

Review

LLM-Generated Feedback in L2 Writing: A Scoping Review

Laurence Craven *  and Daniel R. Fredrick 

Department of English, American University of Sharjah, Sharjah 26666, United Arab Emirates; dfredrick@aus.edu

* Correspondence: lcraven@aus.edu

Abstract

The release of ChatGPT in November 2022 transformed second language (L2) writing instruction and led to rapid growth in research on large language model (LLM)-generated feedback; however, no synthesis has mapped this literature in terms of feedback quality, learner uptake, and pedagogical integration. This scoping review examines 185 empirical studies published between November 2022 and March 2026 that were identified through a Scopus search ($n = 283$ screened) and analysed using a systematic keyword-based charting framework applied to full abstracts, with full-text analysis of 35 studies. The review identifies four major patterns: (1) comparative AI–human feedback research dominates the literature (27.6%); (2) content-level feedback remains underexplored (13.5% of studies); (3) learner uptake is rarely measured as a primary outcome; and (4) LLM feedback is broadly comparable to teacher feedback for surface-level errors but weaker for content and argumentation, while learner perceptions often exceed demonstrated performance outcomes. Hybrid AI–teacher models show promising but underexamined potential, accounting for only 12.4% of the literature. The field shows a focus on perceptions rather than learning outcomes, an apparent tendency toward positive-results reporting, and no clear teaching models. This study proposes a typology of LLM feedback functions and outlines a research agenda focused on uptake, longitudinal outcomes, and hybrid AI–teacher integration.

Keywords: generative AI; large language models; L2 writing; written corrective feedback; learner uptake; feedback literacy; scoping review

1. Introduction

Written feedback is widely recognised as one of the most important mechanisms supporting second language (L2) writing development (Bitchener & Storch, 2016; Ferris, 2003). Research on teacher written corrective feedback (WCF), peer feedback, and automated writing evaluation (AWE) has consistently shown that feedback effectiveness depends not only on the accuracy and specificity of comments but also on how learners interpret, engage with, and apply feedback during revision. Learner uptake is therefore a particularly important construct because it represents the point at which feedback moves from information provided to potential learning realised. Feedback embedded within instructional frameworks and supported by learner engagement has generally been found to produce stronger outcomes than decontextualised correction alone.

The public release of ChatGPT in November 2022 introduced a new form of feedback into L2 writing instruction. Unlike earlier AWE systems, large language models (LLMs) can generate immediate, individualised, and context-sensitive responses across multiple dimensions of writing, including grammar, organisation, content, and argumentation. As a



Academic Editor: Meihua Liu

Received: 30 May 2026

Revised: 29 June 2026

Accepted: 9 July 2026

Published: 27 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

result, researchers have rapidly begun investigating the quality of LLM-generated feedback, its effects on revision and writing performance, learner perceptions of AI-generated comments, and the potential role of AI within classroom instruction. The resulting body of research has expanded quickly and now encompasses a wide range of learner populations, instructional settings, feedback types, and outcome measures.

Despite this growth, the literature remains fragmented. Existing reviews have primarily examined generative AI within language learning more broadly or have focused on automated feedback systems that predate contemporary LLMs. Crosthwaite and Sun (2025), for example, documented the rapid expansion of research on generative AI and written feedback while noting the preliminary nature of the evidence base and the need for further synthesis. To date, however, no review has examined this literature through the integrated lenses of feedback quality, learner engagement with feedback, and pedagogical integration. In particular, the role of learner uptake remains insufficiently synthesised despite its established importance in feedback research and its direct relationship to learning outcomes.

Accordingly, the aim of this scoping review is to map and synthesise empirical research on LLM-generated feedback in L2 writing, with particular attention to feedback types, uptake patterns, comparative effectiveness, and instructional applications.

1.1. Research Questions

This scoping review addresses four research questions:

1. What types of LLM-generated feedback are examined in L2 writing research, and how do they vary by proficiency level and task type?
2. To what extent do learners act on (uptake) LLM-generated feedback, and what factors influence this uptake?
3. How does LLM-generated feedback compare with teacher, peer, and traditional AWE feedback in terms of quality and writing outcomes?
4. What models exist for integrating LLM feedback into L2 writing instruction?

1.2. Theoretical Framework

This review is grounded in sociocultural theory (Vygotsky, 1978) and its application to L2 writing (Lantolf & Thorne, 2006), which frames feedback as a mediational tool that supports learner development within the zone of proximal development. From this perspective, LLM feedback functions as computer-mediated scaffolding, and its effectiveness depends on how learners engage with and act on the feedback they receive. This places learner uptake and agency at the center of analysis.

The review also draws on feedback literacy theory (Carless & Boud, 2018), which treats effective use of feedback as a competence that develops over time. From this view, LLM feedback can either support or limit development depending on how it is used in instruction, either as a substitute for teacher judgment or as a tool that helps learners build their own evaluative skills.

Both sociocultural theory and feedback literacy provide complementary lenses for understanding LLM-generated feedback. Sociocultural theory conceptualizes feedback as a mediational resource whose value depends on learner engagement and interaction, while feedback literacy emphasizes the learner's capacity to interpret, evaluate, and use feedback productively. The combination of these perspectives is particularly relevant to LLM-based feedback environments because access to feedback is no longer the primary constraint; rather, the central issue is how learners engage with and act upon the feedback they receive. These frameworks therefore provide the conceptual foundation for the review's focus on learner uptake, feedback quality, and pedagogical integration.

2. Method

2.1. Scoping Review Design

This study adopts a scoping review methodology following [Arksey and O'Malley \(2005\)](#), with refinements from [Levac et al. \(2010\)](#). This approach is appropriate because the aim is to map the extent, range, and characteristics of research on LLM-generated feedback in L2 writing rather than to evaluate intervention effectiveness. The review is reported in accordance with PRISMA-ScR guidelines ([Tricco et al., 2018](#)).

2.2. Search Strategy

A systematic search was conducted in Scopus in March 2026 using the following Boolean query: (ChatGPT OR “large language model” OR “generative AI”) AND (feedback) AND (“L2 writing” OR “EFL” OR “ESL” OR “second language”). The search was limited to Social Sciences and Arts and Humanities subject areas and to publications from 2022 to 2026. Preliminary searches confirmed that studies referring to GPT-4, GPT-4o, Claude, and Gemini were indexed under at least one of the broader search terms (“ChatGPT,” “large language model,” or “generative AI”) within the title, abstract, or keyword fields. Manual screening also ensured inclusion of studies employing specific LLMs when they met the eligibility criteria.

Scopus was selected due to its broad coverage of applied linguistics and language education journals relevant to this topic, including all major journals in which the LLM feedback literature is concentrated (System, Computer-Assisted Language Learning, Journal of Second Language Writing, Education and Information Technologies). It is acknowledged that supplementary searches of ERIC, Web of Science, or Google Scholar may have identified additional relevant studies, particularly from regional or non-indexed journals, and this constitutes a limitation of the present search strategy.

