p-ISSN: 2338-4174
e-ISSN: 2829-3584

GENERATIVE AI-SUPPORTED FEEDBACK IN EFL WRITING (A SYSTEMATIC LITERATURE REVIEW)

Yetty Tri Putri
Politeknik Negeri Lhokseumawe
yettykirm4n@gmail.com

Abstract: This study aims to systematically review the use of Generative Artificial Intelligence (GenAI)-supported feedback in English as a Foreign Language (EFL) writing. A Systematic Literature Review (SLR) was conducted following the PRISMA 2020 guidelines through the identification, screening, eligibility, and synthesis of relevant studies. Data were collected from peer-reviewed articles indexed in ERIC, Crossref, and major academic databases, resulting in 20 eligible empirical studies published between December 2022 and July 2026. The findings show that GenAI-supported feedback improves revision quality, grammatical accuracy, vocabulary, coherence, and learner engagement. However, teacher feedback remains essential for supporting higher-order writing skills, while combining AI and teacher feedback provides the most effective learning outcomes. Overall, GenAI serves as a valuable tool to support feedback and revision rather than replace teachers in EFL writing instruction.

Keywords: *Generative AI, EFL Writing, Systematic Literature Review*

Abstrak: *Penelitian ini bertujuan mengkaji secara sistematis penggunaan umpan balik berbantuan Generative Artificial Intelligence (GenAI) dalam pembelajaran menulis Bahasa Inggris sebagai Bahasa Asing (English as a Foreign Language/EFL). Penelitian menggunakan metode Systematic Literature Review (SLR) dengan mengacu pada pedoman PRISMA 2020 melalui tahapan identifikasi, penyaringan, penilaian kelayakan, dan sintesis artikel. Data dikumpulkan melalui penelusuran artikel ilmiah terindeks ERIC, Crossref, dan basis data akademik utama, sehingga diperoleh 20 artikel empiris yang memenuhi kriteria inklusi. Hasil kajian menunjukkan bahwa umpan balik berbantuan GenAI mampu meningkatkan kualitas revisi, ketepatan tata bahasa, kosakata, koherensi, dan keterlibatan belajar. Namun, umpan balik guru tetap diperlukan untuk mendukung keterampilan menulis tingkat tinggi. Oleh karena itu, GenAI berperan sebagai pendukung proses umpan balik dan revisi, bukan sebagai pengganti guru dalam pembelajaran menulis EFL.*

Kata Kunci: GenAI, Menulis EFL, Tinjauan Literatur Sistematis

1. Introduction

Writing is one of the most demanding components of English language education because learners must coordinate ideas, organization, vocabulary, grammar, audience awareness, and revision decisions at the same time. For EFL learners, the challenge is intensified by limited exposure to English outside the classroom and by the need to manage both language form and academic meaning. Feedback is therefore central to writing development. Effective feedback clarifies goals, identifies the gap between current and desired performance, and gives learners usable directions for revision (Hattie & Timperley, 2007). Meta-analytic evidence also indicates that feedback has a substantial influence on learning, although its effects vary according to content, timing, and the learner's ability to act on it (Wisniewski et al., 2020).

In practice, high-quality writing feedback is difficult to provide consistently. Teachers must read lengthy drafts, diagnose local and global problems, formulate understandable comments, and respond within a time frame that still allows revision. These demands become especially serious in large classes and in contexts where one teacher is responsible for many students. Automated writing evaluation systems partially address this problem, but conventional tools have historically emphasized grammar, spelling, and pre-programmed categories. Large language models and generative AI systems have expanded these possibilities by producing conversational, context-responsive comments, examples, explanations, and alternative formulations (Barrot, 2023).

Since the public release of ChatGPT in late 2022, the use of GenAI in English language education has grown rapidly. A broad systematic review of 70 empirical ESL/EFL studies found that writing was the most frequently investigated language domain, while rigorous performance-based evidence remained comparatively limited (Lo et al., 2024). Newer classroom studies have since examined AI-generated corrective feedback, AI-supported peer review, model texts, prompt design, hybrid feedback, and learner engagement. This expanding evidence base creates a need for a focused synthesis that does not treat all uses of ChatGPT as equivalent, but instead examines the specific pedagogical act of giving and using feedback in EFL writing.

The central issue is not whether GenAI can generate comments, but whether its feedback helps learners make accurate, meaningful, and transferable revisions. Some studies report gains in writing quality and motivation (Mahapatra, 2024; Song & Song, 2023), whereas others show that human feedback remains superior for accuracy, prioritization, contextual judgment, and higher-order guidance (Steiss et al., 2024). Hybrid feedback studies further suggest that AI and teachers may contribute different forms of value (Zhang et al., 2025). These mixed findings require a structured review of outcomes and implementation conditions.

