

AI Voice-based Chatbots as a Conversational Partner in Practicing English Speaking: A Systematic Literature Review and the Ways Forward

Thi Hien Ho ¹, Nancy Dieu Ngoc Nguyen ², Van Huy Nguyen ^{3*}

¹ Industrial University of Ho Chi Minh City, Vietnam,  0009-0009-7687-8999

² Independent researcher, Vietnam,  0000-0001-9452-9156

³ Hue University of Foreign Languages and International Studies, Vietnam,  0000-0001-6826-6188

* Corresponding author: Van Huy Nguyen (nguyenvanhuy@hueuni.edu.vn)

Article Info

Article History

Received:
28 May 2026

Revised:
29 June 2026

Accepted:
18 July 2026

Published:
10 September 2026

Keywords

AI voice-based chatbots
Chatbot-human interaction
Speaking skills
AI-mediated speaking
EFL

Abstract

The research on AI-mediated language learning has been growing across different educational contexts. Previous systematic reviews have summarized the existing studies on the affordances and challenges of implementing generative AI tools in language education. Some of these reviews covered a range of AI technologies, or Automatic Speech Recognition tools, which relied on pre-programmed speech. However, the reviews of studies focusing on the integration of AI voice-based chatbots (hereafter, AVCs) in speaking skills remain limited. This study reviewed 28 empirical studies published in English between 2020 and 2025, identified from Scopus, Web of Science, Google Scholar, using the PRISMA flowchart and predefined inclusion/exclusion criteria. It aims to examine the trends of research designs in AI-mediated speaking instruction, speaking tasks between chatbots and learners, the impacts, challenges, and mediating factors of AVCs in speaking instruction. The study reveals a consistent trend in research design in this field with the predominance of quasi-experimental mixed-methods research, particularly in EFL contexts where the need to be proficient in speaking skills is substantial, yet the opportunities to communicate with a competent interlocutor are limited. The results also highlight positive impacts of AVCs on facilitating aspects of speaking proficiency, including fluency, pronunciation, vocabulary, and grammar. Several challenges related to technological issues, user experience, and users' affective and behavioral engagement were identified in the selected sources. Mediating factors include learners' language proficiency, affective factors, chatbot-related features, and pedagogical support. The article concludes with some implications and recommendations for diversifying and strengthening the evidence base, particularly in the teaching and learning of speaking skills. This review protocol (IDESR000187) was registered with the International Database of Education Systematic Reviews.

Citation: Ho, T. H., Ngoc Nguyen, N. D., & Nguyen, V. H. (2026). AI voice-based chatbots as a conversational partner in practicing English speaking: A systematic literature review and the ways forward. *International Journal of Technology in Education (IJTE)*, 9(4), 1309-1333. <https://doi.org/10.46328/ijte.8534>



ISSN: 2689-2758 / © International Journal of Technology in Education (IJTE).
This is an open access article under the CC BY-NC-SA license
(<http://creativecommons.org/licenses/by-nc-sa/4.0/>).



Introduction

Language learners have difficulty communicating in English well, even though they have learned speaking skills for a long time, because of several reasons (Chulee et al., 2023). For example, they do not have sufficient opportunities to practice speaking meaningfully inside the classroom (Gan, 2012; Hwang et al., 2024). Other factors, such as large class sizes, limited teacher attention, and a lack of authentic communication situations, also limit the effectiveness of speaking instruction (Ur, 1996). Nunan (2003) further notes that students often experience high levels of anxiety and low confidence when speaking, especially in traditional teacher-led speaking classes, which can hinder participation and limit oral language development. These challenges underscore the need for alternative approaches that provide a more personalized, low-anxiety, and interactive environment for speaking practice.

Recently, educational technologies such as AVCs have served as potential conversational partners that may help teachers and learners resolve the difficulties in practicing speaking skills (Du & Daniel, 2024; Xing & Saeed, 2025). AVCs are tools that combine artificial intelligence and natural language processing to facilitate speech-based interactions that mimic human conversations (Koç & Savaş, 2025; Pérez et al., 2020). Although research on the integration of AVCs in speaking instruction is still in its infancy (Du & Daniel, 2024), the findings mostly support that AVCs can enhance speaking proficiency (Azizimajd, 2023; Çakmak, 2022). Nevertheless, this novel learning experience is not flawless. Existing studies have discussed practical challenges and limitations of integrating AI-mediated speaking tasks in real teaching contexts.

Previous systematic reviews have synthesized the affordances and challenges of integrating technologies in language education (Katsarou et al., 2023; Kızıl et al., 2025; Koç & Savaş, 2025). While these reviews provide valuable and diverse perspectives on implementing technologies in language teaching, several research gaps were identified. For example, the inclusion of a relatively small number of studies may make it difficult to generalize the results across all language-learning contexts (Katsarou et al., 2023). Moreover, most reviews adopt a broad scope, encompassing both text-based and voice-based chatbots across multiple language skills. While reviewing such broad scope offers a useful overview of technological applications in language education, it often leads to a relatively surface-level understanding of the pedagogical potential and limitations of AI chatbots for language education.

To address those gaps, two recent systematic reviews have focused more explicitly on the use of chatbots in speaking instruction (Du & Daniel, 2024; Xing & Saeed, 2025). Du and Daniel (2024) analyzed 24 articles (2017-2023) to examine the trends, benefits, and challenges of AI chatbots in speaking, teaching, and learning. The findings highlight the early stage of research in this area. It also emphasizes the potential of AI chatbots as conversational partners for speaking practice despite chatbots' technological and speech recognition limitations. Xing and Saeed (2025) synthesized a broader range of AI technologies, including both chatbots and automatic speech recognition (ASR) tools, in 55 empirical studies published from 2017 to 2024. The results yield consistent affordances and challenges of using AI technological tools in speaking practice, as in Du and Daniel's (2024) review. It also added limitations of integrating such technologies in speaking practice, particularly with ASR

technologies, such as inaccurate speech recognition, unnatural, and pre-programmed responses.

These two systematic reviews of AI technologies in speaking instruction reveal several critical gaps. Firstly, the focus on how AVCs were integrated in speaking tasks remains underexplored. This gap is important because it made the pedagogical implications unclear, as educators may not know how to carry out the AI-mediated speaking activities in real lessons. Secondly, although prior reviews identified the challenges of implementing AVCs, they largely overlooked the mediating factors that influence the effectiveness of integrating AVCs in speaking instruction. This omission may lead to poorly designed AI-mediated tasks due to insufficient considerations of these factors. To bridge these gaps, the present systematic review examines how AI-mediated speaking tasks were implemented, as well as its trends and challenges. It also investigates the factors that mediate the use of AVCs in speaking instruction. Notably, this review focuses on a single type of AI technologies (AVCs) and in a single language skill (speaking) by synthesizing evidence from 28 peer-reviewed studies published between 2020 and 2025, a period marked by rapid advancement in AI for education.

By addressing these objectives, the study provides several important contributions. For educators, it provides practical pedagogical implications by identifying a range of AI-mediated speaking tasks that can be adapted or adopted across diverse instructional contexts. For researchers, the review provides a comprehensive synthesis of current implementations of AVCs in speaking instruction, from which future research directions and potential replication studies can be derived. For chatbot developers, particularly those involved in designing AVCs, the findings offer valuable insights into what has proven effective and what aspects require further improvement.

Accordingly, the current article aims to address the following research questions:

1. How are AI voice-based chatbots integrated into speaking instruction?
2. What are the impacts of using AI voice-based chatbots on learners' speaking skills?
3. What are the challenges of integrating AI voice-based chatbots in speaking instruction?
4. What are the mediating factors of integrating AI voice-based chatbots in speaking instruction?

Method

This systematic review was conducted according to a protocol registered in the International Database of Education Systematic Reviews (IDESR, registration number IDESR000187). The review process included three major stages (Page et al., 2021), as shown in Figure 1.

Review Process

Stage 1: Identification

We started by formulating six inclusion and exclusion criteria, as outlined in Table 1, to select relevant studies for analysis. First, we included studies written in English (*criterion 1*); therefore, studies written in other languages were not selected. These empirical studies (*criterion 2*) must involve the implementation of at least one AVC either inside or outside the classroom. Thus, the literature review studies and those that used virtual reality (VR), augmented reality (AR), and mixed reality (MR) tools were not included in this study. Types of evidence sources

are limited to peer-reviewed articles (*criterion 3*) focusing on speaking skills or their components (i.e., vocabulary, grammar, pronunciation, and fluency).

