International Journal of Learning, Teaching and Educational Research
Vol. 25, No. 7, pp. 352-378, July 2026
https://doi.org/10.26803/ijlter.25.7.16
Received Apr 30, 2026; Revised Jun 15, 2026; Accepted Jun 24, 2026

AI-Driven Feedback for University Students' ESL/EFL Academic Writing: A Scoping Review of Pedagogical Implementation and Learning Outcomes

Xiaojun Yuan* **and Nee Nee Chan**
Faculty of Social Sciences and Liberal Arts,
UCSI University, Kuala Lumpur, Malaysia

Abstract. This review maps the recent findings on the AI-driven feedback given to undergraduate-level ESL/EFL academic writing, examining how it is implemented, what learning outcomes it produces, and what shapes those outcomes. Drawing on 39 peer-reviewed empirical studies retrieved from Web of Science, Scopus, and ERIC, and following the PRISMA extension for Scoping Reviews, this review covers 2,796 participants across 18 countries and regions. The findings show that 27 of the 39 studies in the full corpus reported measurable improvement in ESL/EFL writing. Lower-proficiency students showed larger surface-level gains but higher rates of passive uptake, while higher-proficiency learners engaged more critically when appropriately scaffolded. This review also identifies a persistent local-global revision gap: grammar and vocabulary improvements were reliably documented, but gains in argumentation, genre awareness, and critical reasoning remained weak and variable across the corpus. While AI-driven feedback shows genuine promise for addressing the feedback deficit in large-scale ESL/EFL writing instruction — a challenge with direct relevance to SDG 4 (Quality Education) and its advocation for inclusive, equitable access to quality learning at all levels - the conditions that produce durable learning gains remain poorly understood. The field's methodological development has not yet caught up with its empirical output.

Keywords: AI-driven feedback; ESL/EFL academic writing; scoping review; pedagogical implementation; learning outcomes; SDG 4; quality education

Citation :
Yuan, X., & Chan, N. N. (2026). AI-Driven Feedback for University Students' ESL/EFL Academic Writing: A Scoping Review of Pedagogical Implementation and Learning Outcomes. *International Journal of Learning, Teaching and Educational Research, 25*(7), 352–378.
https://doi.org/10.26803/ijlter.25.7.16

*Corresponding author: Xiaojun Yuan; *luckyyuan6@gmail.com*

© The Authors
This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0).

1. Introduction

Writing in a second language at the university level can be challenging. The demands go beyond language: students must also learn academic genres, argument structures, and disciplinary conventions that may differ markedly from those in their first language (Storch, 2018). Feedback is widely recognised as the mechanism through which L2 writers develop these skills (Hyland & Hyland, 2006), and a substantial body of research has established that timely, specific and formative commentary can meaningfully improve both the quality of the written products and the metacognitive processes that underpin sustained writing development (Carless & Boud, 2018). The problem is scale. Providing genuinely useful feedback to large groups of students takes time, and quality tends to decline when marking loads are high. This gap between what effective feedback requires and what large-scale instruction can reliably deliver is directly relevant to Sustainable Development SDG 4 (Quality Education), which calls for inclusive and equitable access to quality learning at all levels, including higher education.

Automated writing evaluation (AWE) systems or platforms such as Criterion and Pigai that use natural language processing to score and comment on student texts have been deployed in L2 writing classrooms since the early 2000s as one response to the scalability challenge. While the literature provides evidence that these tools can be effective at improving the surface-level aspects of students' writing, they offer very limited information on argumentation (Link et al., 2022). The public release of ChatGPT in November 2022 represented a qualitative discontinuity in this landscape. Students can now submit drafts, receive contextually sensitive, multifaceted, and conversational feedback, and ask follow-up questions. The educational technology community responded quickly, and empirical research on LLM-based writing feedback began appearing in peer-reviewed journals within months of ChatGPT's release. By 2025, the literature on AI-driven feedback in ESL/EFL academic writing had become substantial. Studies appeared across multiple disciplines.

The effect sizes range from negligible to implausibly large. Some researchers frame AI feedback as a complement to teacher commentary, while others position it as a potential replacement. Barrot (2023) presented an early and influential analysis of the limitations and opportunities of using ChatGPT for ESL composition instruction. This review identified three primary concerns regarding AI-generated feedback: student overreliance, academic integrity issues, and its tendency to focus primarily on surface-level aspects of the students' writing. Other reviews addressed Generative Artificial Intelligence (GenAI) more broadly in language learning contexts (Kohnke et al., 2023) or focused on the benefits and drawbacks of utilising ChatGPT in ESL settings (Yan, 2023). These reviews were informative; however, each reviewed the literature up to 2024, when the number of studies was relatively small. Scoping reviews are particularly suitable for synthesising evidence to determine what is known within a field and what remains unknown (Tricco et al., 2018).

The review is guided by four research questions:
RQ1: How is AI-driven feedback pedagogically implemented in university ESL/EFL academic writing instruction?
RQ2: What writing and learner outcomes are reported following AI-driven feedback interventions?
RQ3: What factors moderate the effectiveness of the AI-driven feedback on student writing?
RQ4: What methodological limitations characterise the current evidence base?

This review makes four specific contributions to the existing literature. Firstly, it covers the 2023–2026 publication window, which contains most of the empirically substantive work in this area that has not yet been systematically synthesised. Secondly, it addresses four explicit research questions spanning implementation, outcomes, moderators, and methodological characteristics, giving the review a structured account of these four dimensions. Thirdly, by analysing all empirical studies (n=39) from 18 countries and regions across three geographic regions, it provides a broad international mapping of the current evidence base. Fourthly, it describes the shortcomings of the existing body of evidence in sufficient detail to create a clear research agenda, rather than merely stating that additional work is needed.

2. Methodology

2.1 Research Design

This review used a scoping review methodology, following the five-stage framework proposed by Arksey and O'Malley (2005), later refined by Levac et al. (2010). Reporting followed the PRISMA extension for Scoping Reviews (PRISMA-ScR; Tricco et al., 2018). Because they map an area of research where studies use different methodologies and therefore cannot be pooled to identify commonalities, scoping reviews are particularly useful in areas such as AI-based feedback in ESL/EFL composition instruction. This area has grown dramatically since ChatGPT's introduction towards the end of 2022, and it continues to develop a common language and set of methods.

2.2 Eligibility Criteria

A priori criteria were set before the search began and held as fixed throughout. Methodological coherence (Criterion F) was applied as a data-extraction feasibility criterion at the full-text stage, not as a formal quality-appraisal threshold, in line with PRISMA-ScR (Tricco et al., 2018). Five conditions had to be met for a study to be included:

2.2.1 Participant population

University-level learners whose primary academic medium was English as a second or foreign language (ESL/EFL). Studies at primary, secondary, and other non-tertiary levels were excluded, as were those targeting native English-speaking students. Following ESL/EFL field conventions, studies describing the participants as 'university students' or 'college students' without specifying graduate standing were treated as undergraduate.

2.2.2 Intervention

The study had to involve AI-generated or AI-mediated written feedback in response to students' academic writing. This covered LLM feedback (e.g., from ChatGPT), AI-augmented AWE platforms, and hybrid systems that combine AI with teacher feedback. Studies that used AI only for text generation or translation, without a feedback component focused on student writing, were excluded.

2.2.3 Writing focus

The writing task had to be academic. Argumentative essays, research reports, genre-based writing, and comparable formal tasks were included. Creative writing and casual online communication were excluded.

2.2.4 Study type

Peer-reviewed empirical research in English was included. Conceptual articles, editorials, book chapters, conference papers and theses were excluded.

