Not specified	Qualitative; two EFL groups	Generative AI reshaped the composing process; distinct patterns of text production and AI-image prompting emerged versus traditional essay writing.
Pack et al. (2024)	Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability	Not specified	Quantitative (psychometric); 119 essays, 4 LLMs, 2 raters	Among four LLMs, GPT-4 showed the best automated-essay-scoring validity and reliability; all models fluctuated across occasions.
Guo et al. (2024)	Effects of an AI-supported approach to peer feedback on university EFL students' feedback quality and writing ability	China	Quasi-experimental; 124 undergraduates	AI-supported peer feedback (chatbot Eva) significantly improved reviewers' feedback quality and their own writing ability.
Punar Özçelik & Yangın Ekşi (2024)	Cultivating writing skills: the role of ChatGPT as a learning assistant-a case study	Not specified	Pre-experimental; 11 students	ChatGPT aided formal-register acquisition but was viewed as unnecessary for informal writing and less effective for neutral register.
Tsai et al. (2024)	Impacts of ChatGPT-assisted writing for EFL English majors: Feasibility and challenges	Not specified	Experimental (double-blind); 44 students	ChatGPT-assisted revision significantly raised scores (largest gains in vocabulary); disproportionate gains for weaker writers raised fairness concerns.
Algaraady & Mahyoob (2023)	ChatGPT's Capabilities in Spotting and Analyzing Writing Errors Experienced by EFL Learners	Not specified	Quantitative descriptive; EFL error corpus	ChatGPT detected most surface-level errors but missed deep-structure and pragmatic errors; it could not replace human instructors.
Nugroho et al. (2025)	Students' appraisals post-ChatGPT use: Students' narrative after using ChatGPT for writing	Not specified	Qualitative; 12 students	Students valued ChatGPT for translation, accuracy, efficiency, and idea generation but flagged inaccuracy and integrity risks, retaining fact-checking awareness.
Guo et al. (2025)	Using AI-supported peer review to enhance feedback literacy: An investigation of students' revision of feedback on peers' essays	China	Qualitative; 44 undergraduates	An AI-supported peer-review system (EvaluMate/Eva) improved peer-comment quality; varied revision strategies fostered feedback literacy.
Tran (2025)	Enhancing EFL Writing Revision Practices: The Impact of AI- and Teacher-Generated Feedback and Their Sequences	Vietnam	Quasi-experimental; 14 undergraduates	AI feedback (Gemini) prompted more revisions than teacher feedback alone; combined AI-plus-teacher feedback produced the highest-quality revisions.
Alkamel & Alwagieh (2024)	Utilizing an adaptable artificial intelligence writing tool (ChatGPT) to enhance academic writing skills among Yemeni university EFL students	Yemen	Quantitative survey; university EFL learners	Learners perceived ChatGPT positively for fluency, accuracy, and quality but raised integrity and plagiarism concerns; recommended as a supplementary tool.
Gozali et al. (2024)	ChatGPT as an automated writing evaluation (AWE) tool: feedback literacy development and AWE tools' integration framework	Indonesia	Qualitative case study; 18 students	ChatGPT with Grammarly and QuillBot supported feedback-literacy development; the study proposes an AWE Tools Integration Framework.