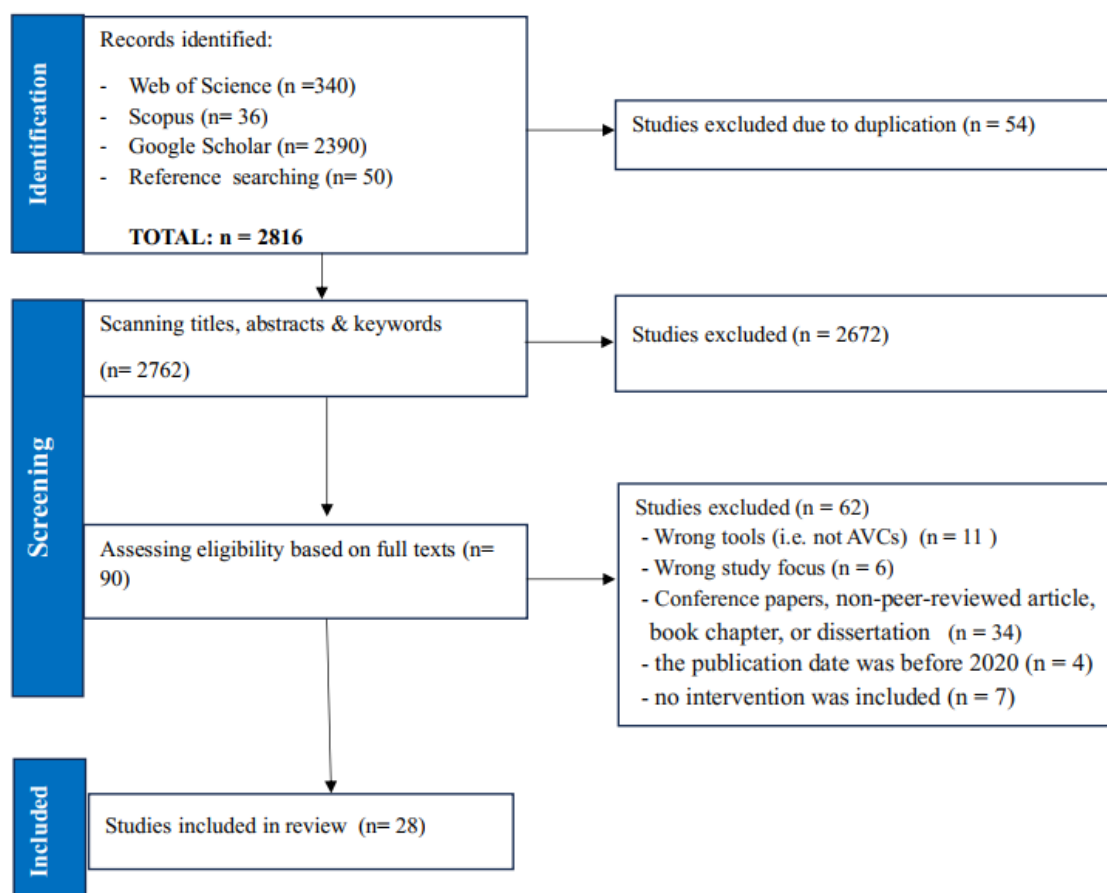


Figure 1. PRISMA Flow Diagram of Review Process

Studies on the impact of AVCs on learners' perception, motivation, engagement, willingness to communicate, or other affective variables were excluded. These articles can be implemented in any educational setting from K12 to higher education (*criterion 4*). The articles must be published between 2020 and 2025 (*criterion 5*). This timeframe was chosen because 2020 marks the advent of GPT-3, a breakthrough in AI conversational capabilities, and 2025 is the most updated point at which this review was conducted. The cut-off point for this current review was December 2025. The studies could be implemented in any country worldwide (*criterion 6*).

The search string includes "AI chatbot" OR "AI voice* chatbot" OR "ChatGPT" OR "virtual conversation* partner" OR "chatbot" OR "AI-generated chatbot" OR "conversation* agent" AND "speak*" OR "speaking performance" OR "oral performance" OR "human-AI conversation" OR "chatbot-human conversation" AND "EFL" OR "English as a foreign language" OR "ESOL" OR "English for speakers of other languages". The study collected data from two academic databases, including Scopus, Web of Science, and one academic search engine (Google Scholar). The selection of sources from multiple databases is commonly used in systematic reviews because each database brings its strengths in terms of subject coverage, indexing standards, and ease of use, which can help reduce the risk of publication bias and the risk of missing relevant studies to be reviewed (Ewald et al.,

2022; Martín-Martín et al., 2018)

Table 1. Inclusion and Exclusion Criteria

Inclusion criteria	
1. Language	Outputs are written in English.
2. Study design	Empirical, primary studies (involving the implementation of at least one AVCs either inside or outside the classroom context)
3. Types of evidence source	Peer-reviewed journal articles Focus on speaking skills or single components of speaking skills such as vocabulary, grammar, pronunciation, and fluency.
4. School levels	K12 to university
5. Time frame	From 2020 to 2025
6. Geographical locus	Outputs drawn from empirical research in any country

We also applied the backward snowballing technique to search for more outputs in the references of the selected studies and added 50 outputs to the list. This step helps ensure comprehensiveness by capturing studies that might have been missed due to indexing issues, variations in keyword usage, or database coverage gaps (Wohlin et al., 2022). The process of searching the sources was conducted from December 1 to December, 2025. We worked together to identify inclusion and exclusion criteria, databases, and strings. After that, we conducted some pilot searches to modify the strings and see how the results matched the inclusion and exclusion criteria. Subsequently, a total of 2816 records were identified, including 340 from Web of Science, 36 from Scopus, 2390 from Google Scholar, and 50 through reference searching.

Stage 2: Screening

Step 1: Screening Titles, Abstracts, and Keywords

The focus of this stage was to screen titles, abstracts, and keywords of exported sources. We conducted a pilot screen by screening the first 10 sources to discuss the inclusion and exclusion criteria. Once the agreement was reached, we continued scanning the rest of the sources independently. We shortlisted 90 outputs after this stage. The outputs were removed because of duplication, not including the use of chatbots, systematic review articles, conference papers, or book chapters, focusing on the use of AI chatbots in writing, reading, or listening, or impacts of using AVCs on affective domains such as willingness to communicate, confidence, motivation, or learning engagement.

Step 2: Assessing Eligibility based on Full Texts

We read full texts of the included records after the initial screening to determine whether they meet the inclusion criteria. When any of us was unsure about the relevance of any study, we would discuss until we reached an agreement on evaluating the relevance of the included studies. After full-text screening, 62 records were excluded for the following reasons: the tool used was not an AVC but only an automatic speech recognition system ($n =$

11); wrong study focus (n = 6); the source was a conference proceeding, non-peer-reviewed article, book chapter, or dissertation (n = 34); the publication date was before 2020 (n = 4); no intervention was included (n = 7). The number of retained outputs for assessing quality was 28.

Step 3: Extracting Data and Assessing the Quality of Evidence

Before assessing the quality of the retained outputs, each study was read closely and its key information was extracted into an Excel file. This spreadsheet includes columns in accordance with the inclusion criteria and other relevant variables aligned with the research questions. To evaluate the quality of the selected studies, we used the appraisal tools devised by the Joanna Briggs Institute (JBI) (Gorard, 2024). The checklist includes nine explicit and straightforward items for evaluating the quality of the selected sources. Each item is framed as a Yes/No question with four alternatives “yes”, “no”, “unclear”, or “not applicable” (Please contact the authors to get the checklists). When assessing article quality using the critical appraisal checklists, we would then provide an overall judgment to either “include” or “exclude” the study. For unsure assessment, we can select “seek further information”, thus one of the other two reviewers can double-read and confirm the assessment.

After selecting the checklists, a pilot quality assessment was done by two researchers. We began this phase by trying to understand the checklists and read five articles together, and agree on how to evaluate the quality of the interventions or the articles as a whole. After that, we continued to assess the quality of the remaining articles. To access inter-rater reliability, 20% of the shortlisted articles were randomly selected and independently double-screened by both raters. High level of agreement (90%) was achieved, indicating strong reliability in the screening process. Any differences in the ratings were resolved through moderation meetings to reach a consensus among raters.

Stage 3: Included

In the final stage of the selection process, 28 studies met all the inclusion criteria and were retained for qualitative synthesis. These records were thoroughly assessed for relevance, methodological quality, and alignment with the review objectives.

Data Analysis

Each of the 28 articles was coded from J01 to J28 and entered into an Excel spreadsheet to extract relevant data for analysis. We employed descriptive statistics and trend visualization techniques to identify bibliometric patterns, such as time frame, research contexts, chatbot types, intervention settings, and methodological approaches. We also applied an inductive thematic analysis approach (Naeem et al., 2023) to identify recurring themes across the records. This bottom-up approach, which does not rely on pre-existing theoretical frameworks, allowed us to minimize potential researcher bias. GRADE-CERQual was then used to evaluate the confidence in each review finding based on four elements, namely methodological limitations, coherence, adequacy of data, and relevance (Lewin et al., 2018). This helped ensure the credibility and overall quality of our review.

Results

Overview of the Evidence Base

This section provides a general description of the overview of the selected sources. First, the number of studies on AVCs in speaking instruction increased substantially recently (Figure 2). During the first four years (2000 - 2023), research on the use of AVCs on speaking skills remained limited, with approximately one to four studies published each year. However, the number of studies published increased noticeably in the following period (2024 - 2025), with nine studies published each year.

