



EFL Students' Error Detection, Correction, and Acceptance Calibration: A Performance-Based Study at the Faculty of Education–Surman

A performance-based investigation with implications for critical evaluation of AI-assisted language

Majdi Sami Ali Awn

Department of English, Faculty of Education–Surman, Sabratha University, Libya

Corresponding author: Majdi.oun@sabu.edu.ly

Received: 05/9/2026 Revised: 10/9/2026 Accepted: 03/10/2026 Publication: 12/10/2026

Abstract

Generative artificial intelligence (GenAI) is increasingly used in English language learning for writing support, grammar correction, feedback, translation, and academic language production. Its educational value depends not only on students' ability to use AI tools but also on their ability to evaluate AI output and make appropriately calibrated reliance decisions. This study investigated EFL students' performance in detecting and correcting linguistic errors in a 30-item English sentence evaluation task and examined the correspondence between their acceptance or rejection decisions and the observed correctness of the items. A quantitative descriptive-correlational design was used with 30 EFL students who evaluated 17 correct sentences and 13 sentences containing targeted errors. For each item, participants judged correctness, provided a correction when necessary, indicated whether they would accept the sentence without further checking, and rated their confidence on a five-point scale. Detection, correction, acceptance calibration, overacceptance, and underacceptance were summarized using means and standard deviations. Spearman's rank-order correlation was used to examine associations between performance and acceptance calibration. Mean detection performance was 92.22% (SD = 4.23), mean correction performance was 91.28% (SD = 8.01), and mean acceptance calibration was 75.33% (SD = 8.51). Mean overacceptance was 17.18% (SD = 10.24), whereas mean underacceptance was 30.39% (SD = 13.56). No statistically significant association was observed between detection and acceptance calibration ($p = .056$, $p = .770$) or between correction and acceptance calibration ($p = .003$, $p = .987$). These findings indicate that accurate linguistic evaluation and calibrated acceptance decisions should not be treated as interchangeable outcomes. Within the limits of the small, non-probability sample and the single task used in this study, the results support the inclusion of explicit verification and critical evaluation activities in EFL instruction.

Keywords: generative artificial intelligence; EFL; error detection; correction; acceptance calibration; critical AI literacy

1. Introduction

1.1 Background of the Study

Generative artificial intelligence has rapidly become part of contemporary educational practice. Tools based on large language models can generate explanations, examples, summaries, translations, grammar corrections, and academic prose within seconds. In English language education, these capabilities create opportunities for individualized support and immediate language assistance. At the same time, they create a new pedagogical problem: AI-assisted language can be fluent and persuasive while still containing grammatical, lexical, pragmatic, or factual problems.

This problem is especially relevant in English as a Foreign Language (EFL) contexts. Learners may use AI to obtain corrections, improve sentences, generate ideas, or prepare academic

assignments. If students assume that fluent AI output is necessarily correct, they may reproduce errors without verification. Conversely, if students distrust AI indiscriminately, they may reject useful language assistance. The educational goal is therefore not simply high trust or low trust, but appropriately calibrated trust.

Recent scholarship increasingly treats critical AI literacy as a multidimensional competence involving the ability to understand AI, use it strategically, evaluate its outputs, and negotiate its limitations and ethical implications. Zhang et al. (2026) synthesized empirical research in English language education and identified critical and ethical evaluation as an important component of GenAI literacy. They also cautioned that self-reported literacy does not, by itself, demonstrate actual performance. Within this emerging literature, a particularly important issue is the relationship between performance and trust. A learner may identify an error when explicitly asked to evaluate a sentence but nevertheless accept similar output during ordinary AI use. Thus, linguistic evaluation and acceptance calibration should not automatically be assumed to represent the same competence.

1.2 Statement of the Problem

The increasing availability of generative AI in EFL learning has created a need to understand not only whether students use AI, but also whether they can critically evaluate its output. Much of the early discussion of AI in language education focused on opportunities, perceptions, adoption, and the usefulness of AI-generated feedback. More recent research has emphasized critical AI literacy, verification, and responsible reliance (Cao & Mao, 2026; Zhang et al., 2026). Nevertheless, performance-based evidence of students' ability to identify and correct AI-assisted English and their simultaneous decisions to trust or reject that output remains comparatively limited. The problem addressed by this study is therefore the possible mismatch between students' linguistic ability to evaluate AI-assisted English and their actual trust decisions.

