International Journal of Linguistics, Literature and Translation
ISSN: 2617-0299 (Online); ISSN: 2708-0099 (Print)
DOI: 10.32996/ijllt
Journal Homepage: www.al-kindipublisher.com/index.php/ijllt

Artificial Intelligence in Second Language Learning: A Critical Systematized Reassessment of Empirical Research, 2023–2026

Mohammad Bani Mohammad

Department of English Language Learning Skills, Language Institute, University of Tabuk, Tabuk, Saudi Arabia

Corresponding Author: Mohammad Bani Mohammad**, E-mail:** mnasser@ut.edu.sa

ABSTRACT

Language practice and feedback can be increased with the help of artificial intelligence, but enhanced assisted performance does not necessarily prove second language acquisition. The current syntheses focus less on what each result that has been mentioned can be in support of. The review investigated the functions and contexts of various artificial intelligences that a re-evaluated corpus represents, the findings provided by the reports on assisted performance, independent performance and retention, the relation between affective outcomes and measured performance and the design and reporting features that limit interpretation. On 10 September 2026 frame of 40 DOI-linked published records, published since 2023, was re-assessed by targeted publisher and repository retrieval. Outcomes were coded as assisted, independent, retained or transferred and reports analyzed at domain level. PRISMA 2020 and PRISMA-S were used to inform reporting. Twenty two reports were evaluated, 20 were retained and two were discarded and 18 could not be located. It was predominantly writing (nine reports) and speaking or pronunciation (eight). Artificial intelligence was not necessarily compared more favourably with active human feedback, multiple interventions were co-modulated with other support, and three reports incorporated a late assessment. In one assessment-condition study, anxiety levels were lower but had no higher scores. It was not a reproducible database search, only a part of the retrieval was realized, and all the stages were completed by one reviewer. The material in the re-evaluated corpus affirms the organized, instructor-assisted applications of synthetic intelligence to training and assessment, although not recommendations of lasting independent scholarly learning. Matched comparators and delayed, no assistance assessment should be the priority of primary research.

KEYWORDS

artificial intelligence; second language acquisition; generative AI; corrective feedback; willingness to communicate; systematised review

ARTICLE INFORMATION

ACCEPTED: 04 October 2026 **PUBLISHED:** 06 November 2026 **DOI:** 10.32996/ijllt.2026.9.11.8

1. Introduction

Artificial intelligence is introduced in language teaching through a variety of different activities: corrective feedback generation, speech recognition, conversation maintenance, and practice adaptation. The activities are labelled with the same name as the technology but are not a single teaching treatment. There is a risk that a conversational agent that promotes repetitive speaking and a writing assistant that suggests sentence corrections alter various features of learning. Their value in education can not be framed on gathering good results under the broad conception of artificial intelligence. Assessment should determine what is actually done by learners, what changes, and whether that change is maintained in the absence of help.

A number of syntheses already exist in this domain, and the current input will have to fit in this range. The article by Huang et al. (2022) measured the language learning assisted by chatbots prior to the publication date considered in this study. Jeon and Lee (2024) summarized the literature on human-chatbot comparisons and highlighted the challenge of comparing an artificial interlocutor with a human. S. Lee et al. (2026) identified 49 articles that covered classroom-based research on generative artificial intelligence and were published between 2023 and 2024, with a focus on higher education, use of self-reported data, and predominantly in English.

Y. Li et al. (2026) meta-analysed 25 task-based studies, and Aljohani (2026) synthesised applications and challenges of one of the most popular models. These reviews determine the magnitude and propagation of the evidence base that the current paper attempts to do to a significantly broader extent.

Copyright: © 2026 the Author(s). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) 4.0 license (https://creativecommons.org/licenses/by/4.0/). Published by Al-Kindi Centre for Research and Development, London, United Kingdom.

The inferential association between a reported result and the learning assertion connected with it is less firmly established than what they do so. This review poses a more restricted question than previous syntheses that search the areas of research interests, instruments and overall effects, assuming the comparator, the state of assessment and measuring time, what shall each reported outcome be able to actually imply? The contribution is thus a critical re-evaluation and not an exhaustive update or bibliometric record. It is based on known descriptions of second language acquisition, since whether a learner has learned something is a question about the developing system of the learner, rather than the tool that was available at the time a performance took place.

The purpose of this review is to reevaluate a specified set of empirical reports on artificial intelligence in second language learning, published between 2023 and 2026, to ascertain the extent to which each reported finding supports the arguments about assisted performance, independent performance, retention, and transfer. Every conclusion is made with reference to the reassessed corpus and should not be applied to the literature of the period, in general.

This is a key difference between assistance and acquisition in writing. A better piece of writing can indicate that a student received some valuable feedback, made an efficient revision, or took personal thoughts and mixed them with generated language. It does not, as such, determine that the learner is in a position to replicate the applicable linguistic options alone. An assistance-free post-test has more direct evidence of independent performance, but can still be indicative of having been familiarised with a practised topic. To maintain retention and transfer, more observations over time, tasks or other communicative contexts are needed. In this review, these consequences are considered related yet analytically distinct.

The same problem of interpretation pertains to speaking. Talking to a responsive system more often could enhance rehearsal, lead to the decrease of the immediate social cost of error, or offer chances that are not possible in a crowded classroom. Its performance in that system, however, may not be similar to interaction with a human interlocutor, and speech recognition comes with its own set of challenges, as learners might modify their speech to what the system can accept. Gains in pronunciation practice, in a scoring speaking task, and willingness to communicate thus respond to various questions. An honest synthesis should not eliminate such differences but instead work out the result of all these favourable outcomes in a general statement about proficiency.

Even where they are not the language test scores, motivation, anxiety, enjoyment and self-regulation have an educational significance which makes the differences in their importance insignificant but not equal to proficiency. A student might feel more comfortable using a chatbot and a teacher to be explained or a peer to have meaningful conversations. Another might play with a tool without practising more outdoor lessons that are necessary. The right question is what are the relationships between affect, behaviour and performance in a specific design. In a report, where only perceptions are involved, they are treated as evidence of experience, but not evidence of learning, in this paper.

Four research questions are inspired by these concerns.

RQ1. What artificial intelligence functions, learner groups, and learning contexts are instantiated in the reassessed corpus?

RQ2. What do the reports show of assisted performance, independent performance, and retention?

RQ3. What are the reported learner-experience outcomes and affective outcomes, and how do such outcomes, or lack thereof, match, or do not match, measured language performance?

RQ4. What are the design and reporting features that limit the interpretation of the reported AI-related outcomes?

It began with a pre-existing 40-record candidate list, not an explicit search of a database that was just done. It is thus named the systematised review, as Grant and Booth (2009) make the distinction between reviews that utilise selected systematic components, versus those that carry a systematic-review assertion. This limit is expressly mentioned here since it conatus omnibus invinitate repercussis: and its effects are discussed once, in the Limitations subtitle.

1.1 Theoretical framework

Second language development and the conditions for learningThe interpretive machinery of such a review is based on four positions known in research on second language acquisition, each of which outlines circumstances under which exposure turns out to be development.

The interaction account is the view that negotiation constitutes meaning, and the altered input and feedback accomplishing this model is what furnishes learners with the evidence that they require to work out their developing systems (Long, 1996). Exposure is not the same thing as the interactional task of sorting out communicative trouble. This creates an immediate question with respect to conversational agents: a system that takes imperfect input without apparent struggle, or that otherwise fixes failures on its own, can eliminate the very negotiation that the account itself claims to be productive.

The output account believes that the production of language causes the learners to stretch out to transform semantic to syntactic processing, hypothesis testing, and identification of differences between desire and means (Swain, 1985). There is a sharp contrast in the amount of output of the tools. The production in a system that writes on behalf of the learner is reduced, whereas in the system that encourages them to make longer speeches, the production rises. Two technologies with the same technological name can thus be opposite to this state.

The noticing account is an account that states features need to be attended to in the input so they can be available to be acquired (Schmidt, 1990). The reason why corrective feedback is applicable here is mostly as a mechanism of controlling attention (Ellis,

2010; Lyster and Ranta, 1997), and the nature of the feedback can take precedence. When the learner reformulates something he/she accepts uninspected, this might create no noticing of whatever interpretation, but when the feedback has the effect of indicating the issue, the learner inspects and infers it; this acts directly on his attention. This position results in the difference between getting feedback and working on it, which is repeated throughout this review.

The sociocultural explanation places development in assisted performance in the zone of proximal development, and regulation eventually transfers off the more competent participant to the learner (Vygotsky, 1978). The most friendly towards assisted performance as a valid object of study, this description, however, is the most demanding because it demands that the assistance be developmentally relevant only on the condition that it be measured against the learner and be increasingly withdrawn. The help we have here is not scaffolding, and the help is constant, uniform, and indefinitely available, and it is just the lack of withdrawal which makes assistance-free measurement required.

