International Journal of Learning, Teaching and Educational Research
Vol. 25, No. 8, pp. 385-407, August 2026
https://doi.org/10.26803/ijlter.25.8.15
Received May 24, 2026; Revised Jul 13, 2026; Accepted Jul 16, 2026

Artificial Intelligence for EFL Speaking Development: A Systematic Review of Empirical Studies (2018–2026)

Paul Gonzalez-Torres[*] and Eliana Pinza-Tapia
Universidad Técnica Particular de Loja
Loja, Ecuador

Abstract. Despite the fast expansion of artificial intelligence (AI) in English as a foreign language (EFL) education, evidence regarding its effectiveness in supporting speaking development remains fragmented, with limited synthesis of AI tool types, methodological quality, learning outcomes, and research trends. This systematic review examined empirical studies on AI-supported EFL speaking development and was published between 2018 and 2026. A search of *Scopus* and *Web of Science* was conducted using keywords related to AI, EFL, and speaking skills, following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. Data from 33 studies that met the inclusion criteria were analyzed according to established coding categories aligned with the research questions. The findings indicate that AI tools include conversational agents, speech recognition systems, intelligent tutoring systems, generative AI, and multimodal applications, which each support different aspects of speaking development. Speech recognition tools mostly improved pronunciation, while conversational and generative AI mainly enhanced fluency, confidence, and autonomous speaking practice. Affective gains, including less speaking anxiety and increased willingness to communicate, were more consistent than improvements in more complex speaking skills. This review also found that the evidence base is methodologically inconsistent, concentrated in higher education and specific geographical regions, and increasingly dominated by generative AI technologies.

Keywords: EFL; artificial intelligence; language learning technologies; speaking skills; systematic review

Citation : Gonalez-Torres, P., & Pinza-Tapia, E. (2026). Artificial Intelligence for EFL Speaking Development: A Systematic Review of Empirical Studies (2018–2026). *International Journal of Learning, Teaching and Educational Research*, *25*(8), 385–407. https://doi.org/10.26803/ijlter.25.8.15

[*]Corresponding author: Paul Gonzalez-Torres; *pfgonzalez@utpl.edu.ec*

1. Introduction

The rapid advancement of artificial intelligence (AI) has significantly transformed educational practices, particularly in the field of English as a foreign language (EFL). AI tools enable learners to engage in interactive, personalized, and feedback-rich environments that were previously difficult to access in traditional classrooms (Crompton & Burke, 2023; Ma et al., 2024; Zawacki-Richter et al., 2019).

These technologies can be broadly classified into traditional AI systems (e.g., speech recognition, automatic speech recognition (ASR), and intelligent tutoring systems) that provide structured instruction and pronunciation feedback, and generative AI tools (e.g., large language models and conversational chatbots), that enable contextually adaptive, open-ended interaction (Li & Zhao, 2026; Rafique et al., 2025). This increase in potential has generated an increasing number of empirical studies (e.g., X. Wang & Wen, 2025; C. Zhang et al., 2024), to position AI as a promising solution for persistent challenges in language instruction.

Speaking is widely recognized as one of the most complex language skills, and requires the integration of linguistic, cognitive, and affective components (Bhar, 2026). EFL learners frequently face hurdles, such as limited authentic oral interaction, communication anxiety, and insufficient feedback (Ghelichli, 2022). In this context, AI-driven tools offer solutions by providing learners with increased opportunities for practice, immediate corrective feedback, and low-anxiety environments that may foster willingness to communicate.

From a theoretical perspective, the integration of AI aligns with several complementary perspectives in second language acquisition. Interactionist approaches emphasize meaningful interaction, feedback, and negotiation of meaning for facilitating language development (Loewen & Sato, 2018; Plonsky & Ziegler, 2016), while the output hypothesis highlights how language production enables learners to notice gaps in their knowledge (Swain, 2005). Socio-cognitive perspectives stress the role of learner engagement, context, and affective factors in shaping language development (Atkinson, 2011), and the affective filter hypothesis suggests that reduced anxiety and increased confidence facilitate language acquisition (Krashen, 2013). Self-determination theory emphasizes that autonomy, competence, and intrinsic motivation support sustained engagement and learning (Godwin-Jones, 2021; Ryan & Deci, 2000). Together, these perspectives suggest that AI may support speaking development through opportunities for increased interaction, providing individualized feedback, and reducing affective barriers.

Despite increasing research attention, the evidence base on AI-supported EFL speaking remains fragmented and methodologically inconsistent. Studies vary considerably in terms of AI tools used, research designs and outcome measures applied, and range from controlled quasi-experimental studies (e.g., Al-Husban, 2025; Behforouz et al., 2026; Elov et al., 2025) to one-group pretest–posttest designs (Farooqi, 2025; Giang et al., 2025). The diversity of AI tools available, including generative chatbots, speech recognition systems, and conversational agents (e.g., Aldosari, 2024; Benaissa et al., 2025; Dakhil et al., 2025; Duong & Suppasetseree, 2024; Phanwiriyarat et al., 2025), complicates comparisons further because these technologies perform different pedagogical functions. Outcome measures are also inconsistent, with some studies employing standardized speaking assessments while others rely primarily on self-reported or affective measures (Alsalem, 2024; Yıldız, 2024). Other limitations include the limited use of control groups, insufficient reporting of instrument reliability, short intervention periods, or an absence of longitudinal evidence (e.g., Mohammadzadeh & Sarkhosh, 2018). Moreover, AI is frequently implemented as a standalone practice tool rather than as part of pedagogically structured learning environments, despite evidence that guided and socially embedded approaches produce stronger learning outcomes (Fathi et al., 2024; Zou, Du et al., 2023).

Although previous systematic reviews examined AI in EFL and language education more broadly, they are limited in scope, technological focus, or publication period. For example, earlier reviews synthesized AI applications in English language teaching or general EFL contexts (Zhai & Wibowo, 2023), and others are centered on dialogue systems, chatbots, or instructional design rather than speaking development specifically (Bhar, 2026; Du & Daniel, 2024). More recent reviews of AI-supported speaking focused on particular technologies or pedagogical dimensions (e.g., Li & Zhao, 2026; Xing & Saeed, 2025).

Therefore, given the rapid expansion of generative AI and speech-enabled technologies (Huang et al., 2026; Kim et al., 2021), an updated systematic review that focuses explicitly on EFL speaking development is needed to synthesize evidence on AI tool types, methodological approaches, learning outcomes, and remaining research gaps. Unlike previous reviews, the present study integrates evidence across multiple AI tool categories and provides a theoretically informed, up-to-date synthesis of recent empirical research on EFL speaking development. This is important not only for advancing theoretical understanding of AI-mediated language learning but also for providing evidence and guidance for educators. Thus, this systematic review addresses the following research questions (RQ):

RQ1: What types of AI tools are used to support the development of English-speaking skills in EFL contexts?

RQ2: What research designs and assessment methods are used to evaluate EFL speaking development with AI tools?

RQ3: What effects do AI tools have on the development of the English-speaking skills of EFL learners?

RQ4: What trends can be identified in the use of AI tools for developing EFL speaking skills over time?

2. Methodology

This systematic review was conducted to synthesize empirical evidence on the effects of AI on the development of EFL speaking skills of learners in secondary and higher education. The PRISMA 2020 statement guidelines (Page et al., 2021) were followed to minimize selection bias, enhance transparency, and improve the reproducibility of the review process. In accordance with these guidelines, a structured approach was implemented, including the explicit specification of search strategies, eligibility criteria, and data extraction procedures.