1.3 Purpose of the Study

The purpose of this study was to investigate EFL students' ability to detect and correct sentence-level linguistic errors and to examine how accurately their acceptance or rejection decisions correspond to the actual correctness of the sentences.

1.4 Research Questions

- RQ1a. To what extent can EFL students correctly detect whether English sentences are correct or incorrect?
- RQ1b. To what extent can EFL students correctly correct the targeted errors in English sentences?
- RQ2. How accurately do EFL students calibrate their acceptance or rejection decisions to the actual correctness of the sentences?
- RQ3a. Is there a statistically significant relationship between detection performance and acceptance calibration?
- RQ3b. Is there a statistically significant relationship between correction performance and acceptance calibration?

1.5 Research Hypotheses

H0a: There is no statistically significant relationship between detection performance and acceptance calibration.

H0b: There is no statistically significant relationship between correction performance and acceptance calibration. Both hypotheses were tested at the .05 level of significance. Because two planned correlations were tested, the results are interpreted separately and no claim is made that a non-significant result proves the absence of an association.

1.6 Significance of the Study

The study contributes to research on critical evaluation in EFL education by distinguishing language-evaluation performance from the calibration of acceptance decisions. Although the task itself did not use AI-generated sentences, its structure has implications for AI-assisted language learning because learners who use AI tools must also evaluate whether generated language is accurate before relying on it. For students, the study highlights verification and informed judgment. For teachers, it offers a classroom activity in which learners identify, explain, and correct linguistic errors. For researchers, it provides a small-scale performance-based approach that can be adapted to future studies using documented AI-generated materials.

1.7 Scope of the Study

The study focused on 30 EFL students from the Department of English, Faculty of Education–Surman, Sabratha University, Libya. The available study record does not document the exact study period, recruitment procedure, or eligibility criteria beyond EFL enrollment; consequently, broader population claims are not warranted. The evaluation task consisted of 30 English sentences, including 17 correct sentences and 13 sentences containing targeted linguistic errors. The study focused mainly on sentence-level grammatical, lexical, and collocational accuracy rather than factual or discourse-level hallucinations, and examined detection, correction, acceptance calibration, overacceptance, underacceptance, and confidence.

2. Literature Review

2.1 Generative AI in English Language Education

Generative AI has expanded the range of language-learning support available to students. Learners can use AI systems for brainstorming, paraphrasing, vocabulary development, grammar explanation, feedback, translation, and academic writing support. Recent systematic reviews of empirical research in language learning and teaching report increased attention to writing, feedback, and individualized support, but also identify concerns about incorrect information, academic integrity, privacy, and the limited number of rigorous performance-based studies (Lo et al., 2024; Li et al., 2025). Empirical studies further suggest that AI-supported feedback can improve aspects of EFL writing when it is integrated with instructional guidance rather than treated as an autonomous authority (Guo et al., 2024; Han & Li, 2024; Song & Song, 2023). These findings make critical evaluation and verification central parts of responsible AI-supported language learning.

2.2 AI Literacy and Critical AI Literacy

AI literacy is broader than technical ability. Long and Magerko (2020) conceptualized it as a set of competencies that enable people to understand and critically engage with AI, while Ng et al. (2021) described it as a multidimensional construct. Reviews of AI-literacy frameworks show recurring links among digital literacy, data literacy, computational thinking, ethical reasoning, and the ability to evaluate AI outputs (Yim, 2024). In language education, ChatGPT literacy has also been framed as a situated competence involving knowledge, use, evaluation, and responsible participation (Ma et al., 2024). The present study therefore distinguishes functional AI use from critical AI literacy: detection and correction represent task-based evaluation of language, whereas acceptance calibration represents the appropriateness of reliance. Recent work in English language education similarly emphasizes conceptual understanding, strategic application, and critical–ethical evaluation (Cao & Mao, 2026; Zhang et al., 2026).