Affective constructs come into this frame in a certain role. The concept of willingness to communicate is viewed as a contingent ability to engage in discourse mediated by state confidence and the interlocutor (MacIntyre et al., 1998), and the self-determination theory can be associated with the motivation in terms of competence, autonomy and relatedness (Ryan and Deci, 2000). Both predictors suggest that a low-threat interlocutor can increase participation; neither predicts that it is participation that leads to development. Affect thus acts in this review as a believable antecedent of practice behaviour rather than acting as a proxy of proficiency.

1.2 Outcome conditions and the instructional contrast

It is against this backdrop that the review isolates four clearly defined, operationally defined evidence items: assisted performance, independent performance, retention, and transfer, which are set out in the Methods. These are coding differences determined by this synthesis as opposed to an acquired hypothesis that is now tested to be sound. They are meant to reveal the inferential gap between a given outcome of observation and a general educational statement. The categories do not necessarily constitute a ladder: assisted performance may be a valid goal whereby usage of tools by professional communication is a regular practice, whereas independent performance will be relevant when the goal is unaided competence.

The second dimension of analysis is the comparison that is estimated. The artificial intelligence feedback vs. artificial intelligence no feedback answers the question of the utility of an instructional package including feedback. The comparison between Artificial intelligence feedback and equally timely teacher feedback is based on a limited comparative question. The combination is tested with a package, adding practice, training, and teacher discussion tests. When one of the conditions has more minutes of practice, is more readily available to the outside, or allows for quicker correction, the competition can be partially due to those characteristics. Review thus considers dose, support, and feedback timing to be parts of the intervention as opposed to small procedural matters.

A third aspect is the role of the learner in the utilization of feedback. An elicited response can then be a point of comparison and debate, a response that is duplicated without much questioning, or a signal to make another suggestion. These applications would be reflective of the noticing condition above. In cases where research presents interviews or revision behaviour, such observations might be able to explain plausible processes but not to test the possibility of that process being the cause of the observed result. Conflicting results are not just considered an inconvenience to be eliminated; they could also be seen as another possible use of tools or the seeming trade-off alters across designs.

2. Methods

PRISMA 2020 (Page et al., 2021) and PRISMA-S (Rethlefsen et al., 2021) were used to inform reporting procedures and limitations and render the procedures of the review transparent. The order of sections is based on the PRISMA 2020 item order, with the availability or unavailability of each feature becoming visible. An accomplished PRISMA 2020 checklist listing the items that were not done or were not applicable.

2.1 Protocol and registration

No protocol was prospectively registered. The rules of eligibility and synthesis have been clarified in the re-evaluation, once the list of candidates was already known, and this poses the risk of eligibility being based on available evidence. The following account is consequently retrodictive and retrospective, and it does not assert that an ancillary protocol controlled the initial search.

2.2 Outcome definitions

Due to the looseness of words like performance, achievement, retention, transfer, and learning, Table 1 definitions were used everywhere during the coding process, and later, these definitions are used throughout the rest of the paper.

Table 1

Outcome definitions used in coding and reporting

Term	Definition used in this review
Assisted performance	Performance on a task completed or revised while the learner had access to the artificial intelligence function.

Term	Definition used in this review
Independent performance	Performance on a task completed without access to the artificial intelligence function or equivalent support, as stated in the report.
Retention	Independent performance measured after an interval following the end of instruction (a delayed or follow-up assessment).
Transfer	Independent performance on a task, topic, genre or mode not directly practised during the intervention.
Achievement	A composite course or programme test score; it is not divided into separate skill effects unless component scores are reported.
Learning	Used only for independent performance, retention or transfer outcomes; never for assisted performance or self-reported gains.
Affective outcome	A measure of anxiety, enjoyment, motivation, willingness to communicate, self-efficacy or a related construct; not treated as evidence of language gain.
Randomised	Used only where a report states that individuals or groups were allocated by a random process; random selection of a sample is not randomisation.
Statistically significant	Used in full whenever a hypothesis-test result is meant.

2.3 Eligibility criteria

The inclusion criterion was journal publication between 1 January 2023 and 10 September 2026 that reported original empirical evidence with identifiable second or foreign language learners. They had to report language performance, a learning process, or an experience related to the use of an identifiable artificial intelligence function and describe how the learners used it. All evidence was eligible, including quantitative, qualitative, and mixed-methods. The reassessed material was published in English, but the publication language was non-restrictive; the same language was also used as the target language (coded independently).

Out of scope were reviews, conceptual discussion, teacher-only opinion studies, and technical model evaluations that did not include a learner-use component. The fact that the study involved an assessment did not rule out an assessment study: a comparison of what learners actually do in terms of speaking assessment could suffice, but a comparison of machine and human ratings of the available essays could not. Intervention also demanded a recognizable artificial intelligence functionality; an artificial intelligence name in a title could not be used alone. Factors and quartiles of journals were not inclusion criteria.

2.4 Information sources and provenance of the frame

The retrieval frame was a list of 40 DOI-linked records that had been compiled on a previous project on artificial intelligence and second language learning. The initial search history of that list was not maintained, and no export of Scopus or Web of Science, search cumulative of ERIC, or any de-duplication sequence was reported. The frame then cannot be reframed as a reproducible search, and these processes are not displayed as fulfilled.

To estimate the overlap of the frame with a database search, the 40 DOIs were paired following the re-evaluation with records exported from Scopus and Web of Science (as of 2023-2026). That export also contained 34 of the 40 records, with 15 out of the 20 maintained records. Overlap is described with this check; this check was not applied to identify, screen, or add studies, and the export is on behalf of the author upon request.

Specific searching with the frame was publisher platforms, which comprise Frontiers, Springer Nature, IRRODL, Elsevier, Taylor and Francis, Wiley and MDPI, and public repository paths, such as the ERIC full-text repository. Retrieval based on item-level queries: DOI resolution, specific DOI strings, title queries in combination with author surname, and title queries that are limited to a repository domain. These are not database-ready Boolean strategies, and no overall number of hits is reported. It was retrieved on 10 September 2026. The procedure is documented in Appendix D.

2.5 Selection process

Full-text availability was determined as allowing enough access to the methods and outcomes of a report to allow reassessment, either by HTML provided by the publisher, or by a PDF provided in a repository; having a landing page or abstract did not constitute full-text availability. Out of 40 candidate records, 22 were re-evaluated at the full-text level, 20 were chosen, and two were rejected due to definite scope considerations. Eighteen could not be retrieved. An online publisher's correction was studied in the context of its original article as a companion notice and not as a second study. The layout of the corpus is shown in Figure 1.

A single reviewer performed selection, extraction, and interpretation, which was assisted by multiple checks with the corresponding full-text sections. There was no second reviewer, and the statistic of inter-rater agreement was not computed. Having the same reviewer repeatedly check is not equal to independent duplicate screening, and the implications are discussed in Limitations.

2.6 Data collection process and data items

Appraisal was structured based on the domains of the Mixed Methods Appraisal Tool (Hong et al., 2018), which can support quantitative, qualitative, and mixed methods designs with a single structure. Based on each retained report, there were allocated and baseline comparability judgements using judgement on allocation and baseline comparability, co-interventions, exposure and adherence, completeness of outcome data, independence of outcome measurement, consistency in reporting, and matches claims and design. There was no overall numerical score, and this was in accordance with the instructions given by the developers. All judgements have been displayed in Appendix E, and Table 4 is based on this appendix.

The productivity difference apparent in the appraisal was the difference between the problem identified and the information lacking. The reason why an intact-class contrast is a design issue is that both the means of baselines are approximately equal; the reason why an unreported version of the model creates a reproducibility divide is that it does not demonstrate that there was no effect on something at all; and the reason why there is a discrepancy between a tabular and narrative context implies that something is wrong is that you must use the estimate cautiously without suggesting that it was found to be ineffective or wrong. Where corrections by publishers were found, they were included; the databases of corrections and retractions were not searched exhaustively.

2.7 Synthesis methods

The synthesis mapped out contexts and technologies first and then clustered evidence into writing, speaking and pronunciation, listening and greater achievement, and experience of the learner. In every group, interpretation was concerned with the actual contrast and the outcome condition. Estimates were made between-group and within-group gains. Statistically non-significant results were retained, and failure to establish a difference was not considered equivalent unless the design of a given study circumstantially warranted the same argument. The same report generated several results which were discussed but not counted as a result of different confirmations.