In addition, ethical considerations were carefully observed throughout the review process. All included studies are appropriately cited to acknowledge original work, and findings are reported accurately. The synthesis aimed to reflect the original authors' interpretations while maintaining transparency and academic integrity. The potential for biases in study selection, data extraction, and interpretation was minimized through the application of predefined criteria and collaborative review between the authors.

2.1 Inclusion and Exclusion Criteria

To ensure consistency, transparency, and replicability throughout the study selection process, standardized inclusion and exclusion criteria were established in advance and applied systematically during the screening and eligibility phases. Studies were considered eligible if they investigated the use of AI technologies in EFL contexts, employed an empirical research design (e.g., experimental, quasi-experimental, mixed methods, or case study), and reported results related to holistic speaking development and/or associated affective variables such as speaking anxiety, confidence, or willingness to communicate. Eligible studies were also required to involve EFL learners in secondary or higher education settings, to have been published in peer-reviewed journals between 2018 and 2026, be written in English, and provide full-text access.

Studies were excluded if they did not involve AI tools, did not address speaking development or oral communication, focused exclusively on language skills other than speaking, or were conducted outside secondary or higher education contexts. Publications such as literature reviews and theoretical papers were also excluded because they did not provide primary empirical evidence. To strengthen the reliability of the evidence base, duplicate records, incomplete reports, studies lacking methodological information, and studies relying on learner perceptions without any connection to speaking outcomes were excluded as well.

The review was limited to secondary and higher education settings, because these contexts are more likely to involve learners who have the cognitive and digital literacy skills required for effective engagement. The publication period (2018–2026) was selected to focus on recent developments in AI, particularly the emergence of generative AI and speech-enabled technologies that are minimally represented in earlier studies. The review was also restricted to English-language publications, because English is the predominant language of international research in educational technology, thereby ensuring consistency in data extraction and analysis.

For a consistent application of eligibility criteria, speaking development was operationalized as a multidimensional construct encompassing fluency, accuracy, lexical and grammatical range, coherence, and communicative effectiveness. Consequently, studies focusing exclusively on isolated pronunciation features were excluded. However, studies in which pronunciation training was part of a broader assessment of speaking performance were retained, provided that multiple dimensions of speaking ability were evaluated.

Once the inclusion and exclusion criteria were defined, electronic databases were systematically searched, study selection was conducted, and eligibility decisions were based on clearly operationalized criteria to ensure consistency. The review process comprised four phases: identification, screening, eligibility, and inclusion.

2.2 Identification

A systematic search was conducted to identify relevant literature. The search targeted titles, abstracts, and keywords across two major academic databases: *Scopus* and *Web of Science*, which were selected because of their extensive coverage of peer-reviewed research in education, linguistics, and educational technology. These two databases are widely recognized for their rigorous indexing standards and inclusion of high-impact journals. Limiting the search to *Scopus* and *Web of Science* ensured a reliable and manageable corpus.

The search strategy employed Boolean operators to combine key concepts related to AI, language learning, and speaking skills (see Table 1). To ensure alignment with PRISMA 2020 guidelines, explicitly defined search strings were developed and applied to titles, abstracts, and keywords in each database. Although the conceptual framework was similar across databases, search terms and syntax were adapted to account for differences in indexing systems and vocabularies. Variations in AI terminology in the two databases reflected both evolving practices and indexing differences, and complementary terms were used to ensure coverage of emerging and established technologies. Despite these variations, the strategy was consistent, and all retrieved records were subjected to the same screening and eligibility criteria.

The selection of keywords was directly aligned with the inclusion criteria of the review. This alignment ensured conceptual consistency between the search strategy and the study selection process. The search in *Scopus* was conducted on March 25, 2026, and initially produced 1,440 records. Filters were then applied to restrict results to peer-reviewed journal articles in the subject areas of arts and humanities and social sciences, which resulted in 203 records. The search in *Web of Science* was carried out on April 6, 2026, and resulted in 277 initial records. Filters were applied to include only English-language publications and document types classified as articles and early access, which yielded 261 records.

In total 464 records were identified in both databases (Scopus = 203; Web of Science = 261). The bibliographic records were exported in BibTeX (.bib) format and imported into Zotero (a reference management tool that enables automated detection of duplicate records based on metadata fields). The search strings for each database, including keywords, Boolean operators, and field specifications, are presented in Table 1.

Table 1: Search strings for Scopus and Web of Science

Database	Search string
Scopus	TITLE-ABS-KEY(("artificial intelligence" OR "AI" OR "AI tools" OR "generative AI" OR "large language models" OR ChatGPT OR "AI chatbot*" OR "intelligent tutoring systems" OR "adaptive learning" OR "learning analytics" OR "automated feedback") AND ("language learning" OR "English learning" OR "English language teaching" OR EFL OR ESL OR "foreign language learning") AND ("communicative competence" OR "communicative skills" OR speaking OR pronunciation OR "oral communication" OR interaction) AND (student* OR learner* OR "secondary education" OR "high school" OR "higher education" OR universit* OR college*))
Web of Science	TS = (("artificial intelligence" OR "AI" OR "machine learning" OR "deep learning" OR "natural language processing" OR "NLP" OR chatbot* OR "intelligent tutor*" OR "AI-powered" OR "speech recognition" OR "automatic speech recognition") AND ("English as a Foreign Language" OR "EFL" OR "English language learner*" OR "second language learner*" OR "L2 learner*") AND ("speaking skill*" OR "oral skill*" OR speaking OR pronunciation OR fluency OR "oral proficiency" OR "spoken English" OR "oral communication") AND (student* OR learner* OR "secondary education" OR "high school" OR "higher education" OR universit* OR college*))

2.3 Screening

During this phase, duplicate entries were identified and removed using Zotero. This process resulted in the removal of 84 duplicate records, yielding 380 unique records for further screening. The remaining records were subjected to a screening process involving the examination of titles and abstracts. After applying standardized criteria, 317 records were excluded, and 63 studies were retained for full-text eligibility assessment.

2.4 Eligibility

After full-text retrieval, 63 studies were assessed for eligibility according to the inclusion and exclusion criteria to ensure transparency, reproducibility, and alignment with the objectives of the review. The study selection process was conducted systematically by evaluating each full text against the eligibility criteria, paying particular attention to the abstract, method, results, and conclusions sections. Any uncertainties regarding eligibility were resolved through discussion between the authors to ensure consistency in decision-making. As a result of this process, 30 studies were excluded, and the remaining studies met all eligibility criteria for inclusion in the review.

2.5 Inclusion

A total of 33 studies met all inclusion criteria and were selected for the final synthesis. Data were systematically extracted using a structured matrix which had been developed in Microsoft Excel, according to which studies were coded sequentially (e.g., A1–A33). The extraction framework was explicitly aligned with the research questions to ensure that key dimensions were consistently captured across all studies. Standardized coding categories were applied to each dimension to enhance transparency and analytical consistency (see Appendix 1).

The coding process consisted of a deductive approach derived from the four research questions. AI tools were coded by technology type and pedagogical function. For instance, based on the empirical results reported, speaking outcomes were coded as positive, mixed/no effect or not directly measured. Affective outcomes were coded according to the aspects examined, including anxiety, confidence, motivation, self-efficacy, and willingness to communicate. Methodological characteristics, such as research design, assessment instruments, sample size, and educational context, were also registered to facilitate cross-study comparison. When a study reported multiple AI tools or outcome variables, all relevant categories were recorded. Doubts were cleared up through author discussion until consensus was reached. The study selection process is summarized in Figure 1.