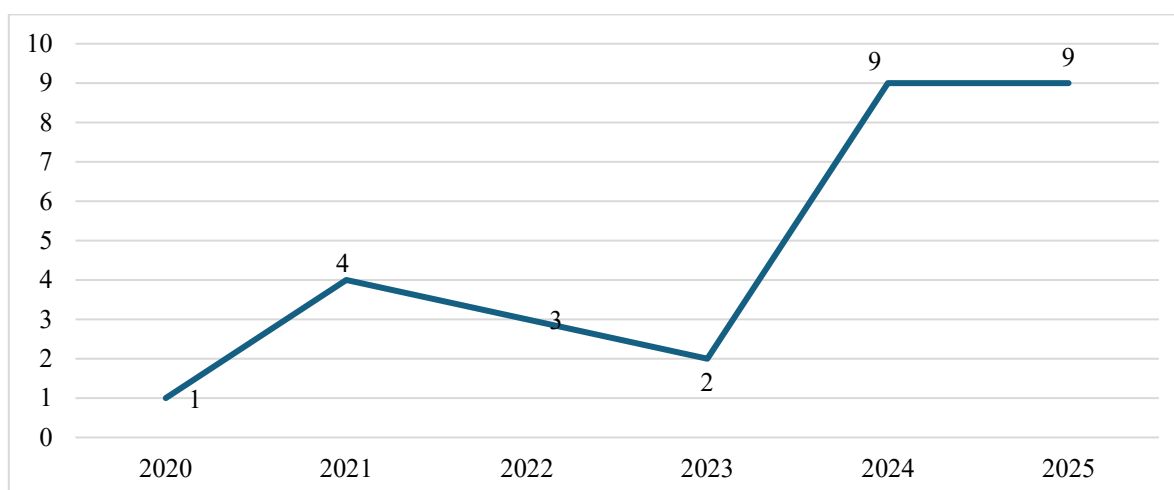


Figure 2. Distribution of the Articles by Year

Second, regarding the geographical features, Figure 3 displays the diverse distribution of research on integrating AVCs in speaking instructions conducted in 13 countries across the continents.

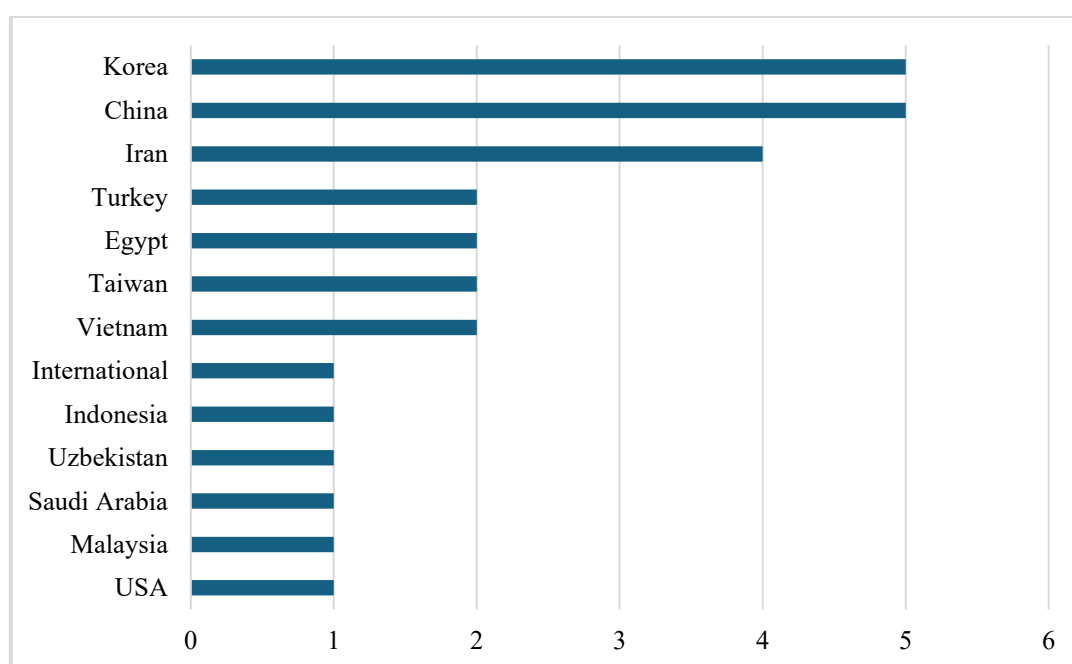


Figure 3. Distribution of the Articles by Countries

The results show predominance of studies conducted in Asian contexts (23 out of 28 studies, over 80% of the selected publications). [Because of Turkey's transcontinental position, studies conducted in Turkey haven't been categorized under Asia or Europe in this analysis.] The remaining articles were based on studies in the African, European, and American continents, with two, two, and one article, respectively. Considering the contributions by country, China and South Korea ($n = 6$) hosted the highest numbers of AI-mediated speaking interventions, followed by Iran ($n = 4$), Vietnam ($n = 2$) and Taiwan ($n = 2$), Egypt ($n = 2$), and Turkey ($n = 2$).

Third, a significant distribution of the AI-mediated speaking interventions was conducted in EFL contexts (89.3%), while the remaining percentages of the publication were in ESL classrooms (11.7%) (see Figure 4).

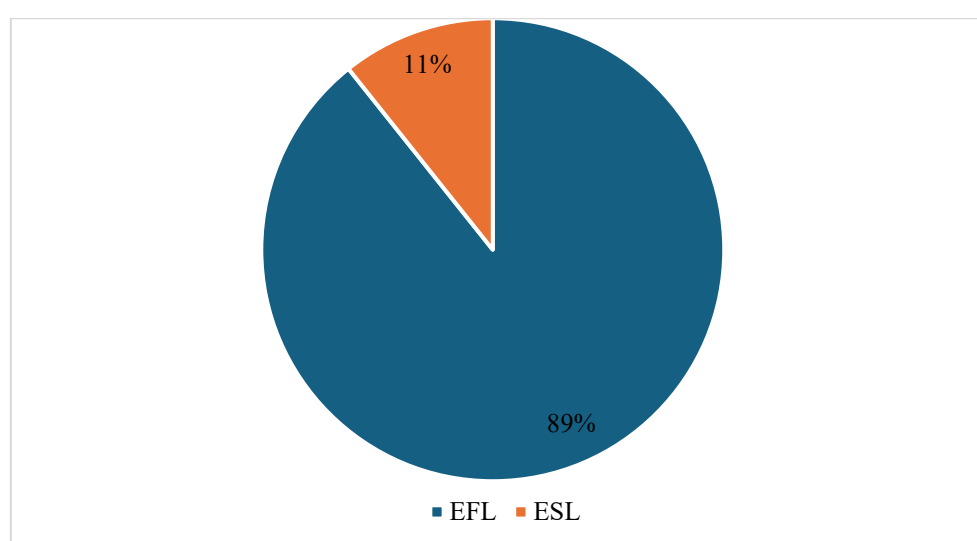


Figure 4. Distribution of the Articles by Context

Fourth, Figure 5 shows the vast popularity of higher education settings where most of the chatbot-mediated speaking interventions were conducted (20 out of 28 articles).

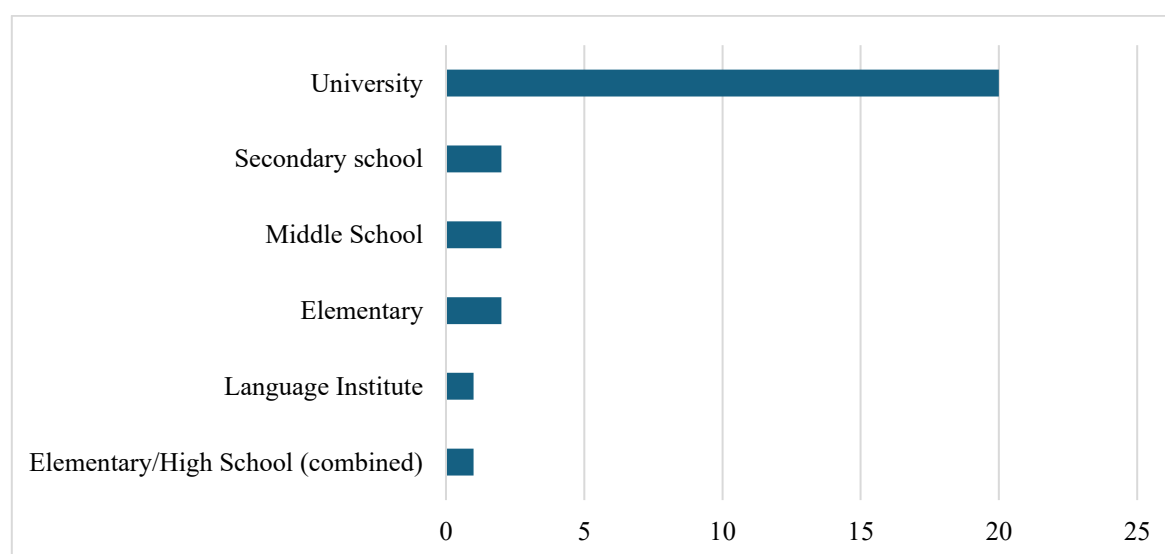


Figure 5. Distribution of the Articles by Educational Settings

The K-12 contexts only took a small number (i.e., Secondary schools ($n = 2$), middle schools ($n = 2$), and Elementary schools ($n = 1$). This uneven distribution of the research on the use of AVCs may reflect the overrepresented sample of university students in applied linguistics research (Andringa & Godfroid, 2020).

Fifthly, regarding the research design, the evidence shows that the mixed-methods approach is the preferred research design (75%), followed by quantitative (25%) (see Figure 6). Notably, twenty-three out of twenty-eight studies (82.1%) employed a quasi-experimental design, including a control or comparison group receiving traditional speaking instruction and an experimental group engaging in AI-mediated speaking tasks.

