

A Systematic Review of Empirical Research about Beyond Automated Correction by Human–AI Feedback Partnership for Developing L2 Writing

Muhammad Imran¹, Zuraina Ali², Dr. Mohammad Musab Bin Azmat Ali²

¹ Senior Language Instructor, Sultan Qaboos University, Sultanate of Oman; PhD Candidate in English, Universiti Malaysia Pahang Al-Sultan Abdullah; Centre for Modern Languages, Universiti Malaysia Pahang Al-Sultan Abdullah, 26600 Pekan, Pahang, Malaysia; Department of English, Daffodil International University, Dhaka, Bangladesh.

² Department of English, Daffodil International University, Dhaka, Bangladesh.

Corresponding Author

Zuraina Ali

Email: zuraina@umpsa.edu.my

ORCID iD: 0000-0002-1527-6320

Abstract: This paper synthesizes empirical studies published from 2023 to 2026 to discuss the growing position of AI-aided writing feedback for L2 students. Drawing upon the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) statement, this study organizes and critically reviews past findings in the context of instructional transformation from existing AWE systems to LLM-empowered collaborative Human-AI writing feedback in L2 education; influence of learner engagement, trust, and AI writing feedback literacy on the adoptability and applicability of AI feedback; and (3) instructional and research limitations that need to be addressed in subsequent research. Our systematic review demonstrates that generative AI facilitates writing quality improvement, revision processes, learner autonomy and increased writing access by providing on-time, tailored and interactive support. Moreover, this review reveals that the contribution of tool-aided feedback to learning lies not just in the features of AI feedback alone, but more in the users' ability to critically understand, judge and utilize the suggested pieces of writing to advance writing competence within the teacher-supervised writing environment. Thus, the results would recommend the Human-AI Partnership framework, in which LLMs would substitute or at least assist the role of the human tutor rather than directly replace; at the same time, it would also shed light upon issues of conducting longitudinal, methodologically robust, and theoretically grounded studies regarding generative AI for L2 writing pedagogy.

Keywords: Artificial intelligence, generative AI, human–AI feedback, L2 writing, systematic review, PRISMA 2020, feedback literacy, large language models

1. Introduction

Generative artificial intelligence (GenAI), along with large language models (LLMs such as ChatGPT, Claude, Gemini, and Microsoft Copilot), has revolutionized second language (L2) writing education through altering the creation, comprehension and uptake of writing feedback in the entire process of writing. Unlike previous automated writing evaluation (AWE) systems, LLMs generate highly contextualized, interactive, and personalized feedback that goes into deep-level of language accuracy, providing insights into revision strategies, alternative phrasings, idea generation, questioning capabilities, and support for multiple drafting and revising cycles. Hence, AI has morphed into an intelligent partner with active participation in learners' writing development beyond a simple corrective tool.



This is a pedagogical shift in the use of AI in L2 writing. Old AWEs targeted at detecting and correcting grammar, vocabulary and punctuation, and sentence structure accuracy were developed to help teachers reduce marking load and students acquire immediate formative feedback. LLMs significantly extend the ability to use AI by enabling learners to establish an interactive dialogue with AI, clarify meanings, justify changes and revise writing through repetitive feedback cycles (Ali et al., 2025; Zhang et al., 2025; Yan & Zhang, 2024).

The interactive nature of LLMs has also redefined feedback in the L2 writing classroom; instead of merely serving as a tool for automated evaluation, it has become a key component of collaborative learning environments where students discuss and negotiate suggestions received from AI, critically analyze alternate versions provided and revise writing through continuous interaction (Ali et al., 2025). This helps develop Learner Autonomy, reflective revision, metacognition, and self-regulation of learning, and frees teachers to focus on higher-order language skills such as discipline-specific writing, critical thinking, ethical consideration, interpreting feedback, and so forth. Writing instruction is moving from a technology-assisted corrective model towards a human-AI feedback partnership.

Since the release of ChatGPT, there's been a surge of empirical work examining AI's impact on academic writing in numerous educational settings. Some studies have investigated the impacts of LLM-generated feedback on various aspects such as writing quality, revisions, learner engagement, feedback literacy, self-regulation of learning, writing confidence and teachers' pedagogical practices (Ali et al., 2025; Cengiz et al., 2025; Pretorius & Thewissen, 2026; Yang et al., 2024). Meanwhile, some scholars have explored challenges such as learner over-dependence, hallucinated feedback, academic integrity, transparency, bias and ethics associated with the use of AI in writing instruction (Chen et al., 2024; Yan & Zhang, 2024).

Despite the growing corpus of research, existing studies vary in educational contexts, participants, AI tools and interventions, methodological designs and outcome measures. Thus, leading to limitations in comparing findings and drawing consensus on best practices. Some studies compare AI-generated feedback versus teacher feedback, while others examine human-AI feedback, collaborative writing and interaction with conversational agents. Outcome measures have significantly diverged from writing accuracy and vocabulary development to student engagement, motivation, feedback literacy, writing revisions and self-regulation of learning. This methodological diversity provided rich insights into how LLMs can be integrated in L2 writing instruction, yet hindered the development of coherent pedagogical recommendations (Ali et al., 2025).

Therefore, a systematic synthesis of recent empirical evidence is much needed. Instead of taking AI as an independent technological tool, the present review views AI-supported feedback as a collaborative process between learners, teachers and intelligent writing systems. Based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines (Page et al., 2021), this review synthesises empirical evidence in the area of AI-supported feedback in L2 writing published between January 2023 and July 2026 by analyzing and integrating studies with respect to the evolution of AI-supported feedback in L2 writing. It further investigates the effects of learner engagement, trust in AI feedback and AI feedback literacy on effective take-up of AI-supported writing feedback. Also, the systematic review looks for and identifying methodological, theoretical and pedagogical lacunae in current research to provide a clear research agenda for the effective and ethical implementation of LLM-supported human-AI feedback partnerships in L2 writing instructions.

1.1 Research Questions

RQ	Research Focus	Key Analytical Dimensions	Expected Synthesis
RQ 1	Evolution of AI-supported feedback in L2 writing (2023–2026)	AWE to LLM feedback; corrective to interactive feedback; human–AI partnership; changing teacher role	Transformation of AI-supported feedback and emergence of human–AI collaborative models

RQ 2	Learner engagement and uptake of AI feedback	Engagement; trust; feedback uptake; AI feedback literacy; critical evaluation; learner agency	Influence of learner-related factors on the uptake and effectiveness of AI-generated feedback
RQ 3	Limitations and future directions	Methodological limitations; pedagogical gaps; theoretical weaknesses; ethical concerns; overreliance	Limitations in current evidence and priorities for future research and classroom implementation

1.2. Rationale

The emergence of AI and generate AI in educational context is opening varied paths for educators. However, due to its recent invasion, there is a dire need to develop a systematic review which encompasses authentic, valid and recent studies for future researchers as a reference and also as an inspiration to explore other domains of learning with AI. Therefore, a period of 2023 to 2026 has been chosen to include latest research so that readers can be sensitized, inspired and aware about latest AI trends to deal with rapid modern transformation.

2. Review of Literature

2.1 Evolution of AI-Supported Feedback in L2 Writing

Following the development of large language models (LLMs), the focus of writing feedback has moved on from surface-level corrective feedback provided by traditional Automated Writing Evaluation (AWE) tools to more sophisticated, conversational systems that guide learners through the writing process. Traditional AIs in the context of writing mainly centered on surface-level features of writing such as grammar, spelling, vocabulary, punctuation and sentence construction.

While these AIs gave learners prompt feedback and lessened the burden of marking for teachers, allowing more accessible formative assessment in a larger class (Shi & Aryadoust, 2024).

Then, feedback generated by traditional AWE tools was usually non-interactive and prescriptive, giving learners little chance to challenge it, seek clarification and learn through the revision process reflectively.

In recent years, we have seen a technological shift toward the use of AI in writing feedback generation. Various review studies of AI-supported written feedback have been published since 2024. The first systematic review by Shi and Aryadoust (2024) examined the effects of AI on written texts and found evidence that AI consistently enhances grammatical accuracy, vocabulary level and writing revision, but that there is still limited evidence to suggest it enhances writing development over time.

Building on recent work, Craven and Fredrick (2026) reviewed 185 empirical studies from late 2022 to early 2026 that investigate the impacts of LLM-based feedback on L2 writing. They recorded an unprecedented amount of research since the announcement of ChatGPT and identified that most studies demonstrated positive effects on writing quality, writing motivation, revision behaviors and accessibility of feedback. However, they found that LLM-based Feedback literature still focuses heavily on classroom-level.

Similar observations were made by Jiao and Wang (2026) in their systematic review of classroom intervention using LLM-based formative feedback in EFL contexts and they again identified that positive learning outcomes included improved learners' writing performance, self-regulated learning, and enhanced Revision Strategy and Feedback Literacy.