There was no calculation of a pooled effect as the reports varied in terms of interventions, comparators, scale, allocation units, and time points, and there were reports that had some gaps in terms of reporting. The article achieves this instead of relying on a numerical effectiveness scale by taking a structured narrative comparison, and does not count statistically significant outcomes in determining whether artificial intelligence is effective.

2.8 Reporting bias and certainty assessment

A pooled analysis was necessary to make it possible to have a statistical evaluation of publication bias. The qualitative perspective on reporting bias: the pattern of retrieval, as stated in the Limitations, is the means of reporting bias qualitatively. No clear certainty framework like GRADE or GRADE-CERQual was used, and no certainty ratings are given a posteriori.

3. Results

3.1 Corpus and study characteristics

The set of retained reports was limited to 20 primary reports, five of which were released in 2023, two in 2024, seven in 2025 and six in 2026 (Table 2). These are the measures of the reconsidered corpus and are not a fashion of publication. There were nineteen reports on English and one report studied the use of Spanish as a foreign language. There were 19 in higher education and one in secondary school, mostly in Asian and Middle Eastern contexts. The corpus hence provides limited reasons as to why one should extrapolate to young children, less commonly taught languages, or unstructured lifelong learning.

The most frequently reported main outcome domains were writing (nine reports), speaking and pronunciation (eight reports), listening or composite achievement (two reports), and learner affect (one report). This distribution is similar to a distribution observed by S. Lee et al. (2026), which provides this profile with some external plausibility, without demonstrating that the corpus is representative.

Table 2

Distribution of retained reports by year, context and outcome domain

Characteristic	Category	n
Publication year	2023	5
	2024	2
	2025	7
	2026	6
Target language	English	19
	Spanish	1
Educational level	Higher education	19
	Secondary	1

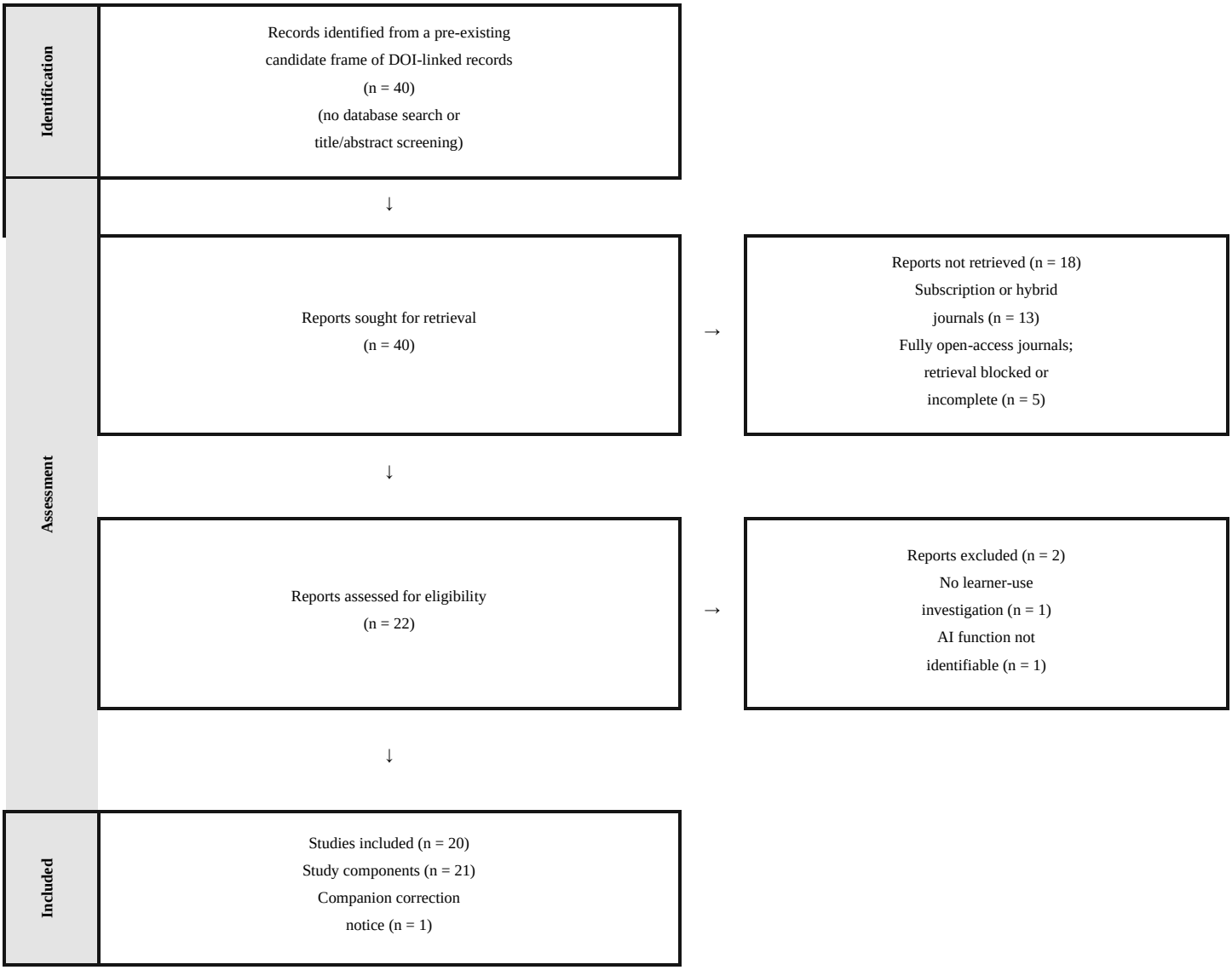
Characteristic	Category	n
Primary outcome domain	Writing	9
	Speaking and pronunciation	8
	Listening or composite achievement	2
	Learner experience or affect	1

Note. Counts refer to the 20 retained reports. Primary outcome domain is the outcome the report foregrounds; several reports also measured affective variables. Each block sums to 20.

According to the results of the study by Escalante et al. (2023), there are two types of distinguishable studies, and the 20 publications have 21 components of the studies instead of 20 similar experiments. Samples, lengths, and evaluations were dissimilar. Other studies involved comparison of instructional groups over a few weeks; others involved revision of one essay or a short series of writing tasks or varied speaking-assessment situations. These design differences have been retained in the Appendix A, and Table 3 shows the difference in the outcome differences along with.

The two exclusions explain the scope of operation. Here, the evidence provided by the learner is needed since an essay is graded by a generative model and human raters based on the essay (J. Li et al. 2024). An assessment-comparison report was used on assessment arrangements, but the artificial intelligence feature of differentiating its intervention was not identifiable enough. These determinations relate to the immediate questions, and not to the overall merit of either of the publications. The 18 records that could not be retrieved have no findings, and none of them can be inferred from their titles.

Figure 1: PRISMA 2020-informed flow diagram showing the disposition of the candidate corpus



Note. The diagram follows the PRISMA 2020 pathway for records identified by other methods, because no database search or title and abstract screening was conducted. Reports not retrieved and reports assessed sum to the frame total (18 + 22 = 40); exclusions and inclusions sum to the assessed total (2 + 20 = 22). The correction notice linked to Song and Song (2023) is shown separately and is not counted as a study.

Table 3

Outcome distinctions illustrated by selected reports

Evidence type	Illustration	Interpretation
Revision gain	Alnemrat et al. (2025): between-feedback gain difference not statistically significant, d = 0.10	No clear advantage for artificial intelligence; not an equivalence test
Delayed targeted pronunciation	Xodabande et al. (2025): advantage retained on target words two weeks after instruction	Retention of practised words; broad transfer untested
Assessment experience	Susoy (2026): anxiety lower under artificial intelligence facilitation; achievement difference not statistically significant	Comfort and demonstrated performance are separate outcomes
Affective response	Miao and Hu (2026): willingness to communicate favoured artificial intelligence	No measured proficiency outcome

Note. Values are as reported by the primary authors and are illustrative contrasts, not pooled estimates. Source locations are listed in Appendix B.

3.2 Writing outcomes and corrective feedback

The reports on the writing take up a conditional account of benefit. Song and Song (2023) presented only a 12-week intervention with 50 participants with English as a foreign language and reported positive writing and motivation effects. The original content was revised by a subsequent corrigendum that changed the interpretation of the completed content (prose) without altering the underlying table (Song and Song, 2026). The fact that this is not revealed under an abstract-level reading supports the way to consider structured feedback, but numerical claims would presuppose addressing the corrected text.

The study by Mahapatra (2024) involved the analysis of academic writing using generative models within one of the Indian universities and involved pre-, post-, and delayed assessments. The sample analysed included 72 out of 134 students who consented, which makes completion-based selection relevant to interpretation. Delayed observation also provides data on persistence, but fails to eliminate selection bias. The corresponding conclusion is related to the perceived completers within the specified arrangement, as opposed to the guaranteed advantage of all students provided with the tool.