This study therefore conducts a focused systematic literature review of empirical research on GenAI-supported feedback in EFL writing education from December 2022 to July 2026. It addresses three research questions: (1) What contexts, participants, tools, and research designs characterize the evidence base? (2) What effects are reported for writing performance, feedback uptake, motivation, and engagement? and (3) What limitations and pedagogical conditions influence the effective use of GenAI feedback? By answering these questions, the review aims to provide an evidence-informed framework for teachers, teacher educators, researchers, and institutions considering responsible integration of GenAI into English writing instruction.

2. Literature Review

2.1. Feedback in Second-Language Writing

Feedback in L2 writing is most useful when it is linked to subsequent action. Comments that merely identify errors may increase awareness, but they do not automatically produce learning. Learners need to interpret the message, evaluate its credibility, decide what to change, and understand how the revision relates to future writing. Carless and Boud (2018) describe these capacities as feedback literacy: appreciating feedback, making judgments, managing affect, and taking action. In GenAI-supported environments, feedback literacy is especially important because the learner receives fluent and confident language that may still be incomplete, overly general, or incorrect.

Writing feedback can address lower-order concerns, such as spelling, grammar, punctuation, and vocabulary, or higher-order concerns, such as purpose, argument, organization, evidence, coherence, and audience. Automated systems have traditionally performed better on lower-order concerns because these features are more readily detected by

language-processing algorithms. Generative models can discuss higher-order issues in natural language, but the quality of those comments depends on prompt specificity, model capability, task context, and the quality of the student's draft. Thus, the apparent breadth of AI feedback should not be confused with consistently reliable pedagogical judgment.

Feedback can also be considered dialogic rather than one-directional. Learners may ask for clarification, request examples, challenge a suggestion, or submit a revised passage for further review. GenAI makes such iterative interaction technically easy and available at any time. The educational value of this interaction, however, depends on whether learners use the dialogue to think, compare, and revise, or simply accept generated text. This distinction between cognitive engagement and passive adoption is central to evaluating AI-supported writing pedagogy.

2.2. Generative AI as a Writing Feedback Resource

GenAI can perform several feedback functions in EFL writing. It can identify language errors, explain grammar, suggest lexical alternatives, comment on coherence, compare a draft with a rubric, generate model paragraphs, and support repeated revision. Guo and Wang (2024), for example, found that ChatGPT generated substantially more feedback than teachers and distributed attention more evenly across content, organization, and language. Quantity, however, was not equivalent to pedagogical quality. Teachers and ChatGPT used different feedback types, leading the authors to recommend human-machine collaboration rather than substitution.

The immediacy and scalability of AI feedback are particularly attractive in large classes. Mahapatra (2024) demonstrated that ChatGPT-supported formative feedback could improve academic writing in a large undergraduate ESL setting when students received preparation and used the tool for self- and peer-assessment. Meyer et al. (2024) similarly found small but meaningful positive effects on Grade 10 EFL learners' revision performance, task motivation, and positive emotions. These findings suggest that rapid access to feedback can strengthen both the cognitive and affective conditions for revision.

At the same time, concerns include hallucinated explanations, inappropriate rewriting, changes in intended meaning, homogenized style, privacy risks, and academic integrity. Barrot (2023) argues that the potentials of ChatGPT for L2 writing must be considered alongside these limitations. A broader scoping review of large language models in education likewise identified practical and ethical challenges involving reliability, transparency, bias, privacy, and over-reliance (Yan et al., 2024). Consequently, effective classroom use requires instructional design, not merely access to a chatbot.

2.3. Rationale for the Present Review

Existing reviews provide valuable broad perspectives. Lo et al. (2024) mapped the early empirical literature on ChatGPT across all ESL/EFL domains, while Jiao and Wang (2026) synthesized 34 studies of large language models for formative writing feedback across educational contexts through February 2026. The present review is narrower in population and broader in pedagogical interpretation: it focuses specifically on EFL/ESL writing feedback, includes both performance and engagement evidence, and compares AI-only, human-comparison, peer-supported, and hybrid designs. This focus makes it possible to identify what GenAI feedback does well, where human mediation remains necessary, and which learner capacities shape effective uptake.