2.3 AI-Assisted Language, Errors, and Verification

Generative AI systems can produce output that is fluent yet inaccurate. Research on large-language-model hallucination describes a recurrent reliability problem: systems may generate plausible, coherent text that is not adequately supported by evidence (Huang et al., 2025). Language-specific evidence is also relevant. Alsaweed and Aljebreen (2024) found that

ChatGPT performance varied across writing-error categories, with weaker performance for unclear messages, connectors, sentence structure, and non-idiomatic expressions. The present study narrows this broader reliability problem to sentence-level grammatical, lexical, and collocational evaluation in EFL learning. Students should therefore compare AI output with linguistic knowledge, course materials, dictionaries, reference grammars, teacher feedback, or other reliable sources.

2.4 Error Detection and Error Correction

Error detection and correction are related but distinct processes. Detection requires a judgment about whether an output is acceptable, whereas correction requires the learner to generate an appropriate alternative. A student may detect an error but be unable to formulate a fully correct correction. For this reason, the present study measures both abilities separately. Successful correction demonstrates a deeper level of engagement with the language form than merely recognizing that a sentence sounds wrong.

2.5 Acceptance Calibration

Trust in automation provides a useful theoretical background for understanding reliance decisions. Lee and See (2004) argued that the key issue is not simply whether people trust automation, but whether they rely on it appropriately. Their framework distinguishes appropriate reliance from misuse, in which people rely on automation when they should not, and disuse, in which people fail to rely on automation when reliance is appropriate. In the present study, however, no AI system was presented to participants. Accordingly, the study operationalizes acceptance calibration rather than direct trust in AI: acceptance of an incorrect sentence is classified as overacceptance, whereas rejection of a correct sentence is classified as underacceptance. Glikson and Woolley (2020) similarly emphasize that trust in AI is shaped by characteristics of the system, the task, and the interaction context; these factors are relevant to future studies using actual AI-generated materials.

2.6 Recent Research and Research Gap

Recent research has begun to examine acceptance calibration and verification directly in L2 writing. Li et al. (2026), for example, reported that clearer classroom norms, learning-oriented accountability, and teacher positioning of AI as contingent input were associated with more frequent verification and deeper revision. Other work has emphasized critical and ethical AI literacy among language learners and pre-service English teachers. These studies suggest that AI instruction is most useful when students are taught to question, evaluate, verify, and negotiate AI output rather than merely generate text with AI.

The reviewed literature demonstrates substantial interest in GenAI use, AI literacy, academic writing, feedback, and critical evaluation. However, performance-based evidence that simultaneously examines EFL learners' error detection, error correction, and calibration of acceptance decisions remains limited. The present study addresses this narrower gap by examining the relationship between language-evaluation performance and decision calibration, while treating the findings as relevant to—but not direct measures of—AI literacy.

2.7 Conceptual Framework

The framework is descriptive and also specifies the two associations tested in the study. The 30 English sentences provide the input. Students make a linguistic judgment (detection), produce a repair when needed (correction), and make an accept/reject decision. Acceptance calibration is defined as correspondence between the decision and the item's observed correctness. Acceptance of an incorrect item is classified as overacceptance, whereas rejection of a correct item is classified as underacceptance. The framework does not claim that detection causes calibration or that correction causes calibration.

EFL Students' Error Detection, Correction, and Acceptance Calibration: A Performance-Based Study at the Faculty of Education–Surman

Figure 1. Descriptive model of task performance and acceptance calibration

English sentence	Student response	Derived indicator
Correct / incorrect sentence	Correctness judgment	Detection score
Incorrect sentence	Open-ended repair	Correction score
All 30 items	Accept or reject without checking	Reliance decision
Incorrect + accept	Decision conflicts with item status	Overacceptance
Correct + reject	Decision conflicts with item status	Underacceptance
Correct + accept or incorrect + reject	Decision corresponds to item status	Calibration score

3. Methodology

3.1 Research Design

The study employed a quantitative descriptive-correlational design. The descriptive component summarized performance on the 30-item task and described acceptance-calibration patterns; it does not estimate the proportion of all EFL students who would perform similarly. The correlational component examined whether the two measured performance scores were monotonically associated with acceptance calibration. A non-probability sample was used; the source file does not document whether recruitment was convenience- or purposive-based. Correlation was examined because students who more accurately evaluate language might also be expected to make more accurate acceptance decisions, although the study design cannot establish causation.