2.2.5 Publication window

The search covered January 2017 to February 2026. Although the search window began in 2017 to capture pre-LLM AWE studies, no study published before 2023 met all eligibility criteria. The corpus this review synthesises did not exist as a coherent line of inquiry prior to ChatGPT's public release in November 2022. Studies available online ahead of print at the time of the search were also eligible.

2.3 Information Sources and Search Strategy

The search was conducted across three databases: Scopus, Web of Science (Core Collection), and ERIC (via ProQuest). These platforms together cover applied linguistics, CALL, educational technology, and second language acquisition, which is where empirical work on AI-mediated writing feedback predominantly appears. The search was completed in February 2026.

Table 1: Search Strings used in the three Databases

Database	Search String
Scopus	TITLE-ABS-KEY (("ChatGPT" OR "generative AI" OR "large language model" OR LLM OR "automated writing evaluation" OR AWE OR "AI feedback") AND ("feedback" OR "written corrective feedback" OR "corrective feedback") AND ("academic writing" OR "L2 writing" OR "EFL writing" OR "ESL writing") AND ("university" OR "higher education" OR "undergraduate" OR "tertiary")) AND PUBYEAR > 2017 AND PUBYEAR < 2026
Web of Science (Core Collection)	TS = (("ChatGPT" OR "generative AI" OR "large language model" OR LLM OR "automated writing evaluation" OR AWE OR "AI feedback") AND ("feedback" OR "written corrective feedback" OR "corrective feedback") AND ("academic writing" OR "L2 writing" OR "EFL writing" OR "ESL writing") AND ("university" OR "higher education" OR "undergraduate" OR "tertiary")) Refined by: PUBLICATION YEARS: 2017–2026; DOCUMENT TYPES: Article
ERIC (via ProQuest)	("ChatGPT" OR "generative AI" OR "large language model" OR "LLM" OR "automated writing evaluation" OR "AWE" OR "AI feedback") AND ("feedback" OR "written corrective feedback" OR "corrective feedback") AND ("academic writing" OR "L2 writing" OR "EFL writing" OR "ESL writing") AND ("university" OR "higher education" OR "undergraduate" OR "tertiary") Filters: Peer-Reviewed; Date: 2017–2026
Total records retrieved	n = 636 (Scopus: 418; Web of Science: 36; ERIC: 182). Search was conducted in February 2026.

2.4 Study Selection

All records from the three databases were loaded into Rayyan (Ouzzani et al., 2016) after duplicates were removed. The search yielded a total of 636 records (Scopus: n=418; Web of Science: n=36; ERIC: n=182). The first stage (screening) of the title and abstract review process for the 612 unique references yielded 24 duplicates. Any disagreements as to whether an item was included for evaluation in the next stage were resolved through discussion. In total, 559 items were excluded for failing to meet one or more inclusion criteria. The next step in this process was to retrieve the full texts of all studies that met the incorporation criteria, for a total of 53. After reviewing the study details, the authors determined that an additional 14 studies did not qualify for inclusion in the review. Full citations and specific exclusion reasons for all 14 full-text excluded studies are reported in Appendix. The process outlined above is shown graphically in Figure 1 in a format consistent with a PRISMA-ScR flow diagram.

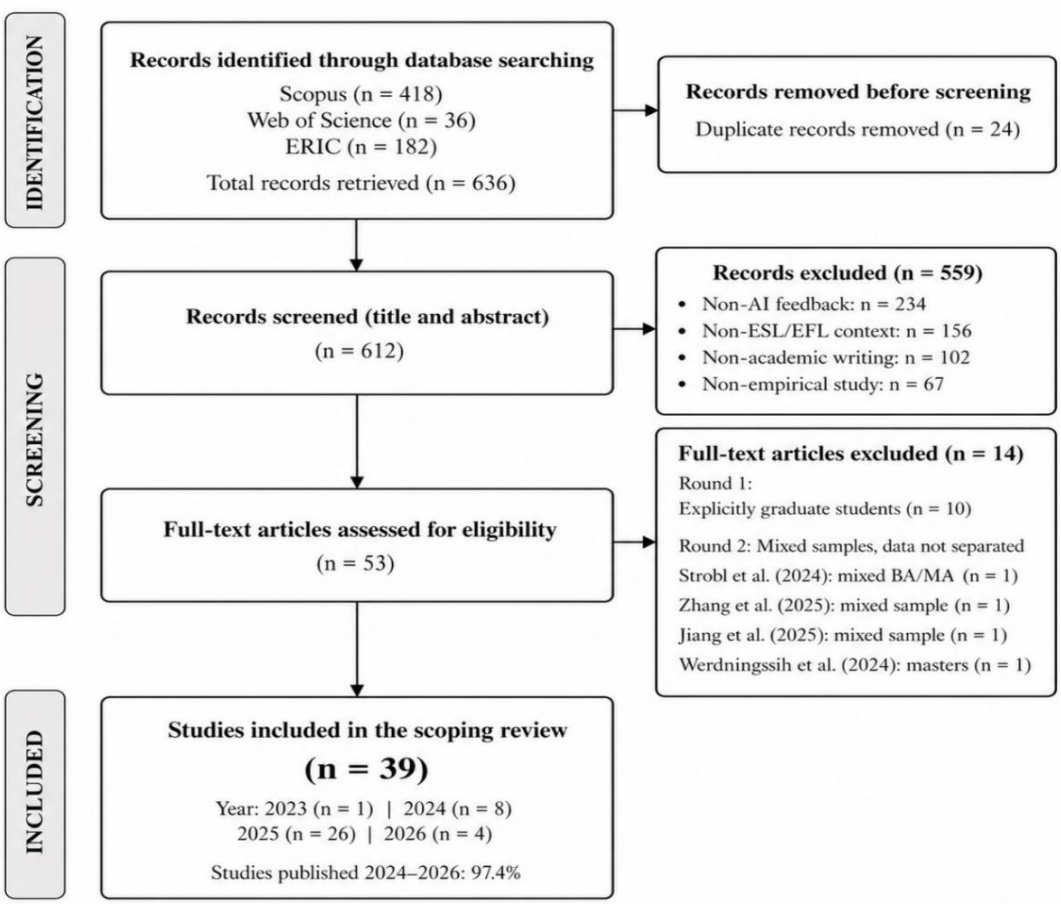

Figure 1: PRISMA 2020 Flow Diagram used for the Study Selection

2.5 Data Extraction

A structured data extraction form was developed in Microsoft Excel and piloted on five studies before full implementation. The template used 38 fields across seven clusters: study identification, participant characteristics, AI tool specifics, pedagogical implementation, research design and methods, learning outcomes, and study quality indicators. The first author extracted all relevant data from the sources. The second author double-checked each field against the source articles, and any discrepancies were resolved by consulting the source articles. No conversions were made to the effect size values as they were reported in their existing form.

2.6 Synthesis

Due to the differences among the studies regarding design, tools and outcome measures, a statistical meta-analytic approach could not be employed. Instead, the studies were synthesised using narrative synthesis principles (Popay et al., 2006), enhanced with thematic analysis techniques (Braun & Clarke, 2006). Descriptive statistics were utilised to provide an overview of the bibliometric data. Narrative synthesis strategies were applied to address RQ1 (pedagogical implementation) and RQ2 (writing and learner outcomes). Cross-study comparisons were conducted to assess whether the outcome patterns varied by proficiency level, teaching method, treatment duration, type of AI tool, and student training, as per Research Question 3 (moderators of effectiveness). Finally, Research Question 4

(methodological limitations) involved conducting an examination of the research methodologies, measurement approaches, and areas of educational achievement that were either missing or underexplored.