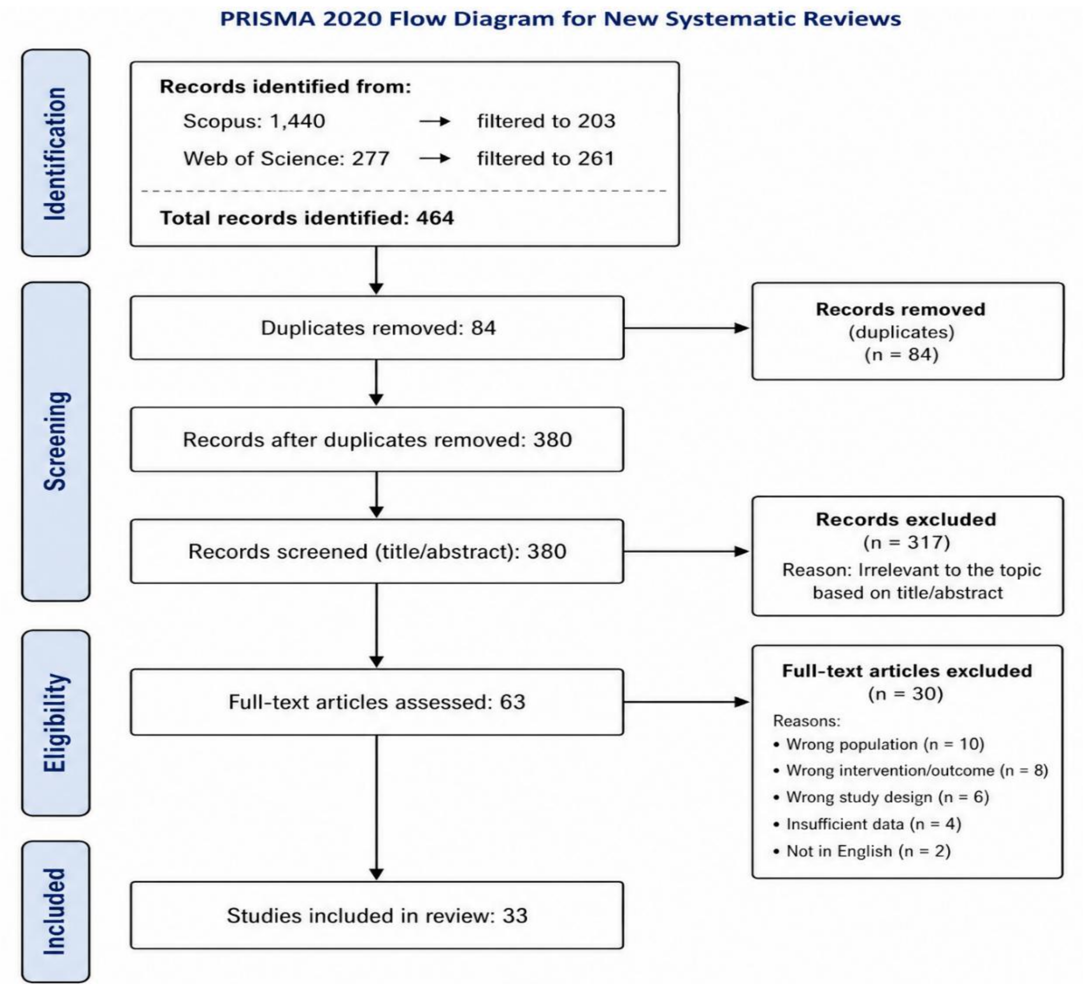

Figure 1: PRISMA flow chart of articles search and selection process

After the coding stage, a combined quantitative and qualitative synthesis was done. Quantitative analysis used descriptive statistics (frequencies, percentages) to identify patterns in AI tool types, research designs, and outcomes. Qualitative analysis involved coding and interpretation of implementation approaches and the pedagogical mechanisms through which AI tools influenced speaking development. This integrative approach facilitated both the identification of generalizable trends and the development of a deep understanding of contextual and functional variations across studies. The findings were organized into thematic categories corresponding to the four research questions: (1) AI tool types; (2) Research designs and assessment methods; (3) Effects on speaking skills; and (4) Trends over time, enabling a structured interpretation of the evidence base.

2.6 Quality Assessment

A quality assessment was conducted using criteria adapted from the Mixed Methods Appraisal Tool (MMAT; Hong et al., 2018). The assessment considered the clarity of research objectives, appropriateness of the research design, transparency of data collection procedures, validity and reliability of measurement instruments, and rigor of data analysis (see Appendix 2).

These criteria were applied across all included studies, and any uncertainties in the evaluation process were resolved through author discussion to maintain consistency and reliability. Because the MMAT criteria were adapted to support methodological evaluation rather than formal study ranking, no numerical quality scores were assigned, and no studies were excluded or weighted on the basis of quality. Instead, the appraisal was used to contextualize the interpretation of the findings and identify methodological strengths and limitations across the evidence base, thereby ensuring that the synthesis was grounded in credible and methodologically strong evidence.

3. Results

The empirical studies selected (see Appendix 1) reflect the recent emergence and accelerated expansion of AI-driven approaches, particularly following advances in generative AI. Within the dataset, considerable variability was observed in technological focus, research design, and assessment practices. This diversity highlights the multidimensional nature of speaking development. While AI integration is a recurring theme, the strength of findings varies across tool types, outcome measures, and methodological approaches. In the sections that follow, we address the results related to the research questions of this review and their respective variables.

3.1 Types of AI Tools Used in EFL Speaking Development

The analysis identified four main categories of AI tools that support EFL speaking development, classified by tool category, core function, and pedagogical use. This categorization is based on the primary pedagogical function of each tool as implemented in the studies, rather than its underlying computational architecture. Findings show that generative AI tools were the most frequently used (36.4%), followed by speech recognition systems (24.2%) and conversational chatbots (24.2%). Multimodal generative AI, intelligent tutoring systems, and mixed or multiple tools (combined use) were minimally represented (see Table 2).

Table 2: Distribution of AI tool categories

Pedagogical AI tool category	Number of studies	%	Representative studies
Generative AI (ChatGPT, CallAnnie, Doubao, Gemini, etc.)	12	36.4	A4, A7, A10, A15, A16, A17, A19, A21, A22, A26, A11, A12
Speech recognition/ASR	8	24.2	A2, A3, A8, A13, A14, A23, A28, A33
Conversational AI (Chatbots)	8	24.2	A1, A5, A20, A25, A27, A29, A31, A32
Generative AI + VR/Multimodal	2	6.1	A6, A18
Intelligent Tutoring System	1	3	A9
Mixed/Multiple tools	2	6.1	A11, A12
Total	**33**	**100**	

Note: The category Mixed/Multiple tools refer to studies that employed more than one AI-based system, typically combining different functionalities such as conversational agents, speech recognition, or automated feedback tools, or integrating multiple AI platforms to support speaking development.

Across the selected studies, three dominant strands can be identified: generative AI, ASR, and conversational chatbots. Generative AI is the most widely adopted, while ASR tools are consistently used in pronunciation-focused tasks, and chatbots are commonly employed for interactive speaking practice. The distribution of tool types indicates a strong emphasis on interaction-oriented technologies, with studies implementing AI to support speaking through feedback, dialogue, and simulated communication contexts.