3.2 Participants

The dataset consisted of 30 completed responses from EFL students enrolled in the Department of English, Faculty of Education–Surman, Sabratha University, Libya. The source material identifies the participants as students with varied years of study and prior AI experience but does not document the recruitment dates, exact inclusion criteria, or whether all participants were adults. The original year-of-study categories, including “Other,” are retained in Table 1 because the source dataset does not define that category. Age was not used in the revised report because the source note requested that the mean age and obtained data be omitted.

Table 1. Participant Characteristics

Variable	Category	n	%
Year of study	Second year	9	30.0
Year of study	Other	8	26.7
Year of study	Fourth year	6	20.0
Year of study	Third year	4	13.3
Year of study	First year	3	10.0
AI use for English	Never	9	30.0
AI use for English	Rarely	7	23.3
AI use for English	Sometimes	6	20.0
AI use for English	Often	5	16.7
AI use for English	Very often	3	10.0
AI experience	Basic	9	30.0
AI experience	Intermediate	7	23.3
AI experience	No experience	5	16.7
AI experience	Beginner	5	16.7
AI experience	Advanced	4	13.3

3.3 Research Instrument and Procedure

The main instrument was an English Sentence Evaluation Task rather than a conventional attitude questionnaire. It contained 30 English sentences: 17 correct sentences and 13 sentences containing targeted errors. For each sentence, participants indicated whether it was correct or incorrect, supplied a correction when necessary, stated whether they would accept it without further checking, and rated their confidence on a five-point scale. The task items were

EFL Students' Error Detection, Correction, and Acceptance Calibration: A Performance-Based Study at the Faculty of Education–Surman

researcher-prepared; they were not generated by an AI system. The source record does not document formal piloting or independent expert validation of the item bank. Participants were instructed not to use search engines, ChatGPT, Grammarly, dictionaries, or other external resources. Because administration-monitoring procedures were not recorded, compliance cannot be independently verified.

3.4 Item Construction

The 17/13 distribution provided a majority of correct items while retaining enough erroneous items for correction analysis; it was a practical design choice rather than a validated difficulty balance. The erroneous items targeted subject–verb agreement, article use, prepositions, verb complementation, tense, noun-phrase structure, and lexical/collocational choice. Thus, the task measures sentence-level language evaluation broadly, not grammatical accuracy alone.

Table 3. Error Types Targeted by the Erroneous Items

Items	Targeted feature
2, 20, 30	Subject–verb agreement
6, 22	Article use
8, 18	Preposition or complementation
14	Verb complementation
24	Modal/verb construction
26	Tense
28	Lexical/collocational choice

Table 2. Item Key and Detection Accuracy

Item	Actual status	Correct judgment (%)
1	Correct	93.3
2	Incorrect	90.0
3	Correct	96.7
4	Incorrect	93.3
5	Correct	100.0
6	Incorrect	93.3
7	Correct	100.0
8	Incorrect	86.7
9	Correct	90.0
10	Correct	93.3
11	Correct	93.3
12	Correct	80.0
13	Correct	96.7
14	Incorrect	96.7
15	Correct	96.7
16	Incorrect	76.7
17	Correct	90.0
18	Incorrect	86.7
19	Correct	80.0
20	Incorrect	93.3
21	Correct	93.3
22	Incorrect	96.7
23	Correct	93.3
24	Incorrect	90.0
25	Correct	96.7
26	Incorrect	100.0
27	Correct	86.7
28	Incorrect	96.7
29	Correct	100.0
30	Incorrect	86.7

3.5 Scoring and Statistical Analysis

The detection score was calculated as correct judgments divided by 30×100 . The correction score was calculated from the 13 erroneous items as acceptable corrections divided by 13×100 . A correction received one point when it repaired the target error, preserved the intended meaning, and did not introduce a new error. Stylistically different answers were accepted when they were grammatical and meaning-equivalent; partially correct, ambiguous, missing, or meaning-changing answers received zero. The source manuscript does not identify a second coder or report inter-rater agreement, so this is recorded as a coding limitation rather than inferred. Acceptance calibration was calculated from all 30 equally weighted accept/reject decisions. Acceptance of a correct item and rejection of an incorrect item were both treated as calibrated decisions because calibration concerns correspondence between the decision and observed item correctness; this is not a claim about internal psychological trust. Acceptance of an incorrect item was classified as overacceptance, and rejection of a correct item was classified as underacceptance.

