HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

BEYOND AI CORRECTION: THE METACOGNITIVE RECALIBRATION OF EFL LEARNERS THROUGH CONFIDENCE–EVIDENCE–REVISION CYCLES

Temirxonova Shaxzoda Qobilovna
Student of Samarkand State Institute of Foreign Languages
doi:https://doi.org/10.5281/zenodo.23466852

Abstract. This article examines how artificial intelligence feedback can support the development of metacognitive skills among English as a Foreign Language (EFL) learners. Using the Confidence–Evidence–Revision (CER) model, it explores learners' self-assessment, evidence-based evaluation of AI suggestions, and independent revision skills. Proposed evaluation indicators and directions for future empirical research are also presented.
Keywords: artificial intelligence, EFL, metacognition, confidence calibration, evidence-based revision, writing skills, feedback.

Annotatsiya. Mazkur maqolada ingliz tilini xorijiy til sifatida o‘rganuvchilarning (EFL) metakognitiv ko‘nikmalarini rivojlantirishda sun’iy intellekt fikr-mulohazalaridan foydalanish masalasi tahlil qilinadi. Ishonch–Dalil–Tahrir (CER) modeli asosida o‘quvchilarning o‘z bilimlarini baholashi, AI tavsiyalarini dalillar asosida tekshirishi va mustaqil tahrirlash ko‘nikmalarini rivojlantirish mexanizmlari yoritiladi. Modelning samaradorligini baholash mezonlari va kelgusidagi empirik tadqiqot yo‘nalishlari taklif etiladi.
Kalit so‘zlar: sun’iy intellekt, EFL, metakognitsiya, o‘ziga ishonch, dalillarga asoslangan tahrir, yozma nutq, fikr-mulohaza.

Аннотация. В статье рассматривается использование обратной связи искусственного интеллекта для развития метакогнитивных навыков учащихся, изучающих английский язык как иностранный. На основе модели «Уверенность–Доказательства–Редактирование» (CER) анализируются способы оценки собственных знаний, проверки рекомендаций ИИ и развития самостоятельного редактирования. Предлагаются критерии оценки эффективности модели и направления дальнейших эмпирических исследований.
Ключевые слова: искусственный интеллект, EFL, метакогниция, уверенность, доказательная проверка, редактирование текста, обратная связь.

Introduction

Generative AI has changed the practical conditions under which second-language writers receive feedback. A learner can now request a grammar explanation, compare alternative sentences, or obtain comments on coherence within seconds. This availability creates an instructional opportunity, but it

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

also creates a conceptual problem: a corrected text may look more accurate even when the learner has not developed the knowledge or judgement required to produce a similar text independently. The central educational question is therefore not simply whether AI can correct language, but under what conditions feedback becomes a resource for learning rather than a substitute for thinking.

In EFL contexts, this distinction is especially important because learners often work with limited exposure to English outside the classroom. They may have difficulty distinguishing a genuine grammatical error from a stylistic preference, judging the reliability of an AI explanation, or deciding whether a suggested revision preserves intended meaning. When a system presents fluent language with high apparent certainty, learners may over-trust it; when a learner assumes that their own first answer is correct, they may reject useful feedback. Both patterns reflect a mismatch between confidence and evidence.

This article proposes that metacognitive recalibration can be strengthened through a Confidence–Evidence–Revision (CER) cycle. The learner records an initial judgement, checks it against linguistic and contextual evidence, makes a reasoned revision decision, and then evaluates the accuracy of the original judgement. The cycle shifts the learning target from error removal alone to improved monitoring: learners should gradually become better at knowing what they know, identifying uncertainty, and choosing when external support is needed.

Conceptual foundations

Metacognition involves knowledge about one's cognitive processes and the monitoring and regulation of those processes. In writing, it includes planning, checking whether a sentence communicates the intended meaning, detecting uncertainty, and deciding which revision strategy is appropriate. Confidence calibration refers to the alignment between a learner's expressed confidence and the correctness or quality of the learner's judgement. A learner who is highly confident in an incorrect grammar decision is miscalibrated; a learner who expresses uncertainty about a correct decision may be underconfident. Both cases provide information that a final score alone cannot reveal.