3. Results

3.1 Overview of the Included Literature

Despite a search window extending back to January 2017, no studies published before 2023 met all five eligibility criteria. The 39 included studies were published between 2023 and early 2026. However, the number of studies each year is uneven. There was only one study in 2023, and eight in 2024. In addition, 26 studies were released in 2025, and four were released in early 2026 prior to conducting the search. The four 2026 publications include both formally published articles and advance online articles retrieved through the February 2026 search date. Taken together, nearly 97.4% of the studies included in this review (n=38) appeared during the years 2024-2026.

There is a heavy geographic clustering in East and Southeast Asia. A total of 21 studies is from East Asia. Six studies are from Southeast Asia, including Vietnam, Thailand and Indonesia. The Middle East and North Africa contribute seven studies, mainly from Iran (n=3) and Saudi Arabia (n=2). The final five studies are distributed across Turkey (n=2), India, Chile, and a cross-institutional Egypt/UK collaboration. No included study was conducted in a Western anglophone context, a geographic skew that plausibly reflects the concentration of EFL university writing instruction in Asia and the Middle East. It also raises questions about the generalisability to ESL settings in countries such as Canada, the United Kingdom or Australia. Together, the 39 studies included in this review involved 2796 participants. The sample sizes ranged from four (Yeung, 2025) to 452 (Teng, 2025) participants, with a median of 50 (M=71.7). In terms of duration, approximately half of the studies were designed to run for anywhere from 9 to 16 weeks within a single course.

Table 2: Characteristics of the Included Studies

Characteristic	Category	n	Percentage	Notes
Publication year	2023	1	2.6%	
	2024	8	20.5%	Majority published 2024–2026: 97.4%
	2025	26	66.7%	
	2026	4	10.3%	
Study design	Mixed methods	26	66.7%	
	Qualitative	8	20.5%	
	Quantitative (incl. experimental / quasi-experimental)	5	12.8%	
Usage mode	Teacher-directed	23	59.0%	Instructor-structured or assigned AI use
	Student-initiated / self-directed	15	38.5%	Students used AI autonomously
	Mixed	1	2.6%	Both modes present in the same study
Feedback type	Multiple types	33	84.6%	Grammar, vocabulary, content, structure, organisation, etc.
	Content only	3	7.7%	Ideas, argumentation, logical structure
	Grammar and mechanics only	3	7.7%	Surface-level error correction
Human feedback combined	No (AI only)	15	38.5%	AI feedback without teacher or peer input
	Yes – teacher feedback	15	38.5%	AI + teacher written corrective feedback
	Yes – teacher + peer	5	12.8%	All three feedback sources combined
	Yes – peer feedback only	4	10.3%	AI + peer review; no teacher feedback
Writing pedagogy	Process writing	33	84.6%	Drafting, revision, and feedback cycles
	Product-based	3	7.7%	Focus on final written product
	Genre-based	1	2.6%	Explicit genre instruction (Tai et al., 2025)

	NR	2	5.1%	
Learning outcomes	Writing quality – Improved	27	69.2%	Pre–post gains or positive holistic quality ratings
	Writing quality – Not measured / NR	11	28.2%	Qualitative, perceptual, or process-outcome studies only
	Writing quality – Mixed	1	2.6%	Inconsistent findings across groups or criteria
	Student acceptance – Positive	24	61.5%	Favorable attitudes toward AI feedback
	Student acceptance – Mixed	5	12.8%	Some positive, some negative reactions
	Student acceptance – NR	10	25.6%	
	Engagement / motivation – Increased	23	59.0%	Higher task motivation or writing confidence
	Engagement / motivation – Mixed / negative / NR	16	41.0%	
	Revision behavior – More revisions	25	64.1%	Greater number or depth of text revisions
Learning outcomes	Learner autonomy – Developed	23	59.0%	Self-regulation or independent writing gains
RQ relevance	RQ1 (Implementation) – High	29	74.4%	
	RQ1 – Medium	8	20.5%	
	RQ1 – Low	2	5.1%	
	RQ2 (Outcomes) – High	32	82.1%	
	RQ2 – Medium	5	12.8%	
	RQ2 – Low	2	5.1%	
RQ relevance	RQ3 (Moderators) – All studies	39	100.0%	Cross-study synthesis; all studies contribute
RQ relevance	RQ4 (Methods) – All studies	39	100.0%	Methodological features charted across the full corpus

Note: The RQ1 and RQ2 relevance ratings reflect individual study-level contributions (High/Medium/Low). RQ3 and RQ4 are addressed through cross-study synthesis. All 39 studies contribute to both.

3.2 RQ1: Pedagogical Implementation of AI-Driven Feedback

Across the 39 included studies, the implementation of AI-driven feedback took strikingly varied forms—in the tools selected, the degree of teacher mediation, the duration and sequencing of activities, and the extent to which AI feedback was integrated with or treated as a substitute for human commentary. While these dimensions are not mutually exclusive, they illustrate an ongoing process among researchers as to what AI-driven feedback is and what role it should take in the pedagogy of writing instruction.

The near ubiquity of ChatGPT is the most immediately striking feature of the corpus. It appeared in some form in 29 of the 39 studies (74.4%), typically without a comparative rationale. Studies rarely justified why they used ChatGPT rather than an alternative. The principal exception was Hong and Shin (2025), who explicitly contrasted restricted, guided, and free ChatGPT integration conditions. A smaller cluster of studies moved beyond ChatGPT. Derakhshan (2025) worked with Gemini across a 16-week intervention, finding significant gains in writing proficiency and reductions in writing anxiety.

Tran (2025) also utilised Gemini to examine the impact of sequencing AI and teacher feedback during revision cycles. Wang and Xie (2025) utilised ERNIE Bot (Baidu's Chinese language model), because institutional restrictions prohibited researchers from utilising OpenAi tools. Researchers based in mainland China report having to obtain their students' drafts through researchers to collect ChatGPT feedback from the students where the students themselves could not access the platform directly (Chen et al., 2025). The tool landscape described in this body of literature is shaped by both the implemented research design and geopolitical infrastructure.

Most implementations were conducted under the direction of teachers. In 23 studies (59%), instructors developed the prompts for AI, located the AI within the writing cycle, and either reviewed the output prior to providing it to the students or organised discussions around the output after providing it. An example of this would be in Wang et al. (2026): this study involved a semester-long framework that included developing structures for using ChatGPT with the goal of creating active and inquiry-based learning environments. Student-initiated use of AI was present in 16 studies (41.0%).

This distinction is important from a practical perspective. Implementations in teacher-directed environments limited what the students could ask AI to assist with in terms of producing complete and grammatically correct sentences, limiting the potential for the students to simply use AI to create correct answers to the writing assignments. Voluntary and self-directed participation when using AI was highly dependent upon individual levels of proficiency and pre-existing beliefs regarding feedback as found by Yeung (2025). These factors are far less controllable once AI use is allowed to occur without instruction from teachers.

Four forms of human-AI interaction occurred throughout the literature. Fifteen studies (38.5%) relied solely on AI as a source of written feedback due to the nature of the study design or the lack of opportunity for teacher commentary.

Another 15 studies (38.5%) integrated AI with teacher feedback, treating them as complementary. Five studies (12.8%) paired AI with peer review, and four studies used all three sources. Zou et al. (2025) provided one of the most direct comparisons: students demonstrated greater utilisation and accuracy in their revisions when receiving teacher feedback overall, and they were more responsive to AI comments pertaining to organisational characteristics specifically. Bozorgian and Rahimi (2025) reported a similar finding: while AI produced technical consistency in grammar and coherence, peer feedback was more subjective and often prompted self-regulatory reflection. Neither form can accomplish everything independently.