With regard to other feedback types, Ding et al. (2026) reviewed existing literature on using LLMs to provide written corrective feedback for L2 writers and said that "teachers may want to 'intervene and guide' the LLM and 'use AI-generated feedback in a more meaningful pedagogical context with learners' needs in mind.'" In brief, these review studies confirm technology's power to reshape writing pedagogy and instruction. But a number of issues, such as the

quality of generated text, ethical considerations like plagiarism and accountability, learners' adaptation to and use of the feedback, teachers' role, the optimal interaction design, and the impact of AI on the development of Higher-Order thinking skills remain to be clarified and discussed.

2.2 Human-AI Feedback Partnership

. This human–AI feedback partnership differs significantly from merely providing a score with Automated Writing Evaluation systems by integrating AI computational ability with teachers' expertise in pedagogy, content knowledge, situational and subjective judgment, and emotional communication skills. Several recent literature reviews provide evidence of collaborations between AI and teachers in supporting learners.

Karagz (2025), for example, conducted a systematic review of 22 empirical studies on AI-generated Feedback in English Writing Instruction and found evidence suggesting that Gen AI can help improve grammar, vocabulary, sentence construction and the overall revision process as a tool providing immediate and individualized feedback.

Therefore, AI-generated feedback should serve as an adjunct to teaching and this finding is echoed by Ding et al.'s (2025) systematic review of LLM-generated corrective feedback in writing for L2 Learners. They highlighted that LLMs stimulate Learning Agency by assisting Learners with iterative Revision, explanations, and writing dialogue throughout the composition process; however, concerns over LLMs providing learners with hallucinatory feedback, inconsistent feedback quality, and insufficient pedagogical judgment result in the recommendation to use teachers' oversight as necessary to support and legitimise and to encourage the critically evaluate LLM suggestions

Other works by Gomez et al. (2025) also provided evidence to demonstrate the value of hybrid learning approaches in learning beyond writing contexts, e.g., within general collaborative learning environments and decision-support scenarios, that effective human–AI interaction consists of seamless interaction, joint effort of different parties involved, and transparency & reciprocal adaptiveness. The partnership and interdependence of humans and intelligent systems enable collaboration where no single actor completely holds authority or capability.

A systematic review by Hussien (2025) investigating Human-GenAI Collaboration in Writing also suggested that Learners achieve the maximum possible benefits in their learning outcomes by viewing AI as a collaborator instead of an autonomous author, while acknowledging that teachers' feedback is irreplaceable to maintain learner agency, Authorship, original Writing, Academic Integrity, and authorial identity.

In corroboration, Singh et al. (2025) who conducted a review on the use of AI in promoting human-AI Co-creativity, found evidence indicating that when writing is facilitated by a learner-empowering AI system that respects learner control, enhances transparency, and adjusts its actions adaptively, trust, student satisfaction, and the feeling of ownership are enhanced- more than in a fully automatic setting.

2.3 Learner Engagement, Trust and AI Feedback Literacy

A final set of findings to consider concerns factors beyond technology that mediate learning outcomes associated with AI feedback. Learners' ability and willingness to engage meaningfully with, critically evaluate, and accurately utilize LLM suggestions are key determinants of the learning value that this technology may impart. Specifically, researchers have begun to explore learner engagement, calibrated trust in the system, and AI feedback literacy as interconnected constructs that explain variation in learning outcomes across and within instructional contexts.

Recent studies consistently find that engagement with AI feedback is a requisite of learning.

Unlike human-generated feedback that may inherently lead to greater learning via such as its complexity, accuracy, specificity, or social presence, LLM Feedback does not work as a magic tool-it only helps the writing process evolve when learners spend time interpreting it, reflecting on it and questioning the quality of the message. The systematic review by Cengiz et al. (2025) identified how L2 writers could leverage LLM-based Feedback to revise the writing-and even more importantly-their revision strategies and self-regulation skills as they engage with feedback

more actively. They demonstrate that active Learners' engagement can have considerable effects in revising quality and self-regulated learning strategies. Yan and Zhang (2024) reviewed and provided additional details that explained that an active process is needed that goes beyond the automatic acceptance of LLM suggestions and that includes an effort by students to question, test, and select them according to their current pedagogical purpose and writing needs."

Further, most studies evaluate immediate effects on writing quality after relatively short interventions. Few examine how learner engagement, calibrated trust, and AI feedback literacy develop over time and affect sustained performance. Moreover, these constructs tend to be examined individually, leading to disconnected theoretical explanations for AI-enhanced learning, with an almost complete absence of longitudinal and cognitive-behavioral-affective approaches to AI feedback literacy.

2.4 Research Gap

A significant deficiency in the current literature is the lack of short-term intervention studies. Most empirical studies have tested the immediate effects on writing. Also, the range of measured outcomes in the literature are also restrictive. While studies predominantly measure writing quality (accuracy and grammatical correctness, development of vocabulary, overall writing quality), far less research has sought to analyse the impact on other, higher-order writing skills (e.g., rhetorical competence, argument development, disciplinary writing, critical thinking, authorial voice). Similarly, learner-centric variables, such as learner engagement, calibrated trust, AI feedback literacy, self-efficacy, motivation, and writing anxiety, remain largely ignored despite awareness of their impact on how learners can make successful use of AI-generated feedback.

Various conceptualizations of AI support feedback exist in the literature. For example, some works compare teacher and AI-generated feedback; others look into hybrid (human-AI) feedback; some investigate communicative and discursive practices with LLMs; or scrutinize the benefits and challenges of specific generative models.

In the context of ethical issues related to LLMs such as academic integrity, algorithmic bias, transparency, learner dependence, privacy, or responsible AI use, there has been little uptake in research into the actual implementation and evaluation of educational implications of these issues; much of the discussion has taken place at a theoretical level and less through empirically sound educational measurement strategies. Moreover, many literature reviews have compiled studies predating the advent of LLMs OR studies dealing with diverse AI technologies under a unified umbrella perspective on feedback. Limited insight has been given on synthesizing pedagogical shifts such as human-AI feedback partnerships, OR interlinked effects of learner engagement, trust, and AI feedback literacy on pedagogical designs and effective utilization.

The present systematic review addresses these shortcomings by focusing solely on papers that appeared after the advent of LLMs- from January 2023 to July 2026- adhering to the PRISMA 2020 guideline. It offers a current synthesis of recent research evidence in this rapidly evolving landscape, tracks pedagogical development trends in AI-assisted writing feedback practices, explores new patterns of human-AI feedback partnership, and makes both pedagogical and methodological recommendations for how the generation of AI text output in classroom instruction can be best facilitated as a practice for enhancing L2 writing success.

3. Methodology

3.1 Review Design

The study adopted a systematic literature review (SLR) methodology to synthesize contemporary empirical evidence on AI-assisted feedback and human-AI feedback partnerships in second language (L2) writing. SLRs provide a transparent, rigorous, and reproducible approach to searching, evaluating, and synthesizing empirical studies while minimizing reviewer bias through predefined criteria (Page et al., 2021). Accordingly, the review followed the PRISMA 2020 statement to guide literature identification, screening, eligibility assessment, quality appraisal, data extraction, and evidence synthesis. Unlike prior systematic reviews focusing primarily on automated written corrective feedback, Automated Writing Evaluation (AWE), or generative AI in education (Shi & Aryadoust, 2024; Craven &

Fredrick, 2026), this review examined the transition from conventional AWE to human–AI feedback partnerships in L2 writing, including learner engagement, trust, AI feedback literacy, and teacher–AI interactions throughout the writing process.

The review included empirical, peer-reviewed studies published between January 2023 and July 2026, a period characterized by the growing educational adoption of large language models (LLMs), including ChatGPT, Claude, Gemini, and Microsoft Copilot. The synthesis was guided by the three research questions presented in Section 1.1: the evolution of AI-assisted feedback from traditional AWE to human–AI feedback partnerships; the influence of learner engagement, trust, and AI feedback literacy on feedback uptake; and methodological, pedagogical, and theoretical limitations informing future research and classroom practice. Given substantial variation in participants, educational contexts, AI tools, intervention duration, methodologies, and outcome measures, a narrative synthesis rather than a quantitative meta-analysis was employed to accommodate methodological and contextual differences (Page et al., 2021; Shi & Aryadoust, 2024).

A systematic search was conducted across Scopus, Web of Science Core Collection, ERIC, Linguistics and Language Behavior Abstracts (LLBA), and Google Scholar. Following PRISMA 2020 recommendations, keywords represented three interconnected groups: L2 writing contexts, AI technologies, and feedback-related pedagogical concepts. Terms included “L2 writing,” “EFL writing,” “generative AI,” “large language model,” ChatGPT, Claude, Gemini, Copilot, “automated writing evaluation,” “feedback uptake,” “learner engagement,” “AI feedback literacy,” and “human-AI feedback partnership.” Boolean operators (AND, OR), phrase searches, and database-specific variations were used to improve relevance. Retrieved records were managed using EndNote 21 and duplicates were manually checked and removed. Backward reference checking and forward citation tracing were additionally conducted on eligible studies to minimize omissions (Page et al., 2021).

3.2. Eligibility criteria

To help enhance transparency, reduce subjectivity in the selection process, ensure consistency, maximize search completeness and increase replicability the eligibility criteria, which were pre-established by the reviewers, include in the study both quantitative, qualitative and mixed methods evidence on how artificial intelligence is being employed.