AI output	Participant decision	Classification
Correct	Accept	Calibrated acceptance
Correct	Reject	Underacceptance
Incorrect	Accept	Overacceptance
Incorrect	Reject	Calibrated skepticism

Means and standard deviations were calculated for detection, correction, acceptance calibration, overacceptance, underacceptance, and confidence. The Shapiro–Wilk test was used as a small-sample diagnostic; detection scores did not significantly depart from normality ($W = 0.942$, $p = 0.104$), whereas correction scores ($W = 0.859$, $p < .001$) and acceptance-calibration scores ($W = 0.884$, $p = .003$) did. The source file does not report Q–Q-plot inspection or outlier diagnostics. Spearman's rank-order correlation was selected because the scores are bounded percentages, the sample is small, and the planned question concerns monotonic association rather than a strictly linear relationship. Detection–calibration and correction–calibration associations were examined separately. The two tests were reported separately at $\alpha = .05$; no multiplicity adjustment was applied, and this decision is stated explicitly.

3.6 Ethical Considerations and Limitations

Participation was voluntary, and the analysis used response codes rather than student names. Results are reported in aggregate form. Ethics approval and informed-consent procedures should be reported here exactly as documented by the institution before submission (e.g., committee name, approval/reference number, and consent procedure, if applicable); these details are not available in the present study record and must not be fabricated. The sample was small and non-probability-based, limiting precision and transferability beyond the participating program. The task focused on sentence-level language evaluation, and compliance with the no-external-resources instruction could not be independently verified. Acceptance calibration was operationalized through accept/reject decisions, and no second correction coder or inter-rater agreement was reported. The cross-sectional design does not establish causal effects.

4. Results

4.1 Participant Profile

The 30 participants represented multiple years of study, with the largest group being second-year students ($n = 9$), followed by students classified as Other ($n = 8$) and fourth-year students ($n = 6$). Reported AI-use frequency ranged from never to very often.

4.2 Descriptive Statistics

Table 3. Descriptive Statistics of Main Study Variables

Variable	Mean	SD	Scale
Detection	92.22	4.23	%
Correction	91.28	8.01	%
Acceptance calibration	75.33	8.51	%
Overacceptance	17.18	10.24	%
Underacceptance	30.39	13.56	%
Mean confidence	4.57	0.15	/5

The sample mean was 92.22% for detection and 91.28% for correction, while acceptance calibration had a lower sample mean of 75.33%. These descriptive differences were not tested with a direct inferential comparison and should not be interpreted as statistically significant. Confidence had a mean of 4.57 out of 5.

4.3 Results for Research Question One

RQ1a and RQ1b concerned performance on the study task. The sample mean detection score was 92.22% (SD = 4.23), and the sample mean correction score was 91.28% (SD = 8.01). Item-level detection accuracy ranged from 76.7% to 100.0%. Item 16, which targeted the collocational expression “reached to an agreement,” had the lowest observed detection rate (76.7%). This item reflects lexical/collocational evaluation as well as grammar, so it should not be used to infer that the collocation itself caused lower performance; wording, familiarity, item difficulty, and chance may also have contributed. Several items were identified with 100% accuracy, including Items 5, 7, 26, and 29.

Table 4. Selected Item-Level Detection Results

Item	Actual status	Detection accuracy (%)
16	Incorrect	76.7
12	Correct	80.0
19	Correct	80.0
5	Correct	100.0
7	Correct	100.0
26	Incorrect	100.0
29	Correct	100.0

4.4 Results for Research Question Two

The second research question concerned acceptance calibration. The mean calibration score was 75.33% (SD = 8.51). The mean overacceptance score was 17.18% (SD = 10.24), indicating that participants sometimes accepted erroneous sentences. The mean underacceptance score was 30.39% (SD = 13.56), indicating that participants also sometimes rejected correct sentences. The mean confidence rating was 4.57/5 (SD = 0.15).

