International Journal of Language and Literary Studies
Volume 8, Issue 6, 2026
Homepage : http://ijlls.org/index.php/ijlls

Artificial Intelligence in English Language Teaching: Applications, Benefits, and Challenges — A Critical Integrative Review

Samia Mahraz[1] & Abdelhamid Nfissi[2]
[1]Ph.D. Candidate, Afro-Mediterranean Research Laboratory of Discourse, Communication, and Cultural Studies, Faculty of Arts and Human Sciences, Saïs–Fès, Sidi Mohamed Ben Abdellah University, Morocco
[2]Senior Professor, Afro-Mediterranean Research Laboratory of Discourse, Communication, and Cultural Studies, Faculty of Arts and Human Sciences, Saïs–Fès, Sidi Mohamed Ben Abdellah University, Morocco
Email: abdelhamid.nfissi@usmba.ac.ma

Corresponding Author Email: samia.mahraz@usmba.ac.ma

DOI: https://doi.org/10.36892/ijlls.v8i6.2873

APA Citation: Mahraz, S., & Nfissi, A. (2026). Artificial Intelligence in English Language Teaching: Applications, Benefits, and Challenges — A Critical Integrative Review. *International Journal of Language and Literary Studies*, *8*(6).442-453. https://doi.org/10.36892/ijlls.v8i6.2873

Received: *10/09/2026*

Accepted: *12/11/2026*

Keywords: Artificial Intelligence, English Language Teaching, Integrative Review, Pedagogical Applications, Language Learning

Abstract

Artificial intelligence (AI) offers opportunities for feedback, spoken interaction, reading support, and listening practice in English language teaching (ELT). This critical integrative review examines selected scholarship to distinguish AI capabilities, learner perceptions, task performance, and evidence of independent language development. Ten purposively chosen illustrative publications were categorized by their primary language-skill focus: six writing (60%), two speaking (20%), one reading (10%), and one listening (10%). Writing or speaking therefore accounted for eight of these ten examples (80%). These proportions describe only this selected illustrative set, not the prevalence of skills in ELT research as a whole. The literature highlights possibilities for feedback, rehearsal, and task-specific support alongside concerns about accuracy, overreliance, academic integrity, privacy, and unequal access. An evidence-to-claim framework is proposed to connect AI function, instructional task, measured outcome, and limitations of inference. The review does not claim a systematic database search or a PRISMA-screened study sample. Future systematic studies should use archived searches, explicit eligibility decisions, and quality appraisal to test the broader distribution and effectiveness of AI applications in ELT.

1. Introduction

1.1 Background and context

Artificial intelligence (AI) is increasingly used in English language teaching (ELT), but the label covers instructional systems with different functions and evidentiary implications. Intelligent tutoring platforms adapt exercises to learner responses; automated writing evaluation (AWE) systems identify and comment on written language; spoken-dialogue and speech-recognition systems support oral rehearsal; and generative models can compose text, explain language choices, and simulate conversation. These technologies should not be regarded as interchangeable. Their educational value depends on the language objective, the role assigned to the learner, the quality of feedback, and the outcome used to judge success (Klimova et al., 2023; Zhu & Wang, 2025).

Since conversational generative AI became widely accessible in late 2022, the literature has grown across writing, speaking, teacher support, and language-learning practice. At the same time, an increase in applications does not establish equivalent improvement in language proficiency. A learner may find a chatbot engaging without acquiring greater conversational competence; an automated system may generate detailed comments without improving the writer's subsequent independent performance. Differentiating instructional possibilities, observed behavior, learner perceptions, and tested attainment is therefore essential for teachers and institutional decision-makers (Li et al., 2024; Steiss et al., 2024).

Research has also expanded beyond university classrooms. For example, lower-secondary learners have reported sustained engagement with spoken-dialogue agents, while studies in higher education have examined AI-supported peer feedback and course-specific listening-decoding instruction (Ericsson & Johansson, 2023; Guo, Pan, et al., 2024; Huang et al., 2024). Context matters because access to devices, learner proficiency, teacher mediation, task duration, and assessment procedures vary. The present article accordingly treats learning outcomes as situated findings, not as properties that transfer automatically from one AI tool to another.