Criteria for inclusion were devised such that only articles directly addressing research questions were selected. Only peer-reviewed empirical studies. These studies were included in the review.

Second, studies focus exclusively on AI-supported feedback or human-AI feedback partnerships in second language (L2) writing instruction.

Table 1
Inclusion and Exclusion Criteria

Criterion	Inclusion	Exclusion
Publication type	Peer-reviewed journal articles	Editorials, books, book chapters, dissertations, conference abstracts, reports
Study design	Quantitative, qualitative, mixed methods	Conceptual papers without empirical data
Publication period	January 2023 – July 2026	Before January 2023
Language	English	Non-English publications
Educational context	L2, ESL, EFL writing	Reading, speaking, listening, translation, programming

AI technology	Generative AI, LLMs, ChatGPT, Claude, Gemini, Copilot, human–AI feedback	Grammar checkers, spell checkers, traditional AWE only
Data availability	Full-text empirical studies	Abstract-only or inaccessible studies

Utilising such eligibility criteria means that the resulting dataset provides a state-of-the-art review of AI-driven feedback for L2 writers. Given the fact that LLMs emerged towards the end of 2022, the synthesis represents a temporal slice of current research (post-LLM era) about generative AI’s impact on writing feedback practices, learner autonomy, writing behaviours, human-AI collaboration practices without compromising the methodological rigor.

3.4 Study Selection Process

A transparent, systematic and reproducible methodology for screening studies was used to minimize selection bias and ensure only relevant studies pertaining to the review topic were included (Page et al., 2021).

The screening procedure followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines.

The screening procedure consisted of four sequential steps: identification, screening, eligibility review, and final inclusion, which are all similar to recent systematic reviews studying automated written corrective feedback, artificial intelligence writing support, and generative AI in language education (Craven & Fredrick, 2026; Shi & Aryadoust, 2024). Following identification of relevant records using the chosen electronic databases, retrieved records are imported into EndNote 21. Identical literature searches between various searches are identified by automatic duplicate detection in the search software and later supplemented by manual searching. The combination of automatic duplicate detection in the search engine with manual searching reduces the chances of duplicate papers appearing in the search(es), which guarantees all studies were included only once in the selection of database studies (Page et al., 2021). Once duplicates have been omitted, the selected set of records proceeds to the screening phase, in which reviewers independently screen the identified articles and their associated abstracts for eligibility according to the explicit selection criteria set forth in 3.3 above. Titles of papers deemed clearly irrelevant to AI-based feedback, human-AI feedback alliances or L2 writing are automatically eliminated. Publications on non-writing language skill domains, studies not being of an empirical nature and literature using traditional automatic writing evaluation other than including generative AI are excluded by reviewers. The use of explicit selection criteria for title and abstract selection contributes greatly to the clarity and credibility of studies conducted during the systematic reviews and eliminates subjective selection decisions during these initial screening processes (Page et al., 2021). The screened literature passed on to the quality appraisal is provided with full-text documents and evaluated for its full-text eligibility criteria such as full-text format, design and setting, participant criteria, type of intervention and design features, etc. Non-empirical findings papers, insufficient methodology reports, duplicates of other identified studies, as well as literature on human-written corrected feedback, automated writing evaluation (non-generative AI), and those focused on written L2, ESL or EFL contexts are eliminated at this stage. The rationale is consistent with the PRISMA statement, where the reasons for excluding papers is kept as clear as possible (Page et al., 2021). Finally, the included studies will enter the third selection and synthesis phase in which they will be selected and reviewed for synthesis. Only if a study meets all relevant inclusion criteria will it enter quality assessment, data abstraction and thematic synthesis. All the procedures from identification to final inclusion of papers in the review are presented by a flow diagram shown in Figure 1, which illustrates the number of records identified, duplicates removed, the number of articles identified and screened, the number of full-text articles screened, the number of articles excluded and the reason for their exclusion and the number of finally included articles. Clear presentation of the screening process improves transparency of review methodology and quality and allows readers of the review to assess the comprehensiveness of the evidence synthesis in the present study (Page et al., 2021).

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

PRISMA 2020 Flow Diagram Template

Systematic review of human–AI feedback partnerships in L2 writing

IDENTIFICATION	SCREENING	ELIGIBILITY	INCLUDED
Records identified from: Scopus (n = <u>15</u>) Web of Science (n = <u>14</u>) ERIC / LLBA (n = <u>Nil</u>) Google Scholar (n = <u>29</u>) Other sources (n = <u>Nil</u>)	→		
Records removed before screening: Duplicate records (n = <u>0</u>) Marked ineligible automatically (n = <u>7</u>) Removed for other reasons (n = <u>7</u>)	↓		
	Records screened by title and abstract (n = <u>29</u>)	→	
	Records excluded (n = <u>10</u>)	↓	
	→	Reports sought for retrieval (n = <u>29</u>)	Full-text reports excluded, with reasons: No L2-writing context (n = <u>16</u>) No meaningful human–AI feedback relationship (n = <u>10</u>) AI used only for text generation (n = <u> </u>) No empirical data (n = <u>5</u>) No feedback/revision component (n = <u>10</u>) Review or conceptual paper (n = <u> </u>) Duplicate dataset / other reason (n = <u>10</u>)
	↓	Reports not retrieved (n = <u>10</u>)	
	→	Full-text reports assessed for eligibility (n = <u>29</u>)	Studies included in qualitative synthesis (n = <u>N/A</u>) Studies included in quantitative synthesis / meta-analysis, if applicable (n = <u>29</u>)

Figure 1. PRISMA flow diagram for the study-selection process.

3.5 Data Extraction

Relevant data were collected from all included studies using a purpose-created systematic data extraction form to ensure consistency, clarity, and methodological transparency and to facilitate systematic comparison across diverse evidence (Shi & Aryadoust, 2024). Descriptive information included authors, publication year, country, educational context, participant characteristics, sample size, demographics, research methodology, AI application, and treatment duration. Analytical information included study purpose or research questions, data collection methods, key findings, reported strengths, limitations, and recommendations for future research. These variables enabled systematic comparison and thematic analysis of AI-supported feedback research in L2 writing.

Data extraction was directly aligned with the three research questions. For RQ1, information concerned developments in AI-assisted feedback, particularly the transition from conventional AWE tools to LLM-based feedback and human–AI collaboration. For RQ2, data focused on learner engagement, trust in AI, AI feedback literacy, feedback uptake, revision practices, and writing outcomes. For RQ3, information included methodological limitations, theoretical perspectives, pedagogical contributions, ethical and equity implications, and recommendations for future research. Extraction entries were cross-checked against full-text studies, with unclear or missing information reviewed where necessary. The resulting database was organized into comparative tables and provided the foundation for the narrative thematic synthesis presented in the Results section.

Table 2

Data Extraction Framework

Category	Extracted Information
Bibliographic information	Author(s), year of publication, country
Educational context	L2, ESL, EFL, higher education, school level
Participants	Sample size, learner characteristics
Research methodology	Quantitative, qualitative, mixed methods
AI technology	ChatGPT, Claude, Gemini, Copilot, other LLMs
Feedback approach	AI feedback, teacher feedback, hybrid human–AI feedback
Intervention	Duration, instructional activities
Outcome variables	Writing quality, revision behaviour, learner engagement, trust, AI feedback literacy
Key findings	Major empirical findings
Limitations	Methodological and pedagogical limitations
Future directions	Authors' recommendations

3.6 Quality Assessment

Quality assessment is an important component of systematic reviews because it enables researchers to evaluate the trustworthiness, methodological rigor, and potential risk of bias in the included evidence. Following PRISMA 2020 guidelines (Page et al., 2021), each included study was appraised before evidence synthesis. Similar quality assessment procedures have been adopted in recent systematic reviews of AI-assisted writing, automated written corrective feedback, and generative AI (Shi & Aryadoust, 2024; Craven & Fredrick, 2026). Given the methodological diversity of the included quantitative, qualitative, and mixed-methods studies, the Mixed Methods Appraisal Tool (MMAT), version 2018, was employed (Hong et al., 2018).

Five criteria were considered: clarity of objectives, appropriateness of research design, participant selection and sampling strategy, data collection and processing, and appropriateness of reporting findings. Studies satisfying at least

four criteria were classified as high quality, those meeting three as moderate quality, and those meeting fewer than three as low quality (Hong et al., 2018). No study was excluded solely because of its quality rating; instead, appraisal results informed the interpretation and synthesis of evidence, consistent with PRISMA recommendations (Page et al., 2021).

Overall, most studies demonstrated high or moderate methodological quality. However, recurring limitations included small samples, short intervention periods, absence of control groups or longitudinal follow-up, inconsistent outcome measures, and insufficient reporting of instrument validity and reliability. Ethical and privacy considerations surrounding generative AI also received limited attention. These methodological limitations were considered when interpreting findings and developing recommendations for future research (Hong et al., 2018; Page et al., 2021).