The review also follows the growing use of systematic and structured literature-review methods in Indonesian scholarship, where explicit search boundaries, transparent screening, thematic categorization, and research-gap mapping are used to synthesize heterogeneous evidence (A'zizah, 2026; Arifai et al., 2025; Mariana et al., 2025). Although the reviewed

subject areas differ, these studies reinforce the methodological value of making selection logic and synthesis procedures visible. In a rapidly evolving field such as GenAI-supported language learning, systematic synthesis is important because claims can quickly outpace evidence and because changes in models, prompts, and classroom policies can affect the generalizability of results.

3. Research Method

3.1. Review Design and Search Procedure

The review followed the identification, screening, eligibility, and inclusion logic of PRISMA 2020 (Page et al., 2021). Searches were conducted between 28 and 31 July 2026 across ERIC-indexed records, Crossref, and official publisher platforms, including SpringerLink, ScienceDirect, Taylor & Francis, Wiley, Sage, Frontiers, and open-access journal portals. Backward and forward citation tracking was used to locate additional empirical studies. The principal search string combined terms for the technology, population, and pedagogical function: (ChatGPT OR generative AI OR large language model OR GPT OR AI writing assistant) AND (EFL OR ESL OR second language OR foreign language) AND (writing OR composition OR academic writing) AND (feedback OR corrective feedback OR formative feedback OR revision).

Search results were logged in a review worksheet and deduplicated manually by DOI, title, and author. Title-and-abstract screening was followed by assessment of full publisher records or full texts. Twenty empirical studies were retained because they directly addressed the review questions and provided verifiable information on participants, feedback arrangements, and outcomes. This study was designed as a focused SLR for pedagogical synthesis rather than an exhaustive bibliometric census. Exact hit totals from open search interfaces were not treated as reproducible because rankings and displayed result counts change over time; this constraint is reported to avoid presenting artificial precision.

Because this review was conducted without direct institutional access to subscription indexing services such as Scopus and Web of Science, it prioritized ERIC-indexed records, Crossref metadata, accessible publisher records, and traceable peer-reviewed sources. This limitation is reported explicitly and should be considered when interpreting coverage. Nevertheless, citation chaining and searches across several major publishers were used to reduce the risk of missing influential studies.

Table 1. Inclusion and Exclusion Criteria

Criterion	Included	Excluded
Publication period	December 2022-July 2026	Published before the public emergence of ChatGPT/GenAI
Population and context	EFL/ESL learners or teachers in English writing education	Non-language-learning populations or writing outside an educational context
Intervention/focus	GenAI-generated or GenAI-supported feedback, revision, peer review, or model-text feedback	AI used only for detection, scoring, translation, or text generation without a feedback dimension

Criterion	Included	Excluded
Study type	Peer-reviewed empirical research with qualitative, quantitative, or mixed data	Reviews, editorials, conceptual papers, theses, and non-peer-reviewed reports
Outcome	Writing quality, revision, feedback uptake, engagement, motivation, perceptions, or feedback quality	No outcome relevant to writing feedback or learning
Language	English-language publication	Publication not available in English

3.2. Data Extraction and Quality Appraisal

A structured extraction sheet was used to record author and year, country, educational level, sample, study design, GenAI tool, feedback arrangement, outcomes, and limitations. Studies were grouped by feedback model: AI-only feedback, AI compared with teacher or peer feedback, AI-supported peer feedback, and hybrid AI-human feedback. Outcome categories included writing performance, revision quality, feedback quality, engagement and motivation, learner perceptions, and ethical or practical concerns.

Methodological quality was considered using design-appropriate questions adapted from the Mixed Methods Appraisal Tool (Hong et al., 2018). Rather than producing a single numerical score, the appraisal examined the appropriateness of the design, clarity of sampling, validity of measures, completeness of outcome reporting, integration of mixed-method components, and alignment between evidence and conclusions. Common limitations were short interventions, single-institution samples, small qualitative samples, incomplete reporting of model versions and prompts, and limited delayed measurement of independent writing. No study was excluded solely because of quality; instead, methodological limitations were incorporated into the interpretation of findings.

3.3. Synthesis Method

The studies were too heterogeneous in intervention length, assessment rubrics, participant proficiency, comparison conditions, and reported statistics for a defensible meta-analysis. A narrative thematic synthesis was therefore conducted. Descriptive patterns were first mapped across settings, tools, and designs. Findings were then coded into four themes: (1) improvement in language accuracy and revision, (2) engagement and affective outcomes, (3) differences between AI and human feedback, and (4) conditions for responsible and effective implementation. Contradictory findings were retained and examined in relation to task type, feedback arrangement, learner characteristics, and study design.