1.2 Problem statement and contribution

Existing syntheses have examined the development of AI in English language teaching (ELT), including broad AI-in-ELT surveys, chatbot-specific reviews, and research published during the early expansion of generative AI (Sharadgah & Sa'di, 2022; Du & Daniel, 2024; Li et al., 2024; Zhu & Wang, 2025). More recent publications consider EFL/ESL instruction and school settings (Uygun & Demir, 2026; Imran et al., 2026). Despite these contributions, teachers still need to know whether a reported benefit reflects a tool's technical capability, a learner's perception, a task-specific performance measure, or sustained learning that transfers to unaided settings. These distinctions matter for comparing studies across tools, language skills, and populations.

This critical integrative review contributes an evidence-to-claim comparison linking AI technology, pedagogical task, target English skill, study design, measured outcome, and implementation risks. Unlike a catalogue of technologies, it asks what each selected source can legitimately establish and where transfer to classroom practice requires additional evidence. The proposed framework is an interpretive aid and has not been empirically validated as a classroom assessment instrument.

The present synthesis draws on selected illustrative publications and relevant prior reviews. It does not report a reproducibly retrieved, comprehensively screened sample from Scopus, Web of Science, and ERIC. Accordingly, the numerical distribution in this paper is restricted to ten purposively chosen cases and must not be generalized to the wider literature.

1.3 Aim and research questions

Five research questions guide this critical integrative review:

RQ1. Which AI technologies and pedagogical applications are represented in the selected sources?
RQ2. Which English language skills and tasks do these publications address?
RQ3. What benefits are reported, and which measurements support them?
RQ4. What pedagogical, ethical, and implementation challenges are documented?
RQ5. Which evidence limitations and future research priorities are suggested by the selected literature?

2. Literature Review and Analytical Orientation

2.1 Technologies and pedagogical roles

AI-supported language learning builds on earlier computer-assisted language learning, but contemporary systems differ in how they produce feedback and respond to learner input. Rule-governed tutoring commonly provides bounded exercises, while adaptive systems change task

difficulty in response to performance. AWE can identify grammatical and rhetorical features of student texts. Automatic speech recognition turns oral production into text or scores, and spoken-dialogue agents provide opportunities to rehearse interaction. Machine translation and generative models can support comprehension, composition, or language comparison, but each also raises questions about learner ownership of language choices (Klimova et al., 2023; Zhu & Wang, 2025).

The pedagogical unit of analysis should be the task rather than the commercial tool name. The same conversational model might provide grammar explanations during revision, simulate a job interview, or answer reading questions. These activities involve different language skills and should be evaluated with different instruments. A writing rubric may assess coherence or accuracy, an oral task may assess interactional competence, and a vocabulary test may assess retention. Treating every positive outcome as "improved English" conceals these distinctions (Zhai & Wibowo, 2023; Chen, 2026).

Table 1. ***Conceptual map of AI technologies and evidence requirements***

Technology family	Illustrative teaching function	Target outcome and caution
Conversational agents / chatbots	Dialogues, rehearsal and role-play	Distinguish willingness to communicate from tested speaking proficiency.
Intelligent tutoring and adaptive systems	Sequence reading, grammar or lexical practice	Require aligned pre/post and delayed assessments.
Automated writing evaluation (AWE)	Error feedback and revision prompts	Measure feedback uptake and unaided writing, not comment volume alone.
Speech recognition and oral feedback	Transcripts and pronunciation indicators	Compare automated scores with human intelligibility ratings.
Machine translation	Compare lexical and syntactic options	Evaluate independent understanding and appropriate attribution.
Generative language models	Drafting support, explanations and simulated interaction	Specify what the learner authored and what the model generated.

2.2 Outcome constructs and inferential boundaries

Language-learning studies can report at least four substantively different outcome types. First, an affordance claim describes what a tool can provide, such as immediate response or repeated practice. Second, a perception or affective claim concerns how learners experience a tool, including confidence, motivation, anxiety, or engagement. Third, a task-performance claim reports observable performance on a specific instrument under defined conditions. Fourth, a learning or transfer claim requires evidence that a learner performs better independently or after a delay. These levels are not a hierarchy in which perceptions are unimportant; rather, each answers a different question.