Author(s), Year	Title	Context	Design & sample	Key findings
Meniado et al. (2024)	Using ChatGPT for second language writing: Experiences and perceptions of EFL learners in Thailand and Vietnam	Thailand & Vietnam	Mixed-methods; 357 survey + 16 interviews	Learners perceived ChatGPT positively and used it for brainstorming, organising, and editing; Vietnamese learners were more favourable than Thai learners.
Kurt & Kurt (2024)	Enhancing L2 Writing Skills: ChatGPT as an Automated Feedback Tool	Türkiye	Qualitative (focus groups); pre-service teachers	ChatGPT offered practicality, interactivity, and adaptability but showed inconsistencies and prompt-dependence; it could not fully replace peer or teacher feedback.
Seo (2024)	Exploring the Educational Potential of ChatGPT: AI-Assisted Narrative Writing for EFL College Students	South Korea	Quasi-experimental; 44 students	ChatGPT-assisted narrative writing significantly improved fluency and overall performance, while syntactic complexity decreased.
Abduljawad (2024)	Investigating the Impact of ChatGPT as an AI Tool on ESL Writing: Prospects and Challenges in Saudi Arabian Higher Education	Saudi Arabia	Mixed-methods; 130 + 6 (focus group)	ChatGPT significantly benefited ESL writing via personalised feedback and vocabulary growth but raised concerns over creativity, individuality, and ethics.
Wang et al. (2025)	Optimizing self-regulated learning: A mixed-methods study on GAI's impact on undergraduate task strategies and metacognition	China	Mixed-methods; 40 undergraduates	A generative-AI chatbot improved task-strategy diversity, metacognitive awareness, and writing performance; participants noted over-reliance and accuracy concerns.
Gao et al. (2025)	Do AI chatbot-integrated writing tasks influence writing self-efficacy and critical thinking ability? An exploratory study	Not specified	Quasi-experimental; 80 vocational students	AI-chatbot writing tasks yielded no significant overall gain in critical thinking or self-efficacy but a significant gain in language construction.

Table 4. Thematic and methodological classification of the included studies

Author(s), Year	Research design	Primary theme (RQ)	AI tool / intervention	Reported outcome
Escalante et al. (2023)	Quasi-experimental	Feedback & assessment (RQ2)	ChatGPT (GPT-4)	Learning outcomes; feedback preference
Liu et al. (2024)	Qualitative	Pedagogical transformation (RQ1)	ChatGPT + Bing Chat/Image Creator	Composing process; multimodal text production
Pack et al. (2024)	Quantitative (psychometric)	Feedback & assessment (RQ2)	GPT-4, GPT-3.5, PaLM 2, Claude 2	AES validity and reliability
Guo et al. (2024)	Quasi-experimental	Feedback & assessment (RQ2)	AI chatbot 'Eva'	Feedback quality; writing ability
Punar Özçelik & Yangın Ekşi (2024)	Pre-experimental	Pedagogical transformation (RQ1)	ChatGPT	Register-knowledge acquisition
Tsai et al. (2024)	Experimental	Feedback & assessment (RQ2)	ChatGPT	Essay scores; assessment fairness
Algaraady & Mahyoob (2023)	Quantitative descriptive	Feedback & assessment (RQ2)	ChatGPT	Error-detection accuracy
Nugroho et al. (2025)	Qualitative	Academic literacy & concerns (RQ3)	ChatGPT	Perceptions; academic-integrity awareness
Guo et al. (2025)	Qualitative	Academic literacy & concerns (RQ3)	EvaluMate ('Eva')	Peer-review quality; feedback literacy
Tran (2025)	Quasi-experimental	Feedback & assessment (RQ2)	Gemini (+ teacher)	Revision frequency and quality
Alkamel & Alwagieh (2024)	Quantitative survey	Academic literacy & concerns (RQ3)	ChatGPT	Perceptions; writing quality; integrity

Author(s), Year	Research design	Primary theme (RQ)	AI tool / intervention	Reported outcome
Gozali et al. (2024)	Qualitative case study	Academic literacy & concerns (RQ3)	ChatGPT + Grammarly + QuillBot	Feedback-literacy development
Meniado et al. (2024)	Mixed-methods	Pedagogical transformation (RQ1)	ChatGPT	Perceptions and writing practices
Kurt & Kurt (2024)	Qualitative	Feedback & assessment (RQ2)	ChatGPT	Perceived feedback affordances/limits
Seo (2024)	Quasi-experimental	Pedagogical transformation (RQ1)	ChatGPT	Writing fluency and complexity
Abduljawad (2024)	Mixed-methods	Academic literacy & concerns (RQ3)	ChatGPT	Writing proficiency; opportunities/challenges
Wang et al. (2025)	Mixed-methods	Pedagogical transformation (RQ1)	GAI chatbot (Tongyi.ai)	Task strategies; metacognition; performance
Gao et al. (2025)	Quasi-experimental	Pedagogical transformation (RQ1)	AI chatbot	Critical thinking; writing self-efficacy