2.3. Inclusion and Exclusion Criteria

Studies were included if they: (1) reported original empirical data; (2) used an LLM tool (e.g., ChatGPT, GPT-4, Claude, Gemini) to generate feedback on written texts; (3) involved L2 learners of any proficiency level or instructional context; (4) examined writing outcomes; and (5) were published between November 2022 and March 2026. Studies were excluded if they: (1) were reviews, meta-analyses, or theoretical articles without original data; (2) did not focus on writing feedback outcomes; (3) did not involve L2 learners; (4) did not use an AI/LLM tool; or (5) were articles in press with a 2026 publication date (indicating post-search publication).

2.4. Study Selection

A total of 283 records were retrieved. Systematic title and abstract screening was conducted for all 283 records against the inclusion criteria: (1) original empirical data; (2) use of an LLM tool to generate feedback on written texts; (3) L2 learners; (4) writing outcomes; and (5) published between November 2022 and March 2026. Articles in press with a 2026 publication date were excluded as post-search ($n = 22$), along with studies not focused on writing feedback ($n = 61$), non-empirical articles including reviews and theoretical papers ($n = 8$), studies without an L2 focus ($n = 6$), and one study without an AI/LLM tool, resulting in a final sample of 185 studies. The selection process is summarised in Table 1.

Table 1. PRISMA-ScR Flow: Records Identified, Screened, and Included.

Stage	Records (n)	Notes
Identification—Scopus	283	Boolean search, March 2026, Social Sciences & Arts & Humanities
Excluded at screening	98	Post-search in-press 2026 (22); no writing/feedback focus (61); non-empirical (8); no L2 focus (6); no AI/LLM tool (1)
INCLUDED IN SCOPING REVIEW	185	Final corpus for narrative synthesis; full-text analysis conducted for 35 studies
Charting note	—	Charting conducted from Scopus full abstracts for all 185 studies; keyword-based coding applied systematically; see Section 2.5
<i>a</i>	—	<i>Exclusions: post-search articles in press 2026 (n = 22); no writing/feedback focus (n = 61); non-empirical (n = 8); no L2 focus (n = 6); no AI/LLM tool (n = 1). Total excluded = 98. 283 – 98 = 185 included.</i>

2.5. Data Charting

Data charting involved extracting information from included studies to map the breadth and distribution of the literature. A charting instrument capturing 20 variables across five categories was applied to all 185 studies: study identity (authors, year, journal), context (learner population, L2, instructional level), LLM tool characteristics (tool, version, feedback type), outcomes (primary outcome, outcome direction, uptake measurement), and methodology (research design, MMAT risk of bias). Charting was conducted using full abstracts exported directly from Scopus, providing complete abstract-level data for all 185 included studies. Systematic keyword-based coding was applied consistently across all abstracts for each variable. Because this charting was automated rather than manually coded, individual category assignments were not independently verified against each abstract, and the resulting frequencies should be interpreted as an approximate mapping of the field rather than as manually validated classifications. This is acknowledged as a limitation of the charting process. Variables requiring full-text access, including prompting strategy, effect sizes, and specific sample characteristics, could not be reliably coded from abstracts and are acknowledged as a limitation of the charting process. However, abstracts provided sufficient information to identify study purpose, feedback type, research design, and primary outcomes. More granular methodological variables (such as prompting strategies, effect sizes, and participant characteristics) were less consistently reported and in those instances could not be analyzed in detail. As a result, certain fine-grained comparisons were not always possible, although the available data were sufficient to identify broader patterns across the corpus. Importantly, the abstract-level approach remains consistent with the mapping objectives of scoping review methodology and enabled systematic coverage of a rapidly expanding body of research. This limitation may affect the precision of some findings but does not alter the broader patterns identified across the corpus.

To support analysis of key themes such as learner uptake, comparative feedback quality, and hybrid models, full-text data were additionally examined for 35 open-access studies identified via DOI resolution. These sources are used in the narrative synthesis to provide detail beyond abstract-level data. Full-text retrieval prioritised two categories of

studies: those reporting uptake as a primary or secondary outcome, and those addressing comparative feedback quality or hybrid AI–teacher models that were theoretically relevant to RQ3 and RQ4. Open-access availability constrained the retrieval pool to 35 studies; selection within that pool was guided by theoretical relevance rather than availability alone. Findings derived primarily from full-text sources are identified as such throughout the synthesis (Appendix A).

2.6. Synthesis Approach

A narrative synthesis was conducted and organised around the four research questions. This approach is appropriate for mapping a heterogeneous literature spanning experimental, quasi-experimental, case study, survey, and mixed methods designs. The synthesis does not include pooled effect sizes; where effect sizes or uptake rates are reported, these are drawn from individual studies and used as illustrative examples. To mitigate this limitation, claims about feedback quality, uptake patterns, and pedagogical models draw primarily on the full-text studies, while abstract-level charting is used to identify broad patterns rather than to support causal conclusions.

3. Results

3.1. Descriptive Overview of the Corpus

The 185 included studies were published across a range of applied linguistics, language education, and educational technology journals. Publication increased rapidly over time: 1 study (0.5%) was published in 2023, 38 in 2024 (20.5%), 98 in 2025 (53.0%), and 48 in the first quarter of 2026 (25.9%). Post-search articles in press were excluded. This accelerating trajectory reflects both the novelty of the research area and the rapid expansion of LLM-based research designs following ChatGPT’s public release. The substantial proportion of 2026 Q1 studies warrants a note of caution: papers published in early 2026 had completed peer review within a rapidly expanding field with few established methodological norms, and findings from this portion of the corpus should be interpreted with awareness of this context. Table 2 presents the full descriptive breakdown, and Table 3 lists the most frequently publishing journals.

Table 2. Descriptive Characteristics of Included Studies (N = 185). † Two uptake figures are reported: “Uptake/revision behaviour” reflects studies where uptake was the primary outcome focus (n = 9); “Uptake measured” reflects the broader set of studies where uptake was assessed in any capacity (n = 11). Both figures reflect abstract-level keyword coding; manual examination of full texts identified additional studies measuring uptake that were not captured at abstract level, so these values are lower-bound estimates (see Sections 3.3 and 4.5).

Characteristic	Category	n (%)
Publication year	2023	1 (0.5%)
	2024	38 (20.5%)
	2025	98 (53.0%)
	2026 (Q1)	48 (25.9%)
Primary feedback type	Corrective/form-focused	34 (18.4%)
	General/mixed	75 (40.5%)
	Comparative (AI vs. human)	51 (27.6%)
	Content/holistic	25 (13.5%)

Table 2. *Cont.*

Characteristic	Category	n (%)
Outcome focus	Perceptions	96 (51.9%)
	Writing performance	80 (43.2%)
	Uptake/revision behaviour	9 (4.9%) †
Outcome direction	Positive/significant	105 (56.8%)
	Descriptive/no comparison	70 (37.8%)
	Mixed/no significant difference	10 (5.4%)
	Challenges/limitations foregrounded	0 (0.0%)
Research design	Quantitative/other	50 (27.0%)
	Survey	33 (17.8%)
	Quasi-experimental	30 (16.2%)
	Qualitative	41 (22.2%)
	Mixed methods	29 (15.7%)
	Experimental (RCT)	2 (1.1%)
LLM tool	ChatGPT (unspecified version)	132 (71.4%)
	Generative AI (unspecified)	34 (18.4%)
	GPT-4/GPT-4o	16 (8.6%)
	LLM (unspecified)	3 (1.6%)
Risk of bias (MMAT)	Low	32 (17.3%)
	Moderate	153 (82.7%)
	High	0 (0.0%)
Hybrid AI–teacher model	Yes	23 (12.4%)
Uptake measured	Yes	11 (5.9%) †

Table 3. Top Journals by Number of Included Studies.