For example, Steiss et al. (2024) compared the quality of human and AI-generated feedback on student essays. Their results inform the quality of comments teachers may receive, not the later writing proficiency of the students. Zhang et al. (2024) studied speaking enjoyment, anxiety, and willingness to communicate; positive changes in these indicators cannot be re-described as measured gains in pronunciation or conversational accuracy. By contrast, a structured listening-decoding assessment can speak to its targeted receptive process, although its generalizability is limited by the course design and population (Huang et al., 2024).

2.3 Relationship to previous reviews

Previous reviews supply useful numerical context, but their totals belong to their own search protocols. For example, Zhu and Wang (2025) analyzed 125 articles and reported writing as a topic in 52 papers (41.6% of 125) and speaking in 30 (24.0% of 125). Because publications may address more than one skill, these percentages cannot simply be added to estimate the distinct share studying writing OR speaking. Uygun and Demir (2026) analyzed 55 studies in a separately defined EFL/ESL corpus. None of those totals is adopted as the present review's included-study count. Table 2 compares prior reviews on their own terms.

Table 2. ***Existing review contributions and the present article's distinct task***

Prior review	Principal focus	Distinct question addressed here
Sharadgah & Sa'di (2022)	AI in ELT, 2015–2021	What can the reported indicator actually justify for an instructor?
Du & Daniel (2024)	Chatbots in EFL speaking practice	Do affective outcomes support or differ from proficiency claims?
Li et al. (2024)	Initial ChatGPT-language education literature	How do outcomes from different tasks and designs compare?
Zhu & Wang (2025)	Broader language-education AI trends	What must be checked before translating trends into classroom claims?
Uygun & Demir (2026)	EFL/ESL AI education, 2023–2025	Which design and measurement limits restrict practical inferences?
Imran et al. (2026)	School language-education settings	How do learner age and setting affect transferability?

The contribution is not a new meta-analysis, comprehensive map of the field, or validated assessment scale. Instead, the evidence-to-claim framework asks readers to check which claims are warranted by the designs and measures of the selected publications. The illustrative source set is heterogeneous and cannot establish comparative effectiveness or field-wide frequencies.

3. Methodology

3.1 Review design and scope

This article uses a critical integrative review design with narrative and thematic comparison of selected empirical publications and relevant prior reviews. Ten previously cited illustrative publications were retained to contrast AI functions, educational settings, targeted English skills, and types of evidence. They were selected purposively, not through a complete reproducible search-and-screening process. Previous systematic reviews inform context but are not counted among the ten focal publications. The analysis therefore offers interpretive comparisons rather than a systematic synthesis of all eligible ELT studies.

No archived database exports, duplicate-removal register, full-text screening log, or completed study-level quality appraisal is available for this source set. Consequently, no PRISMA flow, comprehensive included-study count, or field-wide prevalence estimate is claimed. This reporting boundary follows the distinction between a critical integrative review and a completed systematic review.

3.2 Source identification and selection

The focal publications were selected from sources cited in the prior manuscript because each provided a recognizable AI-supported task, language-skill application, or relevant evaluation, and the set allowed comparison across writing, speaking, reading, and listening. Several additional reviews were consulted to situate the discussion. This was a purposive selection, not a comprehensive search conducted in Scopus, Web of Science, and ERIC. The review makes no claim to have executed the proposed database queries or applied a verified publication-year filter.

For transparency, the ten focal publications are listed individually in Table 3 with their principal skill classifications and context qualifications. They include studies from EFL settings alongside contextual evidence that cannot be treated as directly equivalent to EFL learning outcomes. This selection is appropriate for illustrative comparison but not for generalizing proportions to the overall ELT literature.

3.3 Interpretive selection and credibility limits

For interpretive comparison, sources were prioritized when they stated an identifiable AI function, educational task, and observed outcome or evaluative question. Previous review articles were used as secondary context rather than counted as independent empirical studies.

The set is not exhaustive; no claim is made that all potentially eligible publications were identified or that two reviewers independently screened records.

Each focal example was considered in relation to its participants and setting, instructional task, research design, outcome measure, and the limits of its conclusions. Comparisons of feedback quality are distinguished from evidence of learning; perception measures are distinguished from language-proficiency measures; and exploratory system designs are not treated as proof of durable educational effects. No standardized risk-of-bias score was calculated.