4.5 Results for Research Question Three

The third research question examined whether error-detection and correction performance was significantly related to acceptance calibration. Spearman's rho was used because correction and acceptance-calibration scores were non-normally distributed.

Table 5. Spearman Correlations Between Performance and Acceptance Calibration

Variables	Spearman's ρ	p	Decision at $\alpha = .05$
Detection–acceptance calibration	0.056	0.770	Not significant
Correction–acceptance calibration	0.003	0.987	Not significant

The relationship between detection and acceptance calibration was near zero and not statistically significant ($\rho = 0.056$, $p = 0.770$). The relationship between correction and acceptance calibration was also near zero and not statistically significant ($\rho = 0.003$, $p = 0.987$). Given $n =$

30, the study has limited power to detect small associations, and confidence intervals for rho were not available in the source output. There was insufficient evidence to reject H_{0a} or H_{0b} at $\alpha = .05$; neither null hypothesis was proven.

5. Discussion

5.1 Error Detection and Correction

The observed sample means were 92.22% for detection and 91.28% for correction on the present 30-item instrument. Without an external proficiency benchmark, these figures should be described as task performance rather than general linguistic ability or definitively “high” ability. The separate scores provide performance evidence that complements, but does not replace, self-report measures of AI literacy. Detection and correction also warrant separate treatment because recognizing a problematic sentence and producing an acceptable repair are different task demands.

5.2 Acceptance Calibration

The mean acceptance-calibration score of 75.33% was lower than the detection and correction scores. This difference suggests that the ability to evaluate language accurately does not necessarily translate into perfectly calibrated acceptance decisions. Both forms of miscalibration were present. Overacceptance averaged 17.18%, showing that participants sometimes accepted incorrect sentences, while underacceptance averaged 30.39%, showing that they sometimes rejected correct sentences. This pattern is consistent with the broader concept of appropriate reliance proposed by Lee and See (2004), while the present study itself does not measure direct trust in an AI system.

The result can be interpreted in relation to Li et al. (2026), who examined GenAI feedback and verification in L2 writing, and to Lee and See's (2004) distinction between appropriate reliance, misuse, and disuse. In the present study, however, the task did not involve an AI system; therefore, the terms overacceptance and underacceptance are not used as direct measures of trust in AI. Instead, overacceptance refers to accepting an incorrect sentence and underacceptance refers to rejecting a correct sentence. This distinction is important because the present findings provide evidence about decision calibration, not direct evidence of trust in an AI system. English proficiency was not measured here, so its possible role in calibration remains a question for future research.

5.3 Relationship Between Performance and Acceptance Calibration

In this sample and on this instrument, no statistically significant association was detected between either measured performance score and calibration: detection–calibration yielded $p = 0.056$, $p = 0.770$, and correction–calibration yielded $p = 0.003$, $p = 0.987$. The cross-sectional design and small non-probability sample do not establish a causal or general relationship. The pattern is compatible with a multidimensional account of AI literacy in which language-evaluation performance and critical reliance may require separate instructional attention, but this interpretation requires testing in larger and more diverse samples.

5.4 Confidence and Judgment

Participants reported a mean confidence of 4.57/5, but the available summary does not provide a confidence–accuracy cross-tabulation or correlation. Therefore, the data cannot establish whether high confidence co-occurred with errors. Confidence is not equivalent to accuracy and should not, by itself, be treated as evidence of reliable AI judgment. Future analyses should examine confidence–accuracy correspondence directly.

6. Conclusions and Implications

6.1 Conclusion

The findings suggest that participants performed well on the selected detection and correction task, while acceptance calibration was less complete and included both overacceptance and underacceptance. The two planned correlations were not statistically significant, although the small sample limits the precision and power of these estimates. Within the limits of one institution, one non-probability sample, and one 30-item instrument, the findings suggest that task-based language evaluation and calibrated acceptance decisions should be examined as related but distinct aspects of critical AI literacy. Future studies should replicate the task with documented item-generation procedures, independent coding, larger samples, and direct measures of confidence–accuracy correspondence.