3.2 Research Designs and Assessment Methods

The methodological profile of the reviewed studies shows variation in research designs and assessment approaches. As shown in Table 3, most studies employed quasi-experimental (54.5%) or mixed methods designs (33.3%). Other designs, including survey-only, case study, cross-sectional, and one-group pre/post designs, are minimally represented.

Table 3: Distribution of research designs

Research design	Number of studies	%	Representative studies
Quasi-experimental	18	54.54	A1, A2, A6, A9, A10, A13, A14, A15, A18, A19, A20, A21, A23, A25, A27, A29, A30, A31
Mixed methods	11	33.33	A3, A4, A5, A8, A11, A17, A22, A24, A28, A32, A33
Quantitative (survey only)	1	3.0	A7
Case study	1	3.0	A12
Cross-sectional survey	1	3.0	A26
One-group pre/post test	1	3.0	A16
Total	**33**	**100**	

Assessment practices primarily include speaking tests (75.8%) and questionnaires (72.7%), often used in combination. Interviews are also employed in nearly half the studies (48.5%) as a qualitative complement. Standardized speaking measures are used inconsistently, with many studies relying on researcher-designed instruments. Affective constructs (e.g., motivation, anxiety, and willingness to communicate) are widely measured. However, only 27.3% of studies integrate both linguistic and affective outcomes within a single design. Moreover, reporting practices vary across studies. Inter-rater reliability for speaking assessments is reported in 12.1% of studies (A13, A14, A15, A17), while questionnaire validation is inconsistently documented. Where reported, internal consistency measures are generally acceptable.

3.3 Effects of AI Tools on Speaking Skills

The reviewed studies report a range of outcomes related to EFL speaking performance and affective variables. A majority of studies (60.6%) report improvements in directly measured speaking performance, most often in fluency, followed by pronunciation and grammatical accuracy. A smaller proportion (15.2%) report mixed or non-significant results. In 24.2% of studies, speaking performance is not directly measured, and outcomes are based on self-reported perceptions (see Table 4).

Table 4: Distribution of speaking outcome measurements

Speaking outcome	Number of studies	%	Representative studies
Directly measured (positive effect)	20	60.6	A2, A3, A4, A6, A9, A10, A11, A12, A13, A14, A15, A16, A20, A22, A23, A24, A27, A28, A29, A30
Directly measured (mixed/no effect)	5	15.2	A18, A19, A21, A25, A31
Not directly measured (self-report/affective only)	8	24.2	A1, A5, A7, A8, A17, A26, A32, A33
Total	**33**	**100**	

Reported outcomes vary according to the type of AI tool implemented. Studies using speech recognition systems commonly focused on pronunciation and oral accuracy, whereas those involving conversational chatbots and generative AI more often emphasized fluency, confidence, and willingness to communicate. Multimodal and mixed AI environments report outcomes across both linguistic and affective domains. Affective variables are included in 75.8% of studies, with commonly reported constructs including anxiety, confidence, motivation, and willingness to communicate. Both linguistic and affective measures are included in 45.45% of studies.

3.4 Trends in AI Use for EFL Speaking Development

Research output increased markedly in recent years, with most publications concentrated between 2023 and 2026. Earlier studies appeared sporadically (e.g., 2018, 2021), followed by a sharp rise from 2023 onward, peaking in 2025 (n = 15) and remaining high in early 2026 (n = 6). Figure 2 shows these trends in publication represented by the reviewed studies.

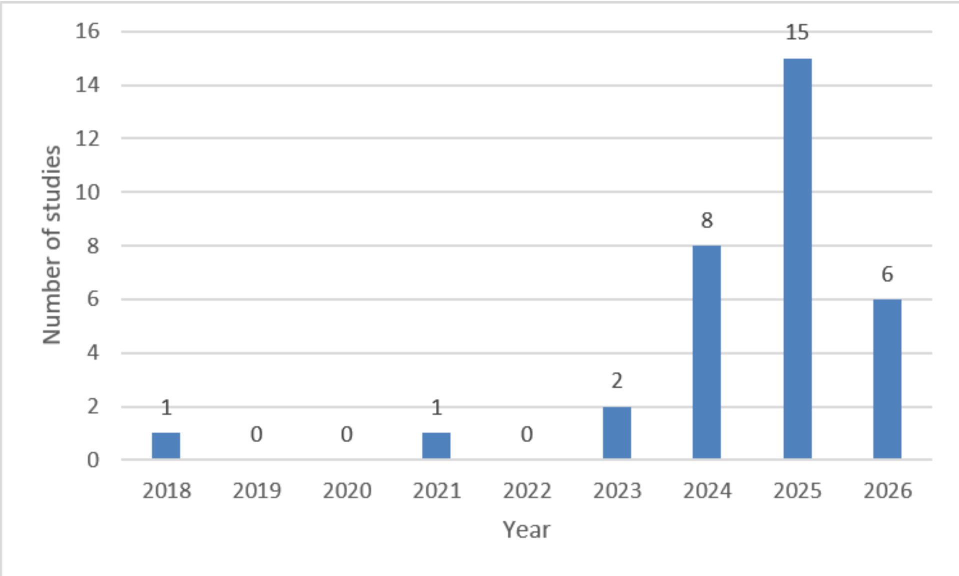

Figure 2: AI in EFL speaking studies: Publications by year

Across studies, AI tools were often implemented in combination. While generative AI tools account for a substantial proportion of studies (A4, A7, A10, A11, A12, A15, A16, A17, A19, A21, A22, A26), in several studies, AI tools were integrated with conversational or speech-based functionalities. Speech recognition tools are also consistently represented across the dataset (A2, A3, A8, A13, A14, A23, A28, A33). Geographically, studies are concentrated in a limited number of countries. China accounts for the largest share (n = 12), followed by Saudi Arabia and Vietnam (n = 4 each), with the remaining studies distributed across a small number of contexts (see Figure 3). There is no representation of Europe or North America, and limited representation from Africa (A29).

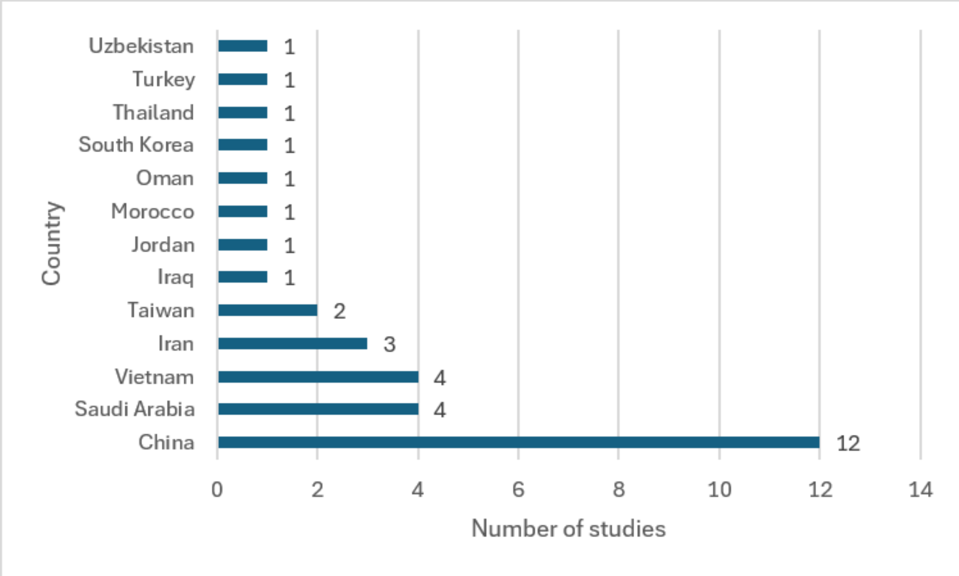

Figure 3: Geographic distribution of studies

An even stronger imbalance is observed in education levels. Nearly all studies are situated in higher education contexts (n = 32), with only one study conducted in secondary education. This situation suggests that current evidence is largely restricted to higher education learners, which leaves the applicability of AI tools to younger populations substantially underexplored.