Table 3
Quality Assessment Criteria (MMAT 2018)

Assessment Domain	Evaluation Criterion
Research objectives	Clearly stated and aligned with the study purpose
Research design	Appropriate for answering the research question
Participants	Adequate sampling procedures and participant description
Data collection	Valid and reliable instruments and procedures
Data analysis	Appropriate analytical techniques with coherent interpretation
Overall quality	High (5/5), Moderate (3–4/5), Low ($\leq 2/5$)

3.7 Data Synthesis

The studies included in the review were analyzed using narrative thematic synthesis, following the selection and quality appraisal stage. As the included studies exhibited considerable heterogeneity in methodology, sample size and characteristics, educational setting, AI technology, intervention design, intervention length, and outcomes, a narrative synthesis approach seemed most appropriate. Similar synthetic methods, using narrative synthesis and no meta-analysis, have been chosen in previous systematic reviews of AI-supported writing instruction and automated written feedback as a result of substantial heterogeneity among participating studies with regard to methodology and the reported outcomes (Craven & Fredrick, 2026; Shi & Aryadoust, 2024). In accordance with the methodological guidance provided by Popay et al. (2006), repeated scrutiny of included data occurred during the synthesis process to obtain familiarity and a full understanding of each study's characteristics and findings. Later on, the studies were critically compared as per the research questions/aims, the methodology and techniques they used to execute the research and the sample size and characteristics of participants; the purpose and context in which the students were trained; the types of AI and LLMs that have been employed in feedback generation, and the final outcome which is attained from writing task(s). Through such iterative comparative analysis, similarities and regularities in the common themes, trends, and discrepancies concerning the topic can emerge and are brought out.

To address the goals of the research directly, themes were developed following the three research questions stated in 1.1. I, evidence relevant to research question 1 (RQ1) will be synthesized to shed light on the pedagogical shift in the field of AI-powered writing feedback, from conventional automated writing evaluation (AWE) tools to human-AI feedback in large language model (LLM)-supported writing tasks. II, Evidence that answers research question 2 (RQ2) was summarised to investigate how learners' motivation, attitude towards AI-generated writing feedback, and AI literacy affect their usage of feedback, revision practices and the quality of writing. III) Evidence answering research question 3 (RQ3) will be presented to examine trends in methodologies, challenges in pedagogical implementations of AI technologies in feedback generation, and possible research directions in the area.

During the entire analysis course of this review, individual findings from different sources were compared, contrasted and combined so that any area of convergence and divergence would come into view. In particular, similarities and differences were found in relation to specific methodological features or outcomes, or to the kinds of instructional techniques and participant groups that made use of them. This comparative procedure is pivotal and enables the researcher not only to be a descriptor, but also an interpreter by integrating disparate evidence bases and arriving at coherent findings regarding AI-assisted feedback. Recent reviews in LLMs in language teaching using this sort of synthesizing also reported: "the heterogeneous nature of literature makes systematic quantitative reviews difficult and necessitates the use of narrative syntheses to integrate the literature on LLMs in language education." (Ding et al., 2026, pp. 1-2). In addition, study quality was considered throughout the analysis of the literature. Greater weight has been attributed to studies exhibiting higher quality as opposed to studies with notable methodological concerns. Studies with strong quality include those with a greater number of participants, clearly stated aims and objectives and the use of AI technology. This form of integrated review also contributes to the validity of the results as the most robust evidence should carry greater influence than the most weakly designed (Page et al., 2021).

4. Results

4.1 Characteristics of the Included Studies

Twenty-nine empirical studies on AI-supported feedback on second language (L2) writing were identified and selected from scholarly databases searched from January 2023 to July 2026 using a systematic literature search and the PRISMA guidelines. Duplicates were deleted, then titles/abstracts were screened, and full texts were reviewed for their relevance and applicability based on inclusion criteria (e.g., peer-reviewed; empirical study; topic is the pedagogical implementation of generative AI, large language models (LLMs), or human–AI feedback partnerships; study focused on the English language).

Five major themes emerged from the analysis, organizing the results section:

1. Characteristics of the Included Studies
2. The Evolution of AI-Supported Writing Feedback in L2 Writing
3. Human-AI Collaboration in Feedback Practice
4. Learner engagement, Trust and AI Feedback Literacy
5. Methodological trends and Research Gaps.

These themes aim to build coherence across evidence of research into AI assistance in L2 writing feedback and are embedded and presented through the research questions of the study.

Thus, directly addressing the stated research aims, and to facilitate comparison with and integration into the conceptual framework presented through a review of relevant literature.

It is evident that research has sharply increased since the introduction of ChatGPT to the public in late 2022, particularly from 2024 to 2026, signalling a paradigm shift from conventional automated writing evaluations to the LLM-supported feedback characterized by conversational interaction, adaptive responses, and collaborative writing (Craven & Fredrick, 2026; Jiao & Wang, 2026). The distribution of literature is dominated by studies conducted in China, Saudi Arabia, Türkiye, Malaysia, and South Korea/other Asian countries, whereas research has been initiated in Europe, North America, and Australia, with significant recent activity in many locations. Nevertheless, studies have been relatively scarce from developing nations, with limited diverse contextual representation.

Most research was conducted in a higher education setting, involving university-level EFL students, but was limited across primary, secondary, and adult learning.

Both quantitative (mostly pre-test/post-test designs) and quasi-experimental methodologies were broadly employed to ascertain the effects of AI feedback interventions. Mixed methods design (interviews, questionnaires, narrative descriptions) were additionally common, to capture students' experiences with and perceptions of these technologies. Qualitative research methods also provided in-depth understandings of interactional processes. Such pedagogical and methodological tendencies conform to several recent systematic reviews (Shi & Aryadoust, 2024).

ChatGPT was the most examined platform (n=16), followed by Claude (n=6), Gemini (n=5), and Microsoft Copilot (n=1), together outweighing the pre-LLM era of Automated Writing Evaluation (AWE) techniques.

In comparison with other standard AWE technologies, LLMs could offer responsive, individualized, personalized, contextual and adaptive answers. This is useful in facilitating LLM collaborative and adaptive writing contexts. (Craven & Fredrick, 2026). Overall, regardless of different forms of intervention, durations, and sample sizes, a host of findings demonstrate significant positive effects of AI-supported feedback on writing quality, mechanical accuracy, textual development, revisional effort, strategy use for autonomous learning, attitude towards AI, and AI feedback literacy under conditions where pedagogical considerations were made.

Table 4
Summary Characteristics of the Included Studies

Characteristic	Summary
Publication period	2023–2026
Educational contexts	EFL, ESL, L2 writing
Participants	Mainly university students
Research designs	Quantitative, mixed methods, qualitative
Most investigated AI	ChatGPT, Claude, Gemini, Copilot
Common outcome measures	Writing quality, revision behaviour, learner engagement, AI feedback literacy, trust
Overall trend	Rapid growth in LLM-supported writing research after 2023

4.2 The Evolution of AI-Supported L2 Writing Feedback (RQ1)

The literature reviewed confirms a clear evolution in the use of AI-supported feedback in L2 writing between 2023 and 2026. While previous developments centered on Automatic Writing Evaluation (AWE) systems designed to mark and grade students' essays automatically, the current paradigm is more holistic: a human-AI learning partnership where LLMs serve as intellectual scaffolding to support a learner-centred approach to planning, drafting, revising, reflecting on and continuously improving the L2 writing product, while also enhancing the development of higher order writing abilities (Shi & Aryadoust, 2024; Jiao & Wang, 2026). A significant outcome from the research reviewed here is LLM's superior capacity to providing more contextual, personal, adaptive and rich feedback than earlier AWE technologies, generating explanations and alternative ways of articulating ideas to encourage learners to engage in iterative exchanges and the multiple rewriting of their texts, thereby transforming feedback from being purely error focused, product-oriented product to process oriented development process of enhancing student's autonomy and reflectedness (Pretorius & Thewissen, 2026). The reviewed literature also indicates that as LLMs' capabilities widen to enhance the generation of ideas, argument organization, coherence and rhetoric, not just basic grammar (Craven & Fredrick, 2026; Ding et al., 2026), they appear to be transitioning from solely functioning as error tools to becoming cognitive learning partners in improving writing practices.

However, one dominant finding across all studies in this review is that the efficacy of AI-assisted writing development depends primarily on the student. Critical feedback reception is related to active engagement, critical evaluation of AI outputs and multiple rounds of revision. A student's engagement with LLM-provided feedback and

revisions leads to far greater positive learning outcomes when they are able to critically receive feedback (Cengiz et al., 2025; Yan & Zhang, 2024) than when they receive feedback from a teacher. While L2 writing feedback given by LLMs will inevitably improve the learner's proficiency but it lacks in higher order writing conventions in terms of critical argumentation, rhetorical organization, audience awareness and writing in one's field, therefore they always show weaknesses in generating hallucinated or inaccurate contents; as such human intervention in validation and guidance, including a teacher's support in higher order critical reading and analysis, is still paramount to the effective use of LLMs in L2 writing (Chen et al., 2024; Ding et al., 2026; Karagz, 2025; Pretorius & Thewissen, 2026).