Journal	n	% of Corpus
Computer-Assisted Language Learning	10	5.4%
System	9	4.9%
Innovation in Language Learning and Teaching	8	4.3%
Education and Information Technologies	6	3.2%
International Journal of Applied Linguistics	6	3.2%
Journal of Second Language Writing	6	3.2%
Computers and Education: Artificial Intelligence	5	2.7%
Assessment and Evaluation in Higher Education	5	2.7%
All other journals	147	70.3%

ChatGPT (unspecified version) was the most commonly reported tool (71.4%), followed by generative AI tools not identified by model version (18.4%). Only 8.6% of studies specified GPT-4 or GPT-4o. This finding has significant methodological implications for the synthesis as a whole. Grouping them all under “ChatGPT” limits model-level comparison because GPT-3.5, GPT-4, GPT-4o, Claude, and Gemini differ substantially in capability,

reasoning ability, and feedback generation quality. Most studies, however, did not report model version with sufficient precision to support meaningful model-specific analysis. Nevertheless, the absence of consistent version reporting does not obscure the broader trends identified across the corpus. The consistency of findings across a large and diverse body of research suggests that the patterns identified in this review reflect recurring characteristics of the LLM feedback literature.

Table 3 presents the journals contributing the largest number of studies to the corpus.

3.2. RQ1: Types of LLM-Generated Feedback in L2 Writing Research

The most prevalent feedback type studied was general or mixed feedback (75 studies; 40.5%), followed by comparative AI–human feedback research (51 studies; 27.6%). Corrective and form-focused feedback was examined in 34 studies (18.4%), including grammatical error correction, vocabulary accuracy, and written corrective feedback (WCF) in the tradition of Truscott (1996) and Ferris (2003). Content and holistic feedback was the least frequently examined type (25 studies; 13.5%).

Full-text analysis of selected studies provides further detail. Shi et al. (2026) found that ChatGPT outperformed a traditional AWE system and a control condition in an eleven-week study with 150 university students in China. Saricaoglu and Bilki (2025) reported variation in feedback specificity across dimensions, with higher specificity for grammatical and lexical features than for task response and coherence. Guo and Wang (2024) found that ChatGPT generated a greater volume of feedback across content, organisation, and language, while teacher feedback was more contextually grounded. Hyland (2025) noted that LLMs are effective for surface-level correction but less suited to feedback on voice, audience awareness, and disciplinary conventions.

Few studies examined feedback effectiveness by learner proficiency level. Alnemrat et al. (2025) found no statistically significant interaction between feedback type (AI vs. teacher) and proficiency level, although Intermediate-Low learners showed the greatest within-group gains in both conditions.

3.3. RQ2: Learner Uptake of LLM-Generated Feedback

Learner uptake was the least frequently examined outcome in the corpus. Abstract-level charting initially identified uptake as a primary outcome in 9 studies; manual examination of full texts identified further studies that measured uptake without foregrounding it in their abstracts, bringing the verified count to at least 11, and this should be read as a lower-bound estimate rather than an exact figure. Regardless of the exact count, uptake was measured far less frequently than perceptions or writing performance, representing the most striking gap in the literature given its theoretical centrality to writing development.

Full-text analysis of the uptake-reporting studies and related full-text sources substantially enriches the abstract-level picture. The available evidence converges on three findings: uptake is systematically asymmetric across feedback dimensions, with surface-level corrections acted on at substantially higher rates than meaning-level suggestions; uptake is mediated by individual learner characteristics, particularly proficiency and feedback literacy; and the conditions under which uptake translates into sustained writing development remain underspecified in the literature. The sections below elaborate each pattern.

3.3.1. Uptake Rates: The Surface–Meaning Asymmetry

The most methodologically rigorous uptake study in the corpus is Yu and Xie (2025), who employed a 2×2 counterbalanced experimental design (Latin Square) with 60 EFL secondary students and four teachers in China, drawing on 1200 feedback records (240 written samples across five raters). Their data show a clear surface-to-meaning asymmetry that holds for both feedback sources: ChatGPT surface-level feedback achieved an uptake rate

of approximately 86%, while ChatGPT meaning-level feedback (organisation, logic, ideas) achieved 32%; the corresponding teacher figures were approximately 92% and 29%. These figures should be understood as indicative rather than definitive, deriving as they do from one study with 60 participants in a specific EFL context. Notably, teacher feedback received higher overall uptake than ChatGPT feedback (approximately 78% vs. 70%), and higher surface-level uptake; ChatGPT's only marginal advantage was on meaning-level uptake, where both sources were low. Students used a range of revision strategies in response to both sources. It should be noted that uptake rates reported in this section derive from the full-text subsample and cannot be assumed to be representative of uptake patterns across the broader corpus, the majority of which did not measure uptake at all. Across the full-text uptake studies examined, surface-level feedback was consistently described as more frequently acted on than meaning-level feedback, with no study reporting the reverse pattern. Because most of the corpus did not measure uptake, this asymmetry should be treated as a pattern within the available uptake studies rather than an established feature of the field as a whole.

Cengiz et al. (2025), in a convergent parallel mixed-methods study of 25 B1-level Turkish EFL students, found that language-use feedback was accepted more frequently than content or organization feedback. Organization feedback resulted in no correction in over one-third of instances, and students reported that such feedback often lacked sufficient guidance for revision. Zou et al. (2025) examined uptake across revision cycles and found that learners preferred teacher feedback for nuanced revisions, while LLM feedback was used more often for organizational changes.

3.3.2. Individual Differences in Uptake

Yan and Zhang (2024), in a mixed-method multiple case study of four L2 writers using ChatGPT as an automated written corrective feedback (AWCF) provider, found that behavioural engagement varied by language proficiency and technological competence. High-proficiency learners refined prompts iteratively and engaged with feedback content at a deeper level, while lower-proficiency learners showed more limited engagement and focused primarily on surface-level revisions.

Koltovskaia et al. (2024), examining the screencasted revision behaviour of six Iranian graduate ESL students, found that participants used ChatGPT primarily for lower-order concerns such as paraphrasing and grammatical correction, despite training on more complex prompting strategies. Participants also reported high levels of satisfaction with the feedback.