Another test is active feedback comparisons and one that addresses the noticing condition directly. In a six-week study of 48 learners, Escalante et al. (2023) compared generated feedback and tutor feedback and failed to find a statistically significant difference in the learning trajectory. There was a separate preference investigation that included 43 learners. The lack of a statistically significant difference is not an indication of equivalence, and the integration of the two studies into a single sample would confuse the findings between the performance and preference evidence.

Alnemrat et al. (2025) have compared generated feedback on argumentative revision with teacher feedback on the same in 120 Jordanian learners at two levels of proficiency. Both conditions resulted in better drafts, and the comparison between conditions showed no obvious benefit to artificial intelligence d = 0.10). This testifies to revision following feedback, rather than a long-lasting acquisition. It can also be used to justify positive results achieved when weaker controls are used: the availability of helpful feedback can cause an improvement even in the case when a provider is a human being, but there is no evidence in the study that teachers and automated systems are exchangeable.

Wang (2024) found encouraging results of the feedback generated on writing and anxiety measures. Nevertheless, the given three-group sample size of 75 fails to match all sample sizes in the results tables, which restricts trust in the accurate extraction. The report is still in the qualitative comparison since its intervention and questions are applicable; however, its estimates are not considered safe. Holding a report is one thing, as opposed to ratifying all the numbers contained in it.

In other comparisons, instructional packages were compared as opposed to individual tools. Ekizoğlu and Demir (2025) investigated 60 Turkish students enrolled in secondary classes during a period of eight weeks, utilizing a combination of automated feedback and instructor attention in regard to higher-order writing problems. A larger Pakistani sample was examined by Mujeeb et al. (2026) with the help of two combined tools and differentiated between an eight-week research schedule and a shorter feedback stage. Both reported positive changes in writing; however, neither identifies the role of one model, and the time spent on the calendar must not be confused with the exposure to treatment or a retention interval.

A study of 322 Iraqi learners using the Mohammed and Khalid (2025) writing and affective benefits in a 10-week study. The instrument was not stated in such a way that it could be interpreted specifically in a version, and the relations between the feedback lines add complexity to the attribution. Divya and Alamelu (2026), in turn, pooled two titled tools in a five-day investigation on 40 engineering students. Its final task, which did not require assistance, had the same topic as the first task, and there was no comparison group; hence, improvement is compatible with practice and instructional effects.

These oppositions, combined, indicate that better texts and a technology-specified learning advantage should be kept as two distinct conclusions. The teacher-sustained package can be educationally helpful without technological superiority, and a similar short-term outcome can justify optional application without demonstrating equivalence. In this corpus, performance was better supported by structured feedback leading to revision, and performance was less assured by transferable learning that was independent.

3.3 Speaking and pronunciation

Sun (2023) studied 61 Chinese students in two classes throughout 14 weeks, with a combination of automatic speech recognition and peer correction. The improvement in pronunciation and speaking reported is thus related to that mixed arrangement. Calais individual, two-class assignment provides stronger protection against confounding compared to an assignment at the class level. The research is applicable to the viability of recognition-aided instruction but contains a restricted foundation of breaking apart the technology and peer engagement or other grade-wide effects.

In a 13-week study involving 93 Chinese students, Qiao and Zhao (2023) reported positive speaking and self-regulation results of a platform-based arrangement. The statement of allocation and sequence of courses renders an isolated-treatment interpretation hard, and the evidence relates to an extensive platform-mediated experience instead of a comparison of generative models. Lumping it in with generative-model trials in the name of being artificially labeled with the term artificial intelligence would mask the variations of task structure, adaptivity, and feedback.

Xodabande et al. (2025) also gave a more specific comparison of pronunciation: 60 Iranian students received 3 weeks using voice based generative or an electronic dictionary with a delayed test that followed 2 weeks later. The artificial intelligence team continued to have an advantage on the intended word test. Since the target words were used again, although in sentence form, the outcome supports the short-term retention of the practice material more directly than transfer to spontaneous speaking.

Ding and Muhyiddin B Yusof (2025) reported speaking improvements and less anxiety in some intervention with two intact groups (resulting in 60 learners), in a six-week intervention, although the practice timeline is not present in all parts of the account. Yan and Singh (2026) analysed 68 students in 2 intact classes during 10 weeks and achieved positive results in speaking and emotions. Its blinded examiner reviews enhance measurement, but class-based dispensation continues to impede causal distinction of the instrument compared to the greater teaching juxtaposition.

In a study of 86 Taiwanese undergraduates in an 18-week blended course, Lai (2026) employed one model in oral work and another in planning, with no evident differing advantage to fluency. In a 14-week course, improved speaking was observed by J. Lee (2025), in which all 39 Korean learners received artificial intelligence plus peer practice; thus, the change cannot be attributed to the technology. These reports support instructional integration as opposed to displacement of human interaction, just as the challenges highlighted by Jeon and Lee (2024) have.

Susoy (2026) discussed the conditions of assessment as opposed to instruction. There was a counterbalanced design with artificial intelligence and human facilitation in speaking assessment among forty-eight first-year Turkish university students. The differences in anxiety when talking with artificial intelligence facilitation and the differences in achievement scores were not statistically significant. The tables included in the report indicate that there is a negative correlation with achievement under human facilitation ($r = -.63$) but not under artificial intelligence facilitation ($r = -.09$); the values in the abstract ($r = -.50$ and $r = -.04$) differ; thus, the numbers are cautiously accepted.

In the artificial intelligence condition, digital literacy was positively associated with speaking performance ($r = .35$), which the report refers to in two ways. Themes identified during the interview process were self-reported proficiency, comparison of the two modes of evaluation, and quality of feedback. The research is useful to serve as a negative case to a simple comfort-equals-learning account: the lower the anxiety the higher the score and a construct irrelevant variable was related to the performance under the artificial intelligence condition.

3.4 Listening and broader achievement

Xiao (2025) found positive changes in listening, flow, and anxiety results in an 84-participant study that involved an eight-week intervention and a three-week follow-up. The connection between the speech-recognition label and a description of methods, including text-to-speech functions, complicates the interpretation of the mechanism. The result is stored as an indication of the instructional package that is reported without ascribing its impacts on speech recognition. The informative positive delayed outcome still provides a positive result, albeit with uncertainties regarding what was delivered to restrict replication and tool-specialized findings.

Research by Wei (2023) on a 10-week platform-based intervention with 60 learners in two intact classes in universities reported positive achievement, motivation, and self-regulation outcomes. Its composite achievement value cannot be translated into

individual impact on each of the skills, and the class allocation does not leave any explanatory context. These reports are written and oral reports that expand the corpus beyond writing and speaking, albeit at too unequal a distribution to rank effectiveness among the reading, listening, vocabulary, grammar, and productive skills.

3.5 Learner experiences and mixed findings

The report by Miao and Hu (2026) contained the only language other than English with which the target language would be provided. A 16-week intervention study of their Spanish study included 69 of 79 learners who were enrolled. Findings of willingness-to-communicate and boredom were more biased towards artificial intelligence, but there was no such outcome of anxiety or enjoyment, and no speaking-proficiency outcome was assessed. It indicates that specific emotional situations must not be reduced to a single positive attitude outcome as the willing-to-communicate literature predicts (MacIntyre et al., 1998).

Accessibility and a chance to rehearse are plausible attractions across the experience material, and interpretation of feedback and the importance of human support are consequential. Mujeeb et al. (2026) mentioned blind acceptance of certain feedback and hybrid provision support, and Divya and Alamelu (2026) wrote about the worries regarding the formality generated and personal voice. Of theoretical interest is uncritical acceptance, since they characterize the state where the corrective feedback is assumed to produce minimal noticing (Schmidt, 1990).

This experience synthesis contrasts published reports; it is not a thematic re-publication of interview transcripts, and no prevalence or saturation statement is included. Sampling of interviews, interviewer prompts, and context: in the reports, the sample is selected, and only excerpts are included instead of full datasets. Their greatest contribution is explanatory and warning: they demonstrate why scores, feelings, and desire to receive certain types of support should be explored jointly and without presuming a union with each other.

3.6 Methodological credibility across reports

Small sample sizes were not the only threats to inference. The appraisal domains are summarized in Table 4, with names of the reports provided behind the respective counts in Appendix F. Eight reports had an allocation or baseline comparability issue, nine had extra practice or support in the artificial intelligence condition, and twelve had excessive conclusions which their comparator and assessor could not support. Although they decrease random uncertainty and increase sample size, these design problems are not solved, and a longer teaching period does not support evidence of post-intervention retention.