Table 5
Key Findings Related to the Evolution of AI-Supported Feedback

Theme	Evidence Synthesised from Included Studies
Technological evolution	Transition from traditional AWE systems to conversational LLMs
Feedback characteristics	Context-sensitive, interactive, personalised, and adaptive feedback
Pedagogical shift	From product-oriented correction to process-oriented writing development
Learner role	Increased interaction, reflection, and iterative revision
Teacher role	Shift from primary evaluator to facilitator and feedback mediator
Persistent limitations	Higher-order writing support, hallucinations, contextual judgement, ethical concerns

4.3 Human-AI Feedback Partnership

Human-AI Feedback Collaboration receives the best supporting evidence regarding L2 writing education and practice. AI is a supporter/enhancer but not a replacement for the teacher. Instead, combining the efficiency of AI with human instruction leads to a more balanced and educationally beneficial output than writing pedagogy based only on AI or only on humans, which provides directions that collaborate rather than compete with technology (Karagz, 2025; Pretorius & Thewissen, 2026).

The literature review demonstrates that the distinct, yet complementary expertise of human teachers and artificial intelligence can be employed. Artificial intelligence affords instant, customized and adaptive feedback directed at grammatical, lexical, syntactical and lingual shortcomings, and explains alternative, persistent teaching strategies with explanatory insights (Gomez et al., 2025). Human teachers, however, are able to impart a subject context, offer a specialist understanding and provide analytical and socio-emotional guidance for critical, intelligent task production.

This unique understanding is beyond the artificial intelligence capabilities at present (Ding et al. , 2026). Therefore, the combined learning potential offered by the two feedback modes will offer students both the instant formative feedback they desire and the advanced educational context they require for academic writing (Gomez et al., 2025).

In terms of instructor mediation, Similar to the human instructor, valuable feedback from an AI-supported computer system requires instructor mediation of AI-provided input. Learners improved their academic performance and outcomes when students were asked to pose queries about the AI output, demonstrate why and provide justification of the feedback options, and reflect on constraints and limitations of computer programs (Hussien,2025; Pretorius &Thewissen, 2026). In contrast, learner writing confidence, academic self-regulation, frequency of review and lexical precision are enhanced where the use of the hybrid learning modes; The Learners receive instant AI support

complemented via further detailed instructor mediation and therefore improve upon faster and deeper feedback instructions (Zhang et al., 2025; Karaguz,2025).

Despite its benefits, literature highlights challenges with the implementation process as well, and teachers should still remain the cornerstone of AI integrated writing pedagogy to design AI supported writing activities, to monitor the quality of AI feedbacks, to cultivate AI literacy among learners and to create responsible usage scenarios for the AI technology because overuse of AI may lead to weakening of the learners' critical decision-making and their unconditional receptivity to wrong or hallucinated replies (Chen et al., 2024; Hussien, 2025; Singh et al., 2025).

Table 6

Synthesised Findings on Human–AI Feedback Partnerships

Dimension	Synthesised Evidence
AI contribution	Immediate, personalised, adaptive, and continuous feedback
Teacher contribution	Contextual judgement, disciplinary expertise, ethical guidance, higher-order writing support
Learner outcomes	Improved revision quality, writing performance, learner autonomy, and engagement
Major challenges	Hallucinations, overdependence, academic integrity, privacy, algorithmic bias
Overall conclusion	Hybrid human–AI feedback consistently outperformed AI-only and teacher-only models in most educational contexts

4.4 Learner Engagement, Trust, and AI Feedback Literacy

The synthesis identified learners, trust in AI-produced feedback and AI feedback literacy as three closely related interrelated aspects that together contribute to the effectiveness of teaching AI-supported writing development. Overall, the existing evidence suggested LLM-produced feedback had a positive impact on writing performance, yet such an effect was not solely the result of inherent technological advantage. Instead, learning gains largely depended on how learners engaged with, perceived and incorporated the AI-generated suggestions into their writing process. Results from numerous studies across different learning settings and tasks suggest that students write at a higher level of quality when they are given feedback. These studies depended on learners' engagement with and integration of LLM-produced feedback (Cengiz et al., 2025; Yan & Zhang, 2024).

In addition, the analysis of the literature revealed that learners' active behavioral, cognitive, and metacognitive engagement played an important role in improving their revision quality and development. While behaviors can be defined as students engaging in multiple revisions, asking for more explanation and performing multiple interactions with AIs during their writing process, cognition engaged learners in analyzing feedback suggestions, comparing revisions, making choices based on those comparisons; meta-cognition refers to learners who regulate their learning on writing process by themselves or through reflection about how one writes (Cengiz et al., 2025; Pretorius & Thewissen, 2026). It was found that learners actively engaged in each aspect of engagement would generally lead to a better learning outcome compared to students who passively accept AI-rendered suggestions such as error correction.

This synthesis also highlighted the centrality of learners' trust towards AI in general. Though most studies revealed learners have gained higher trust toward AIs for lower-order aspects of writing, such as grammar, lexical accuracy, word usage and syntactic errors, learners had considerably less trust toward the feedback related to content, argumentative quality, and logical structures. Some participants noted that they would still seek advice and verification regarding higher-level suggestions from either the teacher or other sources, which was a confirmation that Learners hold Calibrated Trust when using AI for their writing, or in other words, their Trust degree depends on the complexity

of the generated advice (Chen et al., 2024; Yan & Zhang, 2024). Therefore, AI-supported learning depends not on the amount of trust toward technology, but on learners' ability to evaluate which suggestions to include in their writing from this AI tool.

Another prominent issue was the growing importance of AI feedback literacy among learners. Researchers in most of the selected studies stressed that students need to gain new competencies to succeed with AI technologies as tools for learning, specifically related to their AI feedback experience. The concept beyond conventional feedback literacy, AI feedback literacy, referred to skills including prompt writing, evaluation of AI-generated comments and feedback, recognition of hallucination or falsified information, maintaining authorial voice, responsible use of generation of AIs, critical considerations on how AI feedback informs choices, etc. Studies suggested that a high level of AI feedback literacy was positively associated with more active engagement, more sophisticated revision strategies, high self-confidence and better learning performance than a low AI feedback literacy level (Teng, 2024; Jiao & Wang, 2026).

Table 7

● Synthesised Findings Related to Learner Engagement, Trust, and AI Feedback Literacy

Theme	Evidence from the Included Studies
Learner engagement	Active revision, iterative interaction, reflective learning, improved writing quality
Behavioural engagement	Multiple revision cycles and frequent AI interaction
Cognitive engagement	Evaluation, comparison, and selective adoption of AI feedback
Metacognitive engagement	Reflection, monitoring, and regulation of writing strategies
Trust	High for linguistic feedback; lower for higher-order writing feedback
AI feedback literacy	Prompt engineering, critical evaluation, ethical AI use, preservation of authorial voice
Common challenges	Overdependence, hallucinations, bias, uncertainty regarding feedback reliability
Overall synthesis	Effective learning depends on active learner engagement supported by teacher guidance rather than AI capability alone.

4.5 Methodological Trends and Research Gaps

You are right. I should have **retained your original references** rather than rewriting the section without them. Here is the corrected version, keeping your citations and addressing the reviewer's concern about **reporting the nature of the reviewed literature versus making the present review's own recommendations**.

Research Question 3: Methodological, Pedagogical, and Theoretical Limitations and Future Directions

The third research question examined the methodological, pedagogical, and theoretical limitations evident in current empirical research on AI-supported feedback in L2 writing and the nature of the future research directions and classroom implications reported across the reviewed literature. Accordingly, the analysis distinguished between limitations identified in the existing evidence and recommendations reported by the authors of the included studies. At this stage, these recommendations were treated as characteristics of the reviewed literature rather than recommendations generated by the present review. This distinction is consistent with systematic review procedures in

which characteristics and limitations of the available evidence are first synthesized before broader implications are developed (Page et al., 2021; Shi & Aryadoust, 2024).

Methodologically, several recurring limitations were identified across the reviewed studies. These included relatively small or context-specific samples, short intervention periods, limited longitudinal evidence, absence of delayed post-tests, and, in some cases, the absence of control or comparison groups. Several studies also provided insufficient information regarding the validity and reliability of their instruments. Considerable variation was evident in research designs, AI tools, intervention duration, feedback procedures, and outcome measures. Such methodological heterogeneity limits direct comparison across studies and indicates that the empirical evidence surrounding AI-supported feedback remains methodologically diverse (Hong et al., 2018; Page et al., 2021). Similar methodological concerns have also been identified in systematic reviews of AI-assisted writing and automated feedback research (Shi & Aryadoust, 2024; Craven & Fredrick, 2026).

Pedagogical and theoretical limitations were also evident. Although AI-generated feedback was increasingly integrated into L2 writing instruction, the role assigned to AI differed considerably across studies. Some studies treated AI primarily as an automated feedback provider, whereas others investigated more collaborative relationships among learners, teachers, and AI. Attention to learner engagement, trust, feedback uptake, and AI feedback literacy was similarly uneven. Moreover, theoretical perspectives explaining how learners interpret, evaluate, accept, reject, and apply AI-generated feedback were not consistently incorporated. Ethical and equity considerations, including privacy, transparency, responsible AI use, and possible overreliance on generative AI, also received inconsistent attention across the literature (Hong et al., 2018; Page et al., 2021).