4. Results and Discussion

4.1. Characteristics of the Included Studies

The 20 studies were published between 2023 and 2026, demonstrating the rapid growth of the field. Higher education dominated the evidence base, with only a small number of studies conducted in secondary education. Research was concentrated in East Asia and the Middle East, particularly China, Hong Kong, Saudi Arabia, Jordan, and Egypt. This concentration reflects active adoption of AI in EFL contexts but limits transferability to primary education, vocational education, rural schools, and under-resourced settings.

ChatGPT or another GPT-based system was used in most studies. One study examined Grammarly's generative and corrective functions, while others used integrated chatbots or purpose-designed prompting arrangements. Research designs included quasi-experiments, randomized or controlled interventions, mixed-method classroom studies, comparative analyses of feedback quality, surveys, interviews, and qualitative studies of learner experience. Sample sizes ranged from small pilot groups to a randomized study of 459 secondary students.

The studies also differed in what counted as feedback. Some asked ChatGPT to correct or comment on drafts, some compared its comments with teacher feedback, some used AI to support peer review, and others used generated model texts or hybrid teacher-AI sequences. These variations are educationally important because a result for AI-assisted peer feedback cannot be assumed to apply to unguided chatbot use. Table 2 summarizes the studies and their central contribution.

Table 2. Summary of Included Empirical Studies

Study	Context/Sample	Design/Tool	Main Finding
Song & Song (2023)	China; 50 university EFL students	Mixed-method controlled intervention; ChatGPT	Twelve-week AI-assisted instruction improved global writing, organization, grammar, vocabulary, and motivation; learners also raised concerns about accuracy and dependence.
Su et al. (2023)	University EFL argumentative writing	Classroom collaboration with ChatGPT	ChatGPT supported idea development and revision, but effective collaboration required purposeful prompting, evaluation, and teacher-designed tasks.
Guo & Wang (2024)	China; 50 essays and five EFL teachers	Comparative feedback-content analysis	ChatGPT produced more feedback and distributed attention across content, organization, and language; teachers provided more selective and context-sensitive feedback.
Mahapatra (2024)	India; 134 undergraduate ESL students	Mixed-method quasi-experiment; ChatGPT formative feedback	Students receiving trained ChatGPT-supported self- and peer-feedback achieved stronger post-test and delayed post-test writing performance and reported positive experiences.
Meyer et al. (2024)	Germany; 459 Grade 10 EFL students	Randomized feedback study; GPT-3.5	LLM feedback produced small positive effects on revision performance and moderate positive effects on motivation and emotions compared with revision without feedback.

Study	Context/Sample	Design/Tool	Main Finding
Steiss et al. (2024)	Secondary essays; 200 human and 200 AI feedback samples	Comparative quality evaluation	Human feedback was stronger on four of five quality dimensions; AI was competitive for criteria-based feedback and offered major time advantages.
Guo et al. (2024)	University EFL learners	AI-supported peer-feedback intervention	AI support improved the quality of peer comments and contributed to writing development by scaffolding feedback generation.
Woo et al. (2024)	Secondary EFL learners	ChatGPT-supported writing instruction	Students reported motivation and satisfaction, but cognitive load and prompt-use demands showed the need for explicit instructional support.
Zeevy-Solovey (2024)	Israel; 30 EFL learners	Pilot comparison of peer, ChatGPT, and teacher WCF	Students considered teacher and ChatGPT feedback effective, but preferred teacher feedback or a teacher-ChatGPT combination.
Polakova & Ivenz (2024)	110 university EFL students	Mixed-method quasi-experiment; ChatGPT feedback	ChatGPT feedback supported writing development and positive attitudes, while students expressed concerns about reliability and the need to verify suggestions.
Alsofyani & Barzanji (2025)	Saudi Arabia; 102 university EFL students	Pretest-posttest comparison; AI vs teacher feedback	ChatGPT feedback was statistically comparable to teacher feedback for overall writing outcomes; student perceptions were positive.
Alnemrat et al. (2025)	Jordan; mixed-proficiency EFL learners	Quantitative AI-vs-teacher feedback study	Both feedback conditions supported argumentative revision; proficiency shaped responsiveness, and guided AI use was recommended as preliminary support.
Mekheimer (2025)	Egypt; 60 postgraduate EFL students	Quasi-experiment; Grammarly with GenAI features	AI-assisted feedback improved overall proficiency, language use, organization, revision frequency, and confidence, but did not show equivalent gains for content.