Four reports demonstrated reporting discrepancies: Corrected writing results, discrepancies between sample counts, discrepancies in description, and varying correlation values between exposure descriptions. Two reports involved assessment content practised in the intervention (Divya and Alamelu, 2026; Xodabande et al., 2025), an assertion that restricts transfer arguments. There were only three reports that comprised a delayed or follow-up assessment (Mahapatra, 2024; Xiao, 2025; Xodabande et al., 2025). These examples are sufficient to caution with respect to precise estimates without suggesting that all substantive conclusions are false.

Table 4: Appraisal domains, frequency and interpretive consequences

Domain	Problem identified (N)	Not reported (?)	Consequence for interpretation
Allocation and baseline	8	2	Class and treatment can remain confounded; baseline equivalence not secured
Free of major co-intervention	9	1	Effects attach to the combined instructional package, not the AI function alone
Exposure and adherence reported	1	2	Dose differences cannot be ruled out
Complete outcome data	2	1	Results may describe completers rather than all allocated learners
Outcome measurement independence	0	12	Assessor knowledge of condition may influence ratings
Reporting consistency	4	1	Exact estimates are unsuitable for pooling without reconciliation
Claim matched to design	12	0	Stated conclusions exceed what the comparator and assessment can support

Domain	Problem identified (N)	Not reported (?)	Consequence for interpretation
No delayed or follow-up assessment (Appendix A)	17	–	Retention claims unsupported

Note. Counts are taken directly from Appendix E (N and ? cells) and Appendix A; Appendix F names the reports behind each count. Counts are not mutually exclusive. These are domain judgements structured by the Mixed Methods Appraisal Tool (Hong et al., 2018), not MMAT scores or certainty ratings.

4. Discussion

4.1 What the evidence supports

The key conclusion to be drawn in this revised corpus is not some mean effect, but a limit to understanding. Instruction provided with the aid of artificial intelligence often underwent positive language-task results, and the best plausible argument hinges on what improved with the tool. Where expanded arrangement was supported by the addition of feedback or practice to the comparison, the outcome supports this addition. In instances where both artificial intelligence and human feedback were in play, the benefits would not be as immediate. In the case where the result was an assisted revision, evidence will be in relation to the revised product.

This alters the educational decision that ensues. The fact that feedback availability enhances writing may support an extension to writing tutorials where students currently get very little feedback; it is not evidence that an adequately supported writing tutorial should be substituted with automated feedback. It is plausible that evidence that a chatbot provides helpful rehearsal is sufficient that optional pre-seminar practice is justified; it does not prove that conversational rehearsal can substitute the unpredictable social work of the seminar itself. The boundary of the decision is determined by the comparison and the result.

The convergence of the findings, which are approached rather differently, becomes one with wider syntheses. S. Lee et al. (2026) include such studies, which come at the expense of self-reported data and an orientation of perceptions as opposed to acquisition; the study by Jeon and Lee (2024) found the challenge of interpreting human-chatbot contradictions where the two talkers should have differed in more than a single aspect. The most favourable-appearing designs in the current corpus tended to be ones with weak comparators or with the most bundled interventions. The risk of pooled estimates like the one conducted by Y. Li et al. (2026) might thus justify reading in such a way as to suggest the quality of the underlying studies is comparable.

4.2 Why assisted performance is an incomplete acquisition measure

Language acquisition deals with change in the ability of the learner; assisted performance is a cooperatively produced performance among the learner, the tool, and the task. This does not render assisted output insignificant, but it determines where ambiguity originates, as the product is used to determine competence. When aids provide the unknown construction, the final text will fail to display the understanding of whether the learner has acquired it, whether he or she would make a choice, or whether he or she would adjust it to a different purpose.

The sociocultural explanation explains why this is important. Assisted performance is developmentally informative in that regard since the assistance is adjusted and gradually reduced (Vygotsky, 1978). Constant help that is the same is not suitable to that requirement. An intervention where learners have retained access and unrestrained observation is made, but no assistance of any form is involved, cannot differentiate scaffolding and substitution. It is not a methodological preference, but merely a theoretical reason why assistance-free measurement should be assisted.

Ambiguity can be minimized by having observations throughout the instructional sequence. The initial unassisted attempt provides existing performance; the assisted task records feedback choice and amendment; the subsequent unassisted task demonstrates what can be obtained in the absence of aid; and time-delayed or altered tasks test persistence and transfer. Retention must also have a well-determined clock: the 18-week-long course will gauge performance at the conclusion of sustained support and not after support is terminated. The most evidential gap to be found in this corpus is that of only three out of 20 reports containing any delayed or follow-up observation.

4.3 The role of human support and learner judgement

The same is not true of the retained evidence regarding human involvement that is viewed as an extraneous factor. In some of the set-goal arrangements, teachers posed higher-order problems or linked classroom learning to feedback, and peers ensured an audience and offered meaning-negotiation opportunities. These actions can be the reason why a package has been successful. Their contribution can be experimentally isolated to better explain, and the implementation of this is not to claim that its removal would leave the observed benefit unchanged, just that their contribution can be removed.

The interaction account indicates a certain cause of care. When it is the negotiation of meaning which renders interaction power, a partial interlocutor, the kind who grasps degraded input without stuttering, is able to minimize the trouble that is the motivation of the change (Long, 1996). This is a testable proposal, not an established result; none in the corpus tested it directly.

The cognitive demands that are on the learner can also be varied by human support. When there is a requirement to explain a revision, this implies the inviting of a comparison between intention and what is proposed, and when there is a requirement to

justify a decision to reject a suggestion, then disagreement is justified as legitimate. These practices in noticing terms transform a given correction into attended input (Schmidt, 1990), and in terms of output, they reinstate production that would be taken over by a drafting tool (Swain, 1985). These are pedagogical suggestions based on the framework, rather than mechanisms validated by the studies contained within.

4.4 Affect as an outcome and a possible mechanism

An emotional outcome assumes two levels in this review. Less anxiety or boredom is important since the learning experience is important. These differences can also affect the practising, persisting, or seeking of interaction by learners because self-determination theory would predict where a particular tool serves to facilitate perceived competence and free will (Ryan and Deci, 2000). The second assertion, however, deals with a mechanism and has to be supported by evidence of behaviour and subsequent performance, and not by a questionnaire difference only.

This caveat is a consequence of the mixed results. Anxiety was also found to be lower in Susoy (2026) with artificial intelligence facilitation, but achievement did not differ, and performance in such a condition was correlated with digital literacy. In Miao and Hu (2026), favourable willingness to communicate did not also correlate with an advantage in anxiety, either. A recreational activity can be complementary and less linguistically challenging, or a challenging human socialization can be less comfortable but provide significant practice. These are alternative explanations to explore, not the reasons to exclude positive emotions.

4.5 Credibility and technological reproducibility

An identity of an intervention cannot be a product name. The conversation, correction, planning, translation, or adaptive exercises can be offered by a platform, sometimes using various models. A reproducible account must describe the functionality they provide to learners, the model or version they are using (where available), what was accessed on the date, what they have been told to do, and the aid provided by the teachers. S. Lee et al. (2026) also found a lack of version and prompt reporting to be a field-level issue. The variables that should be considered as prompting and exposure are related, as the required minutes, use, and attentive engagement logged should vary.

And equal merit is measurement. Human-ranked verbal tasks may ensure that one is not dependent on the tool, but the authority of the score requires that the raters are trained, blinded, and that there is disagreement management. Unless component evidence is reported, composite scores can demonstrate no information about what component changed. Interpretability includes numerical consistency: the sample analysed must be consistent with degrees of freedom and table values, with corrections to be associated with their eventual publications.

4.6 Implications

The theoretical framework and inference differences used in the review give rise to implications as outlined below. The studies that were included are not tested as interventions and thus should be interpreted as suggestions aimed at practice and research and not evidence-based prescriptions.

4.7 Instructional design and assessment

A tool should be chosen with the intended result in mind. When the quality of the revision is the intended outcome, assisted drafts might be evaluated as products as long as the assistance role is apparent. When the aim is unaccompanied grammatical control, one should have an extra unassisted task. When the goal is spontaneous speaking to people, it will at some point involve human interaction and unexpected twists.

An independent attempt, then a series of targeted feedback followed by the need to briefly explain the accepted and rejected changes, and a task on another topic to test the transfer would appear as a writing sequence. In speaking, artificial intelligence may help in supporting a rehearsal and a human dialogue of randomised questions, and assessment can identify how comprehensible, fluent, and correct it is, as well as interactive responsiveness. It should be locally assessed whether access is improved and teacher workload is reduced; there is no universal saving in the corpus.