Thematic Synthesis

RQ1 - Pedagogical Transformation

Across the corpus, generative AI emerged less as a discrete instructional add-on than as a force reshaping the writing process itself. Several studies documented how learners reorganised their composing routines around AI assistance, using it for planning, idea generation, drafting, and revision. Liu et al. (2024) showed that generative AI restructured the cognitive processes of composing, with learners in a multimodal condition constructing bridging texts and orchestrating AI-generated images through iterative prompting, a workflow qualitatively different from traditional essay writing. In a similar vein, Seo (2024) found that ChatGPT-assisted narrative writing produced significant gains in fluency and overall performance, even as syntactic complexity declined, suggesting that the tool may privilege quantity and smoothness over structural sophistication.

This pedagogical reconfiguration extended to learners' self-regulation and dispositions. Wang et al. (2025) reported that a generative-AI chatbot enhanced task-strategy diversity, metacognitive awareness, and writing performance relative to a control group, positioning AI as a scaffold for self-regulated learning rather than a mere text generator. Yet the transformation was uneven: Punar Özçelik and Yangın Ekşi (2024) found ChatGPT useful for acquiring formal register but redundant for informal writing, while Gao et al. (2025) observed no significant overall improvement in critical thinking or writing self-efficacy, with gains confined to language construction. Taken together, these studies indicate that generative AI reshapes the mechanics and strategy of L2 writing more readily than it transforms higher-order rhetorical and critical competencies, echoing broader observations about AI-mediated language learning (Kohnke et al., 2025; Barrot, 2024b; Kartal, 2024).

RQ2 - Feedback and Assessment

The largest cluster of studies examined generative AI as a source of writing feedback and as an automated evaluator, and here the evidence was both encouraging and cautionary. On efficacy, Escalante et al. (2023) found no difference in learning outcomes between students receiving GPT-4 feedback and those receiving human-tutor feedback, with learner preferences split evenly-an outcome that legitimises AI feedback as a viable complement rather than a substitute. Guo et al. (2024) demonstrated that an AI-supported peer-feedback chatbot significantly improved both the quality of students' feedback and their own writing ability, and Tran (2025) showed that AI feedback elicited more revisions than teacher feedback alone, with a combined AI-plus-teacher sequence producing the strongest revision outcomes. These findings converge on a complementary model in which AI handles surface-level, high-frequency concerns while teachers concentrate on higher-order issues (Kurt & Kurt, 2024).

The evidence on capability and validity was more qualified. Algaraady and Mahyoob (2023) found that ChatGPT reliably detected surface-level errors but missed deep-structure and pragmatic problems, a limitation corroborated by Alsaweed and Aljebreen (2024), and Pack et al. (2024) reported that, although GPT-4 achieved the best automated-essay-scoring validity and reliability among four LLMs, all models fluctuated across scoring occasions, undermining confidence in unsupervised deployment. Tsai et al. (2024) added an equity dimension: ChatGPT-assisted revision significantly raised essay scores, but disproportionately for weaker writers, skewing the score distribution and raising fairness concerns because revised texts no longer reflected learners' independent competence. This pattern resonates with wider concerns about assessment validity in the generative-AI era

(Chapelle, 2025; Chuang & Yan, 2025; Teckwani et al., 2024), and suggests that AI feedback is most defensible when embedded in process-oriented, teacher-mediated assessment rather than used as a standalone grader.