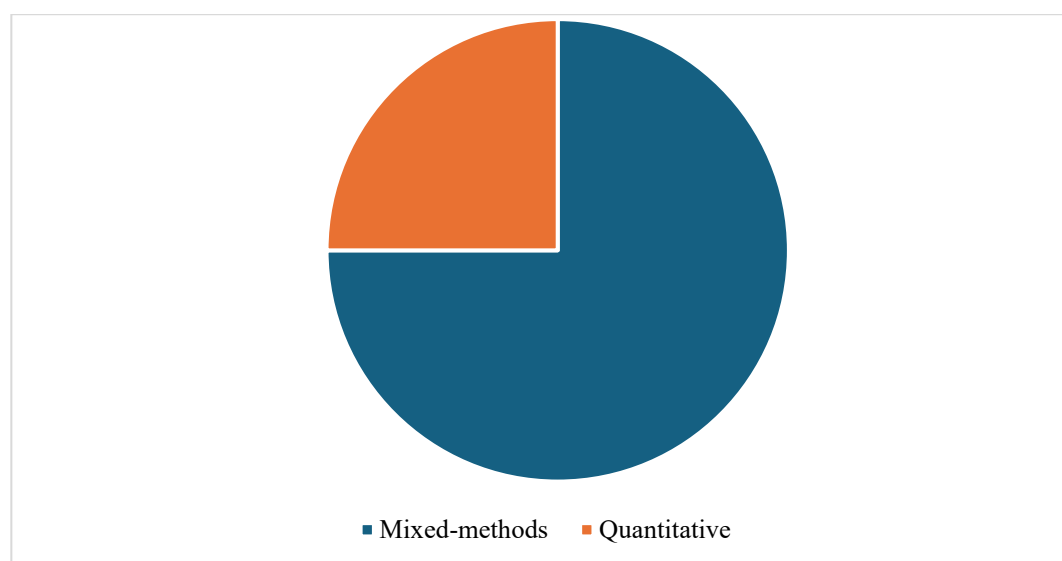


Figure 6. Distribution of the Articles by Research Design

Sample Size of the Research Designs

Most of the selected studies involved either a large or medium scale of participants (14 and 12 studies, respectively) (see Table 2). It means that the interventions involved whole classes or multiple cohorts instead of small groups.

Table 2. Sample Size of the Studies

Scale category	Number of studies
Large-scale ($n > 60$)	14
Medium-scale ($n = 31 - 60$)	12
Small-scale ($n \leq 30$)	2

Note. The classification of sample sizes into small-, medium-, and large-scale categories was adapted from established guidelines in applied linguistics and educational research (Dörnyei, 2007)

Intervention Length

The duration of the interventions varied considerably, ranging from three to 14 weeks (Figure 7). The most

common durations were eight weeks and 12 weeks (seven studies each), followed by 14 weeks (five studies). Shorter interventions, such as three weeks, five weeks, or six weeks, were less frequently implemented. The distribution suggests that most studies align their intervention periods with typical academic schedules, such as half-term or full-semester durations.

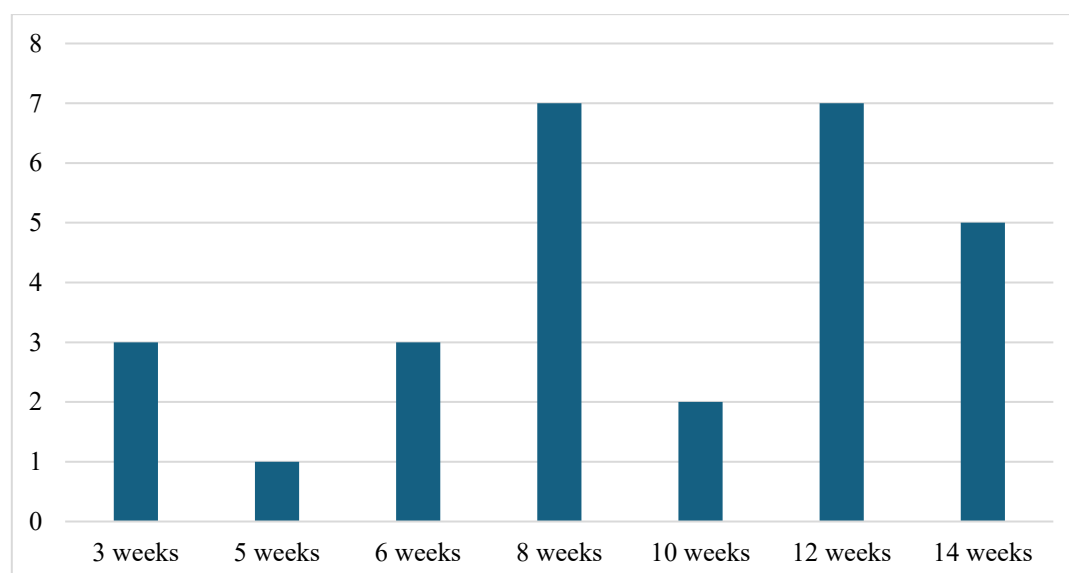


Figure 7. Intervention length

The Implementation of AVCs in Speaking Instruction

Types of AVCs in Speaking Instruction

The selected studies show a wide variety of chatbots used in the 28 interventions (see Figure 8).

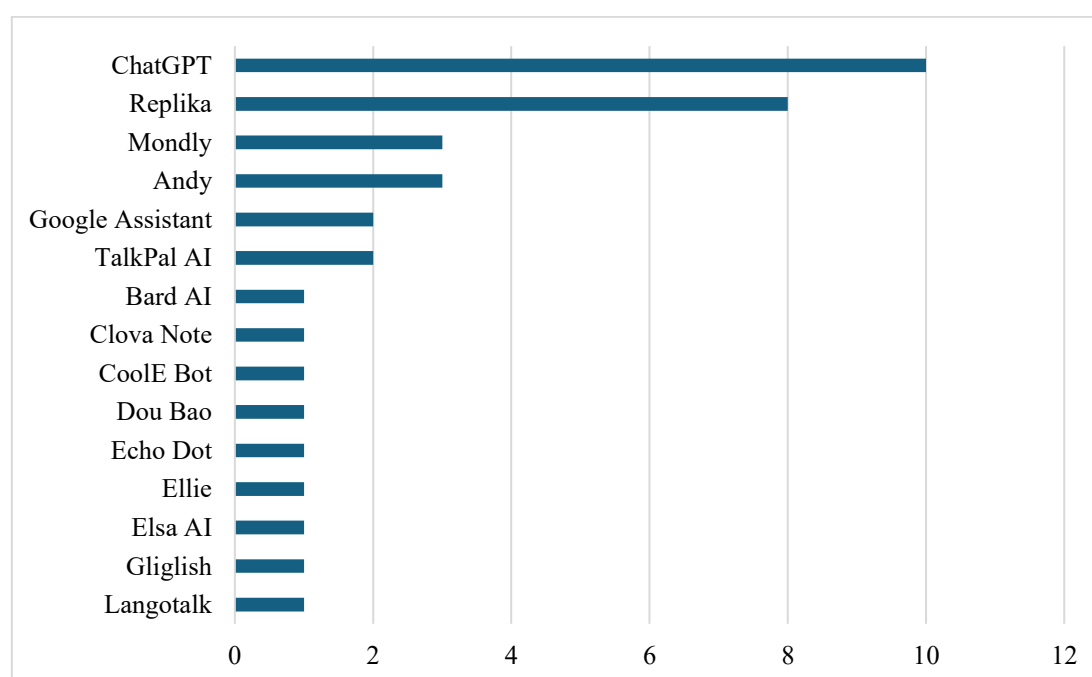


Figure 8. AVCs Used in Speaking Instruction

Among the chatbots, ChatGPT stands out as the most frequently used chatbot (10 studies, e.g., Lim, 2024; Muniandy & Selvanathan, 2024), followed by Replika (eight studies, e.g., Abdullah, 2024; Azizimajd, 2023). Mondly and Andy were each used in three studies, while other AVCs appeared only once. It is important to note that from 2024 onward, generative AI tools, particularly ChatGPT, have become the predominant tool in the interventions, while studies from 2020 - 2023 tended to use pre-existing conversational chatbots such as Replika, Andy, and Mondly. More recently, some studies emerging in 2025 showed a growing trend toward specialized chatbots (e.g., Talk Pal, Gliglish, and Langotalk) to target specific components of speaking proficiency, such as fluency, pronunciation, or vocabulary.

Contexts of the AVC Implementations

Generally, AI-mediated speaking tasks were commonly implemented in formal classroom environment (43%) as shown in Figure 9. Following that, out-of-class interventions were slightly less common, accounting for 32% of the studies. About one-fourth of the studies utilized hybrid interventions with the AI-mediated speaking tasks for both contexts.

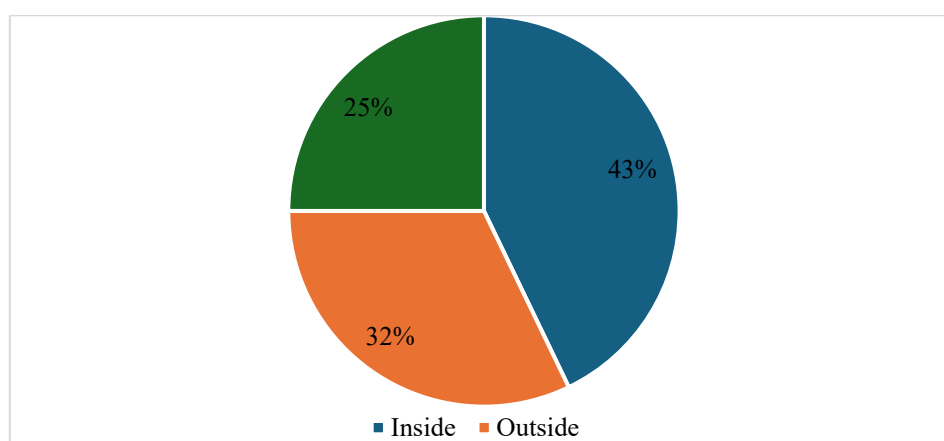


Figure 9. Distribution of the Articles by Contexts of Use

AI-mediated Speaking Tasks

AVCs were integrated into speaking lessons in a diverse range of activities (see Table 3). Among these, role-playing and topic-based chats are the two most popular tasks used for learners to practice speaking skills with AVCs (about 70% of the interventions). Interestingly, research from 2024 onwards has shown that speaking tasks tended to be shifted from free talks to higher-order thinking tasks such as problem-solving or debates (Grab, 2025; Nguyen, 2025).