6.2 Pedagogical Implications

EFL courses can integrate critical AI literacy by using sentence-evaluation activities such as the present task and then extending them to authentic AI-assisted language. Teachers can ask learners to distinguish fluent language from accurate language, explain why they accept or reject a sentence, and verify AI-generated claims or corrections against reliable sources. AI output should be presented as assistance rather than unquestionable authority, and short AI-verification logs can help develop students' metacognitive awareness.

6.3 Recommendations for Further Research and Practice

EFL programs should include explicit instruction on evaluating language and, where AI tools are used, on verifying AI-generated output. Academic writing courses can include activities that require students to detect, explain, and correct AI errors and to justify acceptance or rejection decisions. Future studies should use larger and more diverse samples, include lexical, pragmatic, discourse-level, and factual AI errors, and directly compare performance on researcher-prepared and AI-generated materials. Experimental designs could also examine whether explicit verification training improves decision calibration.

References

- Alsaweed, W., & Aljebreen, S. (2024). Investigating the accuracy of ChatGPT as a writing error correction tool. *International Journal of Computer-Assisted Language Learning and Teaching*, 14(1). <https://doi.org/10.4018/IJCALLT.364847>
- Guo, K., Pan, M., Li, Y., & Lai, C. (2024). Effects of an AI-supported approach to peer feedback on university EFL students' feedback quality and writing ability. *The Internet and Higher Education*, 63, 100962. <https://doi.org/10.1016/j.iheduc.2024.100962>
- Han, J., & Li, M. (2024). Exploring ChatGPT-supported teacher feedback in the EFL context. *System*, 126, 103502. <https://doi.org/10.1016/j.system.2024.103502>
- Huang, L., Yu, W., Ma, W., Zhong, Z., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), Article 42, 1–55. <https://doi.org/10.1145/3703155>
- Li, B., Tan, Y. L., Wang, C., & Lowell, V. (2025). Two years of innovation: A systematic review of empirical generative AI research in language learning and teaching. *Computers and Education: Artificial Intelligence*, 9, 100445. <https://doi.org/10.1016/j.caeai.2025.100445>
- Lo, C. K., Yu, P. L. H., Xu, S., Ng, D. T. K., & Jong, M. S. Y. (2024). Exploring the application of ChatGPT in ESL/EFL education and related research issues: A systematic review of empirical studies. *Smart Learning Environments*, 11, Article 50. <https://doi.org/10.1186/s40561-024-00342-5>

- Ma, Q., Crosthwaite, P., Sun, D., & Zou, D. (2024). Exploring ChatGPT literacy in language education: A global perspective and comprehensive approach. *Computers and Education: Artificial Intelligence*, 7, 100278. <https://doi.org/10.1016/j.caeai.2024.100278>
- Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology*, 14, 1260843. <https://doi.org/10.3389/fpsyg.2023.1260843>
- Yim, I. H. Y. (2024). A critical review of teaching and learning artificial intelligence (AI) literacy: Developing an intelligence-based AI literacy framework for primary school education. *Computers and Education: Artificial Intelligence*, 7, 100319. <https://doi.org/10.1016/j.caeai.2024.100319>
- Cao, Z., & Mao, Z. (2026). Exploring EFL undergraduate students' critical AI literacy in academic writing. *Journal of English for Academic Purposes*, 82, 101733. <https://doi.org/10.1016/j.jeap.2026.101733>
- Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2), 627–660. <https://doi.org/10.5465/annals.2018.0057>
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
- Li, R., Zhang, C., & Ye, H. (2026). When does GenAI feedback support learning? Acceptance calibration and verification in L2 writing assessment. *Assessing Writing*, 69, 101087. <https://doi.org/10.1016/j.asw.2026.101087>
- Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16. <https://doi.org/10.1145/3313831.3376727>
- Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2, 100041. <https://doi.org/10.1016/j.caeai.2021.100041>
- Zhang, Y., Qiu, X., & Wang, J. (2026). A systematic research synthesis of generative AI literacy in English language education. *Discover Artificial Intelligence*, 6, 1074. <https://doi.org/10.1007/s44163-026-02178-z>