Multi-dimensionality in terms of feedback that includes content, organisation, accuracy, and vocabulary simultaneously was by far the most common design occurring in approximately 34 studies (87.2%). For the most part, this multi-dimensionality was a result of how LLMs operate by default rather than being an intentional instructional decision made by educators. Students submitted draft versions of their papers and received broad-spectrum commentary. Whether this multi-dimensionality results in improvements is a separate question.

Qualitative data collected from several studies indicated that students continuously prioritise suggestions made by AI that pertain to surface-level aspects of their work, even when both content and structural feedback are available. Most interventions ran for an entire semester, meaning that there were many opportunities for AI feedback to inform revisions between submissions. Some studies demonstrated how to complete a draft submission before the students began work on a draft (Al Mahmud, 2026; Mahapatra, 2024), while other studies developed multi-session preparatory training for students as part of the course design.

The teachers' roles were identified as facilitative in 27 studies (69.2%): they designed the template prompts, reviewed the AI-generated outputs, facilitated post-AI feedback discussions, and planned the placement of AI comments within a broader course plan. Teacher involvement was limited in three studies. One of the clearer comparisons was offered by Alnemrat et al. (2025), in that AI-driven feedback led to a smaller increase in argumentative writing quality than teacher feedback alone. However, AI-driven feedback consistently produced better results and reduced the time spent grading by the teachers.

Table 3: Bibliographic Overview of the Included Studies

No.	Study	Country / Region	Journal	AI tool
1	Wang et al. (2026)	China	Comput. Educ.: Artif. Intell.	ChatGPT
2	Nguyen et al. (2025)	Vietnam	Educ. Inf. Technol.	ChatGPT
3	Ozdere (2025)	Turkey	IJTE	ChatGPT + other
4	Alnemrat et al. (2025)	Jordan	Front. Educ.	ChatGPT
5	Wei (2023)	China	Front. Psychol.	Other
6	Huang & Wang (2025)	China	System	ChatGPT
7	Zhang et al. (2025)	Hong Kong	System	ChatGPT
8	Allen & Mizumoto (2024)	Japan	RELC J.	ChatGPT
9	Xu & Jumaat (2024)	China	iJIM	ChatGPT
10	ElEbyary & Shabara (2024)	Egypt / UK	TET	ChatGPT
11	Muñoz Muñoz et al. (2025)	Chile	System	ChatGPT
12	Hong et al. (2025)	Thailand	Forum Linguist. Stud.	ChatGPT
13	Fan et al. (2024)	Hong Kong	Aust. J. Appl. Linguist.	ChatGPT + Grammarly + QuillBot
14	Hong & Shin (2025)	South Korea	System	ChatGPT
15	Tran (2025)	Vietnam	Educ. Sci.	Gemini
16	Tseng & Lin (2024)	Taiwan	Electron. J. eLearning	ChatGPT
17	Kurt & Kurt (2024)	Turkey	JITE:R	ChatGPT
18	Aljohani (2025)	Saudi Arabia	Forum Linguist. Stud.	ChatGPT
19	Wang & Xie (2025)	China	Educ. Psychol.	ERNIE (Baidu)
20	Sawaki et al. (2025)	Japan	Lang. Test.	GPT-4 turbo (API)
21	Al Mahmud (2026)	Saudi Arabia	World J. Engl. Lang.	ChatGPT

22	Han & Li (2024)	China	System	ChatGPT
23	Hao et al. (2026)	China	SAGE Open	Pigai + ChatGPT
24	Kohnke et al. (2025)	Hong Kong	Comput. Educ.: Artif. Intell.	ChatGPT + other
25	Ren et al. (2025)	China	Cogent Educ.	ChatGPT
26	Alyasin & Shah (2026)	Qatar	Ampersand	ChatGPT
27	Hongxia & Razali (2025)	China	Int. J. Instr.	ChatGPT
28	Mahapatra (2024)	India	Smart Learn. Environ.	ChatGPT
29	Tai et al. (2025)	Taiwan	Int. J. Appl. Linguist.	ChatGPT
30	Tanharat & Phootirat (2025)	Thailand	rEFLections	ChatGPT
31	Zare et al. (2025)	Iran	ReCALL	ChatGPT
32	Zou et al. (2025)	China	Comput. Assist. Lang. Learn.	ChatGPT
33	Bozorgian & Rahimi (2025)	Iran	J. Lang. Educ.	ChatGPT
34	Derakhshan (2025)	Iran	Asia-Pac. Educ. Res.	Gemini
35	Teng (2025)	Macau	System	ChatGPT
36	Yeung (2025)	Hong Kong	J. Second Lang. Writ.	Learnalytics (AWE)
37	Chen et al. (2025)	Macau	ECNU Rev. Educ.	ChatGPT
38	Puteri et al. (2025)	Indonesia	Indones. J. Appl. Linguist.	ChatGPT
39	Stofiana et al. (2025)	Indonesia	Ampersand	Grammarly + ChatGPT

Note. AI tool: 'ChatGPT + other' = ChatGPT combined with one or more additional platforms. NR = not reported.

3.3 RQ2: Writing and Learner Outcomes

Two categories of outcome are reported here. Writing outcomes concern changes in the texts that students produce, regarding quality ratings, accuracy and complexity measures and changes in specific textual features. Learner outcomes concern changes in the students themselves such as perceptions, revision behaviour, motivation, autonomy, and writing-related anxiety.

3.3.1 Writing Outcomes

Interpreting the outcomes report requires starting with an awareness of the measurements made. In addition to using different methods of assessment to measure the quality of writing and varying degrees of evaluation for instrument

use and the time to evaluate the outcomes, 11 studies report no measures of pre- to post-writing quality. Instead, these studies measured the writers' processes, their perceptions, or their writing strategies. While this variation among the studies does not limit our interpretation of the results, we need to consider the degree of variability when combining the data from these studies to draw general conclusions.

The most consistent and analytically important finding across the corpus is not that AI feedback improves writing in general, but that improvement is systematically asymmetric across linguistic levels. Across the 28 studies reporting any pre-post writing measure, the pattern is recurrent: gains in grammatical accuracy, lexical range, and sentence-level clarity are reliably documented and frequently statistically significant; gains in argumentation, genre awareness, and critical reasoning are weaker, more variable, and in several cases statistically null. Zou et al. (2025) found that students preferentially uptook organisational and surface-level AI suggestions over content-level ones; Stofiana et al. (2025) reported that students accepted AI's structural suggestions more readily than its rhetorical ones; ElEbyary and Shabara (2024) documented that AI-generated commentary on argument structure was often technically plausible but contextually shallow. Hao et al. (2026) provided a large effect size for the application of higher-order thinking.

3.3.2 Learner Outcomes

Of the 29 studies on how students viewed AI-driven feedback, the vast majority (24) reported largely positive views. The students felt that AI-driven feedback provided them with timely, easily available, and broad-based feedback on their assignments which produced less anxiety than receiving feedback from an instructor or peer (Kurt & Kurt, 2024; Kohnke et al., 2025). The other five studies identified a mix of both positive and negative views towards AI-driven feedback. In one study (Fan et al., 2024), it was found that many EAP instructors were concerned that AI-driven feedback could adequately identify students' writing needs in relation to discipline-specific expectations. Another study (Stofiana et al., 2025) demonstrated that students generally accepted the suggestions made by AI for structuring content better than they did for addressing how effectively the content was written rhetorically.