3.3.3. Feedback Literacy as Uptake Moderator

Feedback literacy (Carless & Boud, 2018) is examined in several full-text studies as a factor associated with uptake quality and depth. Yang et al. (2025) found that evaluative judgement, metacognitive skills, and ethical decision-making dimensions of feedback literacy were significant predictors of both uptake rate and perceived usefulness of ChatGPT feedback. Yan and Zhang (2024) similarly found that higher-proficiency learners who engaged more effectively with LLM feedback demonstrated behaviours consistent with stronger feedback literacy: iterative prompt refinement, critical evaluation of suggestions before implementing them, and selective rejection of feedback deemed inaccurate or contextually inappropriate. Koltovskaia et al. (2024) found the converse pattern: high affective satisfaction with LLM feedback was associated with reduced critical evaluation, suggesting that feedback literacy and satisfaction do not co-vary reliably and that satisfaction may in some cases substitute for rather than accompany genuine engagement. In sum, these studies suggest that feedback literacy functions as a moderator of the quality rather than

merely the quantity of uptake: learners with higher feedback literacy act on LLM feedback more selectively and more productively, while those with lower feedback literacy are more likely to accept suggestions wholesale or engage only with surface-level corrections. This pattern has direct implications for instructional design: feedback literacy development cannot be assumed and should be explicitly scaffolded in any pedagogical model that incorporates LLM feedback.

3.4. RQ3: LLM Feedback Compared to Teacher, Peer, and AWE Feedback

The 51 comparative studies (27.6%) represent a substantial portion of the corpus. Across these studies, findings vary by feedback type, outcome measure, and study design.

3.4.1. Accuracy and Form-Focused Feedback

LLM-generated corrective feedback is widely reported as comparable to or exceeding teacher feedback in surface-level accuracy. [Saricaoglu and Bilki \(2025\)](#) found ChatGPT-4 highly accurate across grammatical and lexical dimensions. [Li et al. \(2024\)](#), comparing ChatGPT 3.5 and 4 with four college English teachers on CET-4 essays using G-theory analysis, reported that ChatGPT-4 demonstrated higher scoring reliability than human raters.

However, some studies identify limitations in corrective performance. [ELEbyary and Shabara \(2024\)](#) reported that suggested revisions were not always applied consistently across drafts and that new errors occasionally appeared during complex sentence restructuring.

3.4.2. Content and Argumentation Feedback

LLM performance on content-level feedback was less consistent than on surface-level features. [Guo and Wang \(2024\)](#) found that ChatGPT generated a greater volume of feedback distributed across content, organization, and language, while teacher feedback provided more context-specific explanation and learner-focused scaffolding. LLM-generated content feedback tends to be less context-sensitive and lacks interpersonal and relational dimensions characteristic of teacher feedback ([Hyland, 2025](#)). [Bozorgian and Rahimi \(2025\)](#) found that ChatGPT-4o produced predominantly task-level feedback (94% of instances), with limited variation across learners, whereas peer feedback included more self-level and process-level comments.

3.4.3. The Perception–Performance Divergence

A recurring finding across the comparative literature is a divergence between learner perceptions of LLM feedback and its measured effectiveness. Learners consistently rated LLM feedback highly in terms of clarity, availability, and reduced anxiety, while studies measuring writing performance reported more modest and often non-significant differences. [Alnemrat et al. \(2025\)](#) found no statistically significant difference between AI and teacher feedback on argumentative writing performance (Cohen's $d = 0.10$), despite significant within-group improvement in both conditions. [Tran \(2025\)](#) found that AI-first feedback sequences (AIGF–TGF) produced stronger local-level revisions, whereas teacher-first sequences (TGF–AIGF) better supported higher-order content revision.

3.4.4. LLM Versus Traditional AWE

Studies comparing LLMs to traditional AWE systems (e.g., Grammarly, Turnitin, Pigai) report differences in feedback range, naturalness, and learner engagement. [Shi et al. \(2026\)](#) found that the ChatGPT group outperformed the AWE group on overall writing quality after an 11-week intervention. LLM feedback addressed content and organisation as well as surface-level errors, whereas AWE systems focused primarily on grammatical correction.

3.5. RQ4: Pedagogical Models for LLM Feedback Integration

Only 23 studies (12.4%) involved a structured hybrid AI–teacher feedback model, making pedagogical integration a relatively underexplored area of the corpus despite its practical significance. Full-text evidence from these and related studies provides substantially richer detail than charting alone could capture.

Across these studies, three design dimensions appear to matter: the sequencing of AI and teacher feedback, the degree of teacher mediation in framing AI feedback before learners receive it, and the scaffolding provided to support learner engagement with combined feedback (Table 4).

Table 4. Summary of Scoping Review Findings by Research Question.

RQ	Key Finding	Key Sources (Full-Text Where Noted)
RQ1	General/mixed feedback dominates (40.5%); comparative AI–human research prominent (27.6%); content-level underexplored (13.5%). Differential feedback specificity by dimension: high for grammar/lexis, lower for task response and coherence.	Saricaoglu and Bilki (2025, FT) ; Shi et al. (2026, FT) ; Guo and Wang (2024, FT)
RQ2	Uptake is the least-charted outcome (measured as a primary outcome in a small minority of studies; abstract-level count is a lower-bound estimate). Surface-level uptake far exceeds meaning-level uptake for both AI and teacher feedback (ChatGPT ~86% vs. ~32%; teacher ~92% vs. ~29%; Yu & Xie, 2025 ; single study, $n = 60$). Teacher feedback received higher overall uptake than ChatGPT. Uptake is selective, proficiency-mediated, and feedback-literacy dependent.	Yu and Xie (2025, FT) ; Cengiz et al. (2025, FT) ; Zou et al. (2025, FT) ; Yan and Zhang (2024, FT)
RQ3	Comparable to teacher feedback for surface errors ($d = 0.10$ in best-powered study). Lower quality for content and argumentation. Learner perceptions diverge positively from measured outcomes. ChatGPT superior to traditional AWE.	Alnemrat et al. (2025, FT) ; Guo and Wang (2024, FT) ; Shi et al. (2026, FT)
RQ4	Hybrid AI + teacher models outperform either alone; sequence matters. Prompting strategy is a critical mediating variable. Teacher agency is essential. 12.4% of corpus studies examined hybrid models, though systematic implementation remains limited.	Asadi et al. (2025, FT) ; Soori et al. (2025, FT) ; Tran (2025, FT) ; Luo et al. (2025, FT)

Note. FT = full-text source retrieved via open-access DOI resolution.

3.5.1. Evidence for the Hybrid Model

[Asadi et al. \(2025\)](#), in an explanatory sequential mixed-methods study with 68 Iranian EFL learners and 8 teachers, found that combined ChatGPT-plus-teacher feedback resulted in higher scores than teacher-only feedback across IELTS Task 2 criteria, including task achievement, coherence and cohesion, vocabulary, and grammatical range and accuracy. Teacher interviews reported both recognition of ChatGPT’s value (customization, promptness, adaptability) and concerns about limitations, including lack of contextual understanding and potential overreliance. Teachers also identified the need for professional development to support integration of AI tools.

[Soori et al. \(2025\)](#), comparing teacher-only, AI-only, and hybrid feedback conditions, found that the hybrid condition produced higher performance outcomes than the other

conditions. Models in which teachers retained control over feedback framing showed different outcomes from models in which AI feedback was used without teacher mediation. [Tran \(2025\)](#), using a sequential design, found that the AIGF–TGF sequence (AI first, teacher second) was associated with stronger local-level revisions, while the TGF–AIGF sequence was associated with stronger global-level content revision.