3.4 Illustrative coding and synthesis

For each of the ten focal publications, a single primary language skill was assigned for a transparent descriptive illustration. The classifications were writing, speaking, reading, or listening. Each publication was counted once, yielding six writing, two speaking, one reading, and one listening examples. The classification concerns the focal task rather than every possible secondary skill addressed by a study. The analysis then compared observed indicators, design constraints, and educational implications across cases.

The calculation is 6/10 = 60% writing, 2/10 = 20% speaking, 1/10 = 10% reading, and 1/10 = 10% listening. Writing or speaking accounts for (6 + 2)/10 = 80% of this illustrative set. Because cases were chosen purposively, these figures do not estimate the proportion of writing- or speaking-related research across all AI-in-ELT publications. The ten examples also include contextual evidence beyond directly comparable EFL intervention studies.

3.5 Boundaries of the review

A systematic review would require dated searches in specified databases, archived exports, transparent deduplication and eligibility screening, a documented included-study dataset, and formal quality appraisal before reporting PRISMA counts or prevalence estimates. Those procedures were not completed for this manuscript. The present review therefore retains the narrower integrative designation, and all numerical claims are restricted to the ten illustrative cases.

4. Results

Selected Illustrative Evidence

4.1 Characteristics of the ten illustrative publications

Table 3 describes ten purposively selected publications that illustrate different instructional functions, language skills, and kinds of evidence. Six were categorized as primarily writing-related, two speaking-related, one reading-related, and one listening-related. These counts concern only the listed examples; they are not the included-study total from a systematic search. Several cases have explicit transferability limitations, as noted in the table.

Table 3. ***Descriptive coding of ten purposively selected illustrative publications***

Publication	Primary skill	Population / study focus	Context / evidence caveat
Guo, Li, et al. (2024)	Writing	Qualitative chatbot-assisted argumentative writing; 5 Chinese university EFL learners	Direct EFL
Guo, Pan, et al. (2024)	Writing	AI-supported peer feedback and writing outcomes; 124 university EFL students	Direct EFL
Yang et al. (2024)	Writing	AWE feedback uptake; 5 EFL university students	Direct EFL
Steiss et al. (2024)	Writing	Human vs. AI feedback on secondary-school essays; NOT an EFL intervention	Contextual
Teng (2024)	Writing	Writing-feedback perceptions; 45 EFL learners in Macau	Direct EFL
Kim et al. (2025)	Writing	LLM-supported writing assessment in Korean school context	Direct EFL

Publication	Primary skill	Population / study focus	Context / evidence caveat
Zhang et al. (2024)	Speaking	AI speaking assistant, enjoyment, anxiety and willingness to communicate; 131 EFL students	Direct EFL
Ericsson & Johansson (2023)	Speaking	AI-assisted spoken practice; Swedish lower-secondary English learners	English L2/FL
Huang et al. (2024)	Listening	Online EFL listening-decoding activities	Direct EFL
Wang et al. (2024)	Reading	ChatPRCS reading-support system and evaluation; transferability requires appraisal	English-focused system

Note. Descriptive result: Six writing + two speaking + one reading + one listening = ten illustrative publications. Writing or speaking: (6 + 2) / 10 × 100 = 80%. These figures should not be confused with the proportion of skill-specific research in the wider ELT literature.

4.2 RQ1: AI technologies and pedagogical applications

The selected publications illustrate several overlapping AI functions: conversational agents support rehearsal, automated writing evaluation (AWE) supplies writing feedback, task-specific systems address receptive-skill objectives, and generative models can serve multiple roles. These functions are discussed by instructional task rather than brand name. No ranking or prevalence estimate across the wider AI-in-ELT field can be inferred from the selected cases (Du & Daniel, 2024; Zhu & Wang, 2025).

For example, Yang et al. (2024) examined interaction with AWE feedback, Guo, Pan, et al. (2024) studied AI-supported peer feedback, and Kim et al. (2025) addressed automated and human writing assessment. All concern writing but differ in design and measured outcomes; they should not be combined as equivalent evidence of proficiency improvement.

4.3 RQ2: Published examples across language skills

4.3.1 Writing

Existing writing research shows why coding must distinguish task performance, revision behavior, and learner perceptions. Guo, Pan, et al. (2024) reported improved peer-feedback quality and writing ability in a defined three-week EFL intervention. Guo, Li, et al. (2024) investigated five EFL undergraduates' chatbot-assisted argumentation using qualitative methods, while Yang et al. (2024) studied interaction with AWE feedback. These are research examples; their presence in this discussion does not establish their final eligibility or a field-wide effect.