The motivational and emotional responses of the sampled students were documented in 28 studies; student motivation, interest, and confidence in writing were documented positively in 22 studies. Writing anxiety as a construct was assessed in just one study, where Derakhshan (2025) measured writing anxiety using Gemini over 16 weeks and found a statistically significant decrease. The data also indicated a downward trend according to Latent Growth Curve Modelling. Writing anxiety has been widely recognised as a barrier to English language learners' acquisition of academic writing skills in their second language (Horwitz et al., 1986).

3.4 RQ3: Moderators of Effectiveness

Across all of the qualitative and quantitative studies reviewed, learner ability has been shown to have the largest impact on the amount of learning from an

aggregate perspective. Notably, less able students benefit the most from increases in overall writing quality yet are the least likely to engage in higher-level thinking when responding to feedback. Yeung (2025) noted that the student with the highest proficiency demonstrated a greater proportionate response to AWE feedback and used a higher proportion of content-based suggestions than their lower-proficiency peers, who primarily sought corrective language-level feedback. Finally, Teng (2025) provided what may well be the most detailed analysis regarding how student metacognitive awareness mediates the quality of autonomous engagement with ChatGPT.

The degree to which usage was controlled and how much teachers structured the student's responses influenced the magnitude of treatment effects for both modalities. Using their three-level design, Hong and Shin (2025) demonstrated the same thing experimentally. Students who received guidance had better scores on the quality of their argumentative writings compared to those receiving only free usage and/or restricted usage. There was a high degree of variability among the group using free usage due to the varying degrees of metacognitive awareness and previous AI experience. The training provided to the students emerged as another moderating factor. The studies that included training on prompt usage (Tseng & Lin, 2024), evaluating AI-generated outputs (Wang et al., 2026), and ethically using AI commentary demonstrated more selectivity and criticism when it came to the use of AI commentary relative to studies that did not provide such training or that only provided a short demonstration.

3.5 RQ4: Methodological Limitations and Research Gaps

The 39 included studies provided valuable evidence related to both implementation and outcome issues. When viewing this body of knowledge as a methodological continuum, the evidence is limited by descriptive breadth while lacking in causality depth and time horizon length.

3.6 Methodological Characteristics of the Included Studies

Table 4 summarises the methodological characteristics of the 39 included studies across six domains: study design, sample size, presence of a control or comparison group, intervention duration, outcome measurement type, and evidence quality indicators. These characteristics have a direct impact on the strength of the effectiveness claims that can be drawn from the corpus.

Table 4: Methodological Characteristics of the Included Studies

Criterion	Category / metric	n (%)	Notes
Study design	Mixed methods	26 (66.7%)	Predominant design across the corpus.
	Qualitative	8 (20.5%)	Interview-, case-, or thematic-analysis-based studies.
	Quantitative, non-comparison	3 (7.7%)	Pure quantitative studies without a formal experimental/quasi-experimental group comparison.
	Experimental	1 (2.6%)	True experimental design with a control/comparison condition.
	Quasi-experimental	1 (2.6%)	Non-randomised comparison-group design.
Sample size	Range: 4-452; Median: 50	39 studies	Total N across the corpus = 2,796.
	Small (N < 30)	8 (20.5%)	Range: 4-20.
	Medium (30 <= N <= 100)	23 (59.0%)	Actual range in the corpus: 30-97; median = 50.
	Large (N > 100)	8 (20.5%)	Range: 102-452. Largest study N = 452.
Formal control / comparison group	Present	2 (5.1%)	Includes the experimental and quasi-experimental studies identified in Table 2.
	Absent or not formally designed as a group comparison	37 (94.9%)	Includes mixed methods, qualitative, quantitative non-comparison, single-group, or descriptive designs.
Intervention duration	Single session (< 1 day)	3 (7.7%)	One-off writing task, AI-feedback activity, or rating/feedback session.
	1-4 weeks	6 (15.4%)	Short-term interventions, including approximately one month.
	5-8 weeks	6 (15.4%)	Medium-short interventions, including approximately two months.
	9-12 weeks	3 (7.7%)	Medium-term interventions.
	>= 13 weeks / full semester	15 (38.5%)	Longer-term or semester-length interventions.

Criterion	Category / metric	n (%)	Notes
	Session-based, not directly convertible to weeks	1 (2.6%)	Reported as nine sessions rather than calendar weeks.
	Not reported (NR)	5 (12.8%)	Duration unspecified in the article.
Outcome measurement type	Writing quality assessed	28 (71.8%)	Includes 27 clearly positive writing-quality findings and one mixed writing-quality finding in Table 4.
	Perception / process only or writing quality NR	11 (28.2%)	Studies without reported/measured writing-product quality outcomes.
Reported quantitative evidence	Key statistics reported	24 (61.5%)	Includes effect sizes and other quantitative indicators reported in Table 4, such as d, eta-squared, r, beta, M, MD, chi-square, kappa, t, and uptake percentage.
	No key statistic reported	15 (38.5%)	Indicated by an em dash in the Table 4 key-effect-size column.

Note. Percentages are calculated from the total corpus (N = 39) and rounded to one decimal place.

The research gaps identified across the corpus can be grouped into four domains. Design gaps include that the field needs preregistered studies with active comparators and random or quasi-random allocation, and no such study exists in the current corpus. The measurement gaps are that a shared battery of writing-quality instruments is needed before cross-study comparisons become reliable, and that the current diversity of rubrics, indices, and self-reported measures make accumulation nearly impossible. The outcome gap is that writing transfer to non-AI contexts, long-term retention, writing anxiety and academic integrity as a behavioural variable are each absent or near-absent from the evidence base and represent the highest-priority targets for future work. The geographic gaps is that anglophone ESL/EFL contexts, Latin America, sub-Saharan Africa, and South and Central Asia remain entirely unrepresented; whether the patterns documented here generalise to these settings is unknown.

4. Discussion

The most consistent finding across the corpus was not that AI-driven feedback improves writing quality overall, but that its benefits are unevenly distributed across the different levels of writing: surface-level accuracy gains are the most robustly documented outcome, while higher-order skills such as argumentation, genre awareness, and critical reasoning remain comparatively underexplored and, where measured, show more variable and condition-dependent gains. This surface–higher-order asymmetry is the central conceptual contribution of the

review, and it has direct implications for both pedagogical implementation and the research agenda.

4.1 Pedagogical Implementation: AI as a Configured Resource

The best-supported theoretical finding relates to how teachers can structure an implementation to produce better outcomes. Compared to other AI implementation methods, research has shown that teacher-driven implementations have led to greater improvements in higher-level writing characteristics and more deliberate student engagement with both the AI-generated content and the AI's suggested revisions. Students also engaged less passively with AI-generated responses. These results support Vygotsky's (1978) idea of a "zone of proximal development," which holds that tools such as AI are used optimally when learners receive some form of mediation or scaffolding to engage with them at a level of difficulty beyond what they could achieve independently.

4.2 Writing and Learner Outcomes: An Asymmetric Picture

It is essential to interpret the persistent differential in the effective levels of surface-level and higher-order feedback. Barrot (2021) and Link et al. (2022) have previously noted this phenomenon in the AWE literature, attributing it to the limitations of rule-based NLP. It was anticipated that LLMs would overcome such limitations. Only partial success occurred. The corpus shows that while students can receive LLM-based feedback on their argument structure, content adequacy and genre conventions, they tend not to rely on it. It seems that the bottleneck has moved from the tool to the learner. To effectively engage with the feedback at the content level, learners need to possess the ability to assess whether the AI's suggestions are accurate, contextually relevant, and in alignment with their intended communication, all of which abilities that require learners to develop writing knowledge and feedback literacy (Carless & Boud, 2018).