4. Discussion

With respect to the first research question, the findings indicate a functionally diverse ecosystem of AI tools, differentiated by their pedagogical affordances rather than their technological architecture. Generative AI tools, particularly those based on large language models, dominate recent research and are often portrayed as transformative for speaking practice (Aldosari, 2024; Elov et al., 2025; Yıldız, 2024). However, their prominence appears to be driven more by accessibility and rapid adoption than by consistently demonstrated effectiveness. In contrast, speech recognition systems show more stable links to improvements in pronunciation (Al-Husban, 2025; Ebadi et al., 2025), while conversational chatbots primarily support fluency, confidence, and willingness to communicate rather than broader communicative competence.

These patterns suggest that the effectiveness of AI depends on the alignment between tool capabilities and learning objectives. From an interactionist perspective (Loewen & Sato, 2018; Plonsky & Ziegler, 2016), tools differ in their capacity to provide meaningful interaction, feedback, and opportunities for negotiation of meaning. Speech recognition technologies are effective for pronunciation because of their immediate corrective feedback, whereas conversational AI encourages extended language production but offers less targeted support for complex skills. From a socio-cognitive perspective, language development is shaped by interaction, context, and guided support (Atkinson, 2011). This implies that AI tools are most effective when they are integrated into broader instructional environments rather than used in isolation.

In relation to the second research question, the methodological profile of the reviewed studies limits the strength and consistency of the evidence base. Although quasi-experimental and mixed methods designs are common, the field remains characterized by short intervention durations, little longitudinal evidence, and inconsistent measurement practices. These methodological differences likely contribute to varied findings across studies, which make it difficult to determine whether differences in outcomes reflect the effectiveness of AI tools or the diversity of research designs.

Quasi-experimental studies with control groups (e.g., Aldosari, 2024; Dakhil et al., 2025; Elov et al., 2025) allow for partial causal inference. However, other studies rely on weaker designs, such as one-group pre/posttests or self-report surveys (e.g., Alsalem, 2024 Farooqi, 2025; X. Wang & Wen, 2025), which limit internal validity. Mixed methods approach (e.g., Juan et al., 2025; Zou et al., 2025) offer valuable insights into learner experiences, but integration across data types is often incomplete.

Measurement practices constrain the evidence base further. While some studies employed validated or rubric-based speaking assessments (e.g., Al-Husban, 2025; W. Zhang, 2025), many rely on self-reported data (e.g., Mudawy, 2025; Yıldız, 2024; Zou et al., 2024), which could introduce potential biases such as social desirability and novelty effects. In addition, inconsistent reporting of reliability and validity reduces comparability across studies. These limitations indicate that, although the field has moved beyond exploratory research, it has not yet achieved the methodological consistency required for strong causal claims or broad generalization.

In relation to the third research question, the findings indicate generally positive but inconsistent effects of AI on EFL speaking development (e.g., Le & Do, 2025; The et al., 2026). Improvements are most consistently observed in fluency and pronunciation, particularly in studies using interactive and speech-based tools (e.g., Farooqi, 2025). These patterns are broadly connected to the output hypothesis (Swain, 2005), which emphasizes the role of production and feedback in language development. However, evidence of more complex speaking abilities, including coherence, grammatical accuracy, and interactional competence, remains limited and inconsistent. This suggests that, while AI tools can support foundational and practice-oriented aspects of speaking, their impact on more complex communicative skills is less clear.

Interpretation of these findings is complicated further by measurement practices. A substantial number of studies rely on self-reported outcomes rather than direct assessments of speaking performance, which makes comparisons across studies difficult and limits the strength of conclusions. In contrast, affective outcomes are consistently positive. Reductions in anxiety and increases in confidence and willingness to communicate are reported widely (Alsalem, 2024; C. Zhang et al., 2024). These findings align with the affective filter hypothesis (Krashen, 2013) and socio-cognitive perspectives that emphasize the role of emotional and contextual factors in language development (Atkinson, 2011). They also resonate with self-determination theory (Ryan & Deci, 2000), because AI-supported environments often promote autonomy, individualized pacing, and a sense of competence.

Notably, affective gains often occur without corresponding linguistic improvements. This pattern suggests that AI tools primarily facilitate conditions for speaking (e.g., by reducing anxiety and increasing engagement) rather than directly producing gains in complex communicative competence. This finding also highlights a potential gap between perceived and actual learning outcomes (C. Wang et al., 2024) and emphasizes the need for clearer distinctions between engagement, confidence, and measurable language development. Finally, addressing the fourth research question, the findings point to a field characterized by rapid growth but also structural concentration. The recent increase in publications reflects the prominent visibility of AI in language education, particularly in the wake of advances in generative AI (X. Wang & Wen, 2025; C. Zhang et al., 2024). However, the pace of research expansion appears to exceed the consolidation of robust empirical evidence.

At the technological level, AI tools are increasingly implemented as hybrid systems, by combining generative, conversational, and speech-based features (Al-Husban, 2025). While this broadens pedagogical possibilities, it also introduces variability in implementation and complicates comparisons across studies. At the contextual level, the evidence base is geographically and educationally concentrated. Research is concentrated in a small number of countries (X. Wang & Wen, 2025; C. Zhang, 2025) and largely situated in higher education, with minimal representation of secondary education (F. Wang et al., 2025) and other learner groups. These concentrations limit the generalizability of findings and restrict understanding of how AI functions across diverse educational settings. These patterns also suggest that the field is advancing faster technologically than it is diversifying empirically, thereby highlighting the need for broader geographic coverage, greater inclusion of non-tertiary learners, and more context-sensitive research designs.

Taken together, this review makes several contributions. It frames AI as a diverse set of tools whose effectiveness depends on pedagogical alignment, identifies a consistent gap between linguistic and affective outcomes, and provides a theoretically informed interpretation grounded in second language acquisition frameworks. Importantly, it emphasizes that AI extends rather than replaces human interaction in language learning, while identifying key methodological and contextual gaps for future research.

4.1 Implications

From a pedagogical perspective, these findings suggest that AI technologies are most effective when used as targeted, function-specific support rather than comprehensive instructional solutions. Their strengths lie in providing frequent, flexible, and low-anxiety speaking practice, particularly for developing fluency and pronunciation (Ebadi et al., 2025). However, the limited evidence for higher-order speaking skills (Lai, 2026; C. Wang et al., 2024; F. Wang et al., 2025; Zhou et al., 2025) reinforces the continued importance of human interaction in fostering discourse-level competence and socio-pragmatic awareness. AI should, therefore, be integrated as a complementary tool within structured pedagogical frameworks (Fathi et al., 2024; Tang, 2026; Zou, Du et al., 2023, Zou, Guan et al., 2023) rather than positioned as a replacement for teacher-led instruction.