RQ3 - Academic Literacy, Affordances, and Concerns

A third strand foregrounded how students appropriate generative AI and the literacies and risks this entails. Multiple studies reported broadly positive learner perceptions alongside a consistent set of caveats. Alkamel and Alwagieh (2024) and Abduljawad (2024) found that learners credited ChatGPT with improving fluency, accuracy, vocabulary, and autonomy, while simultaneously voicing concerns about plagiarism, loss of individuality, reduced creativity, and ethics. Meniado et al. (2024), surveying learners in two countries, similarly documented favourable perceptions and pragmatic uses-brainstorming, organising ideas, and editing-together with cross-national variation in enthusiasm. Crucially, Nugroho et al. (2025) showed that students retained a fact-checking, integrity-aware stance even as they leaned on the tool, indicating that critical engagement can coexist with heavy use. Relatedly, learners' emerging prompt-engineering expertise shapes what the tool can deliver (Woo et al., 2025), while novice teachers' limited ability to detect AI-generated text complicates the enforcement of integrity (De Wilde, 2024).

Beyond perceptions, several studies reframed generative AI as an object and driver of feedback literacy. Gozali et al. (2024) demonstrated that ChatGPT, used alongside Grammarly and QuillBot, supported the development of learners' feedback literacy and proposed an integration framework for automated-writing-evaluation tools, while Guo et al. (2025) showed that an AI-supported peer-review system prompted students to enact varied revision strategies, thereby strengthening their capacity to interpret and act on feedback. These results position academic literacy in the generative-AI era not as the avoidance of AI but as the cultivated ability to use it judiciously, verify its output, and integrate it with human judgement-an orientation consistent with calls to treat integrity and critical use as teachable competencies (Perkins, 2023; Tafazoli, 2024; Dong, 2024).

Comparative and Critical Analysis

Methodologically, the corpus was dominated by short, single-cohort interventions and perception-oriented studies. Seven studies used experimental or quasi-experimental designs, five were qualitative, four used mixed methods, and two were quantitative descriptive or psychometric studies. Samples were frequently small: several studies involved fewer than 50 participants and two involved fewer than 15, constraining statistical power and generalisability (Punar Özçelik & Yangın Ekşi, 2024; Tran, 2025). Across the intervention studies, Escalante et al. (2023) found no learning-outcome difference between GPT-4 and human-tutor feedback, whereas Guo et al. (2024), Tsai et al. (2024), Tran (2025), Seo (2024), and Wang et al. (2025) reported positive effects on selected outcomes. In contrast, Gao et al. (2025) found no significant overall improvement in critical thinking or writing self-efficacy. This contrast indicates that positive effects are outcome-specific rather than uniform across the construct of writing competence.

The feedback literature also shows important differences by task and evaluator. Algaraady and Mahyoob (2023) and Alsaweed and Aljebreen (2024) indicate that ChatGPT is more dependable for surface-level error detection than for deeper structural or pragmatic problems. Pack et al. (2024) found that GPT-4 had the strongest automated-scoring performance among the models tested, but model scores varied across occasions. Guo et al. (2025) and Gozali et al. (2024) shift the focus from AI as an evaluator to AI as a scaffold for feedback literacy, while Kurt and Kurt (2024) similarly emphasise practicality but caution against replacement of peer or teacher feedback. Perception-oriented studies (Alkamel & Alwagieh, 2024; Abduljawad, 2024; Meniado et al., 2024; Nugroho et al., 2025) converge on usefulness for fluency, accuracy, vocabulary, brainstorming, and editing, but also report concerns about integrity, creativity, over-reliance, and accuracy. Liu et al. (2024) further shows that AI changes the composing process itself, demonstrating that the intervention is pedagogical as well as evaluative. Taken together, evidence from more than 15 studies supports a complementary rather than substitutive model of generative AI use, while the divergence in higher-order outcomes remains a key empirical uncertainty.

Interpreted as a whole, the evidence suggests that generative AI is changing English L2 writing primarily at the level of process support and product refinement rather than established independent competence. Its most consistent contribution is as a responsive feedback and drafting aid that can increase revision activity and surface-level accuracy; its weakest area is the rhetorical, pragmatic, and critical dimension of writing, where human expertise remains important (Algaraady & Mahyoob, 2023; Kurt & Kurt, 2024). This asymmetry has theoretical implications: models of feedback and self-regulated learning must accommodate a non-human agent that is continuously available but variably reliable, while feedback literacy becomes central to explaining why similar tools produce different outcomes across learners (Guo et al., 2025; Gozali et al., 2024).