Steiss et al. (2024) found that trained human raters gave stronger feedback on four of five evaluated dimensions than ChatGPT. This result qualifies any assumption that rapid generation automatically means higher feedback quality. Moreover, their comparison examined feedback rather than subsequent independent writing. Teng (2024) reported favorable EFL learner perceptions of ChatGPT-assisted writing feedback, but such perceptions should be reported separately from attainment. For a stronger learning claim, future work should assess unassisted new writing and delayed transfer tasks.

4.3.2 Speaking and interaction

Conversational AI can lower the cost of repeated oral rehearsal and provide opportunities unavailable in large classes. Ericsson and Johansson (2023) reported that lower-secondary learners sustained speaking practice with conversational agents and reflected on their educational experiences. Zhang et al. (2024) observed gains in enjoyment and willingness to communicate, alongside lower foreign-language anxiety, during a six-week university EFL intervention. These are meaningful outcomes for participation, but they do not by themselves measure grammatical accuracy, pronunciation intelligibility, or interactional competence.

Zhai and Wibowo (2023) synthesize research on dialogue systems and interactional competence, including issues of negotiation, conversational context, and cultural appropriateness. Their findings suggest that simulated conversation should be treated as

preparation for human interaction rather than as its complete substitute. An appropriate next assessment would compare independent oral performances after tool-assisted practice with unassisted peer or teacher interaction using clearly defined rubrics.

4.3.3 Vocabulary and grammar

AI can generate contextual vocabulary examples and adaptive practice tasks, but a model's ability to produce examples is not evidence that words will be retained. Chen (2026), in a systematic review of 25 higher-education EFL vocabulary studies, reported that the reviewed evidence emphasized vocabulary breadth more than depth, delayed retention, or transfer. That count belongs to Chen's publication and is not an included-study total for this article. Classroom assessment should differentiate immediate recognition, productive use, and later recall.

Grammar corrections embedded in writing feedback require similar caution. A student may accept an automatically corrected article or verb form while remaining unable to apply the rule in a new sentence. The focal writing studies help identify feedback processes but do not support the former manuscript's unsupported count that only two studies worldwide tested grammar gains. More persuasive grammar research would measure independent production on new items and use delayed assessments (Yang et al., 2024).

4.3.4 Reading and listening

AI reading applications may offer tailored questions and explanatory scaffolds. Wang et al. (2024) describe ChatPRCS, a personalized support system for English reading comprehension. Such work helps clarify a design approach, but its technology alone should not be represented as a measured gain across all reading tasks or learner populations. Future evaluation should separate comprehension achieved with hints from unassisted inference and transfer.

Huang et al. (2024) studied chatbot-supported listening-decoding in an online EFL course and reported results from a defined teaching comparison. The study provides evidence about a particular receptive-skill target rather than a broad claim that listening receives the least attention or is inherently less suited to AI. Different listening objectives—segmental decoding, inferencing, and comprehension of unfamiliar speech require distinct instruments and comparisons.

4.3.5 Pronunciation

Speech-recognition tools can provide low-stakes pronunciation feedback and repeated practice. However, an ASR transcript or automated score may reflect accent-specific recognition performance and recording conditions rather than a human listener's judgment of intelligibility. The reviewed source set does not include a complete pronunciation-intervention corpus and cannot support claims about the comparative effectiveness of pronunciation tools. More direct evidence should include independent listener ratings, clearly described learner varieties of English, and follow-up beyond the immediate practice period (Zhu & Wang, 2025).

4.4 RQ3: Benefits and the measures supporting them

The selected publications and relevant prior reviews describe potential benefits including individualized support, timely feedback, repeated practice, and learner confidence. These benefits should be interpreted at the level actually measured: Guo, Pan, et al. (2024) address particular feedback tasks; Steiss et al. (2024) compare feedback quality; Zhang et al. (2024) consider affective measures; and Kim et al. (2025) concerns automated evaluation. Together they do not establish a single generalizable effect on English proficiency.