Beyond teaching benefits, the use of AI in language learning raises ethical and practical concerns that require closer attention. Key issues include data privacy risks, learner overdependence on AI at the expense of authentic communication, and the potential for bias and unequal access to advanced tools. These challenges highlight the need for more critical research and clear pedagogical guidelines to ensure that AI is applied responsibly and equitably.

4.2 Limitations

Several limitations should be noted. The studies reviewed were characterized by short intervention periods, inconsistent measurement practices, and limited diversity in participant populations. Additionally, this review was restricted to two databases, included only English-language, full-text publications and focused on overall speaking skills rather than specific subskills, which may have narrowed the scope of analysis and probably excluded some relevant studies.

Moreover, the evidence base is concentrated in higher education and a small number of geographic regions, thereby reducing its relevance to other contexts such as secondary education or diverse learner groups. Given the variability in study designs, AI tools, and outcome measures, the findings should be interpreted with caution. Rather than indicating universally effective outcomes, the results reflect the context-dependent potential of AI to support speaking development.

4.3 Future Research

Future research should place greater emphasis on longitudinal and experimental designs, the use of standardized and validated speaking assessments, and the systematic integration of linguistic and affective measures. More attention should also be given to comparative analyses across different AI tool types to clarify their differential effectiveness. Expanding research into more diverse education contexts and underrepresented regions will be essential for improving external validity and developing a more comprehensive understanding of AI-supported speaking development. In addition, future studies should emphasize greater methodological consistency, transparent reporting, and more robust validation of assessment instruments. Additionally, research should investigate the mechanisms through which affective improvements translate into sustained linguistic development, particularly through longitudinal designs that can capture the dynamics of this relationship.

5. Conclusion

This systematic review examined empirical evidence on the role of AI in EFL speaking development across four research questions. First, AI tools are diverse and serve distinct pedagogical functions rather than constituting a unified instructional approach. At present, generative AI prevails in recent research because of its accessibility and fast development, while speech recognition tools show more consistent effects on pronunciation. Conversational chatbots mainly support fluency, confidence, and autonomous practice, but their contribution to broader communicative competence remains limited. In general, AI is used primarily as a supplementary rather than integrated instructional or assessment tool.

Second, methodological quality has improved, though some problems remain. Although quasi-experimental and mixed methods design prevail, the evidence base is constrained by short interventions, little longitudinal data, and inconsistent assessment practices. Heavy reliance on self-reported measures, often without validated speaking instruments, weakens causal claims and limits comparability across studies.

Third, AI is generally associated with positive outcomes in EFL speaking. The strongest gains appear in pronunciation and fluency, while evidence for more complex speaking abilities remains inconclusive. Affective outcomes are more robust and often exceed linguistic ones, which suggest that AI primarily facilitates conditions for practice rather than directly developing advanced communicative competence. However, these benefits must be interpreted carefully. Increased dependence on AI might reduce opportunities for critical thinking, problem-solving, and creative language use.

Finally, the field is expanding rapidly but remains structurally imbalanced. The fast development occurs mostly because of generative AI and is mostly used in higher education and is only taking place in a few geographical areas. Although increasingly hybrid tools expand pedagogical possibilities, technological innovation continues to outpace empirical validation. Therefore, AI technologies should be treated as a context-dependent resource that works as a complementary rather than a primary pedagogical resource. Future research should prioritize longitudinal, theory-informed studies, standardized assessments, and more diverse learner populations, to strengthen the evidence base.

Conflict of Interest
The authors declare that there are no conflicts of interest regarding the publication of this paper.

6. Acknowledgments

The authors wish to acknowledge the use of ChatGPT in the writing of this paper. This tool was used to support proofreading and to improve the language and grammar of the manuscript. Moreover, the AI-based tool ConceptViz was used to generate Figure 1. The authors confirm that the manuscript is an accurate representation of their work and intellectual contributions. The authors also gratefully acknowledge the support of Universidad Técnica Particular de Loja for the development of this manuscript.

7. References

Aldosari, M. S. (2024). Another world with artificial intelligence in speaking classes: To delve into the influences on willingness to communicate, personal best goals, and academic enjoyment. *Computer-Assisted Language Learning Electronic Journal, 25*(4), 439–463. https://callej.org/index.php/journal/article/view/470

Al-Husban, N. (2025). The impact of AI-assisted language learning tools on augmenting university EFL students' speaking skills in Jordan. *Journal of Applied Learning & Teaching, 8*(1), 116–127. https://jalt.open-publishing.org/index.php/jalt/article/view/2341

Alsalem, M. S. (2024). EFL students' perception and attitude towards the use of ChatGPT to promote English speaking skills in the Saudi context. *Arab World English Journal, 15*(4), 73–84. https://dx.doi.org/10.24093/awej/vol15no4.5

Atkinson, D. (Ed.). (2011). *Alternative approaches to second language acquisition*. Routledge. https://doi.org/10.4324/9780203830932

Bhar, S. K. (2026). Artificial intelligence in EFL speaking instruction: A systematic review of pedagogical design, affective conditions and instructional input. *Encyclopedia, 6*(4), Article 74. https://doi.org/10.3390/encyclopedia6040074

Behforouz, B., Poorghorban, A., & Al Ghaithi, A. (2026). Enhancing oral language skills with Pi AI: A study on speaking and pronunciation gains. *Online Learning*, 30(1), 507–530. https://doi.org/10.24059/olj.v30i1.4928

Benaissa, M., Attou, Y., Seddik, M., & Nfissi, A. (2025). Can AI-based teacher-bots (T-Bots) enhance students' speaking skills? A case study of "My AI" on Snapchat in Moroccan higher education. *Arab World English Journal, 11,* 144–159. https://dx.doi.org/10.24093/awej/call11.9%20%20

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: the state of the field. *International Journal of Educational Technology in Higher Education, 20*(1), Article 22. https://doi.org/10.1186/s41239-023-00392-8

Dakhil, T. A., Karimi, F., Al-Jashami, R. A. U., & Ghabanchi, Z. (2025). The effect of artificial intelligence (AI)-mediated speaking assessment on speaking performance and willingness to communicate of Iraqi EFL learners. *International Journal of Language Testing, 15*(2), 1–18. https://www.proquest.com/docview/3311653540?pq-origsite=gscholar&fromopenview=true&sourcetype=Scholarly%20Journals

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, Article 100230. https://doi.org/10.1016/j.caeai.2024.100230

Duong, T., & Suppasetseree, S. (2024). The effects of an artificial intelligence voice chatbot on improving Vietnamese undergraduate students' English-speaking skills. *International Journal of Learning, Teaching and Educational Research*, *23*(3), 293–321. https://doi.org/10.26803/ijlter.23.3.15

Ebadi, S., Velayati, S., Ramezanzadeh, A., & Salman, A. R. (2025). Exploring the impact of AI-powered speaking tasks on EFL learners' speaking performance and anxiety: An activity theory study. *Acta Psychologica*, *259*, Article 105391. https://doi.org/10.1016/j.actpsy.2025.105391

Elov, B., Kholikov, A., Abdullayeva, I., Mamadiyarov, Z., & Ruzikulova, A. (2025). Effectiveness of AI chatbots in promoting informal speaking proficiency and social pragmatic skills in Uzbekistan: A multi-analysis study. *Computer-Assisted Language Learning Electronic Journal, 26*(7), 137–161. https://doi.org/10.54855/callej.261234567