That AI plus human feedback generally outperforms AI-only configurations is both practically significant and theoretically predictable. Feedback in the SLA literature is not simply a mechanism for correction; it is a dialogical process that involves how informationally rich a comment is mediated by the social relationship between the provider and recipient, as well as the recipient's ability to process and utilise the feedback, and the instructional environment (Hyland & Hyland, 2006).

4.3 Moderators: What Conditions Support Effectiveness

Moderator evidence in this corpus identifies when and where AI feedback is more likely to be effective; it does not explain why. Learner-related moderators have been studied less often than context-related ones, and most research has focused on a single construct per study. Most studies have not examined interaction effects to determine whether learner-related variables should be classified as moderators or simply as covariates. As a result, claims about when AI "works for" or "does not work for" certain types of learners are supported by less empirical evidence than claims about combinations of contextual factors. Researchers need to design studies that examine at least two learner-related constructs alongside at least two contextual constructs.

4.4 The Methodological Maturity of the Evidence Base

Empirical output has grown sharply: 30 of 39 studies (76.9%) appeared in 2025–2026. Methodological development has not kept pace. However, the growth in empirical output far exceeds that of the methodological structure. There are no studies employing pre-registration. Additionally, there are no studies utilising active comparators with random or quasi-random assignments. Operationalising writing quality has utilised rubrics, automated indexes, and self-report scales. As such, there is no single measure that is suitable for general use. The above describes the evidence maturity of the research field but does not describe a research field devoid of evidence. Rather, it describes a field where evidence of maturity lags behind the rapid increase in production.

4.5 The East Asian Concentration: A Structural Reading

The geographic concentration among the included studies can be justified and should be acknowledged. Specifically, 21 of 39 studies (53.8%) were conducted in East Asia—primarily mainland China —followed by research from Japan, South Korea, and so on. Several structural conditions may plausibly contribute to this pattern, and these are offered below as candidate explanations rather than established causal mechanisms.

There are three possible contributing patterns to this study's dynamic. The first pattern is related to "policy." In mainland China, national initiatives have supported the integration of artificial intelligence into higher education to advance educational priorities in AI-assisted teaching and to provide institutional incentives for tertiary English instructors to utilise AI in their classrooms. The second contributing pattern is referred to as "industrial proximity." Both China and South Korea are leading players in the international technology industry, and there are active relationships between technology companies, universities, and education researchers. The final contributing factor is the size and visibility of the ESL/EFL learner population.

5. Limitations and Conclusion

5.1 Limitations

Several limitations should be acknowledged before drawing conclusions. The first is that only research written in English has been used in this study; it is possible that additional research exists in Chinese, Spanish, Arabic, and possibly other languages that may have been overlooked because of this limitation, further compounding the geographical skew identified by RQ4. In addition, since a formal critical assessment of the methodology was not conducted as part of the scoping review (Peters et al., 2020), all studies were based on their ability to present coherent methodologies rather than on an evidence-based evaluation. Therefore, the conclusions derived from this review reflect the full breadth of methods utilised within this body of literature. The diversity of the design formats, measurement tools, and outcomes reported across the entire collection of studies precludes the possibility of developing a quantitative summary of results through pooled data or meta-analysis. All comparisons made among these studies will remain descriptive.

5.2 Conclusion

Overall, across 39 studies carried out over 18 countries and regions during the LLM era of language education, the findings were generally positive but systematically asymmetric: gains were reliably observed at the surface level of writing, while higher-order outcomes — argumentation, genre awareness, and critical reasoning — remained the weakest dimension of the evidence base. First, teacher-structured AI use is more effective for student-AI interactions than for students working autonomously. Second, when both human and AI inputs are used together, AI-plus-human configurations tend to produce better results than either AI-alone or human-alone configurations. Third, explicit student training in interpreting, evaluating, and selectively applying AI-generated feedback is not supplemental scaffolding. Fourth, student proficiency impacts all of the previous conditions. AI feedback tools will continue to improve, and the specific platforms documented in this review will be superseded.

The most conceptually consequential finding of this review is the systematic gap between surface and higher-order gains. Across the corpus, AI-driven feedback reliably improved grammatical accuracy and surface-level fluency; its contribution to argumentation, genre awareness, and critical reasoning was documented in a minority of studies and consistently depended on teacher-directed implementation, explicit genre scaffolding, and the use of purpose-built AI tools designed for higher-order feedback. Finally, one last theme emerged from this review — that realising the potential of AI feedback to help achieve SDG 4 will require research and implementation efforts extending far beyond those occurring primarily in university ESL/EFL settings.

6. Conflicts of Interest

The authors declare there to be no conflicts of interest. No financial, personal, or professional relationships have influenced the design, conduct, analysis or reporting of this scoping review.

7. Acknowledgements

The authors would like to acknowledge the use of Claude (Anthropic, Claude Opus 4.6) for assistance with language editing, formatting verification, and refining the manuscript's clarity. All AI-generated outputs were critically reviewed, verified, and edited by the authors, who take full responsibility for the content, interpretation, and intellectual contributions of this paper.

8. References

Al Mahmud, F. (2026). Examining the efficacy of ChatGPT and human-derived corrective feedback in addressing grammatical errors in Saudi EFL students' compositions. *World Journal of English Language, 16*(2), 258–270. https://doi.org/10.5430/wjel.v16n2p258

Aljohani, N. J. (2025). Enhancing writing accuracy and empowering students: The transformative influence of ChatGPT's informative feedback on students' writing. *Forum for Linguistic Studies, 7*(7), 1109–1128. https://doi.org/10.30564/fls.v7i7.9849

Allen, T. J., & Mizumoto, A. (2024). ChatGPT over my friends: Japanese English-as-a-foreign-language learners' preferences for editing and proofreading strategies.

RELC Journal. Advance online publication. https://doi.org/10.1177/00336882241262533

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*, Article 1614673. https://doi.org/10.3389/feduc.2025.1614673

Alyasin, A., & Shah, M. A. (2026). Human-AI collaboration in academic writing: Exploring university students' agency through a sociocultural lens. *Ampersand, 16*, Article 100256. https://doi.org/10.1016/j.amper.2026.100256

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology, 8*(1), 19–32. https://doi.org/10.1080/1364557032000119616

Barrot, J. S. (2023). Using automated written corrective feedback in the writing classrooms: effects on L2 writing accuracy. *Computer Assisted Language Learning, 36*(4), 584–607. https://doi.org/10.1080/09588221.2021.1936071

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

Bitchener, J., & Ferris, D. R. (2012). *Written corrective feedback in second language acquisition and writing*. Routledge.

Bozorgian, H., & Rahimi, H. (2025). Peer e-feedback and ChatGPT-4o in EFL writing: A cognitive-interpersonal comparison based on EFL students. *Journal of Language and Education, 11*(4), 54–65. https://doi.org/10.17323/jle.2025.27195

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

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

Chen, Z., Wei, W., Zhu, X., & Yao, Y. (2025). Unpacking the rejection of L2 students toward ChatGPT-generated feedback: Explanatory research. *ECNU Review of Education, 9*(1). https://doi.org/10.1177/20965311241305140

Derakhshan, A. (2025). Tracking the effects of Gemini as a GenAI tool on L2 learners' writing proficiency and anxiety: Latent growth curve modeling approach. *The Asia-Pacific Education Researcher*, 1–11. Advance online publication. https://doi.org/10.1007/s40299-025-01042-5

ElEbyary, K., & Shabara, R. (2024). ChatGPT-generated corrective feedback: Does it do what it says on the tin? *Teaching English with Technology, 24*(3), 68–89. https://doi.org/10.56297/vaca6841//BFFO7057/MYEH4562

Fan, Y., Tan, S., & Lim, G. Y. W. (2024). EAP teacher feedback in the age of AI: Supporting first-year students in EFL disciplinary writing. *Australian Journal of Applied Linguistics, 7*(3), Article 1943. https://doi.org/10.29140/ajal.v7n3.1943

Han, J., & Li, M. (2024). Exploring ChatGPT-supported teacher feedback in the EFL context. *System, 126*, Article 103502. https://doi.org/10.1016/j.system.2024.103502

Hanemann, U. (2019). Examining the application of the lifelong learning principle to the literacy target in the fourth Sustainable Development Goal (SDG 4). *International Review of Education, 65*(2), 251–275. https://doi.org/10.1007/s11159-019-09771-8

Hao, H., & Razali, A. B. (2025). Impact of ChatGPT on English academic writing ability and engagement of Chinese EFL undergraduates. *International Journal of Instruction, 18*(2), 323–346. https://doi.org/10.29333/iji.2025.18218a [Published in source as Hongxia, H., & Razali, A. B.]