Table 3. AI-mediated Speaking Tasks

Task type	Sample tasks	Task description	Sources
Communicative tasks	Role- playing	Learners engaged in role plays with AVCs to simulate daily life situations, such as ordering	El Shazly (2021) Yuan (2023),

Task type	Sample tasks	Task description	Sources
		food, making appointments, participating in job interviews, or shopping.	Fathi et al. (2024) Nguyen (2025)
		Topic-based chats	Learners and chatbots have free talks on daily life topics such as hobbies, travel, family, or food. Kang (2022), Azizimajd (2023), Shikun et al. (2024),
	Expressing opinions	Learners shares opinions on books and / or answer open-ended questions from the AVCs	Kim, Kim & Cha (2021), Duong & Suppasetseree (2024)
	Problem-solving	Learners and chatbot discuss the solutions for given situations such as reducing plastic waste.	Grab (2025), Nguyen (2025)
	Debates	Learners and chatbot have debates on a given topic to improve critical thinking and persuasive language.	Grab (2025), Nguyen (2025)
Structured practice	Q&A sessions	Learner collaborates with the AVC to take turns asking and answering questions on a topic	Aliakbari (2025) Shikun et al. (2024)
	Picture description	An image is provided for both learners and the chatbot, each describes the image(s) in turns using the target vocabulary, and grammatical structures to describe details.	Kim, Kim & Cha (2021), Aliakbari (2025), Rad & Roohani (2025)
	Pronunciation drills	Learners read aloud the given text, chatbot recognize the speech and give feedback on pronunciation and intonation.	Abdullah (2024), Ding et al. (2024), Nguyen (2025)
	Oral presentations	Learners collaborate with chatbot to prepare a speech.	Fathi et al. (2024), Lim (2024).
		Learners discuss with the chatbot to generate ideas, then present for the chatbot to give feedback.	

Impacts of Integrating AVCs on Speaking Skills

The reviewed articles provide evidence on the positive impacts of AVCs on students' fluency development, pronunciation accuracy, vocabulary expansion, and grammatical accuracy, contributing to their improved speaking performance.

AI-mediated Speaking Practice Enhances Learners' Fluency

Fluency is the most frequently improved speaking component across the selected sources as indicated in 23 out of 28 studies (e.g., Chen et al., 2025; Duong & Suppasetseree, 2024). Integrating AVCs in speaking instruction was particularly effective in enhancing speaking speed. For example, Shikun et al. (2024) measured the number

of words per second, and asserted that intermediate-level students' fluency score rose from 2.12 to 2.31 post-intervention. Similarly, some other researchers found that students who practiced speaking with AVCs such as Andy, ChatGPT, or Replika spoke faster (e.g., Duong & Suppasetseree, 2024; Kang, 2022; Lin & Mubarak, 2021). In addition, it is evident that students could speak with fewer hesitations and pauses (e.g., Abdullah, 2024; Azizimajid, 2023; Duong & Suppasetseree, 2024). It is important to note that fluency improvement was found among both intermediate-level and lower-level students (Chen et al., 2025; Shikun et al., 2024).

Pronunciation Improvement

The second most frequently reported benefit of practicing speaking with AVCs lies in pronunciation enhancement. Results from 20 out of 28 studies indicate that after practicing speaking with AVCs, participants showed better articulation of individual sounds and more accurate intonation and stress patterns (e.g., Grab, 2025; Kang, 2022; Kim, Cha & Kim, 2021; Wu et al., 2025). Pronunciation improvement after AI-mediated speaking practice was found in both intermediate- and low-level students improve pronunciation, intonation, and stress, as chatbots provide a model for correct pronunciation (Kim, Kim & Cha, 2021; Shikun et al., 2024).

Vocabulary Expansion

The reviewed articles reveal that with the help of AVCs, students notably expanded their lexical repertoire (Fathi et al., 2024; Shikun et al., 2024; Yuan, 2023; Zou et al., 2024). AVCs facilitated lexical development by introducing contextually relevant vocabulary and encouraging its application in dialogues. Thus, learners could also ask AVCs for an explanation or clarification of a new word (Duong & Suppasetseree, 2024). Learners exposed to AVCs, such as ChatGPT demonstrated increased lexical range, with some studies noting their use of advanced vocabulary from the BNC-COCA 25,000 corpus (Yang et al., 2022; Zou et al., 2024). Dynamic platforms such as CoolE Bot further supported vocabulary acquisition through topic-tailored interactions (Tai & Chen, 2024) that helped students acquire natural expressions and then use them in real life (Kim, Kim & Cha, 2021).

Grammatical Accuracy

The studies consistently suggest that practicing speaking with AVCs facilitates learners' ability to use accurate grammatical structures. To illustrate, Shikun et al. (2024) highlight that AVCs provide instant and effective feedback and offer grammatical corrections for errors. This immediate support and real-time exposure to correct forms help learners notice their errors and improve their grammar. Moreover, AVCs were found to serve as reliable models of correct language use (Duong & Suppasetseree, 2024), allowing learners to learn grammar in context through interactive communication with AVCs.

Challenges of Implementing AI-mediated Speaking Tasks

The present systematic review categorizes the challenges of implementing AI-mediated speaking tasks into three

themes: technological limitations, user experience challenges, and affective and behavioral engagement limitations.

Technological Limitations

The successful conversations between learners and AVCs were commonly interfered with by technological issues such as inaccurate speech recognition, irrelevant chatbot responses, connectivity problems, and accessibility issues. First, a predominant concern among learners was the chatbots' inability to recognize speech accurately, especially in the presence of pronunciation errors or non-native accents. At least six studies (e.g., Çakmak, 2022; Duong & Suppasetseree, 2024) highlighted students' frustration when their utterances were not correctly understood, even after repetition attempts for clarification. Many chatbots (e.g., Replika, Andy, Dou Bao, or Ellie) were not very good at recognizing students' accents due to various reasons.

Another limitation reported in the selected studies was chatbots' inconsistent responses, which significantly affected the quality and coherence of conversations. Such issues as off-topic replies, repeated or predictable answers, and misunderstandings made conversations feel unnatural and caused students to lose interest (Kim, Kim & Cha, 2021; Shikun et al., 2024; Yuan, 2023). Learners expressed disappointment when the chatbots failed to understand what they meant or suddenly changed the topic, even though they were trying to stay on track in the conversation (Çakmak, 2022).

Next, challenges regarding accessibility and connectivity were highlighted in at least six articles. Several examples of these barriers include slow response times, poor internet connections, and technological constraints hindering effective use. For instance, Mondly, ChatGPT, and CallAnnie may discourage learners' interaction due to their delays in giving responses (Muniandy & Selvanathan, 2024; Yuan, 2023; Zou et al., 2024). Moreover, the student-chatbot interaction can be undermined by a poor Internet connection. In order for such interventions to be implemented, learners should have a suitable technological device, such as a computer or mobile phone, and a high-speed internet connection (Fathi et al., 2024). Another technological constraint was advertisements shown in chatbots (e.g., Replika, Andy, Google Assistant) that hinder learners' concentration and continuance of the communication (Kim, Kim & Cha, 2021).

User Experience Challenges

The sources discussed three types of challenges of implementing AI-mediated speaking tasks related to learners' characteristics, including lack of prompting skills, proficiency levels, and prior familiarity with AI-mediated speaking practice. First, students seem to have struggled to construct effective prompts. Learners tended to formulate either too short prompts that led to vague or unhelpful answers from AVCs (Muniandy & Selvanathan, 2024), or too complex ones that resulted in complicated responses (Zou et al., 2024). Second, students' proficiency-level differences significantly influenced their AVC interactions. To illustrate, lower-proficiency students would find it difficult to understand chatbots' responses and express their own ideas clearly due to their limited vocabulary and sentence structures (Muniandy & Selvanathan, 2024). Moreover, their poor pronunciation

may make it difficult for chatbots to understand their linguistic input and provide relevant responses (Shikun et al., 2024). As for advanced students, chatbots often pointed out small mistakes without offering detailed feedback on intonation or stress, which would cause frustration (Zou et al., 2024). Finally, learners' limited prior experience with AVCs can be a barrier for successful implementation of AI-mediated speaking tasks. For instance, learners who had never used AVCs needed more time to familiarize themselves with the interfaces of chatbots and how to maintain smooth conversations with AVCs such as Replika (Kang, 2022) or Mondly (El Shazly, 2021).