3.5.2. Prompting Strategy as Mediating Variable

Full-text evidence identifies prompting strategy as a variable associated with differences in learner engagement and revision behaviour. [Luo et al. \(2025\)](#) found that domain-specific prompts, particularly those incorporating examples, substantially improved the quantity and quality of ChatGPT-generated corrective feedback compared with generic prompts. [Hwang et al. \(2025\)](#), examining learner prompting behaviour and revision quality, found that more complex prompt construction was associated with deeper engagement, with variation observed across proficiency levels.

4. Discussion

4.1. The Uptake Gap: A Field-Level Blind Spot

The interpretation of findings in this section prioritizes evidence from full-text studies where available, particularly for claims regarding uptake, feedback quality, and pedagogical effectiveness. The most consequential finding in this synthesis concerns not what LLMs can do, but what the field has largely not examined: learner uptake. With uptake measured as a primary outcome in only a small minority of studies—and the abstract-level count likely understating even that—the literature remains heavily weighted toward feedback delivery rather than how feedback is interpreted and used. This is a significant limitation, as uptake is widely treated as the variable most closely linked to learning. From the sociocultural perspective that grounds this review ([Vygotsky, 1978](#); [Lantolf & Thorne, 2006](#)), this gap is particularly consequential: if LLM feedback functions as computer-mediated scaffolding, then its developmental value depends entirely on whether and how learners engage with it. Measuring delivery without measuring uptake is like measuring input without measuring interaction, and the field has, to date, focused far more on the former than the latter.

The surface-to-meaning uptake split reported by [Yu and Xie \(2025\)](#)—while derived from a single study with 60 participants—provides a potentially important indication of how feedback functions in practice. In their data, surface-level feedback was acted on at far higher rates than meaning-level feedback for both AI and teacher sources (ChatGPT ~86% vs. ~32%; teacher ~92% vs. ~29%), with teacher feedback receiving somewhat higher overall uptake than ChatGPT. This asymmetry suggests that reported gains in writing quality may reflect improvements in accuracy rather than development in higher-order abilities such as argumentation and organisation.

Variation in uptake is also associated with learner characteristics. Higher-proficiency learners tend to engage more selectively with feedback, refining prompts and evaluating suggestions before revising, whereas lower-proficiency learners more often prioritise surface-level corrections and engage less with feedback content. At the same time, high levels of learner satisfaction with LLM feedback may reduce critical evaluation, leading to acceptance without deeper processing.

Across studies, feedback literacy appears to shape how learners use feedback. Learners with stronger evaluative judgement and metacognitive awareness are more likely to revise meaningfully, while those with lower feedback literacy rely more on surface corrections or accept feedback without scrutiny. Differences in uptake therefore reflect not only feedback quality but also learners' capacity to interpret and apply feedback. From the feedback literacy perspective introduced in Section 1.2 ([Carless & Boud, 2018](#)), this pattern suggests that

LLM feedback, like all feedback, is a resource rather than a mechanism: its developmental value depends on the learner's capacity to engage with it critically, which itself must be cultivated rather than assumed. This suggests that improving LLM feedback quality alone will yield diminishing returns unless instruction also develops learners' feedback literacy.

Given the limited number of uptake-focused studies, the field lacks a clear account of when and how LLM feedback supports learning. Without more systematic investigation across proficiency levels, task types, and feedback dimensions, conclusions about effectiveness remain partial and uneven across feedback types. The centrality of uptake in this review therefore reflects its theoretical importance rather than its empirical representation in the current literature, which remains limited.

4.2. Publication Bias and the Positivity Problem

The corpus shows an apparent directional imbalance: the majority of studies were charted as reporting positive outcomes, with relatively few charted as reporting mixed or null findings. This pattern should be interpreted with caution, since outcome direction was coded from abstracts, which tend to foreground positive or headline findings and rarely foreground limitations; the abstract-level charting may therefore overstate the positivity of the underlying literature. With that caveat, the pattern is broadly consistent with the well-documented tendency toward publication bias and positive-results reporting in applied linguistics and education more generally, alongside a strong reliance on perception-based measures in this corpus. Learner satisfaction and learning outcomes, however, are distinct constructs that do not consistently align.

Across comparative studies, LLM feedback is frequently rated highly for clarity, accessibility, and reduced anxiety, while performance-based measures show more modest and often non-significant differences. This divergence is also reflected in sequencing studies, where AI-first conditions tend to support local-level revision, while teacher-first conditions are more associated with higher-order content development.

The strongest comparative study in the corpus found no significant difference between AI and teacher feedback conditions despite measurable gains within both groups ($d = 0.10$; [Alnemrat et al., 2025](#)). This null finding is consistent with the broader WCF literature, where effect sizes for written corrective feedback are consistently modest ([Kang & Han, 2015](#); [Sinha & Nassaji, 2022](#)) and where the conditions under which feedback produces lasting gains remain contested ([Bitchener & Storch, 2016](#)). The LLM feedback literature has largely ignored this parallel, treating the question of whether AI feedback works as if it were new when in fact it closely mirrors decades of inconclusive WCF research. Until the field adopts the methodological standards that WCF researchers have argued for, including adequately powered comparison designs, standardised outcome measures, and pre-registered analyses, its findings will remain as inconclusive as the WCF literature at its worst. Overall, the evidence suggests that positive learner evaluations align more closely with the clarity and availability of feedback than with its contribution to higher-order writing development, a pattern that is also well-documented in the broader feedback literature ([Carless & Boud, 2018](#)).

An important paradox rising from the literature concerns the distinction between feedback satisfaction and learning effectiveness. Learners frequently report that LLM feedback is clear, immediate, supportive, and less anxiety-inducing than teacher feedback. However, studies measuring actual writing performance often report only modest gains. This suggests that positive emotional responses to feedback may not reliably predict deeper learning outcomes. Future research should investigate the relationship between perceived usefulness, uptake behaviour, and long-term writing development.

4.3. A Typology of LLM Feedback Functions

Drawing on recurring patterns identified across the corpus, this review proposes a typology of four LLM feedback functions that provides a useful framework for organizing, interpreting, and comparing findings within a rapidly developing area of research. The typology was derived inductively from patterns in how studies described and used LLM feedback. Categories are distinguished by the LLM's primary relationship to the instructional context, whether it functions as a standalone corrective tool, an evaluative instrument, a revision scaffold, or a complement to teacher feedback. The four categories are not mutually exclusive in practice but reflect meaningfully distinct functional profiles in the literature:

1. **Error Correction Agent:** LLM as an automated corrector of surface-level linguistic errors. This function is the most strongly supported in the evidence base, with consistently high uptake and stable performance in accuracy-focused tasks. Its primary limitation lies in occasional inconsistency during complex restructuring.
2. **Formative Evaluator:** LLM as a provider of rubric-based feedback on drafts. Evidence is moderate, with quality varying across feedback dimensions. This function is strengthened by structured prompting and teacher mediation.
3. **Revision Scaffold:** LLM as an interactive partner in revision that responds to learner queries. Although less extensively studied, this function shows potential for deeper engagement, particularly among higher-proficiency learners who use feedback iteratively and selectively.
4. **Teacher Support Tool:** LLM as a resource that handles surface-level corrections at scale, freeing teacher feedback for higher-order content, argumentation, and contextually sensitive guidance. Evidence from hybrid model studies (Asadi et al., 2025; Soori et al., 2025) indicates that this division of labour produces stronger outcomes than either AI or teacher feedback alone, with teacher agency over feedback framing identified as a critical moderating variable.