These findings extend and complicate existing frameworks. They support socio-cognitive accounts in which tools mediate strategy and metacognition (Wang et al., 2025), while challenging assumptions that automated feedback straightforwardly transfers to independent competence, given evidence that AI-assisted gains may not

reflect learners' unaided ability (Tsai et al., 2024). For practice, the consistent signal is that generative AI should be integrated as a complement within teacher-guided, process-oriented pedagogy-used for lower-order concerns, drafting support, and feedback practice-rather than as an autonomous tutor or grader. Explicit instruction in prompting, verification, and ethical use is warranted, and assessment designs should emphasise process evidence and in-class performance to preserve validity (Chapelle, 2025; Chuang & Yan, 2025).

Situated against prior reviews, the present synthesis extends work that has mainly catalogued opportunities and challenges or concentrated on individual dimensions of the problem (Teng, 2024; Jeon, 2025; Crompton et al., 2024). Its distinctive contribution is to connect pedagogical transformation, feedback and assessment, and academic literacy within one evidence map. The cross-study comparison also shows that methodological design matters: controlled interventions tend to detect improvements in discrete or proximal outcomes, whereas qualitative and perception studies provide stronger evidence about learner strategies, uptake, and concerns. The recurring contradiction between improved and unchanged higher-order outcomes is therefore better interpreted as evidence of conditional effects than as a simple inconsistency.

The review has several limitations that constrain the certainty and external validity of its conclusions. First, only three databases were searched, with English-language journal articles and full-text availability required; therefore, relevant studies indexed elsewhere or published in other languages may have been missed. Second, no hand-searching or citation-chasing was undertaken, increasing the possibility of retrieval and publication bias. Because the synthesis was narrative and no quantitative pooling was performed, statistical publication-bias tests were not applicable. Third, the included evidence is geographically and educationally uneven, with eight studies not specifying a national context and China the most frequently represented specified setting. Fourth, many studies used small samples, short interventions, and single cohorts, and the corpus was heavily concentrated on ChatGPT. Fifth, the original screening materials did not retain item-level independent decisions needed for a retrospective inter-rater reliability calculation. These constraints mean that claims about higher-order competence, transfer, and generalisability should be interpreted cautiously.

The future research agenda is therefore concrete. Priority should be given to (1) multi-site, longitudinal, and adequately powered designs that compare AI, teacher, and blended feedback over time and test transfer to unaided performance; (2) studies extending beyond ChatGPT to multiple LLMs and to English-for-specific-purposes populations, including health and sport-science students; (3) prospective duplicate screening and independent quality appraisal with a prespecified agreement statistic; and (4) validated measures of feedback literacy, integrity behaviours, critical thinking, and equity impacts. Future primary studies should also report intervention dosage, prompting procedures, comparator conditions, scoring rubrics, and assessor agreement so that evidence can be compared across settings.

In direct response to the research questions, the synthesis indicates the following. For RQ1, generative AI is integrated across the writing cycle and reshapes composing processes, strategy use, and metacognition, but effects on higher-order rhetorical and critical competencies remain inconsistent. For RQ2, generative AI can provide useful feedback and increase revision activity, with several studies indicating that combined AI-plus-teacher arrangements outperform AI or teacher feedback alone; however, its reliability as a standalone evaluator is limited and fairness concerns remain. For RQ3, learners generally perceive generative AI positively and use it pragmatically for brainstorming, organising, editing, and language support, but persistent concerns about integrity, over-reliance, creativity, and accuracy make feedback literacy and ethical, critical use central academic-literacy challenges.