Study	Context/Sample	Design/Tool	Main Finding
Zou et al. (2025)	University EFL learners	Teacher-vs-ChatGPT feedback uptake comparison	Teacher feedback led to stronger uptake for some corrections, while ChatGPT contributed useful structural and language revisions; combined use was recommended.
Zhang et al. (2025)	China; 60 university EFL students	Twelve-week GenAI-only vs hybrid feedback experiment	AI-only feedback improved grammar and sentence variety; hybrid feedback produced much larger gains in organization, critical thinking, motivation, and perceived usefulness.
Lu (2025)	University EFL learners	ChatGPT-generated model texts as feedback	Model-text feedback improved content, organization, vocabulary, and grammar; students noted occasional lack of contextual relevance and unnatural style.
Teng (2025)	40 EFL undergraduates	Semester-long mixed-method study	Metacognitive awareness predicted more critical and effective use of ChatGPT feedback; low-awareness use was more likely to involve copying.
Teng & Huang (2026)	China; 174 university EFL students	ChatGPT-supported vs peer-only feedback	ChatGPT-supported learners improved affective and behavioral engagement and writing performance more than the peer-only group.
Han & Wang (2026)	China; 35 first-year university EFL students	Within-subject mixed-method comparison; teacher vs ChatGPT feedback	ChatGPT-supported revisions received higher scores, while teacher feedback was perceived as easier, more emotionally supportive, and more context-sensitive; AI feedback was comprehensive but often lengthy and generic.
Dewi et al. (2026)	Indonesia; 81 undergraduate EFL students	Questionnaire and interviews; AI vs peer feedback perceptions	Learners valued AI for immediate language-level support and peers for interaction, ideas, and contextual meaning; strategic combination of both sources was recommended.

4.2. Effects on Writing Performance and Revision

The most consistent finding is that GenAI feedback can support measurable improvement in writing and revision, especially when learners are trained and required to produce multiple drafts. Song and Song (2023) reported significant gains in organization, coherence, grammar, vocabulary, and motivation after a twelve-week intervention. Mahapatra (2024) similarly found that ChatGPT-supported formative feedback improved academic writing in both immediate and delayed post-tests. These results indicate that AI feedback can contribute to learning rather than merely polishing a single product when it is embedded in repeated assessment, self-review, and peer-feedback routines.

The large randomized study by Meyer et al. (2024) adds important evidence because it compared AI feedback with revision without feedback in authentic Grade 10 EFL classes. Although the effect on revision performance was small, the intervention improved task motivation and positive emotions. In educational practice, a small effect delivered quickly and at scale may still be valuable, particularly when students would otherwise receive no individualized response. However, this benefit should not be interpreted as evidence that AI feedback is equivalent to expert teaching.

Across studies, gains were strongest for linguistic accuracy, vocabulary, sentence variety, local coherence, and the efficiency of revision. Mekheimer (2025) found that AI-assisted feedback improved language use and organization and was associated with greater revision frequency, while content scores showed less improvement. Zhang et al. (2025) likewise found that AI-only feedback benefited grammar and sentence variety but had limited influence on critical thinking and organization. These patterns suggest that GenAI is particularly effective as a first-response mechanism for lower-order concerns, freeing class time for higher-order instruction.

The use of model texts provides an additional feedback pathway. Lu (2025) found that ChatGPT-generated model texts supported improvement across content, organization, vocabulary, and grammar by allowing learners to compare their drafts with an alternative representation of the task. This approach moves beyond error correction and can support noticing of rhetorical patterns. Yet learners also reported that generated texts could feel unnatural or insufficiently connected to the specific context. Model texts are therefore most useful as objects for comparison and critique rather than templates for imitation.

4.3. Motivation, Engagement, and Learner Agency

Many studies reported positive affective outcomes. Immediate feedback reduced the delay between drafting and revision, enabled private questioning, and gave learners repeated opportunities to clarify language. Song and Song (2023) documented increased writing motivation, Meyer et al. (2024) found positive effects on task motivation and emotions, and Teng and Huang (2026) reported improvements in affective and behavioral engagement. Mahapatra (2024) also found that students valued support for content, organization, grammar, and focused revision.

These benefits are not automatic. Learner agency determines whether feedback becomes learning. Teng (2025) showed that students with stronger metacognitive awareness were more likely to evaluate suggestions, connect feedback with goals, and use ChatGPT strategically. Students with lower metacognitive awareness were more likely to copy language without deep processing. This finding explains why the same tool can produce very different outcomes across learners and supports the inclusion of reflection prompts, revision rationales, and verification tasks in classroom design.