Beyond these functional roles, several cross-cutting patterns shape how these functions operate in practice.

First, the dominance of form-focused feedback in the literature reflects the relative ease of operationalising surface-level correction compared to more complex dimensions such as content and argumentation. This emphasis can obscure differences in feedback quality, as distinct feedback types are often grouped into broad categories.

Second, a consistent pattern across the corpus is that LLM feedback performs reliably for surface-level correction but shows greater variability when applied to higher-order revisions. Content and argumentation feedback tends to be less context-sensitive than teacher feedback and does not consistently reflect individual learner needs or task-specific expectations. In addition, LLM feedback lacks interpersonal dimensions present in teacher and peer feedback, including affective support and relational awareness.

Third, prompting strategy shapes both the quality of feedback generated and how learners engage with it. Studies comparing more sophisticated, domain-specific prompts with generic or open-ended prompts report differences in the quantity and quality of AI-generated feedback, as well as in learner engagement and revision behaviour, with additional variation across proficiency levels. This indicates that feedback effectiveness depends not only on model output but also on how prompts are constructed and used within instructional contexts.

Finally, pedagogical integration remains underdeveloped. The small number of studies examining structured hybrid models shows variation in how AI and teacher feedback are combined, including differences in sequencing and instructional design. These differences are associated with variation in revision outcomes, although systematic patterns are not yet

established. This typology is intended as a heuristic framework for organising an emerging evidence base rather than as a validated classification. It builds on and extends established distinctions in the WCF literature (Bitchener & Storch, 2016) and AWE research (Warschauer & Grimes, 2008) by specifically foregrounding uptake and pedagogical integration as organising dimensions, aspects not systematically addressed in prior frameworks developed before LLMs were widely available. The proposed functions provide a framework for interpreting the current literature and organizing future research, while their relationships and boundaries remain a productive area for continued investigation.

4.4. Implications for Practice

Four implications follow directly from the evidence and are non-obvious enough to warrant explicit statement. First, hybrid AI–teacher feedback models should be treated as the default design rather than an experiment. The evidence that combined feedback consistently outperforms either source alone is the most practically actionable finding in this synthesis, yet such models represent only 12.4% of the corpus. AI-first approaches tend to support surface-level correction, whereas teacher-first approaches are more aligned with higher-order revision; outcomes also vary with sequencing and instructional framing. Writing programmes using LLM feedback tools without teacher mediation are leaving demonstrated gains unrealised.

Second, LLM feedback should not be deployed without explicit pedagogical scaffolding. Low uptake of meaning-level feedback indicates that access alone is insufficient without instruction in how to interpret and act on feedback during revision. Feedback literacy—the evaluative judgement and metacognitive awareness that enables learners to engage critically rather than wholesale—cannot be assumed and needs to be explicitly developed in any instructional model incorporating LLM feedback.

Third, prompting should be treated as an instructional skill rather than a technical detail. More sophisticated, domain-specific prompts are consistently associated with higher-quality, more actionable AI feedback. Differences in prompt construction are associated with meaningful differences in revision behaviour, which means that how prompting is taught is itself a determinant of how much students benefit from LLM feedback tools.

Fourth, LLM feedback tool selection and deployment should be matched to instructional goals. Current models perform reliably for accuracy-focused feedback and surface-level correction at scale; they perform less reliably for content, argumentation, and contextually sensitive guidance that requires interpretive judgment. This is not a reason to avoid LLMs—it is a reason to use them strategically, with teacher feedback reserved for the higher-order dimensions where human judgment remains superior.

4.5. Limitations

Several limitations of this scoping review should be noted. Charting was conducted using full abstracts exported from Scopus for all 185 included studies. While full abstract text was available for all studies, variables such as prompting strategy, specific sample characteristics, and effect sizes are rarely reported in sufficient detail in abstracts, and keyword-based coding of these nuanced variables should be interpreted with appropriate caution. In particular, manual checking indicated that the abstract-level coding undercounts studies measuring learner uptake, since some studies that measured uptake did not signal this in their abstracts; the reported uptake frequency should therefore be read as a lower-bound estimate. Full-text retrieval for 35 open-access studies provided additional depth for the narrative synthesis of key themes. Future syntheses incorporating systematic full-text access would substantially improve on the precision of the present review.

A second limitation concerns search scope: the search was conducted in Scopus only. While Scopus provides broad coverage of major journals in applied linguistics, relevant work in non-indexed regional journals may be underrepresented. Third, the field's near-universal failure to specify LLM version (71.4% of studies using "ChatGPT unspecified") limits the interpretability of findings, as different LLM versions have substantially different capabilities and therefore reduces the precision of cross-study comparisons. Fourth, meaningful geographic charting was not possible because institutional affiliation data was not consistently available in abstract metadata. Partial geographic inference was possible only for studies that explicitly named their institutional context in the abstract, representing a minority of the corpus. Based on this non-representative subset, the corpus may be weighted toward East Asian EFL, Gulf EMI, and Turkish EFL contexts, though this inference should be treated as speculative in the absence of systematic affiliation data. Future scoping reviews should prioritise systematic geographic charting using full-text affiliation data to address this gap properly.

5. Conclusions

This scoping review synthesized 185 empirical studies on LLM-generated feedback in L2 writing published between November 2022 and March 2026. Conducted in accordance with established scoping review methodology and PRISMA-ScR reporting guidelines, the review provides the first comprehensive synthesis of this rapidly expanding field through the integrated lenses of feedback quality, learner uptake, and pedagogical integration.

There were four principal findings. First, the literature is dominated by general and comparative AI-human feedback research, while content-level and higher-order feedback remain comparatively underexplored. Second, learner uptake occupies a surprisingly limited position within the literature despite its central role in writing development. The available evidence suggests that learners engage more readily with surface-level feedback than with feedback addressing ideas, organization, and argumentation. Third, LLM-generated feedback appears broadly comparable to teacher feedback for surface-level language accuracy but less consistent for content development and rhetorical concerns. Finally, hybrid-AI-teacher feedback models show considerable promise but remain relatively uncommon within the current evidence base.

These findings suggest that the effectiveness of LLM-generated feedback depends not only on feedback quality but also on how learners engage with feedback and how feedback is integrated into instructional practice. To help organize this growing body of literature, the review proposes a four-function typology consisting of Error Correction Agent, Formative Evaluator, Revision Scaffold, and Teacher Support Tool. The typology provides a framework for interpreting existing findings and guiding future research on LLM-supported writing instruction.