Conclusions

This systematic review synthesised 18 empirical studies on generative AI in English L2 writing across pedagogical transformation, feedback and assessment, and academic literacy. The review used PRISMA 2020 for reporting, appraised studies with MMAT 2018, and employed thematic narrative synthesis rather than meta-analysis because the included studies were heterogeneous in design, intervention, comparator, and outcome measurement. The evidence indicates that generative AI can improve surface-level accuracy, revision activity, engagement, and some self-regulatory or language outcomes, but effects on higher-order rhetorical, critical, and independent competence are inconsistent. AI is most defensible as a teacher-guided complement, particularly when learners are taught to evaluate, verify, and revise AI output rather than accept it uncritically. The conclusions are constrained by the small and short-duration primary studies, concentration on ChatGPT, limited database coverage, English-only retrieval, absence of prospective protocol registration, and the lack of preserved item-level dual-screening records for retrospective inter-rater reliability estimation. Longitudinal, multi-model, multi-site, and equity-sensitive research is needed to determine whether AI-assisted gains transfer to unaided writing performance.

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

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, 14. https://doi.org/10.4018/IJCALLT.367276

Algaraady, J., & Mahyoob, M. (2023). ChatGPT's capabilities in spotting and analyzing writing errors experienced by EFL learners. Arab World English Journal, Special Issue on CALL (9). https://doi.org/10.24093/awej/call9.1

Alkamel, M. A. A., & Alwagieh, N. A. S. (2024). Utilizing an adaptable artificial intelligence writing tool (ChatGPT) to enhance academic writing skills among Yemeni university EFL students. Social Sciences & Humanities Open, 10. https://doi.org/10.1016/j.ssaho.2024.101095

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

Barrot, J. S. (2024a). ChatGPT as a language learning tool: An emerging technology report. Technology, Knowledge and Learning, 29. https://doi.org/10.1007/s10758-023-09711-4

Barrot, J. S. (2024b). Leveraging ChatGPT in the writing classrooms: Theoretical and practical insights. Language Teaching Research Quarterly, 43. https://doi.org/10.32038/ltrq.2024.43.03

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Chapelle, C. A. (2025). Generative AI as game changer: Implications for language education. System, 132. https://doi.org/10.1016/j.system.2025.103672

Chuang, P.-L., & Yan, X. (2025). Language assessment in the era of generative artificial intelligence: Opportunities, challenges, and future directions. System, 134. https://doi.org/10.1016/j.system.2025.103846

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

De Wilde, V. (2024). Can novice teachers detect AI-generated texts in EFL writing? ELT Journal, 78. https://doi.org/10.1093/elt/ccae031

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, 34. https://doi.org/10.1111/ijal.12561

Dong, L. (2024). ChatGPT in language writing education: Reflections and a research agenda for a ChatGPT feedback engagement framework. Language Teaching Research Quarterly, 43. https://doi.org/10.32038/ltrq.2024.43.07

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. 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

Gao, J., Zhang, J., & Li, Y. (2025). Do AI chatbot-integrated writing tasks influence writing self-efficacy and critical thinking ability? An exploratory study. Computers and Education: Artificial Intelligence. https://doi.org/10.1016/j.caeai.2025.100472

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, 20(1). 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. The Internet and Higher Education, 63. https://doi.org/10.1016/j.iheduc.2024.100962

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, 56. https://doi.org/10.1111/bjet.13540

Gupta, R., Nair, K., Mishra, M., Ibrahim, B., & Bhardwaj, S. (2024). Adoption and impacts of generative artificial intelligence: Theoretical underpinnings and research agenda. International Journal of Information Management Data Insights, 4. https://doi.org/10.1016/j.jjimei.2024.100232

Hong, Q. N., Pluye, P., Bujold, M., & Vedel, I. (2018). Mixed Methods Appraisal Tool (MMAT), version 2018. Canadian Intellectual Property Office, Industry Canada.

Jeon, E.-Y. (2025). Artificial intelligence in ESL/EFL education: Evidence from recent reviews (2024–2025). International Journal of Learning, Teaching and Educational Research, 24(10). https://doi.org/10.26803/ijlter.24.10.24

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. https://doi.org/10.1007/s10639-024-12574-6

Kartal, G. (2024). The influence of ChatGPT on thinking skills and creativity of EFL student teachers: A narrative inquiry. Journal of Education for Teaching, 50. https://doi.org/10.1080/02607476.2024.2326502