The CER cycle treats confidence not as a personality trait but as a revisable judgement. Before consulting AI, learners estimate how certain they are about a selected language feature. They then inspect evidence: the surrounding discourse, a grammar rule, parallel examples, corpus examples where available, or a teacher-approved explanation. Finally, they decide whether to accept, adapt, or reject the AI suggestion and explain why. This sequence makes the learner's reasoning visible and creates a record that can be compared with later performance. Feedback is not automatically useful merely because it is accurate. Learners must understand what feedback means, evaluate its relevance, manage emotional responses, and take action. AI may lower the time cost of obtaining feedback, but it cannot guarantee that a learner interprets the feedback correctly. A human-centred approach therefore keeps the learner responsible for the final decision and treats the teacher as a designer of productive feedback routines, not merely a verifier of AI output.

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

UNESCO's guidance on generative AI in education advocates a human-centred approach, including attention to human agency, inclusion, privacy, and the validation of educational use. This is compatible with CER because the learner is required to examine evidence rather than outsource judgement. The model does not assume that AI output is always wrong or always right; it assumes that its educational value depends on how learners interrogate and use it. Research on digital feedback generally indicates positive average effects, while also showing that outcomes vary according to task, design, learner characteristics, and the type of feedback provided. The evidence supports the use of feedback technologies as potentially useful instructional resources, but it does not justify the stronger claim that access to AI alone reliably produces independent learning.

These estimates should not be ranked as if they came from a single controlled comparison. The studies differ in their inclusion criteria, intervention types, outcome measures, publication periods, and analytical models. The smaller pooled effect in the 2026 synthesis may reflect its specific focus on AI-assisted L2 writing and the outcomes included, rather than a contradiction of earlier findings. The practical implication is not that one technology is universally superior, but that implementation and learning mechanisms matter.

More directly relevant to the proposed framework, a 2026 quasi-experimental study involving 141 EFL learners compared metacognitive instruction plus GenAI feedback with GenAI feedback without the same additional scaffolding. The intervention group showed stronger strategy knowledge and metacognitive strategy use, and more elaborate uptake of AI feedback on global writing issues. The reported advantage concerned content and organization revision; the study did not find significant between-group differences in grammar, vocabulary, and mechanics. This pattern supports the argument that explicit metacognitive instruction may change how learners engage with feedback, although one study cannot establish that CER itself is effective.

A further 2026 meta-analysis of AI feedback in higher education synthesized 234 effect sizes from 43 studies and reported a moderate overall effect on writing performance ($g = 0.58$). It also reported stronger outcomes when AI feedback was combined with human feedback. Taken together, the literature suggests a plausible design principle: AI can expand access to feedback, while metacognitive prompts and teacher guidance can help learners interpret it. The specific causal contribution of confidence recording, evidence checking, and revision justification still requires direct testing.

Selected quantitative evidence

Source / synthesis	Evidence base	Reported estimate	Outcome focus	Interpretive caution
Online feedback in ESL/EFL writing (2021)	17 studies; 19 results	Effect size = 0.753	Writing outcomes	Online feedback is broader than generative AI.

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

Automated writing evaluation (2023)	20 studies; 84 effect sizes; N = 2,828	Hedges' g = 0.55	Writing performance	Automated evaluation tools differ from current GenAI systems.
AI-assisted L2 writing (2026)	19 studies; 66 effect sizes; N = 11,875	Cohen's d = 0.271; 95% CI [0.185, 0.358]	Perceptions, emotions, learning outcomes	Pooled effect is modest; high heterogeneity was reported.
AI feedback in higher education (2026)	43 studies; 234 effect sizes	g = 0.58	English writing performance	Hybrid AI–human feedback was reported as more effective.

Table 1. Selected published meta-analytic findings. Estimates are reported as given in the cited sources and are not a direct comparative ranking.