Affective and Behavioral Engagement Challenges

Affective and behavioral engagement challenges included feelings of frustration and anxiety during the AI-mediated speaking tasks, and a tendency to over-reliance on the chatbots. First, students tended to feel anxious and frustrated when engaging in conversations with AVCs. They became frustrated when the AVCs failed to understand what they were trying to say or when communication broke down entirely (Kang, 2022). Additionally, some students experienced technology-related anxiety, such as chatbot's overwhelming feedback or complicated speaking (Zou et al., 2024). Second, learners did not feel comfortable using AVCs (e.g., Replika) due to the lack of natural flow and emotional warmth (Çakmak, 2022), or human qualities such as facial expressions, gestures, and emotional responsiveness (Kim, Cha & Kim, 2021; Shikun et al., 2024). As a result, conversations with AVCs created an artificial, mechanical experience that can be far different from communicating with peers.

Mediating Factors Enhancing the Effectiveness of Integrating AVCs in Speaking Instruction

The reviewed articles provided evidence on the four common types of mediating factors, including students' proficiency levels, social and psychological factors, chatbot-related features, and pedagogical support.

Students' Proficiency Level

Researchers have found that advanced students' ability to rephrase problematic utterances when communication breakdowns occur between students and AVCs can enhance mutual understanding with the AVCs (Çakmak, 2022). As a result, advanced students showed greater improvements in fluency. When it comes to beginners, AVCs identified errors made by them and enabled them to repeatedly practice the corrected utterances, gradually improving their pronunciation and intonation (Abdullah, 2024; Kang, 2022; Shikun et al., 2024). For intermediate-level learners, their awareness of pronunciation errors and their willingness to be exposed to authentic contexts to practice English help them improve both fluency, pronunciation, stress, and intonation after interaction with AVCs (Kim, Kim & Cha, 2021; Shikun et al., 2024).

Affective Factors

Affective factors refer to learners' emotional states, attitudes, motivation, confidence, and anxiety levels, which significantly influence their overall language learning outcomes (Dörnyei & Ushioda, 2011). The review highlights the significance of these factors in influencing the integration of AVCs into speaking instruction. Lower

anxiety levels can enhance language outcomes by making learners more attentive and improving memory (Alenezi & Alenezi, 2025; El Shazly, 2021). Moreover, students were encouraged and motivated to speak English freely without fear of being judged by peers, or interlocutors, regardless of their skill level. This positive affective state helped learners speak more fluently and express their thoughts more smoothly (Fathi et al., 2024).

Chatbot-related Features

Chatbot-related mediation relates to the design and features of the AVCs that enhance communicative effectiveness. Many advanced chatbots, such as Mondly and ChatGPT, incorporate multimodal features such as audio, visual, and textual affordances that support language acquisition by modeling accurate pronunciation and appropriate expressions. These tools serve as reliable language input sources that learners can imitate, repeat, and practice flexibly at any time and in any place (El Shazly, 2021). Another impactful feature of AI chatbots is their ability to provide real-time, personalized feedback. AVCs can detect learners' errors in pronunciation, grammar, and vocabulary usage and provide corrective feedback for learners to improve (Shikun et al., 2024; Yildiz, 2024).

Pedagogical Support

The reviewed studies highlight roles of pedagogical support during AI-mediated speaking lessons in facilitating learners' speaking performance. This review identifies two main categories of this theme: task design and teacher scaffolding. Students performed better when the tasks were interesting or when they had previously experienced success in similar activities (Çakmak, 2022). In addition, the reviewed studies provide evidence of the crucial role of human instructors in guiding students during the learning process. To illustrate, logically sequenced instructional plans allowed learners to follow the flow of activities smoothly, enhancing their participation (Abdullah, 2024). It is also vital for instructors to provide necessary guidance, technical training, and support so that learners can utilize AVCs effectively (Lim, 2024). Instructor support ensured that AI-mediated speaking tasks aligned with learning objectives (Duong & Suppasetsee, 2024).

Discussion and the Ways Forward

The current review of the 28 studies enhances the understanding of the integration of AVCs into speaking instructions in EFL contexts. The reviewed articles have provided evidence on (1) the AVCs intervention implementation, (2) their impacts on EFL learners' speaking performance, (3) challenges of implementing AI-mediated speaking tasks, and (4) mediating factors enhancing the use of AVCs in speaking instruction.

The synthesis from the selected sources shows a growing trend of research on AI-mediated speaking tasks. This finding marks a notable expansion compared to its infancy two years ago as reported by Du and Daniel (2024). A notable result is that most AVC interventions have been conducted in Asian higher educational contexts where English is taught as a Foreign Language (EFL). Geographically, these studies show a strong concentration in East Asia and the Middle East. The predominance of EFL settings (nearly 90%) reflects an urgent need for AVCs to create simulated language practice environments in countries that lack authentic opportunities for direct

interaction with native speakers (Gan, 2012; Ur, 1996). There is thus a need to conduct related studies in other regions, especially those in the Global South, wherein technological advances and applications are still lagging behind other regions. Moreover, the included studies have been mainly conducted in higher education contexts. This highlights the existing imbalance in formal knowledge production, underscoring the need for more empirical research in other educational settings to diversify and reinforce the evidence base on the use of AVCs in English-speaking instruction.

In terms of research design, most of the research in this area employed a mixed-methods approach with quasi-experimental designs (23 out of 28 studies). The popularity of using quasi-experimental designs indicates that researchers seek to establish causal relationships between AI-mediated interventions and speaking performance outcomes. Moreover, the integration of qualitative and quantitative data reflects an effort to capture both measurable learning gains and learners' experiences (Creswell & Clark, 2018). However, there seems to be a lack of reliable qualitative studies to delve into the lived experience of the stakeholders involved in the use of AVCs in a classroom setting. Further inquiries should place more stress on this research approach to shed more light on the issue.

Regarding the overall description of implementing AVCs in speaking instruction, the evidence consistently demonstrates their feasibility and pedagogical potential of implementing AI-mediated speaking tasks in real classrooms. A wide range of both free and paid AVCs are available for educators to choose from. It means that AI-mediated speaking practice is accessible to learners across diverse contexts providing that they have a digital device and internet access. Among these tools, ChatGPT is the most widely used, followed by Replika. The main difference between these two AVC types is their interface. Replika employs an avatar-based design that enhances perceived social presence and creates a sense of human-like interaction (Kang, 2022; Wu et al., 2025). Although ChatGPT's outlook is not embodied, this AVC is more commonly used because of its perceived outstanding features for speaking practice such as human-like conversation engagement (Du & Daniel, 2024), contextual understanding (Xing & Saeed, 2025), easily accessible (Kızıl et al., 2025), ability to support diverse speaking tasks with adaptive feedback which are critical for language learning (Xing & Saeed, 2025). However, with the rapid development and evolution of AI and its application, more advanced versions of chatbots will be introduced with greater efficiency and more sophisticated intelligence. Further research evidence is needed to provide insights into this uncharted territory of AI-based pedagogy in language education.

In terms of the contexts, the current studies show a strong preference for in-class interventions. The explanation for this trend is that classroom-based implementation allows teachers to structure tasks, provide immediate pedagogical support, and monitor learners' engagement in real time, which are important for effective speaking instruction (Brown, 2001). However, the out-of-class and the hybrid implementations also account for a relatively substantial proportion. It indicates the flexible use of AI-mediated speaking tasks in both formal instructional settings and self-directed learning environments. Outside the classroom, AVCs can provide increased opportunities for autonomous, low-anxiety practice and repeated exposure which are often limited in traditional classroom settings due to time constraints (Dörnyei, 2007).

Hybrid models, which combine in-class guidance with out-of-class practice, are also a promising type of implementation because it balances structure support with learners' self-practice which facilitate sustained engagement and individualized pacing (Dörnyei, 2014). Such flexibility aligns with the principles of blended learning, where teaching processes are distributed across multiple contexts to enhance both guidance and autonomy (Garrison & Vaughan, 2008). Future studies might look into the nature of these teaching models to provide more texture to the dynamics of AI integration into classroom settings and to illustrate how AVCs can assist teachers in accomplishing their roles.

Regarding the AI-mediated speaking activities, the sources showed a diversity of task types in which learners engaged in conversations with AVCs. The evidence highlighted the prominence of communicative tasks, such as role-playing, topic-based dialogues. This finding shows chatbot's ability to simulate real-time interaction and sustain dialogue. This preference is consistent with the emphasis on meaning-focused communication in task-based language teaching (Ellis, 2003). In addition, structured practice activities, such as question-and-answer sessions, picture description, and communication drills were also frequently used to organize AI-mediated speaking tasks. These structured activities support learners' linguistic accuracy and provide more controlled opportunities for practice. Future research might continue to examine how different AI-mediated speaking task types differentially influence learners' fluency, accuracy, interactional competence, and engagement over time

With respect to the challenges of voice chatbot-integrated EFL speaking instruction, this pool of evidence identifies barriers relating to technology (e.g., speech recognition issues, irrelevant or inconsistent responses, slow response time, poor Internet connection), user experience (e.g., learners' ability to formulate effective prompts, their familiarity with the used chatbots) and user affective and behavioral engagement (e.g., learners' frustration and anxiety during interactions and potential overreliance on chatbots). Our review corroborates previous findings (Katsarou et al., 2023) that critique the shortcomings in current chatbots, including communication breakdowns in EFL learner-chatbot interactions, the limited ability of chatbots to reproduce the dynamics of human-human interaction, and other technology-related connectivity and accessibility issues. More research attention should be paid to investigating pedagogical strategies and teacher scaffolding techniques that can reduce technological barriers, learner anxiety, and chatbot overdependence in speaking practice.