Prompting is another form of learner agency. Vague requests such as 'improve my essay' may produce extensive rewriting that obscures the learner's intention and makes it difficult to distinguish feedback from authorship. More focused prompts can ask the model to

identify three organization problems, explain a grammar pattern without rewriting, compare a paragraph with a rubric, or ask questions that help the learner elaborate an argument. Explicit prompt instruction reduces cognitive load and increases the chance that output is actionable and aligned with the lesson objective.

A further issue is emotional trust. Learners often appreciate the nonjudgmental tone and constant availability of a chatbot, but fluent output may create excessive confidence in inaccurate advice. Feedback literacy must therefore include calibrated trust: students should neither reject AI automatically nor accept it uncritically. They should compare suggestions with course materials, dictionaries, corpora, rubrics, peer discussion, and teacher guidance.

4.4. AI Feedback Compared with Teacher and Peer Feedback

Comparative studies show that AI and human feedback have different strengths. Guo and Wang (2024) found that ChatGPT generated more feedback and covered content, organization, and language more evenly, whereas teachers were more selective. Steiss et al. (2024) found that trained human evaluators produced higher-quality feedback on accuracy, clarity of directions, prioritization, and supportive tone, with AI performing competitively mainly on criteria-based feedback. These findings indicate that breadth and speed should not be confused with instructional judgment.

Students also distinguish between revision outcomes, usefulness, and preference. Zeevy-Solovey (2024) found that learners regarded both teacher and ChatGPT feedback as effective but preferred teacher feedback or a combination of teacher and AI feedback. Han and Wang (2026) similarly found that ChatGPT-supported revisions received higher scores, while students rated teacher feedback as easier to use and valued its emotional clarity and contextual sensitivity. Teacher comments carry contextual knowledge of the learner, the curriculum, previous drafts, and classroom expectations. AI comments may be immediate and detailed, but they do not possess the teacher's long-term understanding of the writer or responsibility for curricular decisions.

Evidence for hybrid feedback is particularly strong. Zhang et al. (2025) compared GenAI-only feedback with feedback combining GenAI and a human tutor. The hybrid group achieved much greater gains in organization, critical thinking, and sentence variety and reported stronger motivation. This pattern can be explained by complementarity: AI rapidly addresses frequent language and task-level concerns, while teachers interpret meaning, prioritize revision, ask probing questions, and provide emotional and strategic support.

AI-supported peer feedback offers another hybrid arrangement. Guo et al. (2024) found that an AI-supported approach improved students' ability to generate useful peer comments and supported writing development. In Indonesia, Dewi et al. (2026) found that learners valued AI for immediate language-focused support and peer feedback for interpersonal interaction, idea development, and contextual meaning. In this model, AI does not directly replace the teacher or peer; it scaffolds the learner's feedback practice. Such designs may strengthen feedback literacy because students must judge both the draft and the AI suggestion before communicating feedback to a peer.

4.5. Challenges and Risks

The first challenge is reliability. GenAI can provide inaccurate, unnecessary, or contradictory comments and may change intended meaning while presenting the revision as an improvement. This risk is especially serious for lower-proficiency learners who may lack the language knowledge needed to evaluate sophisticated output. Studies therefore consistently recommend teacher mediation and learner training rather than unrestricted independent use.

The second challenge is over-reliance. When students routinely outsource idea generation, drafting, and revision, they may produce polished products without developing independent writing processes. The risk is not limited to academic dishonesty; it also concerns reduced struggle, reflection, and decision making. Effective assignments should preserve evidence of thinking through outlines, drafts, revision notes, oral explanation, in-class writing, and comparison of accepted and rejected feedback.

The third challenge is ethical and institutional. Students may paste personal information, unpublished research, or classmates' writing into external systems without understanding data practices. Institutions need clear guidance on privacy, acceptable use, attribution, and assessment. Yan et al. (2024) emphasize that large language models raise practical and ethical concerns involving transparency, fairness, privacy, and responsibility. These issues should be integrated into writing pedagogy rather than treated as technical warnings outside the curriculum.

The fourth challenge is research validity. Model outputs change over time, yet many studies do not report the exact model version, system settings, prompt wording, conversation history, or whether regenerated outputs were used. Short interventions may also capture novelty rather than stable learning. Future evidence will be more comparable when researchers publish prompts, define the feedback workflow precisely, use common writing measures, and assess independent writing after AI access is removed.