Farooqi, S. ul H. (2025). Efficacy of AI-generated feedback by SmallTalk2Me for improving speaking skill of Saudi EFL learners. *Forum for Linguistic Studies, 7*(3), 714–728. https://doi.org/10.30564/fls.v7i3.8294

Fathi, J., Rahimi, M., & Derakhshan, A. (2024). Improving EFL learners' speaking skills and willingness to communicate via artificial intelligence-mediated interactions. *System*, *121*, Article 103254. https://doi.org/10.1016/j.system.2024.103254

Ghelichli, Y. (2022). Foreign language anxiety and willingness to communicate: High level vs. low level Iranian EFL learners. *Journal of Applied Linguistics Studies*, *1*(1), 1–9. https://oiccpress.com/jals/article/view/5557

Giang, N. T., Nguyen, T. H., & Tran, T. T. (2025). Using AI-powered tools for enhancing speaking skills of EFL learners. *North American Journal of Psychology, 27*(4), 1015−1016. https://openurl.ebsco.com/contentitem/gcd:190683990?sid=ebsco:plink:crawler&id=ebsco:gcd:190683990

Godwin-Jones, R. (2021). Evolving technologies for language learning. *Language Learning & Technology*, *25*(3), 6–26. https://doi.org/10.64152/10125/73443

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

Huang, P. W., Hwang, Y. L., Hsu, J. L., Peng, C. F., Tsai, C. H., & Wang, C. Y. (2026). The effectiveness of an AI-integrated VR oral training application in reducing public speaking anxiety and interview anxiety. *Computers & Education: Artificial Intelligence, 10*, Article 100514. https://doi.org/10.1016/j.caeai.2025.100514

Juan, W., Binti Ismail, H. H., & Mansor, A. Z. (2025). Enhancing Chinese EFL learners' speaking proficiency through AI-integrated POAs. *Educational Sciences: Theory and Practice*, 2452, 44–57. https://doi.org/10.12738/jestp.2024.2.018

Kim, H., Kim, N. & Cha, Y. (2021). Is it beneficial to use AI chatbots to improve learners' speaking performance? *Journal of Asia TEFL*, *18*(1), 161–178. http://dx.doi.org/10.18823/asiatefl.2021.18.1.10.161

Krashen, S. D. (2013). *Second language acquisition: Theory, applications, and some conjectures*. Cambridge University Press.

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*(1), Article 11. https://doi.org/10.1186/s40862-025-00367-4

Le, V. H. H., & Do, M. T. (2025). Using Copilot to foster utterance fluency of undergraduates: A multiple case study. *Computer-Assisted Language Learning Electronic Journal,* 26(4), 396–417. https://callej.org/index.php/journal/article/view/690

Li, D., & Zhao, Y. (2026). Artificial intelligence applications for oral communication skills in EFL contexts: A systematic review. *The Asia-Pacific Education Researcher*, *35*(2), 255–266. https://doi.org/10.1007/s40299-025-01023-8

Loewen, S., & Sato, M. (2018). Interaction and instructed second language acquisition. *Language Teaching, 51*(3), 285–329. https://doi.org/10.1017/S0261444818000125

Ma, H., Ismail, L., & Han, W. (2024). A bibliometric analysis of artificial intelligence in language teaching and learning (1990–2023): Evolution, trends and future directions. *Education and Information Technologies*, *29*(18), 25211–25235. https://doi.org/10.1007/s10639-024-12848-z

Mohammadzadeh, A., & Sarkhosh, M. (2018). The effects of self-regulatory learning through computer-assisted intelligent tutoring system on the improvement of EFL learner' speaking ability. *International Journal of Instruction, 11*(2), 167–184. https://doi.org/10.12973/iji.2018.11212a

Mudawy, A. M. A. (2025). Exploring EFL learners' perceptions on the use of AI-powered conversational tools to improve speaking fluency: A case study at Majmaah

University. *Forum for Linguistic Studies, 7*(1), 589–598. https://doi.org/10.30564/fls.v7i1.7774

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. *BMJ*, Article 372. https://doi.org/10.1136/bmj.n71

Phanwiriyarat, K., Anggoro, K. J., & Chaowanakritsanakul, T. (2025). Exploring AI-powered gamified flipped classroom in an English-speaking course: A case of Duolingo. *Cogent Education*, *12*(1), 2488545. https://doi.org/10.1080/2331186x.2025.2488545

Plonsky, L., & Ziegler, N. (2016). The CALL-SLA interface: Insights from a second-order synthesis. *Language Learning & Technology*, *20*(2), 17–37. https://doi.org/10.64152/10125/44459

Rafique, J., Dhawan, A., Narjis, S., Debbabi, A. S., & Aslam, Z. (2025). Artificial intelligence in English language teaching: Bridging traditional methods with modern technology. *Journal of Posthumanism*, *5*(12), 240–249. https://doi.org/10.63332/joph.v5i12.3763

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

Swain, M. (2005). The output hypothesis: Theory and research. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 471–483). Routledge.

Tang, J. (2026). AI-assisted speaking affordances, flow experience, and willingness to communicate: A moderated mediation model. *System*, *138*, 103986. https://doi.org/10.1016/j.system.2026.103986

The, N.H., Du, T. T., & Hoang, T. T. (2026). Exploring the impact of Artificial Intelligence on English majors' speaking skills: A case study at a Vietnamese university. *Arab World English Journal (AWEJ) Special Issue* (3), 166-179. https://dx.doi.org/10.24093/awej/AI3.11

Wang, C., Zou, B., Du, Y., & Wang, Z. (2024). The impact of different conversational generative AI chatbots on EFL learners: An analysis of willingness to communicate, foreign language speaking anxiety, and self-perceived communicative competence. *System*, *127*, Article 103533. https://doi.org/10.1016/j.system.2024.103533

Wang, F., Zhou, X., Li, K., Cheung, A. C. K., & Tian, M. (2025). The effects of artificial intelligence-based interactive scaffolding on secondary students' speaking performance, goal setting, self-evaluation, and motivation in informal digital learning of English. *Interactive Learning Environments*, *33*(7), 4633–4652. https://doi.org/10.1080/10494820.2025.2470319

Wang, X., & Wen, Y. (2025). The role of artificial intelligence tools on Chinese EFL learners' speaking skills, speaking anxiety, and motivation. *Acta Psychologica, 261*, Article 105957. https://doi.org/10.1016/j.actpsy.2025.105957

Xing, C., & Saeed, M. A. (2025). A Systematic review on artificial intelligence (AI) technologies in ESL/EFL speaking skills. *International Journal of TESOL Studies*, *8*(2), 240-270. https://doi.org/10.58304/ijts.250908

Yıldız, C. (2024). ChatGPT integration in EFL education: A path to enhanced speaking self-efficacy. *Novitas-ROYAL,* *18*(2), 167–182. https://doi.org/10.5281/zenodo.13861137

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education, 16*, Article 39. https://doi.org/10.1186/s41239-019-0171-0

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*, Article 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*, Article 103259. https://doi.org/10.1016/j.system.2024.103259

Zhang, W. (2025). The impact of AI chatbots on EFL learners' oral proficiency and willingness to communicate. *System, 136*, Article 103919. https://doi.org/10.1016/j.system.2025.103919