Finally, this review suggests that four factors (i.e., learners' proficiency level, affective factors, chatbot-related factors, and pedagogical support) mediate the successful implementation and effectiveness of AVCs in EFL speaking instruction. Regarding the pedagogical support from human instructors, our review substantiates the idea in Ji et al.'s review (2023) that human instructors should play a crucial role not only in designing and assigning tasks as expected in traditional lessons but also in facilitating chatbot-based activities and providing students with feedback or assisting them in responding to feedback automatically generated by chatbots. The findings suggest that the effectiveness of AVCs in speaking instruction is not determined by the technology alone, but by the dynamic interplay of affective, cognitive, pedagogical, and technological factors. In the near future, we believe that more research effort is needed to explore how teachers' pedagogical decisions interact with learner variables and chatbot affordances across different instructional contexts and proficiency levels.

Conclusion

The findings from our review, together with those of two previous reviews (Du & Daniel, 2024; Koç & Savaş, 2025), show a range of encouraging evidence indicating the significance of AVCs in developing students' oral performance in EFL contexts. This evidence suggests the need to develop guidelines and/or training programs focusing on chatbot-based task design, assignment, implementation, as well as feedback giving and responses for EFL teachers to integrate AVCs into their speaking lessons effectively. However, the current review is subject to several notable limitations. First, only two databases (i.e., Scopus, Web of Science), and one search engine (i.e., Google Scholar) were included. This means that potentially relevant publications from other databases (e.g., ERIC, LLBA) and the grey literature may have been overlooked. Moreover, after searching the three bibliographic sources, we implemented a backward citation strategy only. Therefore, the review may have missed literature that could have been identified through two other strategies (i.e., forward citation and hand searching on key journals in the field of computer-assisted language teaching and learning). Additionally, this review focuses on research outputs published in English only. As a result, it may miss literature written in other languages, and geographical and cultural biases may therefore influence its evidence base. These limitations should be considered when interpreting the implications and proposing directions for future research.

Statements and Declarations

Acknowledgments: This work was partly supported by the Industrial University of Ho Chi Minh city and Hue University (via research project DHH2026-07-121).

This manuscript made use of Grammarly to assist in paraphrasing and improving the clarity of text and ChatGPT for locating information, articles, and references. The final manuscript was authored and reviewed by the listed authors without additional AI input.

Supplementary Materials: Not applicable.

Author Contributions: All authors contributed equally for the study design, data collection, data analysis, interpretation of results, and drafting and revising the manuscript. The authors read and approved the final version of the manuscript.

Funding: This work was funded by Hue University (via research project DHH2026-07-121).

Data Availability: Not applicable.

Ethics Approval: The study was performed in accordance with the study protocol and ethical guidelines and regulations.

Informed Consent: Not applicable.

Conflicts of Interest: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

- Abdullah, H. M. (2024). Using an AI-driven conversational chatbot to develop EFL oral communication skills of first year secondary school students and reduce their oral communication apprehension. *CDELT Occasional Papers in the Development of English Education*, 88(1), 319-340. <https://doi.org/10.21608/opde.2024.390894>
- Alenezi, A., & Alenezi, A. (2025). Evaluating the effectiveness of chatbot-assisted learning in enhancing English conversational skills among secondary school students. *Education Sciences*, 15(9), 1136. <https://doi.org/10.3390/educsci15091136>
- Aliakbari, M., Barzan, P., Allahveysi, S. P., & Bakhtiarvand, M. (2025). AI-powered chatbots for English conversation practice: Impact on speaking proficiency. *Technology Assisted Language Education*, 33, 45-62. <https://doi.org/10.22126/tale.2025.12621.1138>
- Andringa, S., & Godfroid, A. (2020). Sampling bias and the problem of generalizability in applied linguistics. *Annual Review of Applied Linguistics*, 40, 134-142. <https://doi.org/10.1017/S0267190520000033>
- Azizimajd, H. (2023). Investigating the impacts of voice-based student-chatbot interactions in the classroom on EFL learners' oral fluency and foreign language speaking anxiety. *Technology Assisted Language Education*, 1(2), 61-83. <https://doi.org/10.22126/tale.2023.2732>
- Brown, H. D. (2001). *Teaching by principles: An interactive approach to language pedagogy*. Addison Wesley Longman.
- Çakmak, F. (2022). Chatbot-human interaction and its effects on EFL students' L2 speaking performance and anxiety. *Novitas-ROYAL (Research on Youth and Language)*, 16(2), 113-136.
- Chen, A. X., Jia, J. Y., Li, Y. Z., & Fu, L. Y. (2025). Investigating the effect of role-play activity with GenAI agent on EFL students' speaking performance. *Journal of Educational Computing Research*, 63(7), 99-125. <https://doi.org/10.1177/07356331241299058>
- Chulee, W., Khanom, K., Chesa, N., Sha'ar, M. Y. M. A., & Buddharat, C. (2023). Why still we cannot speak English? Examining internal demotivating factors among Thai tertiary learners. *MEXTESOL*, 47(4), 1-17 <https://doi.org/10.61871/mj.v47n4-5>
- Creswell, J. W., & Clark, V. L. P. (2018). *Designing and conducting mixed methods research*. Sage Publications.
- Ding, X. X., Chen, M., & Xu, Q. (2024). Anxiety and enjoyment in oral presentations: A mixed-method study into Chinese EFL learners' oral presentation performance. *Iral-International Review of Applied Linguistics in Language Teaching*, 63(3), 1973-1997. <https://doi.org/10.1515/iral-2023-0161>
- Dörnyei, Z. (2007). *Research methods in applied linguistics*. Oxford university press.
- Dörnyei, Z. (2014). *The psychology of the language learner: Individual differences in second language acquisition*. Routledge.
- Dörnyei, Z., & Ushioda, E. (2011). *Teaching and researching motivation*. Routledge.
- Du, J., & Daniel, B. K. (2024). A Systematic Review of AI-Powered Chatbots in EFL Speaking Practice: Transforming Language Education. *Computers and Education: Artificial Intelligence*, 100230.

- <https://doi.org/10.1016/j.caeai.2024.100230>
- Duong, T. V. T., & Suppasetsree, 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>
- El Shazly, R. (2021). Effects of artificial intelligence on English speaking anxiety and speaking performance: A case study. *Expert Systems*, 38(3), e12667. <https://doi.org/10.1111/exsy.12667>
- Ellis, R. (2003). *Task-based Language Learning and Teaching*. Oxford University Press.
- Ewald, H., Klerings, I., Wagner, G., Heise, T. L., Stratil, J. M., Lhachimi, S. K., Hemkens, L. G., Gartlehner, G., Armijo-Olivo, S., & Nussbaumer-Streit, B. (2022). Searching two or more databases decreased the risk of missing relevant studies: A metaresearch study. *Journal of Clinical Epidemiology*, 149, 154-164. <https://doi.org/10.1016/j.jclinepi.2022.05.022>
- Fathi, J., Rahimi, M., & Derakhshan, A. (2024). Improving EFL learners' speaking skills and willingness to communicate via artificial intelligence-mediated interactions. *System*, 121, 103254. <https://doi.org/10.1016/j.system.2024.103254>
- Gan, Z. (2012). Understanding L2 speaking problems: Implications for ESL curriculum development in a teacher training institution in Hong Kong. *Australian Journal of Teacher Education (Online)*, 37(1), 43-59. <https://doi.org/10.14221/ajte.2012v37n1.4>
- Garrison, D. R., & Vaughan, N. D. (2008). *Blended learning in higher education: Framework, principles, and guidelines*. John Wiley & Sons.
- Gorard, S. (2024). Judging the relative trustworthiness of research results: How to do it and why it matters. *Review of Education*, 12(1), e3448. <https://doi.org/10.1002/rev3.3448>
- Grab, M. Ö. (2025). Integrated AI chatbot practice: A pathway to improved ESL speaking skills. *Social Sciences & Humanities Open*, 12, 101933. <https://doi.org/10.1016/j.ssaho.2025.101933>
- Han, D. (2020). An analysis of Korean EFL learners' experience on English classes using AI chatbot. *Robotics & AI Ethics*, 6(3), 1-9. doi.org/10.22471/ai.2021.6.3.01
- Hwang, G. J., Rahimi, M., & Fathi, J. (2024). Enhancing EFL learners' speaking skills, foreign language enjoyment, and language-specific grit utilising the affordances of a MALL app: A microgenetic perspective. *Computers & Education*, 214. <https://doi.org/10.1016/j.compedu.2024.105015>
- Ji, H., Han, I., & Ko, Y. (2023). A systematic review of conversational AI in language education: Focusing on the collaboration with human teachers. *Journal of Research on Technology in Education*, 55(1), 48-63.
- Kang, H. (2022). Effects of artificial intelligence (AI) and native speaker interlocutors on ESL learners' speaking ability and affective aspects. *Multimedia-Assisted Language Learning*, 25(2), 9-43. <https://doi.org/10.15702/MALL.2022.25.2.09>
- Katsarou, E., Wild, F., Sougari, A.-M., & Chatzipanagiotou, P. (2023). A systematic review of voice-based intelligent virtual agents in EFL education. *International Journal of Emerging Technologies in Learning (iJET)*, 18(10), 65–85. <https://doi.org/10.3991/ijet.v18i10.37723>
- Kim, H. S., Kim, N. Y., & Cha, Y. (2021). Is it beneficial to use AI chatbots to improve learners' speaking performance? *Journal of Asia TEFL*, 18(1), 161–178. <https://doi.org/10.18823/asiatefl.2021.18.1.10.16>
- Kim, H. S., Cha, Y., & Kim, N. Y. (2021). Effects of AI chatbots on EFL students' communication skills. *영어학*, 21, 712–734.