The distinction is particularly important for efficiency. Automatic generation of comments can be quick, but an actual teacher-time saving requires measurement of checking, error correction, and follow-up. Likewise, personalized output must be judged by whether it helps students complete a defined learning task, not simply by the fact that a model can vary its responses. Positive perceptions can justify investigating a practice further, while claims of durable learning require subsequent independent performance (Steiss et al., 2024; Teng, 2024).

Table 4. ***Proposed evidence-to-claim framework applied to selected cases***

Observation and source	Warranted interpretation	Claim that is not warranted	Needed stronger test
Steiss et al. (2024): rated human/AI feedback	Feedback quality differs by evaluated criterion.	ChatGPT improves later EFL writing proficiency.	Independent post-test and delayed writing.
Zhang et al. (2024): anxiety and willingness scores	Speaking-related affect changed in this setting.	Oral intelligibility or proficiency necessarily rose.	Blinded oral performance ratings.
Guo, Pan, et al. (2024): peer-feedback intervention	AI-supported feedback can assist specific EFL tasks.	The same effect transfers to all writing genres.	Multi-site replication and transfer tasks.
Huang et al. (2024): listening-decoding course	Specific decoding and engagement results are relevant.	Listening as a whole is improved in every context.	Independent comprehension and delayed tests.
Chen (2026): previous review of vocabulary studies	Its screened corpus emphasizes vocabulary breadth.	This article contains 25 included primary studies.	Independent documented retrieval for this review.
Wang et al. (2024): reading support system	Feasibility of a personalized reading-support design.	Generalized gains in reading comprehension.	Controlled assessment and unassisted measures.

4.5 RQ4: Challenges, safeguards, and implementation limits

The selected literature identifies concerns about inaccurate AI output, overreliance, unequal access, and differences in AI literacy. Documented adverse effects should be distinguished from plausible risks discussed in the literature. These examples do not support estimates of how frequently such challenges occur across schools or ELT settings. For context, publications address factual reliability, learner agency, and feedback interpretation (Crompton & Burke, 2024; Teng, 2024; Yang et al., 2024).

Conversational applications may also fail to represent pragmatic nuance, politeness, culturally situated meaning, or extended interaction, making human dialogue and reflection important (Zhai & Wibowo, 2023). Institutional concerns include academic integrity, privacy, fairness, accessibility, and teacher readiness. Teachers need guidance on appropriate AI assistance, personal data protection, and alternative activities where devices or subscriptions are unavailable. These are implications of the selected literature, not measured prevalence estimates (Uygun & Demir, 2026; Imran et al., 2026).

4.6 RQ5: Research gaps as priorities, not field-wide counts

The focal examples and prior reviews suggest priorities for future research, including longer-term language outcomes, stronger measures of receptive skills, more diverse learners and settings, teacher–AI interaction, and equitable access. Because the present sources were not comprehensively screened, these priorities cannot be ranked by their frequency in the field (Sharadgah & Sa'di, 2022; Zhu & Wang, 2025; Uygun & Demir, 2026).

Table 5. ***Evidence limitations translated into future study designs***

Unresolved question	Proposed study design	Outcome to report
Does AI support durable writing learning?	Matched or randomized EFL conditions with delayed tasks.	Independent rubric scores and revision uptake.
Does reduced speaking anxiety transfer to conversation?	Rehearsal intervention with independent interlocutors.	Affective scales and blinded oral ratings.
Can AI improve receptive skills?	Task-specific listening/reading comparisons.	Decoding, inferencing and delayed comprehension.
How can teacher oversight change results?	Compare teacher-only, AI-only and combined workflows.	Learning, feedback quality and verification time.

Unresolved question	Proposed study design	Outcome to report
Who benefits or is disadvantaged?	Multi-site and subgroup-aware implementation studies.	Access, linguistic variety, privacy and outcomes.

5. Discussion

5.1 Relationship to prior reviews and contribution

In the ten purposively chosen examples, eight were classified as primarily writing- or speaking-related (80%). That distribution illustrates the composition of this article's source set, not a field-wide pattern. By comparison, Zhu and Wang (2025) reported 52 writing and 30 speaking publications among 125 articles in their own review; those published figures are attributed to that review and cannot be treated as this article's sample or simply added where skill categories overlap.