Regarding future research, the reviewed studies reported several recurring directions. These included the need for larger and more diverse samples, longer interventions, longitudinal investigations, delayed measurements, stronger experimental and mixed-methods designs, and clearer reporting of instrument validity and reliability. The literature also identified a need for greater attention to learner engagement, trust, AI feedback literacy, feedback uptake, and revision behavior. Such directions reflect the need to move beyond determining whether AI-generated feedback improves writing performance toward understanding how learners engage with, evaluate, and use such feedback during writing and revision (Shi & Aryadoust, 2024; Craven & Fredrick, 2026).

At the classroom level, the reviewed literature reported increasing attention to the role of teachers within AI-mediated feedback environments. Rather than positioning AI as an independent substitute for teacher feedback, the emerging literature points toward human–AI feedback partnerships in which AI-generated suggestions are interpreted, evaluated, and applied with appropriate teacher involvement. The reported classroom implications also emphasize developing learners’ ability to critically evaluate AI-generated feedback rather than accepting it automatically. Importantly, these findings represent the nature and recurring recommendations of the reviewed literature rather than the recommendations of the present review itself. The specific future research directions and classroom implementation recommendations derived from the overall synthesis are therefore developed separately in the Discussion and Implications sections (Page et al., 2021; Shi & Aryadoust, 2024).

Table 8 Major Methodological Trends and Research Gaps

Area	Current Evidence	Research Gap
Research design	Predominantly quantitative and quasi-experimental	More longitudinal, qualitative, and mixed-methods studies
Educational context	Mainly university EFL learners	School, vocational, multilingual, and adult education
AI platforms	ChatGPT dominates	Comparative studies of Claude, Gemini, Copilot, and other LLMs

Outcome measures	Writing quality and revision behaviour	Critical thinking, transfer of learning, creativity, long-term writing development
Learner variables	Limited attention to AI feedback literacy and trust	Integrated models of learner engagement, trust, and AI literacy
Ethics	Mainly conceptual discussions	Empirical investigations of ethical AI implementation
Theory	Limited theoretical integration	Stronger use of SLA, feedback literacy, and self-regulated learning theories

5. Discussion

5.1 The Transformation of AI-Supported Feedback in L2 Writing

5.1 Evolution of AI-Supported Feedback in L2 Writing

The reviewed studies suggest an evolution in AI-supported feedback in L2 writing from automated correction to interactive, process-oriented feedback. Previous AI writing tools mainly aimed to address surface-level error correction such as grammar, vocabulary, sentence structure and spelling; in contrast, recent LLM-based tools address multiple stages of writing such as planning, drafting, revising and reflection. Taken together, it can be said that these changes are not merely technological, but a pedagogical shift of taking feedback as dialogical learning rather than erroneous correction, where learners are enabled to query, assess and revise not only language but also ideas (Shi & Aryadoust, 2024; Craven & Fredrick, 2026; Jiao & Wang, 2026).

Across the reviewed studies, interactivity and learner agency serve as the central benefits for the effectiveness of LLM-supported feedback. Unlike traditional automated feedback which tends to offer relatively stable corrections, LLMs can provide ongoing interactions regarding a learner's writing such as developing an alternative text, seeking clarification and revision iteratively (Ding et al., 2026; Jiao & Wang, 2026). This implies that the pedagogical effect of LLM feedback is not based on the quality of suggestions, but the cognitive and reflective actions they generate around them. Such interaction encourages re-examining word choices, taking control of revision and being the author of one's writing process, enhancing self-awareness and autonomy (Pretorius & Thewissen, 2026).

However, evidence from the studies also shows that learners' active evaluation, repeated revision and teacher support play important roles in making AI feedback effective. Studies reporting positive results typically associate effective feedback with active critique, re-drafting and human interaction. Active engagement with LLM feedback proves to be highly significant compared with passively accepting all suggestions (Cengiz et al., 2025; Pretorius & Thewissen, 2026); Taken together, these studies challenge the hypothesis that the advancement of AI directly leads to good writing performance, highlighting instead the significance of human critical evaluation and teacher guidance that informs which feedback suggestions learners should adhere to.

Inconsistency in providing LLM feedback is a frequently cited limitation by the researchers. Though generative AI demonstrates potential in supporting language revision, its shortcomings persist when dealing with rhetoric, argumentation, organizational and disciplinary conventions (Karagz, 2025; Ding et al., 2026; Jiao & Wang, 2026); The existence of hallucinations and non-reproducible suggestions calls for human vigilance in determining its usability. Thus, the literature does not indicate any replacement model where AI takes over the conventional role of teacher. The emerging trend shows an increasing awareness of human-AI synergy where AI supplements the learner's immediate response feedback needs; teachers maintain authority in making pedagogical choices and contextual judgments (Ding et al., 2026).

Hence, LLM-supported feedback is not just an automated writing assistant but a socio-pedagogical interactive feedback process where its pedagogical effectiveness is not determined by its ability to operate independently but by how it is embedded within a learner-centered and critically oriented writing process (Shi & Aryadoust, 2024).

5.2 Human-AI Feedback Partnerships: Towards Collaborative Writing Pedagogy

Across the reviewed studies, there was consensus that effective AI-supported writing instruction is best conceptualized as a human-AI partnership rather than an opposition or replacement model. Hybrid and blended approaches were considered particularly productive because they combine the speed, personalization and diverse feedback potential of AI tools with teachers' rich contextual knowledge, pedagogical judgment and domain expertise (Karagz, 2025; Ding et al., 2026). Combined, this indicates that the focus should shift from whether the student should receive feedback from an AI or teacher to how their relative strengths should be integrated. AI systems could provide timely and personalized feedback whilst teachers' roles in helping students to think critically about feedback, to consider the specific context of their writing, to address issues of academic integrity, and to argue in discipline-specific ways remained central.

Secondly, the papers confirmed the changing role of the teacher. Far from just the dispenser of corrective feedback, the role of the teacher becomes mediation, evaluation, and facilitation. Supporting students in their interaction with AI, whether it be through providing instruction on how to prompt the AI, on how to critically evaluate its output, and on how to make a choice about whether or not to accept or to adapt feedback that is provided by the AI (Chen et al., 2024; Cengiz et al., 2025; Pretorius & Thewissen, 2026) is a crucial part of effective support. As a result, interactions are more likely to be co-creative rather than merely assistive, supported by transparent processes and decisions made through shared responsibility and human control (Gomez et al., 2025; Karagz, 2025).

This leads to new forms of literacy for both students and instructors. As the authors argue, teachers require prompting and critical evaluation skills, and institutional guidance, to effectively and pedagogically facilitate human-AI learning, whereas students require a particular form of 'AI feedback literacy' that would enable them to make their own judgment in receiving and acting on prompts (Teng, 2024). These are important, given that the benefits of AI are tempered by risks to academic integrity, ethical data use, potential bias within algorithms, and overreliance on the tools provided (Chen et al., 2024; Hussein, 2025).

Overall, these papers lend themselves to a human-centered model of AI-supported writing instruction that pairs technological capability with a blend of critical analysis, feedback literacy and professional guidance and institutional constraints and processes.

Ultimately, whether this combination results in successful outcomes depends on how institutions empower teachers, instructors and students with control, responsibility and clear governance.

5.3. Learner Engagement, Trust and AI Feedback Literacy Beyond the Technological Adoption

The review concluded that the effectiveness of AI-supported feedback relies more on how students interact meaningfully with these AI outputs than on the sophistication of those outputs, i.e., on the underlying LLMs. Learning improved through learners' active use and reflective evaluation and rewriting of the AI's outputs, not by the presence of the AI per se (Cengiz et al., 2025; Teng, 2024). This suggests the primary impact of AI in writing is pedagogical: enabling enhanced learning, not the automation of the writing process, and "a pedagogical intervention to enhance learner involvement, reflection, and learner agency in revision" (Shi & Aryadoust, 2024; Jiao & Wang, 2026).

The review uncovered learner engagement as the main mechanism mediating the impact of AI-assisted writing feedback on writing progress.

Specifically, more apparent improvements in writers' output have been observed through more active ways of engagement with different kinds of AI feedback and multiple revisions in contrast with non-participatory mode, and involving to explore varied kinds of feedback with regard to what and how they receive advice in the response generation to multiple revise, similar to other conclusions, and a positive correlation was found between behavioral

as well as cognitive engagement, and improvements of writing quality (Cengiz et al, 2025: 3055); lity of revisions and improved writing outcomes. This relates to current perspectives on feedback, which stress the learning process as interpretive and transactional, enabling knowledge creation rather than the mere reception of reform statements. (Waring, 2021; Cengiz et al., 2025).

Learner trust was another critical factor influencing students' usage of AI feedback. Effective AI-supported writing therefore requires appropriate, or "calibrated"-levels of trust, not complete trust or mistrust in AI recommendations. The review demonstrated that students generally trusted AI feedback on grammatical and lexical accuracy more than on critical thinking (coherence, argument) and disciplinary writing contexts (Chen et al., 2024; Jiao & Wang, 2026). Such findings demonstrate that learners increasingly recognized both the strengths and limitations of AI tools; educational practice needs to guide students to check and confirm information suggested by the AI instead of passively trusting its output (Chen et al., 2024).