- Kızıllı, Ş. A., Klimova, B., Pikhart, M., & Parmaxi, A. (2025). A systematic review of the recent research on the usefulness of chatbots for language education. *Journal of Computer Assisted Learning*, 41(2), e70001. <https://doi.org/10.1111/jcal.70001>
- Koç, F. Ş., & Savaş, P. (2025). The use of artificially intelligent chatbots in English language learning: A systematic meta-synthesis study of articles published between 2010 and 2024. *ReCALL*, 37(1), 4-21. <https://doi.org/10.1017/S0958344024000168>
- Lewin, S., Booth, A., Glenton, C., Munthe-Kaas, H., Rashidian, A., Wainwright, M., Bohren, M. A., Tunçalp, Ö., Colvin, C. J., Garside, R., Carlsen, B., Langlois, E. V., & Noyes, J. (2018). Applying GRADE-CERQual to qualitative evidence synthesis findings: Introduction to the series. *Implementation Science*, 13(1), 2. <https://doi.org/10.1186/s13012-017-0688-3>
- Lim, Y. (2024). Learning effects of complementary learning methods in speaking skills using AI-supported devices. *Foreign Studies*, 68(6), 91–112. <http://dx.doi.org/10.15755/jfs.2024.68.91>
- Lin, C.-J., & Mubarak, H. (2021). Learning analytics for investigating the mind map-guided AI chatbot approach in an EFL flipped speaking classroom. *Educational Technology & Society*, 24(4), 16-35.
- Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & López-Cózar, E. D. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, 12(4), 1160–1177. <https://doi.org/10.1016/j.joi.2018.09.002>
- Muniandy, J., & Selvanathan, M. (2024). ChatGPT, a partnering tool to improve ESL learners' speaking skills: Case study in a public university, Malaysia. *Teaching Public Administration*, 43(1), 4-20. <https://doi.org/10.1177/01447394241230152>
- Naeem, M., Ozuem, W., Howell, K., & Ranfagni, S. (2023). A step-by-step process of thematic analysis to develop a conceptual model in qualitative research. *International Journal of Qualitative Methods*, 22, 16094069231205789. <https://doi.org/10.1177/16094069231205789>
- Nguyen, N. T. (2025). The impact of AI chatbot integration on Vietnamese EFL learners' speaking performance, fluency, and confidence. *Language, Technology, and Social Media*, 4(1), 70-88. <https://doi.org/10.70211/ltsm.3026-7196.314>
- Nunan, D. (2003). *Practical english language teaching*. McGraw-Hill Education.
- 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. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372. <https://doi.org/10.1136/bmj.n71>
- Pérez, J. Q., Daradoumis, T., & Puig, J. M. M. (2020). Rediscovering the use of chatbots in education: A systematic literature review. *Computer Applications in Engineering Education*, 28(6), 1549-1565. <https://doi.org/10.1002/cae.22326>
- Rad, S. H., & Roohani, A. (2025). AI language alchemists: Unleashing task-based chatbots to enhance speaking proficiency, shape attitudes, and foster a translanguaging space. *Journal of Educational Computing Research*, 63(7-8), 1659-1688. <https://doi.org/10.1177/07356331251363281>
- Shikun, S., Grigoryan, G., Huichun, N., & Harutyunyan, H. (2024). AI chatbots: Developing English language proficiency in EFL classroom. *Arab World English Journal*, 292–305. <https://doi.org/10.24093/awej/ChatGPT.20>
- Tai, T. Y., & Chen, H. H. J. (2024). Improving elementary EFL speaking skills with generative AI chatbots:

- Exploring individual and paired interactions. *Computers & Education*, 220, 105112. <https://doi.org/10.1016/j.compedu.2024.105112>
- Ur, P. (1996). *A course in language teaching* (Vol. 1). Cambridge University Press.
- Wohlin, C., Kalinowski, M., Felizardo, K. R., & Mendes, E. (2022). Successful combination of database search and snowballing for identification of primary studies in systematic literature studies. *Information and Software Technology*, 147, 106908. <https://doi.org/10.1016/j.infsof.2022.106908>
- Wu, T. T., Hapsari, I. P., & Huang, Y. M. (2026). Effects of incorporating AI chatbots into think–pair–share activities on EFL speaking anxiety, language enjoyment, and speaking performance. *Computer Assisted Language Learning*, 39(5–6), 1408–1446. <https://doi.org/10.1080/09588221.2025.2478271>
- Xing, C., & Saeed, M. A. (2025). *A systematic review on artificial intelligence (AI) technologies in ESL/EFL speaking skills* (Pt. 250908). 1-31. <https://doi.org/10.58304/ijts.250908>
- Yang, H., Kim, H., Lee, J. H., & Shin, D. (2022). Implementation of an AI chatbot as an English conversation partner in EFL speaking classes. *ReCALL*, 34(3), 327-343. <https://doi.org/10.1017/S0958344022000039>
- Yildiz, 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>
- Yuan, Y. (2024). An empirical study of the efficacy of AI chatbots for English as a foreign language learning in primary education. *Interactive Learning Environments*, 32(10), 6774-6789. <https://doi.org/10.1080/10494820.2023.2282112>
- Zou, B., Wang, C., Yan, Y., Du, X., & Ji, Y. (2024). Exploring English as a foreign language learners' adoption and utilisation of ChatGPT for speaking practice through an extended technology acceptance model. *International Journal of Applied Linguistics*, 35(2), 689-704. <https://doi.org/10.1111/ijal.12658>

Appendix A. List of the Articles Included in This Review

Code	Author (Year)	Country	School levels	Sample size	Research designs	Contexts of use	AVC	Intervention length
J1	Han (2020)	South Korea	Middle School	44	Quanti (QE)	Outside	Echodot	10 weeks
J2	El Shazly (2021)	Egypt	University	48	Mixed (QE)	Both	Audrey, Charles, Cristal, and Mike, Mondly	8 weeks
J3	Kim, Cha & Kim (2021)	South Korea	University	49	Mixed (QE)	Inside	Replika, Andy, Google Assistant	1 semester
J4	Kim, Kim & Cha (2021)	South Korea	University	110	Quanti (QE)	Inside	Replika, Andy, and Google Assistant	1 semester
J5	Lin & Mubarak (2021)	Taiwan	University	50	Quanti (QE)	Both	Replika	5 weeks
J6	Kang (2022)	USA	University	61	Mixed (QE)	Outside	Replika	14 weeks
J7	Yang et al. (2022)	South Korea	Elementary /High School	314	Mixed	Inside	Ellie	3 weeks
J8	Çakmak (2022)	Turkey	University	90	Mixed (QE)	Outside	Replika	12 weeks
J9	Azizimajd (2023)	Iran	University	60	Quanti (QE)	Inside	Replika	8 weeks
J10	Yuan (2023)	China	Elementary	74	Mixed (QE)	Inside	Mondly	3 months
J11	Abdullah (2024)	Egypt	secondary school	33	Quanti (QE) one group	Inside	Replika	12 weeks
J12	Fathi et al. (2024)	Iran	university	65	Mixed (QE)	Both	Andy English	12 weeks
J13	Lim (2024)	South Korea	University (Military)	39	Mixed (QE)	Inside	Clova Note, ChatGPT	Intensive
J14	Muniandy & Selvanathan (2024)	Malaysia	University	40	Mixed	Both	ChatGPT	1 semester
J15	Shikun et al. (2024)	China	Middle School	90	Quanti	Outside	TalkPal, Bard, Doubao	1 semester
J16	Tai & Chen (2024)	Taiwan	Elementary	85	Mixed (QE)	Inside	CoolE Bot	3 weeks
J17	Tam Duong & Suppasetsee (2024)	Vietnam	University	30	Mixed (QE)	Both	Andy English Bot	8 weeks
J18	Yildiz (2024)	Turkey	University	65	Mixed	Outside	ChatGPT	10 weeks
J19	Zou et al. (2024)	China	University	212	Mixed	Outside	ChatGPT	N/A
J20	Alezeni & Alezeni (2025)	Saudi Arabia	secondary school	30	Mixed (QE)	Inside	GPT-powered bot	3 weeks

Code	Author (Year)	Country	School levels	Sample size	Research designs	Contexts of use	AVC	Intervention length
J21	Aliakbari (2025)	Iran	EFL University	60	Quanti (QE)	Inside	Elsa AI	12 weeks
J22	Chen et al. (2025)	China	University	71	Mixed	Outside	ChatGPT 4o	8 weeks
J23	Ding et al. (2024)	China	University	60	Mixed (QE)	Outside	Mondly	6 weeks
J24	Elov et al. (2025)	Uzbekistan	Language Institute	50	Mixed (QE)	Inside	ChatGPT	6 weeks
J25	Grab (2025)	Brazil, India, Nigeria, Saudi Arabia, Vietnam, and Spain.	University	120	Mixed (QE)	Both	TalkPal Gliglish Langotalk ChatGPT	12 weeks
J26	Nguyen (2025)	Vietnam	University	38	Mixed (QE)	Outside	ChatGPT-4	8 weeks
J27	Rad & Roohani (2025)	Iran	University	80	Mixed (QE)	Both	ChatGPT	12 weeks
J28	Wu et al. (2025)	Indonesia	University	75	Mixed (QE)	Inside	Replika	6 weeks