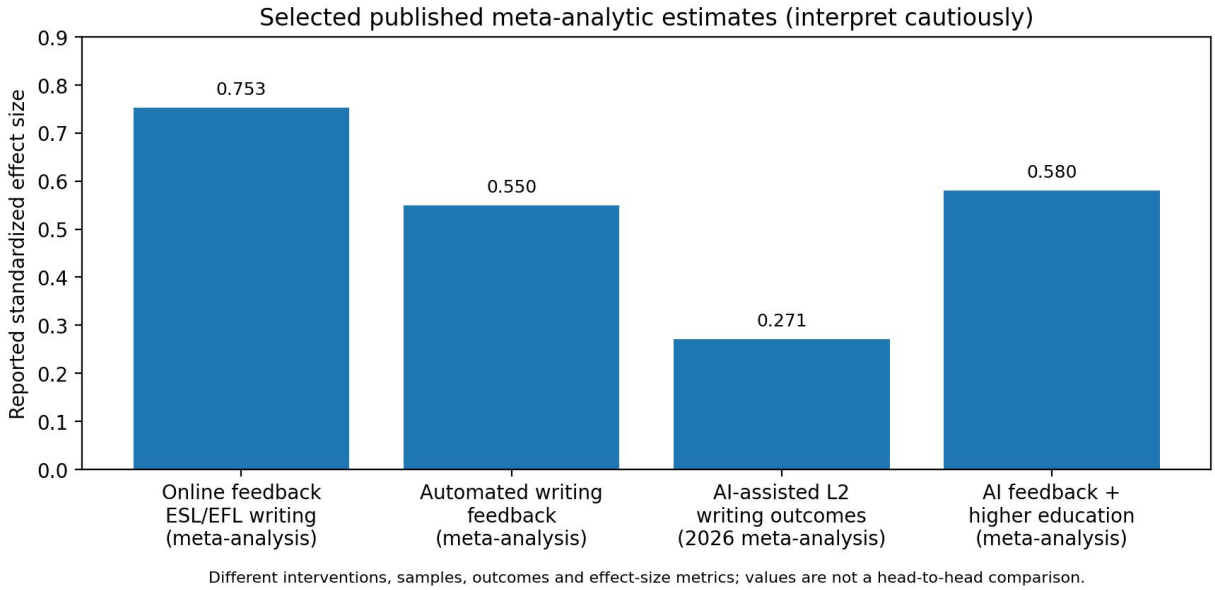


Figure 1. Selected published estimates shown for descriptive orientation only. Differences in research questions and methods prevent direct head-to-head interpretation.

The Confidence–Evidence–Revision model

The proposed CER cycle contains five connected moves. First, the learner states a judgement about a specific language feature before asking AI for an answer. Second, the learner records confidence on a simple scale, for example from 0 to 100, or chooses a verbal category such as low, moderate, or high. Third, the learner requests a focused explanation or compares the sentence with an authoritative rule or example. Fourth, the learner makes a revision decision—accept, adapt, or reject—and provides a short rationale. Fifth, the learner checks the revised text and updates their confidence in

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

the original judgement. The process then recurs in a later task so that recalibration can be observed over time.

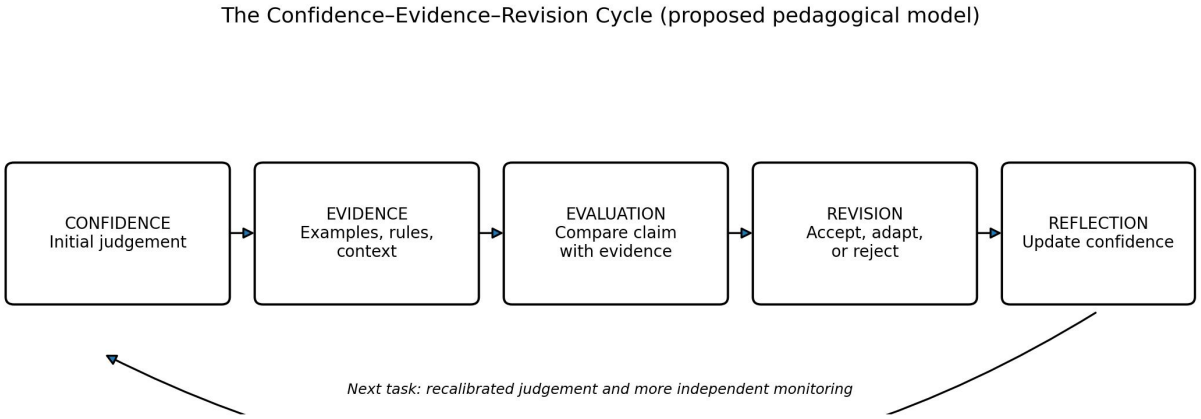


Figure 2. Proposed CER cycle. This is a conceptual model created for this article, not a measured empirical result.