Last but not least, in the realm of writing supported by AI, another essential competence is AI feedback literacy, which is the ability not just to follow instructions for using AI feedback but also to design useful prompts, make evaluative judgments about responses, identify hallucinations, keep authorial authority, and make well-informed choices regarding revisions. It was also demonstrated that, compared with students less adept in AI feedback literacy, those who have developed higher AI literacy achieved greater write quality and also felt more independent and organized in terms of the revision process. (Teng, 2024; Pretorius & Thewissen, 2026).

This suggests prompt training, feedback checking, and validation, awareness of the risks associated with AI, including hallucination and ethical considerations, and maintaining student voice should be included as basic digital literacies in teaching and learning, not as advanced electives. Together, working within a framework of learners' engagement and moderated trust, learner feedback literacy means the ability of student-users of AI tools to leverage AI-generated feedback effectively and avoid its potential pitfalls(Ding et al., 2026; Jiao & Wang, 2026).

5.4 Methodological Implications and Future Research Directions

In the aftermath of the emergence of LLMs, the research field of AI-supported writing has experienced exponential growth, but limitations in methodological designs in several studies continue to limit the interpretability and generalizability of current findings. Given the wide variety in experimental designs, intervention durations, learner profiles, AI tool configurations and writing tasks utilized in prior research, comparability between studies is difficult, yet overall themes regarding the positive impact of AI on writing processes and learner participation are consistent across studies (Shi & Aryadoust, 2024; Craven & Fredrick, 2026). Much of the research to date has relied on short-duration interventions (i.e., for one semester, one assignment), which can provide evidence that an immediate, in-course intervention impacts writing outcomes and learner participation; however, little is known about the maintenance or transfer of this impact into the future (Jiao & Wang, 2026). More follow-up and longitudinal studies need to be designed to determine if recurrent and longitudinal AI-facilitated feedback in writing can contribute to stable and enduring effects on student writing and autonomy.

Further, there is a need to enhance research diversity. While quantitative and quasi-experimental studies can yield robust results on writing development and other outcomes, these approaches tend to lack depth for exploring what and why of learner-AI interactions. To explore the cognitive and instructional processes occurring during interactions between students and AI Feedback systems in detail, qualitative and mixed-methods studies using instruments like interviews, classroom observation, thinking-aloud protocols, learning analytics and writing portfolios are warranted (Pretorius & Thewissen, 2026; Ding et al., 2026). When considering theories of SLA, feedback literacy, self-regulated learning and sociocultural theory to support the investigation of how AI affects writing development and under which conditions, it is equally important to further examine how these theories interact with and shape the effects of AI in the second language context (Jiao & Wang, 2026).

Finally, studies should investigate more than comparisons between AI feedback and human feedback. An in-depth investigation of the various modes of collaboration (e.g., in-depth human-AI teaming with teachers as facilitator,

AI-only use with no teachers in writing, students working entirely alone with AI feedback), the impact of classroom context, and ethical implementation of AI technologies remains essential. Future research should move beyond establishing whether AI improves writing to examine AI's potentially negative aspects, such as algorithmic bias, hallucinations, overdependence, the ethical issues of data privacy, and cheating (Hussien, 2025; Gomez et al., 2025), along with the ways in which students can be empowered to use AI responsibly and critically (Ding et al., 2026). As stated, more sophisticated, rigorously designed studies that provide an enriched understanding of AI-supported writing contexts will benefit teaching and learning practices with AI technologies.

6. Conclusion

This systematic review synthesized empirical research published between January 2023 and July 2026 to examine the evolution of artificial intelligence (AI)-supported feedback in second language (L2) writing, with particular emphasis on the emergence of human–AI feedback partnerships, learner engagement, trust in AI-generated feedback, and AI feedback literacy. Guided by the PRISMA 2020 framework (Page et al., 2021), the review provides an up-to-date synthesis of recent evidence generated during the rapid expansion of large language models (LLMs), including ChatGPT, Claude, Gemini, and Microsoft Copilot. By focusing exclusively on the post-LLM period, the review offers a contemporary perspective on how generative AI is reshaping writing instruction and feedback practices.

The findings demonstrate that AI-supported feedback has evolved substantially beyond traditional automated writing evaluation systems. Whereas earlier technologies primarily supported grammatical correction and surface-level language editing, contemporary LLMs facilitate interactive, adaptive, and personalised feedback throughout the writing process. Across the reviewed studies, AI-supported feedback consistently enhanced writing quality, revision behaviour, learner engagement, and self-regulated learning, particularly when integrated into structured instructional environments. However, the review also revealed that AI remains less effective in addressing higher-order writing dimensions such as rhetorical organisation, critical reasoning, disciplinary writing conventions, and audience awareness, highlighting the continuing importance of teacher expertise (Shi & Aryadoust, 2024; Ding et al., 2026).

A major contribution of this review is the identification of human–AI feedback partnerships as the dominant pedagogical model emerging within contemporary L2 writing instruction. Rather than replacing teachers, generative AI increasingly functions as a collaborative instructional resource that complements teachers' pedagogical knowledge, contextual judgement, and ethical guidance. The evidence consistently indicates that hybrid feedback approaches produce more balanced educational outcomes than AI-only or teacher-only feedback because they combine the efficiency of AI with the interpretative and instructional expertise of teachers (Karagöz, 2025; Pretorius & Thewissen, 2026). These findings reinforce the growing consensus that the future of AI-supported writing lies in collaborative intelligence rather than technological substitution.

The review also demonstrates that the effectiveness of AI-supported feedback depends fundamentally on learner-centred factors. Across the included studies, learner engagement, calibrated trust, and AI feedback literacy emerged as key determinants of successful feedback uptake and writing development. Students benefited most when they critically evaluated AI-generated suggestions, engaged in iterative revision, and used AI as a learning partner rather than an unquestioned authority. These findings indicate that educational institutions should prioritise the development of AI feedback literacy alongside writing instruction to ensure that learners use generative AI responsibly, critically, and ethically (Cengiz et al., 2025; Teng, 2024).

Despite these promising developments, the review identified several limitations within the existing evidence base. Current research is characterised by short-term intervention studies, an overrepresentation of university-level EFL contexts, methodological heterogeneity, limited longitudinal evidence, and insufficient attention to ethical AI implementation and theoretical integration. Addressing these limitations will require more rigorous research designs, stronger theoretical frameworks, broader educational representation, and greater emphasis on sustainable learning outcomes. Future investigations should also examine comparative human–AI collaboration models, institutional AI governance, and the long-term development of AI feedback literacy across diverse educational settings.

The present review has several strengths. By following the PRISMA 2020 guidelines, employing transparent eligibility criteria, and synthesising recent empirical evidence from multiple databases, it provides a rigorous and comprehensive overview of the current state of AI-supported feedback research. Nevertheless, several limitations should be acknowledged. The review included only English-language, peer-reviewed publications and excluded grey literature, dissertations, and conference proceedings. Furthermore, because of the methodological heterogeneity of the included studies, a narrative thematic synthesis was undertaken instead of a quantitative meta-analysis. Consequently, the findings should be interpreted as a qualitative synthesis of current evidence rather than pooled statistical estimates.

In conclusion, this review demonstrates that generative AI is transforming second language writing instruction by redefining feedback as a collaborative process involving learners, teachers, and intelligent systems. The educational potential of AI extends beyond technological innovation; its value lies in supporting meaningful interaction, reflective revision, and learner autonomy while complementing teachers' pedagogical expertise. As AI technologies continue to evolve, future progress will depend on developing ethically responsible, theoretically informed, and pedagogically sound human-AI partnerships that enhance rather than replace effective language teaching. By synthesising the rapidly expanding post-2023 evidence base, this review contributes to a clearer understanding of AI-supported writing instruction and provides a foundation for future research, educational policy, and classroom practice.

This systematic review draws together empirical evidence published between January 2023 and July 2026 to explore the evolution of artificial intelligence (AI)-supported feedback on second language (L2) writing, focusing on the rise of human-AI feedback partnerships, student engagement, trust in AI, and feedback literacy. The review follows the PRISMA 2020 guidelines (Page et al., 2021) to provide a contemporary overview of research generated during the current LLM boom (e.g. ChatGPT, Claude, Gemini, Microsoft Copilot) by solely analysing research conducted following this explosion. Drawing on studies published from 2023 onwards, this review gives insight into the changing landscape of instructional and feedback practices in L2 writing, in post-LLM environments.

This review shows that AI-supported feedback now goes beyond simple correction tools: LLMs offer conversational, contextualized and personalized feedback across the writing process, differing significantly from earlier, static AI approaches focused purely on surface-level language corrections. In almost all cases of AI-supported feedback presented in studies, writing quality improved, as did revision behavior, engagement, and self-regulated learning behaviors, especially when embedded in instruction. However, AI showed less success in handling higher-order concerns of writing, including rhetorical organization, critique and disciplinary writing conventions, requiring the teacher's greater presence, judgment, and input (Shi & Aryadoust, 2024; Ding et al., 2026).