Hao, H., Razali, A. B., & Zuo, R. (2026). Exploring the impact of integrated AWE and generative AI feedback on Chinese EFL undergraduates' higher-order thinking in internation. *Sage Open*. https://doi.org/10.1177/21582440251413884

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hong, H., Vate-U-Lan, P., & Viriyavejakul, C. (2025). Cognitive offload instruction with generative AI: A quasi-experimental study on critical thinking gains in English writing. *Forum for Linguistic Studies, 7*(7), 325–334. https://doi.org/10.30564/fls.v7i7.10072

Hong, S., & Shin, Y. K. (2025). Effects of three levels of AI integration on second language academic writing: Evaluating restricted, guided, and free use of ChatGPT. *System, 134*, Article 103820. https://doi.org/10.1016/j.system.2025.103820

Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal, 70*(2), 125–132. https://doi.org/10.1111/j.1540-4781.1986.tb05256.x

Huang, Y., & Wang, D. (2025). Can ChatGPT serve as a writing collaborator? Insights from Chinese EFL learners. *System, 133*, Article 103775. https://doi.org/10.1016/j.system.2025.103775

Hyland, F., & Hyland, K. (2006). Feedback on second language students' writing. *Language Teaching, 39*(2), 83–101. https://doi.org/10.1017/S0261444806003399

Hyland, K. (2003). *Second language writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511667251

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for Language Teaching and Learning. *RELC Journal, 54*(2), 537-550. https://journals.sagepub.com/doi/10.1177/00336882231162868

Kohnke, L., Zou, D., & Su, F. (2025). Exploring the potential of GenAI for personalized English teaching: Learners' experiences and perceptions. *Computers and Education: Artificial Intelligence, 8*, Article 100371. https://doi.org/10.1016/j.caeai.2025.100371

Kurt, G., & Kurt, Y. (2024). Enhancing L2 writing skills: ChatGPT as an automated feedback tool. *Journal of Information Technology Education: Research, 23*, Article 24. https://doi.org/10.28945/5370

Levac, D., Colquhoun, H., & O'Brien, K. K. (2010). Scoping studies: Advancing the methodology. *Implementation Science, 5*, Article 69. https://doi.org/10.1186/1748-5908-5-69

Link, S., Mehrzad, M., & Rahimi, M. (2022). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning, 35*(4), 605–634. https://doi.org/10.1080/09588221.2020.1743323

Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments, 11*(1), Article 9. https://doi.org/10.1186/s40561-024-00295-9

Muñoz Muñoz, B. C., Nassaji, H., & Bello Carrillo, F. I. (2025). ChatGPT-generated versus human direct corrective feedback on L2 writing. *System, 134*, Article 103805. https://doi.org/10.1016/j.system.2025.103805

Nguyen, L. Q., Le, H. V., & Nguyen, P. T. (2025). A mixed-methods study on the use of ChatGPT in the pre-writing stage: EFL learners' utilization patterns, affective engagement, and writing performance. *Education and Information Technologies, 30*(8), 10511–10534. https://doi.org/10.1007/s10639-024-13231-8

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Shaban, A. (2016). Rayyan—A web and mobile app for systematic reviews. *Systematic Reviews, 5*, Article 210. https://doi.org/10.1186/s13643-016-0384-4

Ozdere, M. (2025). AI in academic writing: Assessing the effectiveness, grading consistency, and student perspectives of ChatGPT and You.com for EFL students. *International Journal of Technology in Education, 8*(1), 123–154. https://doi.org/10.46328/ijte.1001

Peters, M. D. J., Marnie, C., Tricco, A. C., Pollock, D., Munn, Z., Alexander, L., McInerney, P., Godfrey, C. M., & Khalil, H. (2020). Updated methodological guidance for the conduct of scoping reviews. *JBI Evidence Synthesis, 18*(10), 2119–2126. https://doi.org/10.11124/JBIES-20-00167

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: A product from the ESRC Methods Programme*. Lancaster University. https://doi.org/10.13140/2.1.1018.4643

Puteri, C. G., Cahyono, B. Y., Widiati, U., & Suryati, N. (2025). Using ChatGPT as a writing assistant: A study on essay quality development among Indonesian EFL university students. *Indonesian Journal of Applied Linguistics, 15*(2), 332–346. https://doi.org/10.17509/3ebmy158

Ren, B., Dong, Y., & Zhang, B. (2025). From errors to excellence: AI-driven scaffolding for advancing L2 writing skills. *Cogent Education, 12*(1), Article 2588512. https://doi.org/10.1080/2331186X.2025.2588512

Sawaki, Y., Ishii, Y., Yamada, H., & Tokunaga, T. (2025). Examining the consistency of instructor versus large language model ratings on summary content: Toward checklist-based feedback provision with second language writers. *Language Testing, 42*(4), 447–475. https://doi.org/10.1177/02655322251349217

Stofiana, T., Sunendar, D., Mulyati, Y., & Sastromiharjo, A. (2025). Writing with AI, thinking with Toulmin: Metacognitive gaps and the rhetorical limits of argumentation. *Ampersand, 15*, Article 100242. https://doi.org/10.1016/j.amper.2025.100242

Storch, N. (2018). Written corrective feedback from sociocultural theoretical perspectives: A research agenda. *Language Teaching, 51*(2), 262–277. https://doi.org/10.1017/S0261444818000034

Tai, H.-Y., Lin, M.-F., & Chen, Y.-S. (2025). Incorporating ChatGPT into genre-based instruction for argumentative writing among EFL college students. *International Journal of Applied Linguistics, 36*(1), 767–781. https://doi.org/10.1111/ijal.12777

Tanharat, N., & Phootirat, P. (2025). Integrating generative AI in EFL academic writing: Thai English-major students' purposes, perceptions, and experiences with ChatGPT. *rEFLections, 32*(3), 1891–1916. https://doi.org/10.61508/refl.v32i3.285982

Teng, M. F. (2025). Understanding EFL student writers' metacognitive awareness in utilizing ChatGPT. *System, 135*, Article 103848. https://doi.org/10.1016/j.system.2025.103848

Tran, T. T. T. (2025). Enhancing EFL writing revision practices: The impact of AI- and teacher-generated feedback and their sequences. *Education Sciences, 15*(2), Article 232. https://doi.org/10.3390/educsci15020232

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., Straus, S. E. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine, 169*(7), 467–473. https://doi.org/10.7326/M18-0850

Tseng, Y.-C., & Lin, Y.-H. (2024). Enhancing English as a foreign language (EFL) learners' writing with ChatGPT: A university-level course design. *The Electronic Journal of e-Learning, 22*(2), 78–97. https://doi.org/10.34190/ejel.21.5.3329