4.6. A Pedagogical Model for EFL Writing

The synthesis supports a three-stage model. In the first stage, AI acts as a first responder. Students submit a self-written draft and request focused feedback linked to a rubric, such as grammar patterns, paragraph unity, cohesion, or clarity. The model should be instructed not to rewrite the entire text and to explain the reason for each suggestion. This stage provides speed and repeated practice.

In the second stage, the learner acts as an evaluator. Students classify each suggestion as accepted, adapted, or rejected and write a brief reason. They verify uncertain information with trusted sources and identify any changes in meaning. This stage converts AI output into a feedback literacy task and makes learner thinking visible.

In the third stage, the teacher acts as the strategist. Teacher feedback concentrates on argument, evidence, audience, organization, voice, and recurring learner needs rather than repeating every surface correction. The teacher can also review a sample of AI suggestions, discuss prompt quality, and address ethical decisions. This staged model preserves the teacher's pedagogical authority while using AI to increase the frequency and responsiveness of formative support.

For schools with limited connectivity or unequal device access, the model can be implemented through teacher-mediated demonstrations, shared devices, printed AI feedback samples, or group analysis. Equity must be considered because a feedback system that benefits only students with private devices, stronger digital literacy, or paid model access may widen existing differences.

5. Conclusion

This focused systematic literature review synthesized 20 empirical studies on GenAI-supported feedback in EFL writing education published between 2023 and July 2026. The evidence indicates that GenAI can improve writing revision, linguistic accuracy, vocabulary, coherence, motivation, and engagement, particularly by providing immediate and iterative support. The strongest and most consistent benefits concern lower-order and task-level features of writing. Evidence is less consistent for argument development, critical thinking, contextual interpretation, rhetorical organization, and prioritization of revision.

Teacher feedback remains essential because it is grounded in curriculum, learner history, classroom relationships, and professional judgment. The most promising evidence comes from hybrid designs that combine the speed and availability of AI with human guidance. GenAI should therefore be used as a first-response and revision-support tool, not as an autonomous evaluator or replacement for English teachers.

The effectiveness of GenAI feedback depends on prompt quality, feedback literacy, metacognitive awareness, critical verification, ethical guidance, and opportunities to write independently. Institutions should support teachers with professional development and clear policies, while classroom tasks should require students to explain how they used, modified, or rejected AI suggestions. Future research should include longer interventions, delayed post-tests, school-level contexts, diverse proficiency groups, transparent prompt reporting, and measures of unaided writing transfer.

6. Recommendations

English teachers are advised to introduce GenAI feedback gradually through focused, rubric-based tasks rather than unrestricted essay rewriting. Students should be taught to ask specific questions, verify output, preserve their intended meaning, and document revision decisions. Teacher feedback can then focus on higher-order concerns and on patterns that require individualized explanation.

Teacher education programs and schools should develop practical guidance covering prompt design, feedback literacy, privacy, academic integrity, and equitable access. Researchers should report model versions, prompts, task conditions, and feedback sequences in sufficient detail for replication. Journals should also encourage transparent declarations regarding the use of GenAI in research and manuscript preparation.

References

A'zizah, A. (2026). Integrasi pendidikan agama dalam pembentukan karakter peserta didik: A systematic literature review. Jurnal Eksperimental, 15(1), 1-10.

Alnemrat, A., Aldamen, H., Almashour, M., Al-Deaibes, M., & AlSharefeen, R. (2025). AI vs. teacher feedback on EFL argumentative writing: A quantitative study. Frontiers in Education, 10, 1614673. https://doi.org/10.3389/feduc.2025.1614673

Alsofyani, A. H., & Barzanji, A. M. (2025). The effects of ChatGPT-generated feedback on Saudi EFL learners' writing skills and perception at the tertiary level: A mixed-methods study. Journal of Educational Computing Research, 63(2), 431-463. https://doi.org/10.1177/07356331241307297

Arifai, M., Mariana, M., Fahlevi, H., Indriani, M., & Darwanis, D. (2025). Structured literature analysis on sustainability report disclosure in public sector organizations. MIX: Jurnal Ilmiah Manajemen, 15(2), 483-501. https://doi.org/10.22441/jurnal_mix.2025.v15i2.009

Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. Assessing Writing, 57, 100745. https://doi.org/10.1016/j.asw.2023.100745

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. Assessment & Evaluation in Higher Education, 43(8), 1315-1325. https://doi.org/10.1080/02602938.2018.1463354