The central message from the current findings is the emergence of human-AI collaborative feedback partnerships as the defining pedagogical approach in contemporary L2 writing. Contrary to initial fears that the teacher's role would become diminished or redundant. A consensus is building that large language models are significantly valuable instructional agents supporting teacher pedagogies by offering valuable linguistic support, with classroom teachers retaining a vital role to guide learners through contextual considerations and ethics (Karagz, 2025; Pretorius & Thewissen, 2026). Many authors suggest that a combination of technology and human guidance is ideal in the field of automated feedback.

Furthermore, AI-supported feedback relied on students' skills in leveraging the feedback. When students effectively utilized human-AI feedback by engaging in critique and iteration and viewing the AI as a learning tool rather than an infallible expert system, it offered the greatest benefits in both overall writing skill and development of self-regulation, particularly when this occurred in conjunction with instruction on the development of critical feedback skills (Cengiz et al., 2025; Teng, 2024). Institutions need to prioritize training students to utilize feedback from large language models, given the development of critical and effective use of these types of systems. There were still limits within current AI evidence-based research, though, including shorter intervention lengths; a greater incidence of EFL instruction at the college/university level than across all educational levels; variable methodologies, with many qualitative studies and only two quantitative studies (with significant heterogeneity); and a low level of longitudinal evidence. Researchers recommended more rigor in design, such as longer intervention

times, further inquiry into ethical implications and implementation issues; research into not just collaborative and competitive human vs. Automated models, but also institutional issues like feasibility and scalability for establishing and maintaining new feedback practices over time, as well as building feedback literacy within both pedagogic and course curricula.

It is also worth noting that the current review provides certain strengths, in having drawn on relevant information, systematic searching, transparent inclusion criteria and a wide coverage of evidence from different countries in line with the PRISMA 2020 reporting process. On the other hand, limitations exist, such as including only academic papers written in English and excluding non-English literature and grey literature in addition, the research was completed using qualitative rather than quantitative methods. Due to the disparity of study designs, the data was evaluated using a narrative synthesis rather than performing a meta-analysis, which thus gives results that can only be seen as subjective.

Overall, generative AI is set to fundamentally redefine L2 writing Feedback in education, altering the dialogue between learners, teachers and intelligent systems. Its role as an effective instructional tool goes beyond merely enhancing technology by promoting interaction, critical revision, and autonomy while strengthening teacher feedback. Looking ahead, as technology in this field continues to evolve, so too will the focus on thoughtfully and ethically driven human and artificial intelligence that supplements (but not supplants) effective instruction. Using the exponentially expanding pool and depth of evidence available since 2023, the present study provides awareness in the area of AIDE writing and lays groundwork for more work in the theory, practice, and pedagogy of AIDE writing in learning and educational settings.

REFERENCES

1. Ajabshir, Z. F., & Ebadi, S. (2023). The effects of automatic writing evaluation and teacher-focused feedback on CALF measures and overall quality of L2 writing across different genres. *Asian-Pacific Journal of Second and Foreign Language Education*, 8, Article 26.
2. Ali, M. M., Alaa, A. M., Alharbi, W., & Al Qurashi, I. (2025). Investigating the Impact of ChatGPT on Reading Proficiency in EFL Context. *International Journal of Technology in Education*, 8(4), 1222-1244.
3. Ali, M. M., Anwar, M. N., Fazal, S., & Ayyaz, S. (Eds.). (2025). *AI applications for English language learning*. IGI Global.
4. Ali, M. M., Fazal, S., ullah Ubaid, U., Saeed, M. A., & Shah, S. R. A. (2025). Integrating AI-Powered Digital English Textbook at Grade12: Pakistani Students and Teachers' Perceptions. *International Journal of Adult Education and Technology (IJAET)*, 16(1), 1-20.
5. Ali, M. M., Imran, M., Khan, R., Sami, A., & Ali, Y. A. (2025). FROM LANGUAGE CLASSROOMS TO STEAM DISCIPLINES: THE IMPACT OF AI-DRIVEN STEAM LEARNING ON STUDENTS' COMMUNICATIVE PROFICIENCY. *Lex Localis: Journal of Local Self-Government*, 23(10), 750.
6. Cengiz, B. C., Bilki, Z., Ataş, A. H., & Celik, B. (2025). Exploring second language writers' engagement with ChatGPT feedback: Revision behaviors and perceptions. *System*, 134, Article 103837. <https://doi.org/10.1016/j.system.2025.103837>
7. Chen, Z., Zhu, X., Lu, Q., & Wei, W. (2024). L2 students' barriers in engaging with form- and content-focused AI-generated feedback in revising their compositions. *Computer Assisted Language Learning*. Advance online publication.
8. Craven, L., & Fredrick, D. R. (2026). LLM-generated feedback in L2 writing: A scoping review. *Education Sciences*, 16(8), Article 1196. <https://doi.org/10.3390/educsci16081196>
9. Ding, L., Zou, D., Kohnke, L., & Xie, H. (2025). Harnessing generative AI for L2 written corrective feedback: A systematic review of current research and future direction. In *Proceedings of the 2025 2nd International Conference on Artificial Intelligence and Teacher Education (ICAITE)* (pp. 43–48). IEEE. <https://doi.org/10.1109/ICAITE68636.2025.11442331>
10. Ding, L., Zou, D., Kohnke, L., & Xie, H. (2026). Harnessing generative AI for L2 written corrective feedback: A systematic review of current research and future directions. *Computer Assisted Language Learning*. Advance online publication.
11. Ekizoğlu, M., & Demir, A. N. (2025). The role of AI-assisted writing feedback in developing secondary students' writing skills. *Discover Education*, 4, Article 454.
12. Fan, N. (2023). Exploring the effects of automated written corrective feedback on EFL students' writing quality: A mixed-methods study. *SAGE Open*, 13(2).
13. Gomez, C., Cho, S. M., Ke, S., Huang, C.-M., & Unberath, M. (2025). Human–AI collaboration is not very collaborative yet: A taxonomy of interaction patterns in AI-assisted decision making from a systematic review. *Frontiers in Computer Science*, 6, Article 1521066. <https://doi.org/10.3389/fcomp.2024.1521066>

14. Hong, Q. N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., & Vedel, I. (2018). *Mixed Methods Appraisal Tool (MMAT), version 2018: User guide*. Canadian Intellectual Property Office.
15. Hussien, M. (2025). *Understanding human–GenAI collaboration in writing in the age of generative AI: A systematic review of literature*.
16. Hyland, K., & Hyland, F. (2019). *Feedback in second language writing: Contexts and issues*. Cambridge University Press.
17. 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*, Article 1834085. <https://doi.org/10.3389/feduc.2026.1834085>
18. Kao, C.-W., & Reynolds, B. L. (2024). Timed second language writing performance: Effects of perceived teacher versus perceived automated feedback. *Humanities and Social Sciences Communications, 11*, Article 1012.
19. Karagöz, İ. (2025). AI-generated feedback in English writing instruction for language learners: A systematic review. *The Reading Matrix, 25*(1), 51–67.
20. Luo, W., & Zhan, Y. (2026). Effects of AI feedback on students' English writing performance in higher education: A meta-analysis. *Cogent Education, 13*(1).
21. Maleki, A. (2026). Effects of AI-generated and teacher feedback on EFL learners' writing performance and emotional experience. *Discover Artificial Intelligence, 6*.
22. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Thomas, J., 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., McGuinness, L. A., Stewart, L. A., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. <https://doi.org/10.1136/bmj.n71>
23. Park, P., & Kim, T.-W. (2024). The effects of ChatGPT-assisted teacher feedback on Korean college students' English writing. *The Journal of Yeolin Education, 32*(6), 43–65.
24. Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N., Roen, K., & Duffy, S. (2006). *Guidance on the conduct of narrative synthesis in systematic reviews*. ESRC Methods Programme.
25. Pretorius, M., & Thewissen, J. (2026). Mediating L2 writing revision: The impact of teacher-moderated ChatGPT feedback on L2 accuracy and lexical diversity. *Journal of Second Language Writing, 71*, Article 101286.
26. Shi, H., & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. *ReCALL, 36*(2), 187–209. <https://doi.org/10.1017/S095834402300026X>
27. Singh, S., Hindriks, K., Heylen, D., & Baraka, K. (2025). *A systematic review of human–AI co-creativity*.
28. Teng, M. F. (2024). "ChatGPT is the companion, not enemies": EFL learners' perceptions and experiences in using ChatGPT for feedback in writing. *Computers and Education: Artificial Intelligence, 7*, Article 100270. <https://doi.org/10.1016/j.caeai.2024.100270>
29. Yan, D., & Zhang, S. (2024). L2 writer engagement with automated written corrective feedback provided by ChatGPT: A mixed-method multiple case study. *Humanities and Social Sciences Communications, 11*.
30. Yang, H., Gao, C., & Shen, H.-Z. (2024). Learner interaction with, and response to, AI-programmed automated writing evaluation feedback in EFL writing. *Education and Information Technologies, 29*.
31. Yu, H., Verspoor, M., & Peng, H. (2026). Effects of generative AI feedback and hybrid feedback in functional adequacy development in L2 academic writing. *Innovation in Language Learning and Teaching*. Advance online publication.
32. 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*. Advance online publication.