Why confidence must precede AI feedback

If learners see an AI correction before recording their own judgement, the original level of knowledge becomes difficult to observe. A learner may later report that the answer seems obvious, although the answer was not available to them independently. Asking for a confidence estimate first creates a baseline and can reduce hindsight bias. The estimate need not be perfectly precise: its value lies in making uncertainty explicit and enabling comparison across repeated tasks. Confidence should not be treated as a score of ability in isolation. Some learners may systematically use the scale differently, and language proficiency, anxiety, task difficulty, and prior experience may affect ratings. Teachers should therefore interpret confidence together with task accuracy and the quality of the learner's explanation. The most informative cases may be high-confidence errors, because these indicate where a learner's internal rule or heuristic requires reconsideration.

AI feedback can be linguistically plausible without being optimal for the intended context. A suggested expression may be grammatical but alter emphasis, register, politeness, or meaning. Consequently, learners should be prompted to ask: What claim is the system making? What evidence supports it? Does the explanation fit the surrounding text? Is the suggestion required by grammar, or is it only one stylistic option? Can I explain the change without copying the system's wording? This routine is particularly relevant to academic writing, where cohesion, argument structure, and precision matter alongside grammar. The learner should compare alternatives rather than assume that the most fluent version is necessarily the best version. When the AI's explanation conflicts with a teacher-provided source or a reliable reference, the conflict should become an object of inquiry rather than a reason for automatic acceptance.

Revision quality involves more than the number of edits. A learner may accept a correct suggestion, adapt a partially useful suggestion, or reject a suggestion that damages meaning. All three

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

decisions can be evidence of learning if they are supported by a defensible rationale. Conversely, accepting every suggestion may yield a polished paragraph while concealing weak understanding. A revision log should therefore record the original sentence, the proposed change, the learner's decision, the evidence considered, and the reason for the decision. The cycle should also include transfer. After guided revision, learners should complete a short new task without AI and explain one or two language decisions. If performance improves only on the AI-assisted draft, the intervention may have improved the product without producing independent competence. Delayed tasks are needed to determine whether the learner retains the underlying principle.

To test CER rigorously, future classroom research should use a pretest–intervention–posttest design with a delayed follow-up. A suitable comparison would contrast ordinary AI feedback with AI feedback embedded in CER routines, while keeping the writing task, access to the AI system, instructional time, and scoring rubric as similar as possible. If feasible, classes should be assigned at the class level to reduce contamination between conditions. Blind raters should assess writing quality, and the analysis should account for baseline proficiency and clustering within classes.

Indicator	Operational definition	Why it matters
Confidence calibration	Difference between confidence and correctness on targeted language decisions	Shows whether self-judgement becomes better aligned with evidence.
Feedback uptake quality	Proportion of accepted, adapted, or rejected suggestions supported by an accurate rationale	Distinguishes thoughtful engagement from indiscriminate acceptance.
Revision quality	Blind-rated change in content, organization, language accuracy, and coherence	Measures the quality of the revised text rather than edit quantity alone.
Transfer	Performance on a new writing task completed without AI	Tests whether learning extends beyond the assisted draft.
Delayed retention	Performance on a parallel task after a planned interval	Checks whether gains persist rather than disappear after immediate feedback.

Table 2. Proposed evaluation indicators for a future empirical study; no new participant data are claimed in this article.

A useful calibration measure is the absolute difference between a learner's confidence estimate and the correctness of the decision, expressed on a common scale. For example, if confidence is recorded from 0 to 100 and a decision is scored as correct or incorrect, a simple descriptive measure can compare confidence with the corresponding outcome. However, binary correctness may

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

oversimplify writing decisions, because some revisions are matters of register or style rather than clear right–wrong choices. Researchers should separate objectively scorable grammar items from expert-rated discourse decisions and report reliability for the latter.

Quantitative analysis should include effect sizes and confidence intervals, not p-values alone. Qualitative data—think-aloud protocols, revision histories, and short reflective notes—can explain why learners accepted or rejected suggestions. A mixed-methods design would help distinguish superficial compliance from evidence-based decision-making. Researchers should preregister primary outcomes where possible, report attrition, and avoid presenting a small convenience sample as representative of all EFL learners.