Several priorities can be suggested for future research. Greater attention should be given to learner uptake, longitudinal writing development, hybrid feedback models, and the role of feedback literacy in mediating engagement with AI-generated feedback. More consistent reporting of model versions and implementation procedures would also strengthen cross-study comparison and improve the precision of future syntheses. As LLM technologies continue to evolve, research that moves beyond perceptions toward demonstrated learning outcomes will be essential for understanding how AI-generated feedback can best support second language writing development.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not Applicable.

Data Availability Statement: The charting dataset generated for this study is not publicly available, as it forms the basis of an ongoing programme of research, including a planned systematic review by the authors.

Acknowledgments: During the preparation of this work, the authors used an AI-assisted writing tool (ChatGPT, GPT-5.5, OpenAI, 2026) for the purposes of language editing and manuscript refinement.

Conflicts of Interest: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Appendix A. Full-Text Studies Retrieved via Open-Access DOI Resolution (n = 35)

Full bibliographic details are provided in the reference list. Appendix A identifies the subset of studies included in the full-text synthesis (n = 35).

The following studies were retrieved in full text via open-access DOI resolution to supplement abstract-level charting. They are identified with (FT) in the narrative synthesis and in Table 4.

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

Asadi, M., Ebadi, S., & Mohammadi, L. (2025). The impact of integrating ChatGPT with teachers' feedback on EFL writing skills. *Thinking Skills and Creativity*, 56, 101766. <https://doi.org/10.1016/j.tsc.2025.101766>.

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), 51–65. <https://doi.org/10.17323/jle.2025.27195>.

Cengiz, B. C., Bilki, Z., Atas, A. H., & Celik, B. (2025). Exploring second language writers' engagement with ChatGPT feedback: Revision behaviors and perceptions. *System*, 134, 103837. <https://doi.org/10.1016/j.system.2025.103837>.

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>.

Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies*, 29(7), 8435–8463. <https://doi.org/10.1007/s10639-023-12146-0>.

Hwang, M., Jeens, R., & Lee, H.-K. (2025). Exploring learner prompting behavior and its effect on ChatGPT-assisted English writing revision. *Asia-Pacific Education Researcher*, 34(3), 1157–1167. <https://doi.org/10.1007/s40299-024-00930-6>.

Hyland, K. (2025). EFL teachers and feedback fatigue: AI to the rescue? *Language Teaching*. <https://doi.org/10.1017/S0261444825101018>.

Kang, S., & Kim, M. (2026). AI-generated feedback in an EAP writing classroom: The collaborative process of feedback, uptake, and revision quality. *Journal of English for Academic Purposes*, 80, 101639. <https://doi.org/10.1016/j.jeap.2026.101639>.

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, 101130. <https://doi.org/10.1016/j.jslw.2024.101130>.

Li, J., Huang, J., Wu, W., & Whipple, P. B. (2024). Evaluating the role of ChatGPT in enhancing EFL writing assessments in classroom settings. *Humanities and Social Sciences Communications*, 11(1), 1268. <https://doi.org/10.1057/s41599-024-03755-2>.

- Lin, S., & Crosthwaite, P. (2024). The grass is not always greener: Teacher vs. GPT-assisted written corrective feedback. *System*, 127, 103529. <https://doi.org/10.1016/j.system.2024.103529>.
- Luo, N., Wang, Y., Zhang, Z. V., Zhou, Y., & Zhao, R. (2025). Impact of prompt sophistication on ChatGPT's output for automated written corrective feedback. *ReCALL*. <https://doi.org/10.1017/S095834402510044X>.
- Ma, X., Zhang, C., & Lee, I. (2026). L2 learners' engagement with AI-generated feedback on writing. *Assessing Writing*, 68, 101020. <https://doi.org/10.1016/j.asw.2026.101020>.
- Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: a mixed methods intervention study. *Smart Learning Environments*, 11(1), 9. <https://doi.org/10.1186/s40561-024-00295-9>.
- Meng, Y., Cui, Y., Geng, X., & Cao, Y. (2025). Exploring feedback literacy in L2 students' academic writing: Insights from their reported engagement with GenAI for typical revision activities. *The Asia-Pacific Education Researcher*. <https://doi.org/10.1007/s40299-025-01051-4>.
- Muñoz Muñoz, B. C., Nassaji, H., & Bello Carrillo, F. I. (2025). ChatGPT-generated versus human direct corrective feedback on L2 writing. *System*, 134, 103805. <https://doi.org/10.1016/j.system.2025.103805>.
- Polakova, P., & Ivenz, P. (2024). The impact of ChatGPT feedback on the development of EFL students' writing skills. *Cogent Education*, 11(1), 2410101. <https://doi.org/10.1080/2331186X.2024.2410101>.
- 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, 101286. <https://doi.org/10.1016/j.jslw.2026.101286>.
- Saricaoglu, A., & Bilki, Z. (2025). The capacity of ChatGPT-4 for L2 writing assessment: A closer look at accuracy, specificity, and relevance. *Annual Review of Applied Linguistics*, 45, 253–273. <https://doi.org/10.1017/S0267190525100160>.
- Schiller, R., Fleckenstein, J., Höft, L., Horbach, A., & Meyer, J. (2025). On the role of engagement in automated feedback effectiveness: Insights from keystroke logging. *Computers and Education*, 238, 105386. <https://doi.org/10.1016/j.compedu.2025.105386>.
- Shi, H., Chai, C. S., Zhou, S., & Aubrey, S. (2025). Comparing the effects of ChatGPT and automated writing evaluation on students' writing and ideal L2 writing self. *Computer Assisted Language Learning*. <https://doi.org/10.1080/09588221.2025.2454541>.
- Soori, A., Khojasteh, L., & Javed, F. (2025). Comparing teacher E-feedback, AI feedback, and hybrid feedback in enhancing EFL writing skills. *Technology in Language Teaching and Learning*, 7(3), 102626. <https://doi.org/10.29140/tl.v7n3.102626>.
- Tran, T. T. T. (2025). Enhancing EFL writing revision practices: The impact of AI- and teacher-generated feedback and their sequences. *Education Sciences*, 15(2), 232. <https://doi.org/10.3390/educsci15020232>.
- Wang, C., & Wang, Z. (2025). Investigating L2 writers' critical AI literacy in AI-assisted writing: An APSE model. *Journal of Second Language Writing*, 67, 101187. <https://doi.org/10.1016/j.jslw.2025.101187>.
- Wu, Z., & Zhao, J. (2025). Exploring learners' engagement with GenAI-generated feedback in EFL writing: a Chinese middle school case study. *Computer Assisted Language Learning*. <https://doi.org/10.1080/09588221.2025.2600035>.
- 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(1), 1086. <https://doi.org/10.1057/s41599-024-03543-y>.