4.8 Priorities for primary research

Research aimed at isolating the feedback provider needs to be an equivalent of task, timing, dose of practice, and teacher involvement, but package research needs to characterize and assess a complete package. In the distribution of classes, there must be sufficient replication at the class level, and analysis of that replication cannot be done without such allocation. The accounting of participants must report separately on recruitment, allocation, exposure, withdrawal, and analysis.

Performance should be differentiated into immediate assistance-free and delayed performance and transfer, and access to the tools during any follow-up interval recorded. Reported estimations must include uncertainty intervals, and reported comparisons not declared significant need to be an uncertain comparison unless an adequately-powered equivalence design was outlined. Process data include revisions, explanations, and conversational repairs, and can bridge what learners get to what they do with it.

4.9 Limitations

The main limitation is that a 40-record frame that had been previously curated was utilized, with its original search history not being stored, and therefore it cannot be replicated as a database query. Eighteen records could not be retrieved. Thirteen of these

were contained in subscription or hybrid journals containing some of the main sources of computer-assisted language learning and second language writing research, and five in fully open-access journals where access was hindered or incomplete.

The 20 retained reports were all accessed via open access channels. Access conditions thus influenced the success of retrieval and not eligibility alone, and the bias is probably directional and not random. In case the subscription outlets release more restrictive designs, the set retained may under-represent those studies. There is no justification to make a claim of completeness, of the more commonality of positive studies or of the most common distribution of languages worldwide; hence the designation of the paper as a systematised review.

The second limitation is related to the conduct of reviews. One reviewer carried out retrieval, extraction, appraisal, and writing without independent duplication, and no agreement statistics can be reported. Independent duplicate screening and extraction could not be done within the resources of this reassessment. A single-reviewer conductor exposes the risk that undetected errors may be in the eligibility decisions, extracted values, and domain judgements, as well as lacks the mechanism of a second judgement reviewing the borderline decisions.

In connection with this, the evidence provided and the interpretation adopted by the reviewer cannot be completely dislodged. The coded framework, the appraisal domains, and the choice of illustrative contrasts were planned and implemented by the same individual, resulting in the narrative synthesis that bears interpretive loads that cannot be audited by a reader solely by reading the original studies. Appendices A, B, E and F are given, where each of the judgements could be tracked to a location in the source and verified.

Studies included also varied significantly in design and outcome, and some reported that the information they had was inadequate to perform any quantitative synthesis. Narrative comparison does not give up those differences, but is interpretive and gives no pooled magnitude, product ranking, or probability of personal gain. The only available interview evidence was in published accounts. Lastly, technology and pedagogy tended to be uncombustible, incomplete version information, and short follow-up and familiarity with tasks restricted conclusions about long-term independent competence.

5. Conclusion

Three conclusions can be made in the reformed corpus of 20 reports. First, assisted performance does not equate to independent acquisition: most positive findings involve revised products or end-of-intervention performance and only three studies involved measurements of performance post-delay. Second, interpretations depend on comparator quality: few benefits were most consistently found when artificial intelligence was run against human-active feedback, and numerous positive studies were in packages. Third, lower anxiety and positive attitudes are both important in themselves, but did not consistently correlate with better measured performance.

These conclusions are specific to the reassessed corpus and not to the literature of the period, in general. They advocate teacher-scaffolded applications of artificial intelligence to practice and feedback, and direct primary research to matched individuals, transparent intervention reporting, and no-assistance assessment.

5.1 Declarations

Conflicts of interest. The author declares no competing interests.

Ethics approval. Not applicable. This review analysed published aggregate data and involved no human participants.

Data availability. All data supporting the findings are contained in the article and its appendices. The retained-study characteristics (Appendix A), source index (Appendix B), exclusion and non-retrieval log (Appendix C), appraisal record (Appendix E) and the study lists behind each appraisal count (Appendix F) constitute the complete dataset.

Use of generative artificial intelligence. Generative artificial intelligence (Claude, Anthropic) was used during revision to support language editing, internal consistency checking of counts across tables and text, and drafting of revised passages. The author reviewed all AI-assisted text, verified data against the original reports, and takes full responsibility for the content. No artificial intelligence tool is listed as an author.

Funding and author contributions are given on the separate title page.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Aljohani, N. J. (2026). ChatGPT in language learning: A systematic review of applications and challenges. *Social Sciences & Humanities Open*, *13*, Article 102357. https://doi.org/10.1016/j.ssaho.2025.102357

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

[3] Ding, D., & Muhyiddin B Yusof, A. (2025). Investigating the role of AI-powered conversation bots in enhancing L2 speaking skills and reducing speaking anxiety: A mixed methods study. *Humanities and Social Sciences Communications*, *12*, Article 1223. https://doi.org/10.1057/s41599-025-05550-z

[4] Divya, T. J., & Alamelu, C. (2026). Integrating DeepL Write and Claude AI to enhance argumentative writing competence of engineering students in India. *Frontiers in Artificial Intelligence*, *9*, Article 1838463. https://doi.org/10.3389/frai.2026.1838463

[5] Ekizoğlu, M., & Demir, A. N. (2025). The role of AI assisted writing feedback in developing secondary students' writing skills. *Discover Education*, *4*, Article 454. https://doi.org/10.1007/s44217-025-00919-3

[6] Ellis, R. (2010). A framework for investigating oral and written corrective feedback. *Studies in Second Language Acquisition*, *32*(2), 335–349. https://doi.org/10.1017/S0272263109990544

[7] Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, *20*, Article 57. https://doi.org/10.1186/s41239-023-00425-2

[8] Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal*, *26*(2), 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

[9] Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., Vedel, I., & Pluye, P. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for Information*, *34*(4), 285–291. https://doi.org/10.3233/EFI-180221

[10] Huang, W., Hew, K. F., & Fryer, L. K. (2022). Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, *38*(1), 237–257. https://doi.org/10.1111/jcal.12610

[11] Jeon, J., & Lee, S. (2024). Can learners benefit from chatbots instead of humans? A systematic review of human–chatbot comparison research in language education. *Education and Information Technologies*, *29*(17), 23329–23360. https://doi.org/10.1007/s10639-024-12725-9

[12] Lai, Z. C.-C. (2026). AI in EFL blended learning: Impact on speaking performance and learning resilience. *Asian-Pacific Journal of Second and Foreign Language Education*, *11*, Article 11. https://doi.org/10.1186/s40862-025-00367-4

[13] Lee, J. (2025). Speaking English with AI or humans: What engages EFL learners more? *English Teaching*, *80*(5), 41–66. https://doi.org/10.15858/engtea.80.5.202512.41

[14] Lee, S., Choe, H., Zou, D., & Jeon, J. (2026). Generative AI (GenAI) in the language classroom: A systematic review. *Interactive Learning Environments*, *34*(1), 335–359. https://doi.org/10.1080/10494820.2025.2498537

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

[16] Li, Y., Shadiev, R., & Chiu, T. K. F. (2026). Applying generative artificial intelligence to task-based language teaching and learning: A systematic review and meta-analysis. *TechTrends*, *70*, 150–163. https://doi.org/10.1007/s11528-025-01140-7

[17] Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). Academic Press.

[18] Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition*, *19*(1), 37–66. https://doi.org/10.1017/S0272263197001034

[19] MacIntyre, P. D., Clément, R., Dörnyei, Z., & Noels, K. A. (1998). Conceptualizing willingness to communicate in a L2: A situational model of L2 confidence and affiliation. *The Modern Language Journal*, *82*(4), 545–562. https://doi.org/10.1111/j.1540-4781.1998.tb05543.x

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

[21] Miao, Y., & Hu, P. (2026). Artificial intelligence-assisted speaking in Spanish as a foreign language: Effects on foreign language emotions and willingness to communicate. *Frontiers in Psychology*, *17*, Article 1877956. https://doi.org/10.3389/fpsyg.2026.1877956

[22] Mohammed, S. J., & Khalid, M. W. (2025). Under the world of AI-generated feedback on writing: Mirroring motivation, foreign language peace of mind, trait emotional intelligence, and writing development. *Language Testing in Asia*, *15*, Article 7. https://doi.org/10.1186/s40468-025-00343-2

[23] Mujeeb, N., Lodhi, M. A., Ali, A., Mashwani, W. K., Mostafa, S. M., & Ahmadini, A. A. H. (2026). Beyond red ink: AI feedback and ESL academic writing development in higher education. *Language Testing in Asia*, *16*, Article 24. https://doi.org/10.1186/s40468-026-00442-8

[24] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *PLOS Medicine*, *18*(3), Article e1003583. https://doi.org/10.1371/journal.pmed.1003583