The evidence-to-claim framework offers one potential interpretation tool: in a specified learner group, a defined AI function is used for a defined task, measured with an identified instrument over a stated duration, supporting a bounded conclusion. For example, lower speaking anxiety does not automatically demonstrate improved oral proficiency; high feedback quality does not by itself establish retained writing skill. The proposed framework has not been independently validated for reliability or impact, and the completed review should assess its value against the nearest published syntheses.

5.2 Implications for classroom practice

For writing, a defensible sequence is to have learners produce an initial draft, request AI feedback against a clear rubric, assess the validity of suggestions, make selective revisions, and justify important choices. The teacher may then compare the revised product with an unassisted follow-up draft to test whether learning transfers beyond the immediate AI-supported task. This recommendation follows from the distinction between feedback generation and feedback uptake; its effectiveness as a classroom routine remains an empirical question (Guo, Pan, et al., 2024; Yang et al., 2024).

For speaking, instructors may use a chatbot for structured rehearsal of a communicative goal, then require students to discuss register, pragmatic choices, and turn-taking in a peer or teacher-led interaction. Anxiety and willingness to communicate can be monitored without being confused with oral proficiency. For reading and listening, teachers should choose tasks and tests aligned with comprehension or decoding objectives, rather than accept a tool's ability to produce questions or transcripts as proof of learning (Zhang et al., 2024; Huang et al., 2024).

5.3 Teacher education and institutional responsibilities

Teacher education should address pedagogical design, output checking, measurement literacy, and ethical decision-making together. Instructors need to understand both where AI can provide a useful starting point and why a machine-produced response may require correction. Program-level guidance should state what kinds of AI support are permitted for different assignments, how use should be disclosed, and when independent performance must be demonstrated. Policies should also distinguish unverified suspicion from evidence of misconduct and should protect student privacy during any investigation (Crompton & Burke, 2024; Uygun & Demir, 2026).

Equitable implementation requires alternatives when learners cannot access commercial tools or suitable connectivity. A policy may allow controlled AI-assisted practice while preserving an equivalent non-AI pathway to the same learning goal. Institutions should also evaluate how tools handle accents, dialects, disability accommodations, and personal learner data. These proposals are safeguards grounded in the challenges identified by the reviewed literature, not measured evidence that any one policy eliminates bias or misconduct (Imran et al., 2026).

5.4 Limitations

The central limitation is purposive selection of ten illustrative publications. The numerical breakdown (six writing, two speaking, one reading, and one listening) is transparent but does not describe an independently screened or representative corpus. Some examples provide contextual rather than direct EFL learning evidence. Differences in design, participants, and outcome measures further restrict comparison, and the proposed evidence-to-claim framework has not been independently validated.

This article did not complete reproducible searches, deduplication, full-text screening, or formal study-level quality appraisal. Findings therefore support cautious interpretive discussion rather than general effectiveness claims. A future systematic review could test the broader patterns by implementing recorded searches, explicit eligibility criteria, a PRISMA flow, and quality appraisal. Bibliographic details and contextual interpretations should also be checked against primary reports before final submission.

6. Conclusion

This critical integrative review illustrates how AI can support English-language teaching through writing feedback, spoken rehearsal, reading scaffolds, and listening activities, while also raising concerns about reliability, overreliance, privacy, and access. Of ten purposively chosen focal publications, six mainly concerned writing, two speaking, one reading, and one listening. Thus, writing or speaking accounted for eight of the ten examples (80%), a descriptive figure that cannot be generalized to the entire research field.

The article proposes an evidence-to-claim framework that links AI function, instructional task, observed indicator, and design limitation to a proportionate teaching implication. Feedback quality, learner engagement, and independent language proficiency should not be treated as interchangeable outcomes. Future systematic reviews and longitudinal classroom investigations are needed to assess wider research patterns and durable learning effects. In the meantime, AI is most defensibly used as a teacher-guided, task-aligned support rather than an automatic substitute for assessment of independent learning.