Dewi, U., Daulay, S. H., & Ismahani, S. (2026). AI feedback vs. peer feedback in EFL writing: Exploring Indonesian students' perceptions. Asian-Pacific Journal of Second and Foreign Language Education, 11, 49. https://doi.org/10.1186/s40862-026-00424-6

Guo, K., Pan, M., Li, Y., & Lai, C. (2024). Effects of an AI-supported approach to peer feedback on university EFL students' feedback quality and writing ability. The Internet and Higher Education, 63, 100962. https://doi.org/10.1016/j.iheduc.2024.100962

Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. Education and Information Technologies, 29, 8435-8463. https://doi.org/10.1007/s10639-023-12146-0

Han, F., & Wang, Z. (2026). Teacher feedback and ChatGPT feedback on Chinese university EFL learners' English essay revision: A mixed-methods study. Computers and Education: Artificial Intelligence, 10, 100590. https://doi.org/10.1016/j.caeai.2026.100590

Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81-112. https://doi.org/10.3102/003465430298487

Hong, Q. N., Pluye, P., Fabregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., & Vedel, I. (2018). Mixed Methods Appraisal Tool (MMAT), version 2018: User guide. McGill University.

Jiao, Y., & Wang, Q. (2026). Large language models for formative feedback in writing instruction: A systematic review of classroom interventions, feedback quality, and student outcomes. Frontiers in Education, 11, 1834085. https://doi.org/10.3389/feduc.2026.1834085

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

Lu, D. (2025). Exploring the use of ChatGPT-generated model texts as a feedback instrument in EFL writing. Innovation in Language Learning and Teaching. https://doi.org/10.1080/17501229.2025.2525341

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

Mariana, M., Diana, D., Arifai, M., & Jannah, M. (2025). Public sector accounting reform: A systematic literature review. HEI EMA: Jurnal Riset Hukum, Ekonomi Islam, Ekonomi, Manajemen dan Akuntansi, 4(2), 132-141. https://doi.org/10.61393/heiema.v4i2.320

Mekheimer, M. (2025). Generative AI-assisted feedback and EFL writing: A study on proficiency, revision frequency and writing quality. Discover Education, 4, 170. https://doi.org/10.1007/s44217-025-00602-7

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. Computers and Education: Artificial Intelligence, 6, 100199. https://doi.org/10.1016/j.caeai.2023.100199

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., Hrobjartsson, 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

Polakova, P., & Ivenz, P. (2024). The impact of ChatGPT feedback on the development of EFL students' writing skills. Cogent Education, 11(1), 2410101. https://doi.org/10.1080/2331186X.2024.2410101

Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. Frontiers in Psychology, 14, 1260843. https://doi.org/10.3389/fpsyg.2023.1260843

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. Learning and Instruction, 91, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. Assessing Writing, 57, 100752. https://doi.org/10.1016/j.asw.2023.100752

Teng, M. F. (2025). Metacognitive awareness and EFL learners' perceptions and experiences in utilising ChatGPT for writing feedback. European Journal of Education, 60, e12811. https://doi.org/10.1111/ejed.12811

Teng, M. F., & Huang, J. (2026). Assessing ChatGPT feedback for EFL learners' engagement and writing performance. International Journal of Applied Linguistics. https://doi.org/10.1111/ijal.70155

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. Frontiers in Psychology, 10, 3087. https://doi.org/10.3389/fpsyg.2019.03087

Woo, D. J., Wang, D., Guo, K., & Susanto, H. (2024). Teaching EFL students to write with ChatGPT: Students' motivation to learn, cognitive load, and satisfaction with the learning process. Education and Information Technologies, 29, 24963-24990. https://doi.org/10.1007/s10639-024-12819-4

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gasevic, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. British Journal of Educational Technology, 55(1), 90-112. https://doi.org/10.1111/bjet.13370

Zeevy-Solovey, O. (2024). Comparing peer, ChatGPT, and teacher corrective feedback in EFL writing: Students' perceptions and preferences. Technology in Language Teaching & Learning, 6(3), 1482. https://doi.org/10.29140/tltl.v6n3.1482

Zhang, Z., Aubrey, S., Huang, X., & Chiu, T. K. F. (2025). The role of generative AI and hybrid feedback in improving L2 writing skills: A comparative study. Innovation in Language Learning and Teaching. https://doi.org/10.1080/17501229.2025.2503890

Zou, S., Guo, K., Wang, J., & Liu, Y. (2025). Investigating students' uptake of teacher- and ChatGPT-generated feedback in EFL writing: A comparison study. Computer Assisted Language Learning. https://doi.org/10.1080/09588221.2024.2447279