Kohnke, L., Zou, D., & Su, F. (2025). Exploring the potential of GenAI for personalised English teaching: Learners' experiences and perceptions. Computers and Education: Artificial Intelligence, 8. https://doi.org/10.1016/j.caeai.2025.100371

Kurt, G., & Kurt, Y. (2024). Enhancing L2 writing skills: ChatGPT as an automated feedback tool. Journal of Information Technology Education: Research, 23. https://doi.org/10.28945/5370

Lai, X., Lai, Y., Chen, J., Huang, S., Gao, Q., & Huang, C. (2025). Evaluation strategies for large language model-based models in exercise and health coaching: Scoping review. Journal of Medical Internet Research, 27. https://doi.org/10.2196/79217

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions. PLoS Medicine, 6(7), e1000100. https://doi.org/10.1371/journal.pmed.1000100

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

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. https://doi.org/10.1016/j.caeai.2024.100313

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

Nugroho, A., Andriyanti, E., Widodo, P., & Mutiaraningrum, I. (2025). Students' appraisals post-ChatGPT use: Students' narrative after using ChatGPT for writing. Innovations in Education and Teaching International. https://doi.org/10.1080/14703297.2024.2319184

Pack, A., & Maloney, J. (2023). Potential affordances of generative AI in language education: Demonstrations and an evaluative framework. Teaching English with Technology, 23(2). https://doi.org/10.56297/BUKA4060/VRRO1747

Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. Computers and Education: Artificial Intelligence, 6. https://doi.org/10.1016/j.caeai.2024.100234

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

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). https://doi.org/10.53761/1.20.02.07

Punar Özçelik, N., & Yangın Ekşi, G. (2024). Cultivating writing skills: The role of ChatGPT as a learning assistant-A case study. Smart Learning Environments, 11. https://doi.org/10.1186/s40561-024-00296-8

Seo, J.-Y. (2024). Exploring the educational potential of ChatGPT: AI-assisted narrative writing for EFL college students. Language Teaching Research Quarterly, 43. https://doi.org/10.32038/ltrq.2024.43.01

Tafazoli, D. (2024). Exploring the potential of generative AI in democratizing English language education. Computers and Education: Artificial Intelligence, 7. https://doi.org/10.1016/j.caeai.2024.100275

Teckwani, S. H., Wong, A. H., Luke, N. V., & Low, I. C. C. (2024). Accuracy and reliability of large language models in assessing learning outcomes achievement across cognitive domains. Advances in Physiology Education, 48(4), 904–914. https://doi.org/10.1152/advan.00137.2024

Teng, M. F. (2024). A systematic review of ChatGPT for English as a foreign language writing: Opportunities, challenges, and recommendations. International Journal of TESOL Studies, 6. 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

Tica, L., & Krsmanović, I. (2024). Overcoming the writer's block? Exploring students' motivation and perspectives on using ChatGPT as a writing assistance tool in ESP. ELOPE: English Language Overseas Perspectives and Enquiries, 21(1), 129–149. https://doi.org/10.4312/ELOPE.21.1.129-149

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

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, 29. 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, 56. https://doi.org/10.1111/bjet.70018

Wang, Y., & Wang, X. (2024). Artificial intelligence in physical education: Comprehensive review and future teacher training strategies. Frontiers in Public Health, 12, 1484848. https://doi.org/10.3389/fpubh.2024.1484848

Woo, D. J., Guo, K., & Susanto, H. (2025). EFL secondary students' use of ChatGPT for writing task completion pathways. The Journal of Educational Research, 118. https://doi.org/10.1080/00220671.2025.2510382

Zhang, X., & Umeanowai, K. O. (2025). Exploring the transformative influence of artificial intelligence in EFL context: A comprehensive bibliometric analysis. Education and Information Technologies, 30. https://doi.org/10.1007/s10639-024-12937-z

Zhong, Q., Jiang, J., Bai, W., Yin, Z., Liao, Z., & Zhong, X. (2025). Application of digital-intelligent technologies in physical education: A systematic review. Frontiers in Public Health, 13, 1626603. https://doi.org/10.3389/fpubh.2025.1626603