- Yang, Y., Hussin, S., & Hashim, H. (2025). Feedback literacy and EFL learner engagement with ChatGPT feedback: Predicting feedback uptake and perceived usefulness. *Theory and Practice in Language Studies*, 15(11), 3760–3771. <https://doi.org/10.17507/tpls.1511.30>.
- Yeung, S. (2025). University students' engagement with generative AI-supported automated writing evaluation (AWE) feedback. *Journal of Second Language Writing*, 68, 101203. <https://doi.org/10.1016/j.jslw.2025.101203>.
- Yu, H., & Xie, Q. (2025). Generative AI vs. teachers: Feedback quality, feedback uptake, and revision. *Language Teaching Research Quarterly*, 47, 113–137. <https://doi.org/10.32038/ltrq.2025.47.07>.
- Zhang, Z., Aubrey, S., Huang, X., & Chiu, T. K. F. (2025). The role of generative AI and hybrid feedback in improving L2 writing skills: a comparative study. *Innovation in Language Learning and Teaching*. <https://doi.org/10.1080/17501229.2025.2503890>.
- Zhu, R., Wang, H., & Qin, X. (2025). Longitudinal comparison of AI, exemplar, and teacher feedback for sustainable L2 writing development: a latent growth curve analysis. *Innovation in Language Learning and Teaching*. <https://doi.org/10.1080/17501229.2025.2586142>.
- Ziqi, C., Xinhua, Z., Qi, L., & Wei, W. (2024). L2 students' barriers in engaging with form and content-focused AI-generated feedback in revising their compositions. *Computer Assisted Language Learning*. <https://doi.org/10.1080/09588221.2024.2422478>.
- Zohali, L., Storch, N., & Knoch, U. (2026). From red pen to algorithm: How do L2 learners engage with human and automated feedback? *RELC Journal*. <https://doi.org/10.1177/00336882261419844>.
- Zou, S., Guo, K., Wang, J., & Liu, Y. (2025). Investigating students' uptake of teacher- and ChatGPT-generated feedback in EFL writing: a comparison study. *Computer Assisted Language Learning*. <https://doi.org/10.1080/09588221.2024.2447279>.

References

- Alnemrat, A., Aldamen, H., Almashour, M., Al-Deaibes, M., & AlSharefeen, R. (2025). AI vs. teacher feedback on EFL argumentative writing: A quantitative study. *Frontiers in Education*, 10, 1614673. [\[CrossRef\]](#)
- Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology*, 8(1), 19–32. [\[CrossRef\]](#)
- Asadi, M., Ebadi, S., & Mohammadi, L. (2025). The impact of integrating ChatGPT with teachers' feedback on EFL writing skills. *Thinking Skills and Creativity*, 56, 101766. [\[CrossRef\]](#)
- Bitchener, J., & Storch, N. (2016). *Written corrective feedback for L2 development*. Multilingual Matters. [\[CrossRef\]](#)
- 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. [\[CrossRef\]](#)
- Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, 43(8), 1315–1325. [\[CrossRef\]](#)
- Cengiz, B. C., Bilki, Z., Atas, A. H., & Celik, B. (2025). Exploring second language writers' engagement with ChatGPT feedback: Revision behaviors and perceptions. *System*, 134, 103837. [\[CrossRef\]](#)
- Crosthwaite, P., & Sun, S. (2025). Generative AI and L2 written feedback studies: A scoping review. *RELC Journal*, 57(1), 207–219. [\[CrossRef\]](#)
- 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. [\[CrossRef\]](#)
- Ferris, D. R. (2003). *Response to student writing: Implications for second language students*. Lawrence Erlbaum.
- Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies*, 29(7), 8435–8463. [\[CrossRef\]](#)
- Hwang, M., Jeens, R., & Lee, H.-K. (2025). Exploring learner prompting behavior and its effect on ChatGPT-assisted English writing revision. *Asia-Pacific Education Researcher*, 34(3), 1157–1167. [\[CrossRef\]](#)
- Hyland, K. (2025). EFL teachers and feedback fatigue: AI to the rescue? *Language Teaching*, 59, 310–326. [\[CrossRef\]](#)
- Kang, E., & Han, Z. (2015). The efficacy of written corrective feedback in improving L2 written accuracy: A meta-analysis. *The Modern Language Journal*, 99(1), 1–18. [\[CrossRef\]](#)

- 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, 101130. [\[CrossRef\]](#)
- Lantolf, J. P., & Thorne, S. L. (2006). *Sociocultural theory and the genesis of second language development*. Oxford University Press.
- Levac, D., Colquhoun, H., & O'Brien, K. K. (2010). Scoping studies: Advancing the methodology. *Implementation Science*, 5(1), 69. [\[CrossRef\]](#) [\[PubMed\]](#)
- Li, J., Huang, J., Wu, W., & Whipple, P. B. (2024). Evaluating the role of ChatGPT in enhancing EFL writing assessments in classroom settings. *Humanities and Social Sciences Communications*, 11(1), 1268. [\[CrossRef\]](#)
- Luo, N., Wang, Y., Zhang, Z. V., Zhou, Y., & Zhao, R. (2025). Impact of prompt sophistication on ChatGPT's output for automated written corrective feedback. *ReCALL*, 38, 240–254. [\[CrossRef\]](#)
- Saricaoglu, A., & Bilki, Z. (2025). The capacity of ChatGPT-4 for L2 writing assessment: A closer look at accuracy, specificity, and relevance. *Annual Review of Applied Linguistics*, 45, 253–273. [\[CrossRef\]](#)
- Shi, H., Chai, C. S., Zhou, S., & Aubrey, S. (2026). Comparing the effects of ChatGPT and automated writing evaluation on students' writing and ideal L2 writing self. *Computer Assisted Language Learning*, 39, 1148–1175. [\[CrossRef\]](#)
- Sinha, T. S., & Nassaji, H. (2022). ESL learners' perception and its relationship with the efficacy of written corrective feedback. *International Journal of Applied Linguistics*, 32, 41–56. [\[CrossRef\]](#)
- Soori, A., Khojasteh, L., & Javed, F. (2025). Comparing teacher e-feedback, AI feedback, and hybrid feedback in enhancing EFL writing skills. *Technology in Language Teaching and Learning*, 7(3), 102626. [\[CrossRef\]](#)
- Tran, T. T. T. (2025). Enhancing EFL writing revision practices: The impact of AI- and teacher-generated feedback and their sequences. *Education Sciences*, 15(2), 232. [\[CrossRef\]](#)
- 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. [\[CrossRef\]](#) [\[PubMed\]](#)
- Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning*, 46(2), 327–369. [\[CrossRef\]](#)
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
- Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal*, 3(1), 22–36. [\[CrossRef\]](#)
- 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(1), 1086. [\[CrossRef\]](#)
- Yang, Y., Hussin, S., & Hashim, H. (2025). Feedback literacy and EFL learner engagement with ChatGPT feedback: Predicting feedback uptake and perceived usefulness. *Theory and Practice in Language Studies*, 15(11), 3760–3771. [\[CrossRef\]](#)
- Yu, H., & Xie, Q. (2025). Generative AI vs. teachers: Feedback quality, feedback uptake, and revision. *Language Teaching Research Quarterly*, 47, 113–137. [\[CrossRef\]](#)
- 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*, 39, 1062–1091. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.