[25] Qiao, H., & Zhao, A. (2023). Artificial intelligence-based language learning: Illuminating the impact on speaking skills and self-regulation in Chinese EFL context. *Frontiers in Psychology*, *14*, Article 1255594. https://doi.org/10.3389/fpsyg.2023.1255594

[26] Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., Koffel, J. B., & PRISMA-S Group. (2021). PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews. *Systematic Reviews*, *10*, Article 39. https://doi.org/10.1186/s13643-020-01542-z

[27] Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

[28] Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*(2), 129–158. https://doi.org/10.1093/applin/11.2.129

[29] 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*, Article 1260843. https://doi.org/10.3389/fpsyg.2023.1260843

[30] Song, C., & Song, Y. (2026). Corrigendum: Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology*, *17*, Article 1917101. https://doi.org/10.3389/fpsyg.2026.1917101

[31] Sun, W. (2023). The impact of automatic speech recognition technology on second language pronunciation and speaking skills of EFL learners: A mixed methods investigation. *Frontiers in Psychology*, *14*, Article 1210187. https://doi.org/10.3389/fpsyg.2023.1210187

[32] Susoy, Z. (2026). Reducing anxiety and enhancing performance: The impact of AI chatbots versus human facilitation on EFL speaking assessment outcomes. *Frontiers in Psychology*, *16*, Article 1745942. https://doi.org/10.3389/fpsyg.2025.1745942

[33] Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. M. Gass & C. G. Madden (Eds.), *Input in second language acquisition* (pp. 235–253). Newbury House.

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

[35] Wang, D. (2024). Teacher- versus AI-generated (Poe application) corrective feedback and language learners' writing anxiety, complexity, fluency, and accuracy. *The International Review of Research in Open and Distributed Learning*, *25*(3), 37–56. https://doi.org/10.19173/irrodl.v25i3.7646

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

[37] Xiao, Y. (2025). The impact of AI-driven speech recognition on EFL listening comprehension, flow experience, and anxiety: A randomized controlled trial. *Humanities and Social Sciences Communications*, *12*, Article 425. https://doi.org/10.1057/s41599-025-04672-8

[38] Xodabande, I., Shiri, S., & Zohrabi, M. (2025). Exploring the impacts of an AI-driven instructional intervention on Iranian EFL learners' pronunciation skill development. *Discover Education*, *4*, Article 307. https://doi.org/10.1007/s44217-025-00782-2

[39] Yan, H., & Singh, M. K. S. (2026). The impact of AI-mediated instruction on speaking proficiency, enjoyment, anxiety, and emotional engagement: A mixed-methods approach. *Humanities and Social Sciences Communications*, *13*, Article 568. https://doi.org/10.1057/s41599-026-06705-2

Appendix A. Characteristics of retained empirical reports

Table A1: Characteristics of retained empirical reports

Report (ID)	Setting and design	Arrangement and comparator	Timing and outcomes
Sun (2023) P01	China; n = 61; two classes	Speech recognition plus peer correction; comparison class	14 weeks; speaking and pronunciation
Qiao and Zhao (2023) P02	China; n = 93; controlled platform study	Platform package; comparison instruction	13 weeks; speaking and self-regulation
Wei (2023) P03	China; n = 60; two intact classes	Platform package; comparison class	10 weeks; achievement and self-report
Song and Song (2023) P04	China; n = 50; two groups	Generative model feedback; conventional instruction	12 weeks; writing and motivation
Escalante et al. (2023) P06	Asia–Pacific university; n = 48 and n = 43	Generative model vs. tutor feedback; separate preference study	6-week performance study; preference study
Mahapatra (2024) P07	India; n = 72 analysed of 134 consenting	Generative model formative feedback; comparison instruction	1 month; post and delayed writing tests
Wang (2024) P13	China; n = 75 stated, table counts inconsistent	Generative model; teacher feedback; no-feedback condition	14 sessions; writing and anxiety
Mohammed and Khalid (2025) P17	Iraq; n = 322; comparative mixed methods	Tool unspecified; comparison feedback	10 weeks; writing and affect
Xodabande et al. (2025) P18	Iran; n = 60; two assigned groups	Voice-based generative model; electronic dictionary	3 weeks; pronunciation; 2-week delay
Ekizoğlu and Demir (2025) P19	Türkiye; n = 60 secondary learners; two groups	Automated feedback plus teacher support; teacher feedback	8 weeks; writing
Alnemrat et al. (2025) P20	Jordan; n = 120; feedback by proficiency group	Generative model; teacher feedback	Draft and revision; argumentative writing
Xiao (2025) P22	China; n = 84; reported randomised trial	Speech technology package; comparison instruction	8 weeks; listening; 3-week follow-up
Ding and Muhyiddin B Yusof (2025) P23	China; n = 60; two intact groups	Conversation application; alternative speaking practice	6 weeks; speaking and anxiety
J. Lee (2025) P24	Korea; n = 39; shared learning regimen	Conversation application and peer practice for all learners	14 weeks; speaking and engagement
Miao and Hu (2026) P25	China; n = 69 analysed of 79; Spanish	Conversation application; comparison speaking instruction	16 weeks; affect and willingness to communicate
Mujeeb et al. (2026) P26	Pakistan; n = 400; quasi-experiment	Two tools combined; comparison feedback	8-week schedule; feedback weeks 2–4
Lai (2026) P27	Taiwan; n = 86; two intact classes	Two generative models; conventional instruction	18 weeks; speaking and resilience
Divya and Alamelu (2026) P28	India; n = 40; single group	Two tools combined; no separate comparator	5 days; same-topic assistance-free writing tests
Susoy (2026) P29	Türkiye; n = 48; counterbalanced assessments	Artificial intelligence vs. human assessment facilitation	6-week procedure; 2-week washout; no follow-up
Yan and Singh (2026) P30	China; n = 68; two intact classes	Voice assistant; comparison instruction	10 weeks; speaking and affect

Note. Record identifiers correspond to the original 40-record candidate list. English is the target language except in Miao and Hu (2026). Sample sizes are not summed. Duration describes the intervention or assessment schedule, not a retention interval. Susoy (2026) appeared online in December 2025 in volume 16 and is dated by its 2026 publication.

Appendix B. Source locations and interpretation checks

Table B1: Full-text source index

Report	Source location	Interpretation check
Sun (2023)	https://doi.org/10.3389/fpsyg.2023.1210187	Two classes; recognition and peer correction combined
Qiao and Zhao (2023)	https://doi.org/10.3389/fpsyg.2023.1255594	Allocation and sequencing need cautious interpretation
Wei (2023)	https://doi.org/10.3389/fpsyg.2023.1261955	Two intact classes; composite achievement measure
Song and Song (2023)	https://doi.org/10.3389/fpsyg.2023.1260843	Corrected content-result prose; reconcile quantitative reporting

Report	Source location	Interpretation check
Escalante et al. (2023)	https://doi.org/10.1186/s41239-023-00425-2	Distinct samples and questions; non-significance is not equivalence
Mahapatra (2024)	https://doi.org/10.1186/s40561-024-00295-9	Analysed 72 of 134 consenting learners
Wang (2024)	https://doi.org/10.19173/irrodl.v25i3.7646	Stated group sizes conflict with some table counts
Mohammed and Khalid (2025)	https://doi.org/10.1186/s40468-025-00343-2	Tool identity and feedback arrangements limit attribution
Xodabande et al. (2025)	https://doi.org/10.1007/s44217-025-00782-2	Target-word task; two-week delay
Ekizoğlu and Demir (2025)	https://doi.org/10.1007/s44217-025-00919-3	Automated feedback accompanied by teacher support
Alnemrat et al. (2025)	https://doi.org/10.3389/feduc.2025.1614673	Gain-score comparison not statistically significant, d = 0.10
Xiao (2025)	https://doi.org/10.1057/s41599-025-04672-8	Speech-recognition and text-to-speech descriptions differ
Ding and Muhyiddin B Yusof (2025)	https://doi.org/10.1057/s41599-025-05550-z	Intact groups; inconsistent exposure descriptions
J. Lee (2025)	https://files.eric.ed.gov/fulltext/EJ1494238.pdf	All learners received both forms of practice
Miao and Hu (2026)	https://doi.org/10.3389/fpsyg.2026.1877956	Affective outcomes; no proficiency test
Mujeeb et al. (2026)	https://doi.org/10.1186/s40468-026-00442-8	Feedback phase shorter than research schedule
Lai (2026)	https://doi.org/10.1186/s40862-025-00367-4	Bundled tools; fluency finding differs from overall result
Divya and Alamelu (2026)	https://doi.org/10.3389/frai.2026.1838463	Same topic; no separate comparison group
Susoy (2026)	https://doi.org/10.3389/fpsyg.2025.1745942	Achievement difference not statistically significant; correlations differ between abstract and results tables
Yan and Singh (2026)	https://doi.org/10.1057/s41599-026-06705-2	Intact classes; blinded examiner-rated speaking

Note. Section descriptions identify where the relevant methods or results can be checked. The correction linked to Song and Song (2023) is available at https://doi.org/10.3389/fpsyg.2026.1917101; it revises the prose description of the content finding and adds no new sample.