7. References

Chen, X. (2026). Artificial intelligence in EFL vocabulary learning: A systematic review of higher education practices and research trends (2015–2025). *Cogent Education, 13*(1), Article 2667534. https://doi.org/10.1080/2331186X.2026.2667534

Crompton, H., & Burke, D. (2024). The educational affordances and challenges of ChatGPT: State of the field. *TechTrends, 68*(2), 380–392. https://doi.org/10.1007/s11528-024-00939-0

Du, J., & Daniel, B. K. (2024). Transforming language education: A systematic review of AI-powered chatbots for English as a foreign language speaking practice. *Computers and Education: Artificial Intelligence, 6*, 100230. https://doi.org/10.1016/j.caeai.2024.100230

Ericsson, E., & Johansson, S. (2023). English speaking practice with conversational AI: Lower secondary students' educational experiences over time. *Computers and Education: Artificial Intelligence, 5*, 100164. https://doi.org/10.1016/j.caeai.2023.100164

Guo, K., Li, Y., Li, Y., & Chu, S. K. W. (2024). Understanding EFL students' chatbot-assisted argumentative writing: An activity theory perspective. *Education and Information Technologies, 29*(1), 1–20. https://doi.org/10.1007/s10639-023-12230-5

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

Huang, W., Jia, C., Hew, K. F., & Guo, J. (2024). Using chatbots to support EFL listening decoding skills in a fully online environment. *Language Learning & Technology, 28*(2), 62–90. https://doi.org/10.64152/10125/73572

Imran, M., Almusharraf, N., & Shehzad, K. (2026). Artificial intelligence in school language education: A systematic review of trends, challenges, and future directions. *Journal of Interdisciplinary Studies in Education, 15*(6), 47–78. https://doi.org/10.32674/w506dk33

Kim, K., Lee, J. H., & Shin, D. (2025). The potential advantages of using an LLM-based chatbot for automated writing evaluation for English teaching practitioners. *Language Learning & Technology, 29*(1), 1–12. https://doi.org/10.64152/10125/73628

Klimova, B., Pikhart, M., Polakova, P., Cerna, M., Yayilgan, S. Y., & Shaikh, S. A. (2023). A systematic review on the use of emerging technologies in teaching English as an applied language at the university level. *Systems, 11*(1), 42. https://doi.org/10.3390/systems11010042

Li, B., Lowell, V. L., Wang, C., & Li, X. (2024). A systematic review of the first year of publications on ChatGPT and language education: Examining research on ChatGPT's use in language learning and teaching. *Computers and Education: Artificial Intelligence, 7*, 100266. https://doi.org/10.1016/j.caeai.2024.100266

Sharadgah, T. A., & Sa'di, R. (2022). A systematic review of research on the use of artificial intelligence in English language teaching and learning (2015–2021): What are the current effects? *Journal of Information Technology Education: Research, 21*, 337–377. https://doi.org/10.28945/4999

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction, 91*, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Teng, M. F. (2024). "ChatGPT is the companion, not enemies": EFL learners' perceptions and experiences in using ChatGPT for feedback in writing. *Computers and Education: Artificial Intelligence, 7*, 100270. https://doi.org/10.1016/j.caeai.2024.100270

Uygun, E., & Demir, B. (2026). A PRISMA-based systematic review of artificial intelligence in English as a foreign and second language education (2023–2025). *Discover Education, 5*, Article 478. https://doi.org/10.1007/s44217-026-01504-y

Wang, X., Zhong, Y., Huang, C., & Huang, X. (2024). ChatPRCS: A personalized support system for English reading comprehension based on ChatGPT. *IEEE Transactions on Learning Technologies, 17*, 1762–1776. https://doi.org/10.1109/TLT.2024.3405747

Yang, H., Gao, C., & Shen, H. (2024). Learner interaction with, and response to, AI-programmed automated writing evaluation feedback in EFL writing: An exploratory study. *Education and Information Technologies, 29*, 3837–3858. https://doi.org/10.1007/s10639-023-11991-3

Zhai, C., & Wibowo, S. (2023). A systematic review on artificial intelligence dialogue systems for enhancing English as foreign language students' interactional competence in the university. *Computers and Education: Artificial Intelligence, 4*, 100134. https://doi.org/10.1016/j.caeai.2023.100134

Zhang, C., Meng, Y., & Ma, X. (2024). Artificial intelligence in EFL speaking: Impact on enjoyment, anxiety, and willingness to communicate. *System, 121*, 103259. https://doi.org/10.1016/j.system.2024.103259

Zhu, M., & Wang, C. (2025). A systematic review of AI in language education: Current status and future implications. *Language Learning & Technology, 29*(1), 1–29. https://doi.org/10.64152/10125/73606