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wang, R., Yin, Y., & Cao, Y. (2026). An LLM-based pedagogical framework for active, inquiry-based and adaptive learning in L2 writing. *Computers and Education:*

Artificial Intelligence, 10, Article 100535. https://doi.org/10.1016/j.caeai.2025.100535

Wang, Y., & Xie, Q. (2025). ERNIE-assisted feedback and its impact on EFL learners' revision strategies and writing performance. *Educational Psychology*, 1–23. Advance online publication. https://doi.org/10.1080/01443410.2025.2599787

Wei, L. (2023). Artificial intelligence in language instruction: Impact on English learning achievement, L2 motivation, and self-regulated learning. *Frontiers in Psychology, 14*, Article 1261955. https://doi.org/10.3389/fpsyg.2023.1261955

Xiao, F., Zhu, S., & Xin, W. (2025). Exploring the landscape of generative AI (ChatGPT)-powered writing instruction in English as a Foreign Language education: A scoping review. *ECNU Review of Education,* 8(1), 1–19. https://doi.org/10.1177/20965311241310881

Xu, T., & Jumaat, N. F. (2024). ChatGPT-empowered writing strategies in EFL students' academic writing: Calibre, challenges and chances. *International Journal of Interactive Mobile Technologies, 18*(15), 95–114. https://doi.org/10.3991/ijim.v18i15.49219

Yan, D. (2023). Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. *Education and Information Technologies, 28*, 13943–13967. https://doi.org/10.1007/s10639-023-11742-4

Yeung, S. (2025). University students' engagement with generative AI-supported automated writing evaluation (AWE) feedback. *Journal of Second Language Writing, 68*, Article 101203. https://doi.org/10.1016/j.jslw.2025.101203

Zare, J., Al-Issa, A., & Ranjbaran Madiseh, F. (2025). Interacting with ChatGPT in essay writing: A study of L2 learners' task motivation. *ReCALL, 37*(3), 385–402. https://doi.org/10.1017/S0958344025000035

Zhang, R., Zou, D., & Cheng, G. (2025). ChatGPT affordance for logic learning strategies and its usefulness for developing knowledge and quality of logic in English argumentative writing. *System, 128*, Article 103561. https://doi.org/10.1016/j.system.2024.103561

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*, 1–30. Advance online publication. https://doi.org/10.1080/09588221.2024.2447279

9. Appendix

Full-Text Excluded Studies with Reasons for Exclusion

Study	Primary exclusion reason	Evidence supporting exclusion
Cheng, D., Li, M., & Lee, T. (2025). Leveraging ChatGPT for research writing: An exploration of ESL graduate students' practices. *Computers & Composition*, *76*, Article 102934. https://doi.org/10.1016/j.compcom.2025.102934	Explicitly graduate-level sample	Title/abstract: ESL graduate students.
Chen, Z., Zhu, J., McGrane, J., Hopfenbeck, T. N., Wang, Y., & Ji, Y. (2026). More inspiration and attention: How generative AI tools impact graduate students' affective engagement in L2 source-based academic writing. *Computers & Education*, *242*, Article 105495. https://doi.org/10.1016/j.compedu.2025.105495	Explicitly graduate-level sample	Title/abstract: graduate students (RCT, N = 81).
Curado Fuentes, A. (2025). A qualitative approach to EFL postgraduates' GenAI-assisted research writing within social sciences. *Education Sciences*, *15*(11), Article 1521. https://doi.org/10.3390/educsci15111521	Explicitly graduate-level sample	Title/abstract: 7 postgraduate EFL students
Koltovskaia, S., Rahmati, P., & Saeli, H. (2024). Graduate students' use of ChatGPT for academic text revision: Behavioral, cognitive, and affective engagement. *Journal of Second Language Writing*, *65*, Article 101130. https://doi.org/10.1016/j.jslw.2024.101130	Explicitly graduate-level sample	Title/abstract: 6 Iranian graduate ESL students in STEM.
Meng, Y., Cui, Y., Geng, X., & Cao, Y. (2026). Exploring feedback literacy in L2 students' academic writing: Insights from their reported engagement with GenAl for typical revision activities. *The Asia-Pacific Education Researcher*, *35*, 559–570. https://doi.org/10.1007/s40299-025-01051-4	Explicitly graduate-level sample	Abstract: 16 Chinese L2 graduate students.
Vy, P. N. T., & Lap, T. Q. (2026). Reflection from teachers as graduates: How has generative AI affected personal agency in AW? *Reflective Practice*, *27*(1), 65–76. https://doi.org/10.1080/14623943.2025.2533839	Explicitly graduate-level sample	Title/abstract: 53 master's-level TEFL learners

Wan, Y., Wu, C., Moorhouse, B. L., & Wu, M. (2025). Developing L2 postgraduate students' strategic and responsible use of GenAI through a purposefully designed workshop in academic writing. *Innovation in Language Learning and Teaching*. Advance online publication. https://doi.org/10.1080/17501229.2025.2575084	Explicitly graduate-level sample	Title/abstract: L2 postgraduates in a Hong Kong workshop.
Werdiningsih, I., Marzuki, & Rusdin, D. (2024). Balancing AI and authenticity: EFL students' experiences with ChatGPT in academic writing. *Cogent Arts & Humanities*, *11*(1), Article 2392388. https://doi.org/10.1080/23311983.2024.2392388	Explicitly graduate-level sample	Methods: 3 EFL postgraduate students
Werdiningsih, I., Marzuki, Indrawati, I., Rusdin, D., Ivone, F. M., Basthomi, Y., & Zulfahreza. (2024). Revolutionizing EFL writing: Unveiling the strategic use of ChatGPT by Indonesian master's students. *Cogent Education*, *11*(1), Article 2399431. https://doi.org/10.1080/2331186X.2024.2399431	Explicitly master's-level sample	Title/abstract: 16 master's (S2) students at four Indonesian universities.
Zhang, D., Wen, L., & Wu, J. G. (2025). Structured or semi-structured? The use of reflection journals in postgraduates' generative artificial intelligence literacy development in an L2 academic writing context. *European Journal of Education*, *60*(3), Article e70189. https://doi.org/10.1111/ejed.70189	Explicitly graduate-level sample	Title/abstract: postgraduate students
Zou, M., & Huang, L. (2024). The impact of ChatGPT on L2 writing and expected responses: Voice from doctoral students. *Education and Information Technologies*, *29*(11), 13201–13219. https://doi.org/10.1007/s10639-023-12397-x	Explicitly graduate-level sample	Title/abstract: doctoral students
Strobl, C., Ailhaud, E., Benetos, K., Devitt, A., Kruse, O., Proske, A., & Rapp, C. (2019). Digital support for academic writing: A review of technologies and pedagogies. *Computers & Education*, *131*, 33–48. https://doi.org/10.1016/j.compedu.2018.12.005	Review article; level-specific evidence not extractable	Review covered undergraduate and postgraduate contexts; findings not separable by level.

Zhang, R., Zou, D., & Cheng, G. (2026). Non-academic learner socialisation with ChatGPT and its influences on learning English argumentative writing logic. *Education and Information Technologies*, *31*(1), 99–142. https://doi.org/10.1007/s10639-025-13776-2	Degree level not reported	Methods: 40 university students aged 18–27; degree level unreported.
Jiang, Y., Wu, Q., Yang, Y., Jian, C., & Zhao, J. (2026). Learner agency in revising GenAI-generated statements of purpose. *British Journal of Educational Technology*. Advance online publication. https://doi.org/10.1111/bjet.70041	Mixed undergraduate/postgraduate sample; data not disaggregated	Methods: junior/senior undergraduates and first-year postgraduates; results not separated by level.