Appendix C. Excluded and non-retrieved records

Table C1: Full-text exclusions

Record	DOI	Operational reason
P08 – J. Li et al. (2024)	10.1057/s41599-024-03755-2	Essay-scoring comparison without the required learner-use investigation
P12	10.1186/s40468-024-00329-6	Artificial intelligence function distinguishing the assessment intervention insufficiently identifiable

Note. The reason recorded for P12 is insufficient intervention identification, not a claim that the platform lacks artificial intelligence functionality.

Table C2: Records without completed full-text reassessment

Record	DOI	Journal	Access model
P05	10.3390/languages8030212	Languages	Fully open access
P09	10.1016/j.caeai.2024.100313	Computers and Education: Artificial Intelligence	Fully open access
P10	10.1016/j.caeai.2024.100270	Computers and Education: Artificial Intelligence	Fully open access
P11	10.1016/j.rmal.2024.100116	Research Methods in Applied Linguistics	Subscription or hybrid
P14	10.1016/j.caeai.2024.100271	Computers and Education: Artificial Intelligence	Fully open access
P15	10.1016/j.system.2024.103505	System	Subscription or hybrid
P16	10.1016/j.jslw.2025.101185	Journal of Second Language Writing	Subscription or hybrid
P21	10.3390/educsci15020232	Education Sciences	Fully open access
P31	10.1016/j.system.2024.103254	System	Subscription or hybrid
P32	10.1016/j.system.2024.103259	System	Subscription or hybrid
P33	10.1016/j.system.2024.103533	System	Subscription or hybrid

Record	DOI	Journal	Access model
P34	10.1080/09588221.2024.2430761	Computer Assisted Language Learning	Subscription or hybrid
P35	10.1080/09588221.2025.2478271	Computer Assisted Language Learning	Subscription or hybrid
P36	10.1080/09588221.2025.2453191	Computer Assisted Language Learning	Subscription or hybrid
P37	10.1016/j.system.2025.103828	System	Subscription or hybrid
P38	10.1016/j.system.2025.103805	System	Subscription or hybrid
P39	10.1111/flan.12790	Foreign Language Annals	Subscription or hybrid
P40	10.1080/09588221.2026.2613219	Computer Assisted Language Learning	Subscription or hybrid

Note. Thirteen of the 18 records appear in subscription or hybrid journals and five in fully open-access journals, where retrieval was blocked or incomplete through the routes used. All 20 retained reports were obtained through open-access routes. Inclusion of a DOI in this log does not certify eligibility.

Appendix D. Retrieval and preparation record

Retrieval scope and date

Targeted retrieval was completed on 10 September 2026 against the 40-record frame. Direct DOI and publisher routes were attempted, followed where useful by exact DOI strings, title and author queries with a format term, and title queries restricted to a repository domain. These describe an item-level retrieval procedure, not a database-ready Boolean strategy or an archived log of every query result.

Information sources

Publisher routes included Frontiers, Springer Nature, IRRODL, Elsevier, Taylor & Francis, Wiley and MDPI. Repository attempts included the ERIC full-text repository and targeted institutional repository domains. Publisher blocking, rate limits, abstract-only access and unproductive repository results prevented complete reassessment. The disposition reconciles as 40 = 20 retained + 2 excluded + 18 not retrieved.

Review independence

Screening, extraction, appraisal and interpretation were conducted by a single reviewer, and no inter-rater agreement statistic was calculated. Each extraction in Appendices A, B and E was checked against the full-text section identified in Appendix B, but repeated checking by the same reviewer is not independent duplicate screening.

Data and reproducibility

The review uses published aggregate reports and presents no newly collected participant data. No preregistration, complete database export, independent appraisal file or meta-analysis code exists for the reassessment. The appendices support inspection of the defined synthesis but do not establish comprehensive search reproducibility.

Appendix E. Domain-level appraisal of retained reports

Table E1: Domain judgements for the 20 retained reports

Report	Allocation and baseline	Free of major co-intervention	Exposure and adherence reported	Complete outcome data	Outcome measurement independence	Reporting consistency	Claim matched to design
Sun (2023)	N	N	Y	Y	?	Y	N
Qiao and Zhao (2023)	?	N	?	Y	?	Y	N
Wei (2023)	N	N	Y	Y	?	Y	N
Song and Song (2023)	Y	Y	Y	Y	?	N	Y
Escalante et al. (2023)	Y	Y	Y	Y	Y	Y	Y
Mahapatra (2024)	Y	Y	Y	N	Y	Y	Y
Wang (2024)	Y	Y	Y	?	?	N	Y
Mohammed and Khalid (2025)	?	?	?	Y	?	Y	N
Xodabande et al. (2025)	Y	Y	Y	Y	Y	Y	Y
Ekizoğlu and Demir (2025)	Y	N	Y	Y	?	Y	N
Alnemrat et al. (2025)	Y	Y	Y	Y	Y	Y	Y
Xiao (2025)	Y	N	Y	Y	Y	?	N

Report	Allocation and baseline	Free of major co-intervention	Exposure and adherence reported	Complete outcome data	Outcome measurement independence	Reporting consistency	Claim matched to design
Ding and Muhyiddin B Yusof (2025)	N	Y	N	Y	?	N	N
J. Lee (2025)	N	N	Y	Y	?	Y	N
Miao and Hu (2026)	Y	Y	Y	N	Y	Y	Y
Mujeeb et al. (2026)	N	N	Y	Y	?	Y	N
Lai (2026)	N	N	Y	Y	?	Y	N
Divya and Alamelu (2026)	N	N	Y	Y	?	Y	N
Susoy (2026)	Y	Y	Y	Y	Y	N	Y
Yan and Singh (2026)	N	Y	Y	Y	Y	Y	N

Note. Y = domain adequately addressed; N = problem identified; ? = insufficient information reported. A '?' indicates a reporting gap, not evidence that a study was poorly conducted. Judgements were made by a single reviewer against the sections listed in Appendix B. 'Claim matched to design' records whether a report's conclusions stay within what its comparator and assessment can support; N does not mean the results are incorrect. Susoy (2026) is coded N for reporting consistency because its abstract and results tables give different correlation values.

Appendix F. Reports behind each appraisal count

Table F1: Reports behind each appraisal count

Domain	Problem identified (N)	Not reported (?)
Allocation and baseline	8: Sun (2023); Wei (2023); Ding and Muhyiddin B Yusof (2025); J. Lee (2025); Mujeeb et al. (2026); Lai (2026); Divya and Alamelu (2026); Yan and Singh (2026)	2: Qiao and Zhao (2023); Mohammed and Khalid (2025)
Free of major co-intervention	9: Sun (2023); Qiao and Zhao (2023); Wei (2023); Ekizoğlu and Demir (2025); Xiao (2025); J. Lee (2025); Mujeeb et al. (2026); Lai (2026); Divya and Alamelu (2026)	1: Mohammed and Khalid (2025)
Exposure and adherence reported	1: Ding and Muhyiddin B Yusof (2025)	2: Qiao and Zhao (2023); Mohammed and Khalid (2025)
Complete outcome data	2: Mahapatra (2024); Miao and Hu (2026)	1: Wang (2024)
Outcome measurement independence	0: none	12: Sun (2023); Qiao and Zhao (2023); Wei (2023); Song and Song (2023); Wang (2024); Mohammed and Khalid (2025); Ekizoğlu and Demir (2025); Ding and Muhyiddin B Yusof (2025); J. Lee (2025); Mujeeb et al. (2026); Lai (2026); Divya and Alamelu (2026)
Reporting consistency	4: Song and Song (2023); Wang (2024); Ding and Muhyiddin B Yusof (2025); Susoy (2026)	1: Xiao (2025)
Claim matched to design	12: Sun (2023); Qiao and Zhao (2023); Wei (2023); Mohammed and Khalid (2025); Ekizoğlu and Demir (2025); Xiao (2025); Ding and Muhyiddin B Yusof (2025); J. Lee (2025); Mujeeb et al. (2026); Lai (2026); Divya and Alamelu (2026); Yan and Singh (2026)	0: none
Delayed or follow-up assessment present (Appendix A)	3: Mahapatra (2024); Xodabande et al. (2025); Xiao (2025)	–

Note. Generated from the same coded record as Appendix E and Table 4, so that every count in the text can be traced to named reports.