A practical CER lesson can begin with a short paragraph containing a small number of target features. Before using AI, each learner marks two uncertain decisions and records confidence. The learner then asks the AI for an explanation rather than a complete rewrite. Next, the learner checks the explanation against class materials or a trusted reference and annotates the evidence. After making a decision, the learner writes a brief rationale in their own words. The teacher reviews the rationale, not merely the final sentence, and selects examples of strong reasoning for discussion. Teachers should avoid turning the routine into a bureaucratic worksheet for every error. It is more productive to target high-value language features and gradually reduce scaffolding as students develop skill. For beginners, confidence categories may be easier than numeric ratings; more advanced learners can distinguish grammatical certainty from stylistic preference. Pair discussion can also make reasoning visible, provided that learners still submit an individual final judgement.

Assessment should reward justified decisions, accurate explanations, and independent transfer—not simply the number of AI suggestions adopted. Students should be allowed to disagree with AI when they can support their decision. This is essential to preserving agency and preventing the classroom from treating machine output as an unquestionable authority. Institutions should also set clear expectations about privacy, attribution, acceptable use, and the handling of student writing.

The present article is a conceptual synthesis, not a systematic review or a new experiment. It does not claim that the CER cycle has already been validated, and the proposed diagram and indicators should be treated as hypotheses for future testing. The selected meta-analyses differ in scope and methodology; their effect sizes provide context, not a single unified estimate of AI's impact. Publication bias, rapidly changing tools, inconsistent outcome measures, and variation in learner proficiency may limit generalization. AI systems may produce incorrect explanations, culturally narrow examples, or revisions that erase a learner's intended voice. They may also process sensitive student text, making data protection relevant. A responsible implementation should use institutionally approved tools where available, avoid uploading identifiable or confidential material, and teach students to verify claims. AI feedback should supplement, not replace, teacher expertise and peer interaction. The objective is not to eliminate errors immediately but to make errors diagnostically useful and to build learners' capacity to evaluate evidence.

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

Conclusion

The educational value of AI-assisted EFL writing cannot be judged solely by how many errors disappear from a draft. A stronger criterion is whether learners become more capable of detecting uncertainty, evaluating explanations, making defensible revision decisions, and transferring what they learn to later tasks. The Confidence–Evidence–Revision cycle offers a structured way to pursue these outcomes by placing learner judgement before AI feedback and requiring evidence-based reflection after revision. Existing meta-analytic findings suggest that digital, automated, and AI-assisted feedback can support writing outcomes, but effect sizes vary and improvements in immediate performance do not automatically establish durable cognitive gains. Early empirical evidence also suggests that explicit metacognitive instruction can influence how EFL learners engage with GenAI feedback. Future studies should test CER directly with transparent comparison conditions, validated measures of calibration, blind-rated writing, and delayed transfer tasks. Until such evidence is available, CER is best understood as a theoretically motivated and testable pedagogical framework—not a proven universal solution. Its core principle is straightforward: AI should help learners examine evidence and refine judgement, rather than make judgement unnecessary.

References

1. Al-Hassan, M. (2021). The effects of online feedback on ESL/EFL writing: A meta-analysis. [Article available through PubMed Central]. https://pmc.ncbi.nlm.nih.gov/articles/PMC8155179/
2. Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. (2023). PubMed. https://pubmed.ncbi.nlm.nih.gov/37465061/
3. Effects of AI feedback on students' English writing performance in higher education: A meta-analysis. (2026). Cogent Education. https://doi.org/10.1080/2331186X.2026.2665494
4. Miao, F., & Holmes, W. (2023). Guidance for generative AI in education and research. UNESCO. https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research
5. The effects of AI-assisted feedback on students' perceptions, emotions, and learning outcomes in L2 writing: A three-level meta-analysis. (2026). System, 139, 104018. https://doi.org/10.1016/j.system.2026.104018
6. Yan, D. (2026). Metacognitive instruction to enhance EFL writers' metacognition, feedback uptake, and revision performance in a GenAI-assisted feedback context. International Journal of Applied Linguistics. https://doi.org/10.1111/ijal.70328