Zhou, Q., Hashim, H., & Sulaiman, N. A. (2025). Supporting English speaking practice in higher education: The impact of AI chatbot-integrated mobile-assisted blended learning framework. *Education and Information Technologies*, *30*(10), 14629–14660. https://doi.org/10.1007/s10639-025-13401-2

Zou, B., Du, Y., Wang, Z., Chen, J., & Zhang, W. (2023). An investigation into artificial intelligence speech evaluation programs with automatic feedback for developing EFL learners' speaking skills. *Sage Open, 13*(3). https://doi.org/10.1177/21582440231193818

Zou, B., Guan, X., Shao, Y., & Chen, P. (2023). Supporting speaking practice by social network-based interaction in artificial intelligence (AI)-assisted language learning. *Sustainability*, *15*(4), Article 2872. https://doi.org/10.3390/su15042872

Zou, B., Liviero, S., Ma, Q., Zhang, W., Du, Y., & Xing, P. (2024). Exploring EFL learners' perceived promise and limitations of using an artificial intelligence speech evaluation system for speaking practice. *System*, *126*, Article 103497. https://doi.org/10.1016/j.system.2024.103497

Zou, P. B., Xie, S., & Wang, C. (2025). Students' willingness to communicate (WTC) in using artificial intelligence (AI) technology in English-speaking practice. *International Journal of Information and Communication Technology Education, 21*(1), 1–18. https://doi.org/10.4018/IJICTE.375387

Appendix 1

Coding matrix of included studies

ID	Author(s) (publication year)	AI tool	AI category	Research design	Speaking outcome	Key finding
A1	C. Zhang et al. (2024)	Lora	Conversational AI	Quasi-experimental	Not measured	Positive affective outcomes
A2	Al-Husban (2025)	ELSA Speak	Speech recognition	Quasi-experimental	Positive	Significant improvement across five speaking standards
A3	Zou, Du et al. (2023)	Liulishuo, EAP Talk, et al.	Speech recognition	Mixed methods (no control group)	Positive	Large effect size (d = 1.526); no control group
A4	W. Zhang (2025)	CallAnnie	Generative AI	Mixed methods	Positive	Experimental group outperformed control group; $\eta^2 = 0.27$
A5	Mudawy (2025)	Unspecified	Conversational AI	Mixed methods	Not measured	93% positive perceptions; no objective measure
A6	Huang et al. (2026)	Custom AI-virtual reality	Generative AI + VR	Quasi-experimental	Positive	Word count, errors, sentence length improved
A7	Alsalem (2024)	ChatGPT	Generative AI	Quantitative (survey)	Not measured	Positive perceptions; strong preference for humans
A8	Zou et al. (2025)	EAP Talk	Speech recognition	Mixed methods (no CG)	Not measured	Willingness to Communicate (WTC) improved (d = 0.362); no speaking measure
A9	Mohammadzadeh & Sarkhosh (2018)	Autotutor	Intelligent tutoring systems	Quasi-experimental	Positive	Self-monitoring > seeking-help; retention at 1 month
A10	Aldosari (2024)	ChatGPT	Generative AI	Quasi-experimental	Positive	EG outperformed CG on speaking, WTC, academic enjoyment, personal best goals
A11	Giang et al. (2025)	ChatGPT + ELSA	GenAI + automatic speech recognition	Mixed methods (no CG)	Positive	d = 0.58; lexical resources improved most (d = 0.83)
A12	Le & Do (2025)	Copilot	Generative AI	Case Study (n=3)	Positive	Beginner gained + 20 points; very small sample
A13	Dakhil et al. (2025)	ELSA Speech Analyzer	Speech recognition	Quasi-experimental	Positive	4/5 components improved; pronunciation not significant
A14	Farooqi (2025)	SmallTalk2Me	Speech recognition	Quasi-experimental (no CG)	Positive	International English Language Testing System (IELTS) band +12.12%; fluency +11.18%
A15	Elov et al. (2025)	ChatGPT	Generative AI	Quasi-experimental	Positive	d = 1.92; also measured social pragmatics
A16	X. Wang & Wen (2025)	Doubao	Generative AI	One-group (no CG)	Positive	d = 0.97 (speaking); d = 1.93 (motivation)

ID	Author(s) (publication year)	AI tool	AI category	Research design	Speaking outcome	Key finding
A17	Yıldız (2024)	ChatGPT	Generative AI	Mixed methods	Not measured	Self-efficacy improved (η^2 = 0.886); no speaking measure
A18	C. Wang et al. (2024)	Typebot, D-ID Agent	Generative AI	Quasi-experimental	No Effect	Avatar chatbot improved affect, not speaking
A19	Lai (2026)	ChatGPT, Gemini	Generative AI	Quasi-experimental	Mixed	Pronunciation/accuracy improved; fluency no effect
A20	Duong & Suppasetseree (2024)	Andy English Bot	Conversational AI	Quasi-experimental	Positive	Fluency and accuracy improved
A21	Zhou et al. (2025)	ChatGPT-4	Generative AI	Quasi-experimental	Mixed	Fluency/appropriateness improved; accuracy no effect
A22	Juan et al. (2025)	ChatGPT	Generative AI	Mixed methods	Positive	Fluency and accuracy improved
A23	Zou, Guan et al. (2023)	Liulishuo, EAP Talk, et al.	Speech recognition	Quasi-experimental	Positive	EG with social interaction outperformed CG
A24	Phanwiriyarat et al. (2025)	Duolingo	Generative AI	Mixed methods	Positive	Significant gains in general questions and discussion
A25	F. Wang et al. (2025)	LAIX (Liulishuo)	Conversational AI	Quasi-experimental	Mixed	Vocabulary improved; other skills no significant difference
A26	Tang (2026)	Tencent Yuanbao	Generative AI	Cross-sectional	Not measured	WTC mediated by flow experience
A27	Behforouz et al. (2026)	Pi AI	Conversational AI	Quasi-experimental	Positive	EG outperformed CG in speaking and pronunciation
A28	Ebadi et al. (2025)	SmallTalk2Me	Speech recognition	Mixed methods	Positive	All subskills improved; anxiety reduced
A29	Benaissa et al. (2025)	My AI (Snapchat)	Conversational AI	Mixed methods	Positive	Very large effect (d = 8.61)
A30	The et al. (2026)	Unspecified AI tools	AI-powered tools	Mixed methods	Positive	EG improved from 4.6 to 6.08; CG from 5.15 to 5.5
A31	Kim et al. (2021)	Replika, Andy, Google Assistant	Conversational AI	Quasi-experimental	Mixed	Voice-chat > text-chat for opinion expression
A32	Fathi et al. (2024)	Andy English Chatbot	Conversational AI	Mixed methods	Positive	AI group outperformed face-to-face group
A33	Zou et al. (2024)	EAP Talk	Speech recognition	Mixed methods	Positive (self-report)	Self-reported improvement across multiple dimensions

Appendix 2

Quality assessment summary of included studies

Quality rating	Criteria	n	Representative studies
High	Meets all five criteria with strong evidence	13	A2, A4, A6, A9, A10, A13, A15, A18, A19, A24, A25, A28, A32
High-moderate	Meets four criteria fully; one partially	13	A3, A11, A14, A17, A20, A21, A22, A23, A27, A29, A30, A31, A33
Moderate	Meets three criteria fully; two partially	6	A1, A5, A8, A12, A16, A26
Low-moderate	Meets two criteria fully; three partially	1	A7